DISCUSSION PAPER SERIES

IZA DP No. 8280

Direct and Indirect Treatment Effects: Causal Chains and Mediation Analysis with Instrumental Variables

Markus Frölich
Martin Huber

June 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Direct and Indirect Treatment Effects: Causal Chains and Mediation Analysis with Instrumental Variables

Markus Frölich
University of Mannheim
and IZA

Martin Huber
University of St. Gallen

Discussion Paper No. 8280
June 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 8280
June 2014

ABSTRACT

Direct and Indirect Treatment Effects: Causal Chains and Mediation Analysis with Instrumental Variables*

This paper discusses the nonparametric identification of causal direct and indirect effects of a binary treatment based on instrumental variables. We identify the indirect effect, which operates through a mediator (i.e. intermediate variable) that is situated on the causal path between the treatment and the outcome, as well as the unmediated direct effect of the treatment using distinct instruments for the endogenous treatment and the endogenous mediator. We examine different settings to obtain nonparametric identification of (natural) direct and indirect as well as controlled direct effects for continuous and discrete mediators and continuous and discrete instruments. We illustrate our approach in two applications: to disentangle the effects (i) of education on health, which may be mediated by income, and (ii) of the Job Corps training program, which may affect earnings indirectly via working longer hours and directly via higher wages per hour.

JEL Classification: C14, C21

Keywords: direct effect, indirect effect, instrument, treatment effects

Corresponding author:

Markus Frölich
University of Mannheim
L7, 3-5
68131 Mannheim
Germany
E-mail: froelich@uni-mannheim.de

* We have benefitted from comments by Kosuke Imai and Teppei Yamamoto and seminar participants in Zurich (ETH research seminar) and Laax. We are grateful to Thomas Aeschbacher for his excellent research assistance. The first author acknowledges financial support from the Research Center SFB 884 “Political Economy of Reforms” Project B5, funded by the German Research Foundation (DFG). The second author acknowledges financial support from the Swiss National Science Foundation grant PBSGP1 138770.

1 Introduction

The treatment evaluation literature has traditionally focussed on the assessment of the total impact of a treatment on an outcome of interest, such as the average treatment effect (ATE). However, in many economic problems, not only the ATE itself appears relevant, but also the causal mechanisms through which it operates. In this case, one would like to disentangle the *direct* effect of the treatment on the outcome as well as the *indirect* ones that run through one or more intermediate variables, so-called mediators.

We subsequently consider a few examples. When assessing the employment or earnings effects of an active labor market policy, one may want to know to which extent the total impact stems from increased search effort, increased human capital, or other mediators that are themselves affected by the policy. When analyzing the health effect of education one might be interested in whether or how much any impact is driven by the fact that higher educated people also have higher incomes. Considering e.g. the effect on smoking, education on the one hand increases incomes which permits consuming more cigarettes. On the other hand, education might have a direct effect on health-related behavior leading to a reduction of smoking. Similarly, interventions in primary school on adult outcomes may be mediated by higher education. When for instance evaluating the earnings effects of math education as in Rose and Betts (2004), math courses likely affect the decision to obtain further education and may also have a direct effect on earnings (conditional on total years of education). Analogously, reductions in class size during compulsory education could affect the probability of obtaining a college degree while also having a direct effect on productivity not mediated by college attainment. As a final example, it has been noted that migration can have at least two effects on the family left behind. On the one hand, migration often triggers remittances, which are expected to alleviate household budget constraints and thereby reduce child labor and improve schooling. On the other hand, the absence of the migrated family members may increase the need for (child) labor in the left-behind households and could also have negative psychological consequences on children's school outcomes. See e.g. Antman (2011), Bargain and Boutin (2014), Binzel and Assaad (2011) or Mu and van de Walle (2011).

Early work on the evaluation of causal mechanisms, frequently referred to as mediation analysis, includes Cochran (1957), Judd and Kenny (1981), and Baron and Kenny (1986). Thereafter, mediation analysis has become very popular in social sciences, see for instance Heckman, Pinto, and Savelyev (2013) for an example in the field of economics. While earlier studies often relied on tight linear specifications, more recent research focuses on non- and semiparametric identification of causal mechanisms, see for instance Pearl (2001), Robins (2003), Petersen, Sinisi, and van der Laan (2006), Simonsen and Skipper (2006), Flores and Flores-Lagunes (2009), van der Weele (2009), Imai, Keele, and Yamamoto (2010), Albert and Nelson (2011), and Huber (2013). With the exception of studies assessing causal mechanisms within subpopulations based on principal stratification, see Rubin (2004) and Mealli and Rubin (2003), the vast majority of the literature assumes that the treatment and the mediator are conditionally exogenous given observed covariates to obtain identification.

In contrast to most other studies, this paper allows for treatment and mediator endogeneity which cannot be tackled by observed covariates alone. As main contribution we demonstrate nonparametric identification via instrumental variables (IV) when both the treatment and the mediator are endogenous due to the non-observability of important confounders. We use distinct (i.e. at least two) instruments to control for both treatment endogeneity –e.g., due to imperfect compliance with treatment randomization in experiments– and mediator endogeneity. In our heterogenous treatment effect model with a binary treatment, identification relies on particular monotonicity and exogeneity assumptions of the instruments. The latter may only be conditionally valid given a set of observed covariates, which is similar in spirit to Frölich (2007) for the evaluation of the total effect on the compliers (known as local average treatment effect, LATE), the subpopulation whose treatment state reacts on the instrument. Under the imposed assumptions the proposed methods allow disentangling the LATE into the (local) direct and indirect effects on compliers. Our identification strategies consider both continuous and discrete mediators. As special cases, our results also cover the scenarios of a random treatment, which corresponds to a situation with perfect compliance, or of unconditional instrument validity, implying that one

need not control for observed confounders.

The present work appears to be the first one to use (at least) two distinct instruments to tackle the endogeneity of the treatment and the mediator in a nonparametric treatment effect model. In contrast, most of the comparably few IV approaches suggested in the mediation literature use a single instrument and therefore cover less general identification problems. Robins and Greenland (1992) and Geneletti (2007) consider an exogenous treatment and an endogenous mediator for which a 'perfect' instrument is at hand in the sense that it forces the mediator to take a particular (and desired) value. This is equally attractive as directly manipulating the mediator exogenously, see for instance the discussion on perfect manipulation in Imai, Tingley, and Yamamoto (2013). Even though perfect instruments may exist in some clinical trials, they are hard to find in most economic problems including our two applications. Imai, Tingley, and Yamamoto (2013) also discuss nonparametric identification in experiments (again with an exogenous treatment) based on imperfect and discrete instruments for the mediator.[1] Under a very specific experimental design they identify the average indirect effect in the subgroup of individuals whose mediator value reacts on the instrument ("mediator compliers"). In contrast, in our paper we identify the effects on the entire complier population, which is larger than the mediator compliers in Imai, Tingley, and Yamamoto (2013). A further distinction is that we also allow for treatment endogeneity.

Joffe, Small, Have, Brunelli, and Feldman (2008) assume a single instrument for both the treatment and the mediator and discuss identification and estimation in linear models under a particular set of assumptions. However, in a nonparametric framework, a single instrument for both endogeneity problems is generally not sufficient for identification. An exception is Yamamoto (2013), who considers nonparametric identification based on an instrument for the treatment and a latent ignorability assumption similar to Frangakis and Rubin (1999) with respect to the mediator. This allows controlling for the endogeneity of the latter despite the absence of a second

[1] See their Section 4.2 on cross-over encouragement designs or the corresponding discussion in Imai, Keele, Tingley, and Yamamoto (2011). Also Mattei and Mealli (2011) consider a random treatment and a binary instrument for the mediator to derive bounds on direct effects within principal strata defined upon potential mediator states (as a function of the treatment), so called principal strata direct effects.

instrument for the mediator. However, identification generally fails if latent ignorability, i.e. the exogeneity of the mediator conditional on treatment compliance (and possibly further observed covariates), is not satisfied. As an alternative strategy, we therefore base identification on distinct instruments for the treatment and the mediator. Powdthavee, Lekfuangfu, and Wooden (2013) is one of the very rare studies also using two instruments. Considering Australian data, they estimate the indirect effect of education on life satisfaction running via the mediator income by using regional differences in the timing of changes in schooling laws as instrument for education and income shocks (inheritance, severance pay, lottery wins) as instruments for total personal income. However, Powdthavee, Lekfuangfu, and Wooden (2013) consider a fully parametric model with linear equations characterizing the outcome, the mediator, and the treatment, which does not permit treatment-mediator interaction effects and thus, heterogeneity in direct and indirect effects across treatment states. In contrast, the nonparametric identification results of our paper naturally allow for heterogenous direct and indirect effects across treatment states and observed covariates.

We provide a few examples for potential applications where two distinct instruments for the treatment and the mediator may exist, as required by our identification results. In the case of migration, historical migration networks in the origin region, travel costs, and distance to destination countries as pull factors for out-migration of, usually male, household members may serve as instruments for the treatment 'migration' and transaction costs of transferring funds, exchange rate appreciation and changes in labor market conditions in the destination countries as instruments for the mediator 'non-labor income' (including remittances) of the (left-behind) household. When assessing the direct effect of mother's education on child's birth weight as well as its indirect impact via mother's smoking habits, changes in compulsory schooling might be used as instrument for mother's education (treatment), and variation in cigarette taxes as instrument for smoking. In the best case, the first (or even both) instruments stem from randomized assignment (with imperfect compliance). An example in the field of educational interventions is the Project STAR experiment, see for instance Krueger (1999), in which early graders were randomized into

small classes (treatment). To assess its direct and indirect impacts on adult outcomes, one may instrument the mediator 'college degree' by the variation in tuition fees or distance to college, see for instance Card (1995) and Kane and Rouse (1995).

In this paper, we provide empirical illustrations for two further research questions: Firstly, we disentangle the effect of education (treatment) on the health outcome 'social functioning' into a direct component and an indirect impact running via income (mediator). To this end, we instrument education by compulsory schooling laws and income by windfall income (such as lottery wins). While we use similar instruments as in Powdthavee, Lekfuangfu, and Wooden (2013), our data (coming from the British Household Panel Survey), outcome of interest, and methodology differ. Secondly, we analyze experimental data from the U.S. Job Corps program aimed at increasing the human capital of disadvantaged youth. We use randomization into Job Corps as instrument for first year program participation (treatment) to disentangle the earnings effect among female compliers in the third year into an indirect effect mediated by hours worked and a direct effect that likely reflects productivity gains, as it is conditional on working hours. To control for mediator endogeneity, we use the number of children younger than 6 and 15 in the household as instruments for hours worked. For all instruments in either application, we discuss several methods for (partially) testing IV validity.

The remainder of this paper is organized as follows. In Section 2, a nonparametric model for mediation analysis is introduced and the effects of interest are defined: (natural) direct and indirect effects as well as the controlled direct effect. Section 3 discusses different approaches to the identification of these effects based on distinct instruments for the treatment and the mediator. While the treatment and its instrument are always binary (even tough the exposition could be easily extended to non-binary instruments for the treatment), we present various settings with either continuous or discrete mediators and continuous or discrete instruments for the mediator. In Section 4, we analyze the properties of some estimators in a brief simulation study. Section 5 presents two empirical applications: In the first application, we disentangle the health effects of education, which may be mediated via income, based on the identifying assumptions under a

continuous instrument of the mediator (windfall income). In the second application, we separate the direct earnings effect of Job Corps from the indirect one running through hours worked and assume the instrument of the mediator to be discrete (number of children). Section 6 concludes.

2 Model and parameters of interest

2.1 Direct and indirect effects in nonparametric model

We are interested in disentangling the total effect of a binary treatment D on an outcome variable Y into a direct effect and an indirect effect operating through some scalar mediator M.[2] Identification will be based on two instruments Z_1 and Z_2 for the endogenous variables D and M. To this end, we postulate the following structural model consisting of a non-separable nonparametric system of equations characterizing the outcome, the mediator, and the treatment:

$$Y = \varphi(D, M, X, U), \tag{1}$$

$$M = \zeta(D, Z_2, X, V), \tag{2}$$

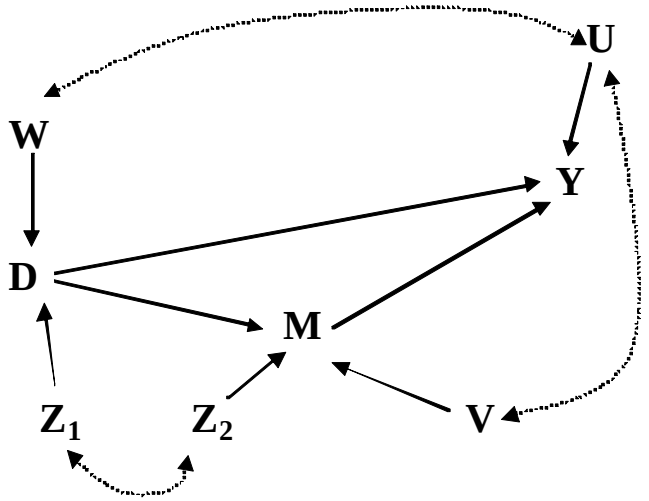
$$D = 1\left(\chi(Z_1, X, W) \geq 0 \right), \tag{3}$$

where φ, ζ, χ are unknown functions. $1(\cdot)$ is the indicator function which is equal to one if its argument is true and zero otherwise. U, V, W comprise unobservables and may be arbitrarily correlated, so that the treatment and the mediator are in general endogenous. X is a (possibly empty) set of observed covariates. The latter are not necessarily always required for identification and for most of the intuition of the main identification channels it is in fact helpful to ignore X, i.e. suppose that it corresponds to the empty set. Yet, some of the assumptions discussed later on may be more plausible after conditioning on observable characteristics. For instance, monotonicity assumptions with respect to V or W or conditional independence assumptions may be more reasonable within subpopulations that have the same values of X. Note that the X variables are

[2] Extensions to vector valued mediators are possible, but would require additional instrumental variables, which might be difficult to find in applications. We thus leave the multivariate mediator extension for future research.

not required to be exogenous. The X variables may be correlated with the unobservables and, they might even be causally affected by D as long as the subsequent assumptions are satisfied (which clearly places some limits on the causal endogeneity of X).

Z_1 is the instrument for treatment D, henceforth denoted as the first instrument. For ease of exposition, we assume Z_1 to be *binary*, albeit the discussion could be extended to multi-valued instruments with bounded support as in Frölich (2007). Z_2 denotes the instrument for mediator M, referred to as the second instrument hereafter, and for most of the paper it is assumed to be *continuous*.



Based on a causal diagram, Figure 1 provides a graphical example of causal relations between observed and unobserved variables that satisfy our structural model and the identifying assumptions, however, omitting any covariates X for ease of exposition.[3] Each one-sided arrow represents a causal impact, two-sided arrows permit causation in either direction. In addition, if X was non-empty, there would be additional arrows from X to D, M, and Y. Further, X could be arbitrarily associated with U, V, W as well as the two instruments.

Identification of the total (local) average treatment effect has been shown in Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). In this paper, we aim at disentangling the total effect into the part which is mediated by M and a remainder which directly affects Y

[3] See for instance Pearl (1995) for a discussion of causal diagrams.

(but could in principle run via further mediators other than M). Two endogeneity problems arise in this context. The first one stems from the permitted association between W and U, even after conditioning on X. The first instrument Z_1 is used to tackle this issue by assuming instrument independence and monotonicity, where we permit that the IV may be valid only conditional on X, as in Abadie (2003) and Frölich (2007). A second issue is that the mediator, which may be discrete or continuous, is confounded by V, which is possibly related to U and W as well. We therefore exploit the second instrument Z_2 to induce variation in M that is independent of variation in D. This requires Z_2 to affect M conditional on X, but to be excluded from the outcome equation. Some form of monotonicity condition is needed for identification and two distinct approaches are considered: We either assume monotonicity of the mediator in the unobservable V, leading to a control function approach, or alternatively, monotonicity of the mediator in the instrument Z_2. While the second approach has the advantage of implying testable restriction, it has the disadvantage of not identifying all parameters of interest.

To ease our discussion of direct and indirect effects, we make use of the potential outcome framework advocated by, among many others, Rubin (1974) and also used in the direct and indirect effects framework for instance by Rubin (2004), Ten Have, Joffe, Lynch, Brown, Maisto, and Beck (2007), and Albert (2008). Let Y^d, M^d denote the potential outcome and the potential mediator state under treatment $d \in \{0,1\}$. We may also express the potential outcome as a function of both the treatment and the potential mediator: $Y^{d,M^{d'}}$. This is useful for the definition of the effects of interest further below. In terms of our model, these parameters are defined as

$$\begin{aligned} M_i^d &\equiv M_i^d \equiv \zeta(d, Z_{2i}, X_i, V_i), \\ Y_i^{d,M^{d'}} &\equiv \varphi(d, M^{d'}, X_i, U_i), \end{aligned}$$

for $d, d' \in \{0,1\}$ and i indexing a particular subject in the population.

Similarly, we define potential treatment states for $z_1 \in \{0,1\}$

$$D_i(z_1) = 1\,(\ \chi(z_1, X_i, W_i) \geq 0\ )\,.$$

As discussed in Angrist, Imbens, and Rubin (1996), the population can be categorized into four subpopulations or types (denoted by T), according to the treatment behavior as a function of the first instrument: The *always takers* ($T_i = at$) take treatment irrespective of Z_1, i.e. $D_i(0) = D_i(1) = 1$. The *never takers* ($T_i = nt$) do not take treatment irrespective of Z_1, i.e. $D_i(0) = D_i(1) = 0$. The *compliers* ($T_i = co$) take treatment only if Z_1 is one, i.e. $D_i(0) = 0, D_i(1) = 1$. Finally, the *defiers* ($T_i = de$) take treatment only if Z_1 is zero, i.e. $D_i(0) = 1, D_i(1) = 0$. We will assume that the last group has probability mass zero, i.e. defiers do not exist. Note that the type T_i is a function of X_i and W_i as it is uniquely determined by $\chi(1, X_i, W_i)$ and $\chi(0, X_i, W_i)$. This further implies that in subpopulations conditional on X, the type is a function of W only.

It would be straightforward to extend the model defined by (1) to (3) to

$$\begin{aligned} Y &= \varphi(D, M, X, U), \\ M &= \zeta(D, Z_2, X, V), \\ D &= 1\left(\chi(Z_1, Z_2, X, W) \geq 0 \right), \end{aligned}$$

so that both instruments entered the treatment equation. This model is more general as it permits the second instrument to also influence treatment choice.[4] The main implication of this extension is that the type T_i is a function of Z_{2i}, X_i and W_i, because the potential treatment states are obtained from $\chi(1, Z_{2i}, X_i, W_i)$ and $\chi(0, Z_{2i}, X_i, W_i)$. There are still three main types (compliers, always- and never takers) and since all subsequent identification approaches only make use of the type identifier but not of the structure of the treatment equation itself, most of the later results would go through for this extended model with few modifications of the assumptions.

[4] This model bears some similarities with the idea of an "included instrument" in D'Haultfoeuille, Hoderlein, and Sasaki (2014), since the instrument Z_2 appears in the choice and in the outcome equation.

On the other hand, a model where

$$\begin{aligned} Y &= \varphi(D, M, X, U), \\ M &= \zeta(D, Z_1, Z_2, X, V), \\ D &= 1\,(\,\chi(Z_1, X, W) \geq 0\,)\,, \end{aligned}$$

i.e. where both instruments Z_1 appear in the mediator equation, is not admissible in this framework. The reason is that M is an (intermediate) outcome variable and identification requires an instrument that shifts D, for a given type, without affecting outcomes (unless we would restrict the outcome equation φ).

After establishing the model, we now define the treatment effects of interest. Our nonparametric IV strategy will allow identifying total, (natural) direct and indirect, as well as controlled direct effects in the subpopulation of compliers. The total average effect among compliers corresponds to the local average treatment effect (LATE), also known as complier average causal effect (CACE):

$$\Delta = E[Y^1 - Y^0 | T = co] = E[Y^{1,M^1} - Y^{0,M^0} | T = co].$$

The (natural) *direct* effect among compliers is given by the mean difference in outcomes when exogenously varying the treatment, but keeping the mediator fixed at its potential value for $D = d$, which shuts down the indirect causal mechanism:

$$\theta(d) = E[Y^{1,M^d} - Y^{0,M^d} | T = co], \qquad \text{for } d \in \{0, 1\}. \tag{4}$$

Analogously, the (natural) *indirect* effect among compliers corresponds to the mean difference in outcomes when exogenously shifting the mediator to its potential values under treatment and non-treatment, but keeping the treatment fixed at $D = d$:

$$\delta(d) = E[Y^{d,M^1} - Y^{d,M^0} | T = co], \qquad \text{for } d \in \{0, 1\}. \tag{5}$$

Because (4) and (5) refer to the compliers alone, they are local versions of the natural or pure direct and indirect effects discussed in Robins and Greenland (1992), Pearl (2001), and Robins (2003), respectively. For convenience, we will simply call them direct and indirect effects in the subsequent discussion.

It is worth noting that the LATE is the sum of the direct and indirect effects defined upon opposite treatment states:

$$\begin{aligned}\Delta &= E[Y^{1,M^1} - Y^{0,M^0}|T = co] \\ &= E[Y^{1,M^1} - Y^{0,M^1}|T = co] + E[Y^{0,M^1} - Y^{0,M^0}|T = co] = \theta(1) + \delta(0) \\ &= E[Y^{1,M^0} - Y^{0,M^0}|T = co] + E[Y^{1,M^1} - Y^{1,M^0}|T = co] = \theta(0) + \delta(1).\end{aligned}$$

The notation $\theta(1)$, $\theta(0)$, $\delta(1)$, $\delta(0)$ makes explicit that direct and indirect effects may be heterogenous with respect to the treatment state, which permits interaction effects between the treatment and the mediator.

Finally, the *controlled* direct effect is the mean difference in compliers' outcomes when exogenously varying the treatment, but (exogenously) setting the mediator to a particular value, say m, rather than the potential mediator state[5]

$$\gamma(m) = E[Y^{1,m} - Y^{0,m}|T = co], \qquad \text{for } d \in \{0, 1\}.$$

That is, contrary to the (natural) direct effect, which is the direct impact conditional on the mediator state that would 'naturally' occur as a reaction to a particular treatment, the controlled direct effect is obtained by forcing the mediator to take a particular value.

Which of these parameters is of primary interest depends on the research question at hand. Suppose we would like to assess the effectiveness of the first program in a sequence of two labor market programs (e.g., a job application training followed by a computer course) aimed at in-

[5] Note that there is no equivalent to the controlled direct effect in terms of indirect effects, see the discussion in Pearl (2001).

creasing employment and earnings. The natural direct effect then assesses the effectiveness of the first program (D) conditional on the participation in the second one (M) that would in the current state of the world follow from participation or (non-)participation in the first program. This may be interesting for assessing the (relative or absolute) effectiveness of the first program under status quo program assignment rules to the second program. However, if these rules can in principle be changed, then also evaluating whether the first program is effective conditional on enforcing (non-)participation in the second one appears interesting in order to optimally design program sequences. Therefore, the controlled direct effect may provide additional policy guidance whenever mediators can be prescribed. In contrast, if prescription is unrealistic, the natural direct effect, which relies on status quo mediator response to treatment, appears to be the only parameter worth considering. We refer to Pearl (2001) for further discussion of what he calls the 'descriptive' and 'prescriptive' natures of natural and conditional effects.

Without further assumptions, neither of these parameters is identified for several reasons. Firstly, only one potential outcome out of Y^{1,M^1} and Y^{0,M^0} is known for any observation. Secondly, Y^{0,M^1} and Y^{1,M^0} are never observable and therefore inherently counterfactual, as a person cannot be treated and non-treated at the same time. Thirdly, the type of any observation cannot be uniquely determined. Therefore, identification of direct and indirect effects hinges on the generation of exogenous variation in the treatment and mediator, in our case based on instrumental variables.

2.2 Potential outcomes, quantile treatment effects and effects on inequality

The following sections focus on the identification of the means of Y^{1,M^d}, Y^{0,M^d}, $Y^{1,m}$, and $Y^{0,m}$, rather than the mean effects $\theta(d)$, $\delta(d)$, and $\gamma(m)$ only, as the mean potential outcomes provide richer information than their differences alone. In particular, one may use the formulae derived in the following theorems to estimate not only average treatment effects, but also distributional effects. To this end, note that the cumulative distribution function of Y^{d',M^d} (with $d, d' \in \{0,1\}$),

denoted by $F_{Y^{d'},M^d}$, can be written as

$$F_{Y^{d'},M^d|T=co}(a) = E\left[1\left(Y^{d',M^d} \leq a\right)|T = co\right]. \tag{6}$$

Even though the subsequent theorems provide identification results for $E\left[Y^{d',M^d}|T = co\right]$, using (6) we immediately obtain the results for $F_{Y^{d'},M^d}$ among compliers by replacing Y with $1\left(Y \leq a\right)$ in those theorems throughout. This allows estimating the distribution functions $F_{Y^{1,M^1}}$, $F_{Y^{1,M^0}}$, $F_{Y^{0,M^1}}$, and $F_{Y^{0,M^0}}$ as well as $F_{Y^{1,m}}$ and $F_{Y^{0,m}}$ for compliers. One may also obtain quantile treatment effects through the inversion of the distribution functions[6] or calculate other inequality measures such as the Gini coefficient, Theil index, etc. These distributional estimates permit assessing e.g. whether direct and indirect effects have offsetting effects on inequality or whether both move in the same direction as well as the relative importance of direct and indirect effects for inequality. Similarly, in applications where the lower tail of Y is of particular interest, e.g. low birth weights in public health, poverty rates or poor school performance, one might be interested in (separating direct and indirect) quantile treatment effects.[7]

3 Identifying direct and indirect effects

In this section we discuss the identification of (natural) direct and indirect effects under endogeneity using instrumental variables. We permit that the instruments are only valid conditional on observed covariates X, which may themselves be endogenous. Our first assumption requires the instruments to be independent of the unobservables U, V, W conditional on X. Such assumptions are rather standard in the literature on the LATE requiring an instrument for the treatment only, see e.g. Imbens and Angrist (1994) or Angrist, Imbens, and Rubin (1996) for unconditional IV independence and Abadie (2003) or Frölich (2007) for IV independence given X. Here, conditional independence needs to hold for both instruments Z_1 and Z_2. For ease of exposition, As-

[6] One could also adopt the approaches of e.g. Frandsen, Frölich, and Melly (2012) and Frölich and Melly (2013) to obtain direct estimators of the quantile treatment effects.

[7] Note that in contrast to mean effects, direct and indirect quantile treatment effects do not add up to the total quantile treatment effect.

sumption 1 is slightly stronger than needed for the various lemmas and theorems to follow. We express the independence condition with respect to type T and not with respect to the unobservable W, as we later only require independence within the types and not for each value of W.

Assumption 1: IV independence

$$(Z_1, Z_2) \perp\!\!\!\perp (U, V) | T, X,$$
$$Z_1 \perp\!\!\!\perp (U, V, T) | Z_2, X,$$

where the symbol $\perp\!\!\!\perp$ denotes statistical independence.

It is worth noting that Assumption 1 would be implied e.g. by the following stronger assumption:

$$(Z_1, Z_2) \perp\!\!\!\perp (U, V, W) | X. \tag{7}$$

The main difference is that Assumption 1 permits Z_2 and W to be dependent, whereas (7) does not. As W determines the type, i.e. whether someone is a complier, always taker, or never taker, permitting dependence between Z_2 and W could be relevant in applications where Z_2 is not fully randomly assigned but possibly dependent on treatment choice. Assumption 1 also allows for an association between Z_1 and W, as long as the dependence vanishes when conditioning on Z_2.

The stronger assumption (7) is not required for the results of our paper. If it were nevertheless imposed it would imply that the probability of being a complier did not depend on Z_2. This condition is testable, because $\Pr(T = co|Z_2, X)$, the proportion of compliers given Z_2 and X, is identified further below. It would further imply $Z_2 \perp\!\!\!\perp D | X, Z_1$. Hence, in applications where both assumptions appear equally plausible, this may be used to construct partial tests for identification.

Note that we permit X to be endogenous, i.e. associated with any of the unobservables, because Assumption 1 only needs to hold conditional on X. Our assumption implies that the

first instrument is conditionally independent of the potential treatment states $D(1), D(0)$ and does not have a direct association with the mediator or the outcome through V or U. Z_1 could for instance represent the assignment indicator in a randomized experiment assessing a training program or any other policy intervention. If randomization is successful, the potential treatment states are independent of Z_1. Furthermore, if it is credible that random assignment itself does not directly affect the mediator, the outcome, or associated unobserved factors (other than through the treatment), then independence of Z_1 and V, U is satisfied. Generally and in particular in observational studies, Assumption 1 may be more plausible once we control for covariates X.

In addition, for some of our results we require the two instruments Z_1 and Z_2 to be independent of each other, again possibly only after conditioning on X as postulated in Assumption 2.

Assumption 2: Conditional independence of Z_1 and Z_2

$$Z_1 \perp\!\!\!\perp Z_2 | X.$$

We note that Assumption 1 and Assumption 2 jointly imply[8]

$$Z_1 \perp\!\!\!\perp (Z_2, U, V, T) | X. \tag{8}$$

Note that while Z_1 has to be independent of W, dependence between Z_2 and W is still permitted.

Assumption 2 is needed for some, but not all identification results and is often satisfied by construction or through a transformation of the instrument. In randomized trials, it holds by construction if both instruments are independently randomized. If only Z_1 is under the control of (i.e. randomized by) the experimenter, it is also satisfied if Z_2 is assigned at the same time as or shortly prior to Z_1, because in experiments, any pre-randomization variable is independent of the randomization indicator Z_1. In observational studies, one needs to be more circumspect.

[8] This follows because $A \perp\!\!\!\perp B | C$ and $A \perp\!\!\!\perp C$ together imply $A \perp\!\!\!\perp (B, C)$.To see this note that $f_{ABC} = f_{B|AC} \cdot f_{A|C} \cdot f_C$ and entering the independence assumptions this equals $f_{ABC} = f_{B|C} \cdot f_A \cdot f_C = f_A \cdot f_{B,C}$. Hence, we thus have shown that $f_{ABC} = f_A \cdot f_{B,C}$ or equivalently: $A \perp\!\!\!\perp (B, C)$.

Note first that Z_1 and Z_2 need to be independent only conditional on X, which is testable. We can, for instance, always write the linear projection $Z_2 = \alpha Z_1 + \beta X + \varepsilon$, where the projection error ε is orthogonal to Z_1 and X by construction, and test whether $\alpha = 0$ and ε is independent of Z_1. Even if Z_1 and Z_2 are not (conditionally) independent, we may attain independence via a transformation of Z_2. Suppose Z_2 is continuously distributed with a *strictly* increasing cumulative distribution function (cdf). Define

$$\tilde{Z}_{2i} = \Phi^{-1}\left(F_{Z_2|Z_1,X}\left(Z_{2i}, Z_{1i}, X_i\right)\right), \tag{9}$$

where Φ is the cdf of the standard normal distribution and Φ^{-1} its quantile function.[9] Note that $\tilde{Z}_2|Z_1, X$ is standard normal with mean 0 and variance 1 and thus *independent* of Z_1 (and also of X). We can thus use $\tilde{Z}_2$ instead of Z_2 as the second instrument throughout and so that Assumption 2 is satisfied. (Assumption 1 also holds true for $(Z_1, \tilde{Z}_2)$ if it is satisfied for the original instruments (Z_1, Z_2).) Hence, Assumption 2 is more a normalization rather than a substantial restriction, although in practice $F_{Z_2|Z_1,X}$ has to be estimated for constructing $\tilde{Z}_{2i}$ via (9), which may complicate the estimation process.

In addition to the independence assumptions, identification requires particular monotonicity assumptions. Assumption 3 imposes monotonicity of the treatment in its instrument, which rules out the existence of defiers, as it is standard in the LATE framework:

Assumption 3: Weak monotonicity of treatment choice

i) Monotonicity

$$D(1) \geq D(0) \qquad \text{with probability 1} \tag{10}$$

ii) Existence of compliers

$$E(D|Z_1 = 1) > E(D|Z_1 = 0). \tag{11}$$

[9] As an alternative to the normal, one could also use the uniform distribution or any other continuous distribution function.

Assumption 3 imposes weak positive monotonicity of D in Z_1 in the entire population, see Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). Part (i) rules out the existence of defiers and is equivalent to assuming

$$\Pr\left(1\left(\chi(1,X,W) \geq 0 \right) \;\geq\; 1\left(\chi(0,X,W) \geq 0 \right) \right) = 1.$$

Part (ii) requires that at least some units in the total population are compliers and is directly testable in the data. In contrast, part (i) is mostly untestable (although see Huber and Mellace (2014) and Kitagawa (2008) for recent methods jointly testing monotonicity *and* IV independence). While Assumption 3 permits the existence of always takers, never takers, and compliers, causal parameters are only identified for the compliers under our assumptions.[10]

Assumptions 1 and 3 permit us to identify the fraction of compliers. To ease notational burden, we will make use of the following expressions for the conditional instrument probabilities (or propensity scores) throughout:

$$\begin{aligned} \Pi &= \pi(X) = \Pr(Z_1 = 1|X) \\ \bar{\Pi} &= \bar{\pi}(Z_2, X) = \Pr(Z_1 = 1|Z_2, X). \end{aligned} \tag{12}$$

Under Assumptions 1 and 3, which implies $Z_1 \perp\!\!\!\perp T|Z_2, X$, the probability mass of compliers is identified as

$$\Pr\left(T = co\right) = E\left[\frac{D}{\bar{\Pi}}\frac{Z_1 - \bar{\Pi}}{1 - \bar{\Pi}}\right]. \tag{13}$$

If in addition $Z_1 \perp\!\!\!\perp T|X$, which would for instance be implied by assumption (7), the fraction of

[10] Alternatively, one could also invoke weakly negative monotonicity (allowing for defiers, but ruling out compliers). As both cases are symmetric, we only consider weakly positive monotonicity in the remainder of the paper.

compliers is also obtained as

$$\Pr\left(T = co\right) = E\left[\frac{D}{\Pi}\frac{Z_1 - \Pi}{1 - \Pi}\right].$$

3.1 Natural effects with continuous mediator

We first consider the case of a continuous mediator M. The identification of natural effects requires separating hypothetical changes in treatment D from those in M. We therefore make use of a control function approach that allows shifting D exogenously, i.e. independent from movements in the density of the mediator.[11] The next assumption restricts the mediator to be monotonic in the unobserved term V which is assumed to be a continuous scalar with a strictly increasing cumulative distribution function (cdf).

Assumption 4: Monotonicity of mediator (control function restriction)
(i) V is a continuously distributed random variable with a cdf $F_{V|X=x,T=co}(v)$ that is strictly increasing in the support of V, for almost all values of x,
(ii) $\zeta(d, z_2, x, v)$ is strictly monotonic in v for almost all d, z_2, x. (We normalize ζ to be increasing in v.)

Assumption 4 is crucial for our control function approach. It restricts V to be a scalar random variable, which appears to be quite strong. However, V may also reflect an index function determined by several unobserved variables, which somewhat eases the severity of this restriction. Then, Assumptions 4 (i) and (ii) need to hold with respect to the index (rather than each of its determinants). Invoking strict monotonicity of the mediator in V allows pinning down the distribution function of V given X among compliers by means of the conditional distribution of

[11] See Ahn and Powell (1993), Newey, Powell, and Vella (1999), Blundell and Powell (2003), Das, Newey, and Vella (2003) and Imbens and Newey (2009) for prominent applications of control functions in semi- and nonparametric models.

M given D, Z_2, X among compliers. To this end, we define the control function

$$C_i = C\left(M_i, D_i, Z_{2i}, X_i\right), \tag{14}$$

with

$$C\left(m, d, z_2, x\right) = \frac{E\left[(d + D - 1) \cdot \left(Z_1 - \bar{\pi}\left(z_2, x\right)\right) \ | M \leq m, Z_2 = z_2, X = x\right]}{E\left[D \cdot \left(Z_1 - \bar{\pi}\left(z_2, x\right)\right) | Z_2 = z_2, X = x\right]} \cdot F_{M|Z_2,X}\left(m, z_2, x\right). \tag{15}$$

Control function C identifies V_i as shown in the following lemma. (Note that in settings where we also impose Assumption 2, i.e. that $Z_1 \perp\!\!\!\perp Z_2 | X$, we have that $\bar{\pi}(Z_2, X) = \pi(X)$ throughout.)

Lemma 1: Under Assumptions 1, 3, and 4 it follows that

a)

$$C_i = F_{M|D,Z_2,X,T=co}\left(M_i, D_i, Z_{2i}, X_i\right) = F_{V|X=X_i,T=co}\left(V_i\right), \tag{16}$$

b)

$$V_i = F^{-1}_{V|X=X_i,T=co}\left(C_i\right), \tag{17}$$

c)

$$M \perp\!\!\!\perp U | C, X, T = co. \tag{18}$$

Part (a) of Lemma 1 shows that the control function corresponds to the distribution function of V conditional on X. Part (b) shows that C_i is a one-to-one mapping of V_i. I.e., conditional on X, V is a one-to-one function of C, and V is thus identified. That means conditioning on C is equivalent to conditioning on V, as long as we control for X throughout, which we always do. Finally, part (c) shows that once we control for C (in addition to X) we can separate the mediator from the unobservable U in the outcome equation, within the complier subpopulation. (In fact, the latter also holds in the always taker and never taker subpopulations, but since we cannot identify all mean potential outcomes for these latter groups, we will only focus on the

compliers alone.)

Before giving the formal results in Theorem 1, we discuss the intuition of the identification approach, which also illustrates the common support assumption needed. The key idea underlying identification is to (hypothetically) vary Z_1 in order to change treatment D, while at the same time keeping M unchanged through a variation of Z_2 that undoes the effect of Z_1 on M. To this end, we need to condition on unobservable V, which is replaced by its control function C. Assume we are interested in the mean potential outcome $E\left[Y^{1,M^0}|T=co\right]$, which can be expressed as

$$\begin{aligned} E\left[Y^{1,M^0}|T=co\right] &= \int \varphi(1,M^0,X,U)\cdot dF_{M^0,X,U,C|T=co} \\ &= \int \varphi(1,M^0,X,U)\cdot dF_{M^0,U|X,C,T=co}\cdot dF_{X,C|T=co} \\ &= \int \varphi(1,M^0,X,U) dF_{U|X,C,T=co}\cdot dF_{M^0|X,C,T=co}\cdot dF_{X,C|T=co}, \end{aligned}$$

where the last equation follows as M^0 is independent of U conditional on X and C by Assumption 1 and Lemma 1. For being able to identify the distribution of M^0, $M^0 \perp\!\!\!\perp Z_1|X,C,T=co$ needs to hold, which is implied by $Z_2 \perp\!\!\!\perp Z_1|X,V,T$. It follows that $F_{M^0|X,C,T=co} = F_{M|Z_1=0,X,C,T=co}$ and thus

$$= \int \varphi(1,M,X,U) dF_{U|X,C,T=co}\cdot dF_{M|Z_1=0,X,C,T=co}\cdot dF_{X,C|T=co}.$$

As $dF_{M|Z_1,X,C,T=co}$ is identifiable (see the appendix), we may multiply and divide by $dF_{M|Z_1=1,X,C,T=co}$ to obtain

$$\begin{aligned} &= \int \varphi(1,M,X,U) dF_{U|X,C,T=co}\cdot dF_{M|Z_1=0,X,C,T=co}\frac{dF_{M|Z_1=1,X,C,T=co}}{dF_{M|Z_1=1,X,C,T=co}}\cdot dF_{X,C|T=co} \\ &= \int \left\{\varphi(1,M,X,U) dF_{U|X,C,T=co}\right\}\cdot \omega(M,X,C)\cdot dF_{M|Z_1=1,X,C,T=co}\cdot dF_{X,C|T=co}, \end{aligned} \tag{19}$$

where $\omega(M,X,C) = \frac{dF_{M|Z_1=0,X,C,T=co}}{dF_{M|Z_1=1,X,C,T=co}}$. Using $U \perp\!\!\!\perp (M,Z_1)|X,C,T=co$ by Assumption 1 and

Lemma 1 we obtain

$$
\begin{aligned}
&= \int \left\{\varphi(1, M, X, U) dF_{U|M,X,C,Z_1=1,T=co}\right\} \cdot \omega(M, X, C) \cdot dF_{M|X,C,Z_1=1,T=co} \cdot dF_{X,C|T=co} \\
&= E\left[YDZ_1 \cdot \frac{\omega(M, X, C)}{\Pr\left(Z_1 = 1|X\right)} | T = co\right]. \qquad (20)
\end{aligned}
$$

The last expression indicates that the counterfactual outcome can be identified based on observable variables in the complier subpopulation. However, since the compliers are unknown, we require an expression for the entire population that is equal to zero in the always and never taker subpopulations. As formally shown in the appendix, the following expression

$$
E\left[\left(\frac{YDZ_1}{\Pr\left(Z_1 = 1|X\right)} - \frac{YD(1 - Z_1)}{\Pr\left(Z_1 = 0|X\right)}\right) \cdot \omega(M, X, C)\right] \qquad (21)
$$

turns out to be zero within the always and never taker subpopulations and therefore equals (20) multiplied with the share of compliers $\Pr\left(T = co\right)$, with the latter being identified by Assumption 1. Hence, by estimating (21) and dividing by $\Pr\left(T = co\right)$, we obtain (20), which thus gives $E\left[Y^{1,M^0} | T = co\right]$.

Note that from equation (19), one can see the support condition that needs to be satisfied for identification. Namely, it must hold that $dF_{M|Z_1=1,X,C,T=co} > 0$ at every m where $dF_{M|Z_1=0,X,C,T=co} > 0$ or in other words, that

$$
Supp\left(M|Z_1 = 0, X, C, T = co\right) \subseteq Supp\left(M|Z_1 = 1, X, C, T = co\right).
$$

On the other hand, for the identification of $E\left[Y^{0,M^1} | T = co\right]$ we would need (by symmetric derivations) that

$$
Supp\left(M|Z_1 = 0, X, C, T = co\right) \supseteq Supp\left(M|Z_1 = 1, X, C, T = co\right).
$$

In order to summarize all identification results succinctly into Theorem 1, we impose both support

conditions jointly, i.e. that

$$Supp\left(M|Z_1 = 0, X, C, T = co\right) = Supp\left(M|Z_1 = 1, X, C, T = co\right).$$

This condition can generally only be satisfied if Z_2 is a continuously distributed variable. In Section 3.5 we discuss the case when both Z_1 and Z_2 are discrete instrumental variables based on somewhat different identifying assumptions.

An alternative way of expressing this support condition is that no values of M given C and X perfectly predict the value of the first instrument. That is, conditional on C, X, the mediator state must not be a deterministic function of the first instrument, otherwise identification is infeasible due to the lack of comparable units in terms of the mediator across values of the first instrument. We summarize this support condition as

$$0 < \Pr\left(Z_1 = 1|M, C, X, T = co\right) < 1 \qquad a.s.$$

Because of the unique mapping between C and V as established in Lemma 1, we may write this restriction equivalently in the following way:

Assumption 5: Common support of M

$$0 < \Pr\left(Z_1 = 1|M, V, X, T = co\right) < 1 \qquad a.s. \tag{22}$$

Assumption 5 is equivalent to requiring the weights $\omega(M, X, C)$ to be neither zero nor infinity. The weights are formally defined below and need to be estimated in applied work. If some of these weights are close to zero or extremely large, this could indicate that the above support condition is not satisfied. In such situations, one may redefine the objects of interest on subsets of the support spaces of M, X, C for which common support holds. An implication of Assumption

5 is $0 < \Pr(Z_1 = 1|X = x) < 1$ for all x in the support of X, as invoked for example in Frölich (2007). Therefore, no values of the covariates may perfectly predict the first instrument, because otherwise observations that are comparable in X would not always exist across $Z_1 = 1$ and $Z_1 = 0$.

Under Assumptions 1 to 5, the potential outcomes and thus also $\theta(d)$ and $\delta(d)$ are identified based on exogenous variation in D and M generated by Z_1 and Z_2 conditional on X.

Theorem 1: Under Assumptions 1 to 5 the potential outcomes are identified as

$$\begin{aligned}
E\left[Y^{1,M^0}|T=co\right] &= E\left[YD\Omega\frac{Z_1-\Pi}{\Pi(1-\Pi)}\right]\frac{1}{\Pr(T=co)},\\
E\left[Y^{1,M^1}|T=co\right] &= E\left[YD\frac{Z_1-\Pi}{\Pi(1-\Pi)}\right]\frac{1}{\Pr(T=co)},\\
E\left[Y^{0,M^1}|T=co\right] &= E\left[\frac{Y(D-1)}{\Omega}\frac{Z_1-\Pi}{\Pi(1-\Pi)}\right]\frac{1}{\Pr(T=co)},\\
E\left[Y^{0,M^0}|T=co\right] &= E\left[Y(D-1)\frac{Z_1-\Pi}{\Pi(1-\Pi)}\right]\frac{1}{\Pr(T=co)},
\end{aligned}$$

where

$$\begin{aligned}
\Omega = \omega(M,C,X) \equiv 1 - \frac{E[Z_1|M,C,X]-\pi(X)}{E[DZ_1|M,C,X]-E[D|M,C,X]\cdot\pi(X)}\\
= \frac{E[(D-1)(Z_1-\pi(X))|M,C,X]}{E[D(Z_1-\pi(X))|M,C,X]}\\
= \frac{E[\pi(X)-Z_1|M,C,X,D=0]\Pr(D=0|M,C,X)}{E[Z_1-\pi(X)|M,C,X,D=1]\Pr(D=1|M,C,X)}
\end{aligned}$$

and $\Pi = \pi(X)$ with $\pi(x) = \Pr(Z_1 = 1|X = x) = E[Z_1|X = x]$. C is identified by Lemma 1 and $\Pr(T = co)$ is identified by (13). The proof is provided in the appendix.

Two remarks are worth noting concerning this identification result. First, the identification of direct and indirect effects hinges on identical assumptions. This also follows from the fact that the direct (indirect) effect on the compliers corresponds to the difference between the total and the indirect (direct) effect defined upon opposite treatment states. Second, note that perfect

treatment compliance with the random assignment Z_1 can be regarded as a special case of the framework underlying Theorem 1. If all individuals comply with their treatment assignment, then $Z_1 = D$ and $\Pr(T = co) = 1$. In this case, the formulae of Theorem 1 simplify by replacing Z_1 with D everywhere, and they represent the average direct and indirect effects on the total population (as everyone is a complier if $\Pr(T = co) = 1$).

3.2 Controlled direct effects with continuous mediator

3.2.1 Control function approach

This section discusses the identification of the controlled direct effect $\gamma_{co}(m)$ for the mediator fixed at m (rather than at its potential value under a particular treatment). In contrast to the natural direct effect, the identification of controlled direct effect does not require knowledge of the distribution of M^d, which allows weakening the independence assumptions. In particular, Assumption 2 is no longer needed so that dependence between Z_2 and Z_1 is permitted, even conditional on X. As before, we focus on the identification of the mean potential outcomes (rather than the effect), from which $\gamma_{co}(m)$ is obtained as their difference. The be concise, we discuss the identification of $E\left[Y^{1,m}|T = co\right]$ alone, while $E\left[Y^{0,m}|T = co\right]$ can be obtained by symmetric arguments.

We present two different approaches for identification. Theorem 2 follows a control function approach and exploits monotonicity of the mediator in V. Alternatively, Theorem 3 does not require monotonicity in V, but instead imposes monotonicity in the instrumental variable Z_2. Both approaches permit Z_1 and Z_2 to be dependent. However, as shown in Theorem 2, the identification expressions are simpler if Z_1 and Z_2 happen to be independent (conditional on X). We thus provide the more general result (without independence) in Theorem 2a, while Theorem 2b provides the simpler expressions when additionally invoking Assumption 2. Before presenting the formal results, we provide some intuition for identification.

Our mean potential outcome of interest, $E\left[Y^{1,m}|T = co\right]$, can also be expressed as

$E[\varphi(1,m,X,U)|T=co]$. Pursuing a control function approach, we can express this parameter as

$$\begin{aligned} E\left[Y^{1,m}|T=co\right] &= \int \varphi(1,m,X,U)\cdot dF_{X,U,C|T=co} \\ &= \int \varphi(1,m,X,U)\cdot dF_{U|X,C,T=co}\cdot dF_{X,C|T=co} \\ &= \int \varphi(1,m,X,U)\cdot dF_{U|M=m,Z_1=1,X,C,T=co}\cdot dF_{X,C|T=co} \end{aligned}$$

because $U \perp\!\!\!\perp (Z_1,Z_2)|X,V,T=co$

$$= \int E[Y|M=m,Z_1=1,X,C,T=co]\cdot dF_{X,C|T=co}. \tag{23}$$

Finally, estimable expressions for $E[Y|M,Z_1,X,C,T=co]$ and $dF_{X,C|T=co}$ based on observable variables can be obtained as outlined in the appendix.

For the previous derivations, we require the support condition

$$Supp\,(X,C|T=co) \subseteq Supp\,(X,C|M=m,Z_1=1,T=co)$$

or equivalently, that the conditional mediator density $f_{M|X,C,Z_1=1,T=co}(m,x,c) > 0$ at every value x,c where $f_{X,C|T=co}(x,c)$ is positive. Given the one-to-one relationship between C and V by Lemma 1, we can also express this as requiring $f_{M|X,V,Z_1=1,T=co}(m,x,v) > 0$ at every value x,v where $f_{X,V|T=co}(x,v)$ is positive. In other words, $f_{M|X,V,Z_1=1,T=co}(m,X,V)$ must be positive almost everywhere. This is summarized in the following support condition: $f_{M|X,C,Z_1=1,T=co}(m) > 0 \quad a.s.$. Equivalently in terms of V, the following has to hold:

Assumption 6: Common support

$$f_{M|X,V,Z_1=1,T=co}(m) > 0 \qquad a.s. \tag{24}$$

In terms of model (2) postulating $M = \zeta(D, Z_2, X, V)$, this assumption requires that for every x, v in the support of X, V in the subpopulation of compliers, there exists (at least) one value z_2, which has positive density, such that $\zeta(1, z_2, x, v) = m$. As Assumption 6 can equivalently be written as $f_{M|X,C,Z_1=1,T=co}(m) > 0$ *a.s.*, it is testable. Under this support condition, the mean potential outcomes are identified.

Theorem 2a: Under Assumptions 1, 3, 4, and 6

$$E\left[Y^{1,m}|T = co\right] = \frac{1}{\Pr\left(T = co\right)} \int E\left[YD\frac{Z_1 - \bar{\Pi}}{1 - \bar{\Pi}}\Omega|X, M = m\right] \cdot dF_X \tag{25}$$

with weights

$$\Omega = \omega(C, X) = f_{M|X}(m)\frac{E\left[\frac{D}{\bar{\Pi}}\frac{Z_1 - \bar{\Pi}}{1 - \bar{\Pi}}|C, X\right]}{\frac{\partial}{\partial m}E\left[1\left(M \leq m\right) \cdot D\frac{Z_1 - \bar{\Pi}}{1 - \bar{\Pi}}|C, X\right]}, \tag{26}$$

where $\Pi = \pi(X) = \Pr(Z_1 = 1|X)$ and $\bar{\Pi} = \bar{\pi}(Z_2, X) = \Pr(Z_1 = 1|Z_2, X)$.

Theorem 2b: Under Assumptions 1, 2, 3, 4 and 6

$$E\left[Y^{1,m}|T = co\right] = \frac{1}{\Pr\left(T = co\right)} \int E\left[YD\frac{Z_1 - \Pi}{1 - \Pi}|C, X, M = m\right] \cdot \Omega \cdot dF_{C,X} \tag{27}$$

with weights

$$\Omega = \omega(C, X) = \frac{1}{\Pi}\frac{E\left[D\left(Z_1 - \Pi\right)|C, X\right]}{E\left[D\left(Z_1 - \Pi\right)|M = m, C, X\right]}. \tag{28}$$

Theorem 2 provides the identification results with and without Assumption 2. While Theorem 2a, which does *not* assume independence of Z_1 and Z_2, is more general, Theorem 2b provides a simpler expression. In particular, we obtain an explicit expression for the complier density when imposing Assumption 2, namely

$$f_{M|Z_1=1,X,C,T=co} = \frac{E\left[D\left(Z_1 - \Pi\right)|M, X, C\right]}{E\left[D\left(Z_1 - \Pi\right)|X, C\right]} \cdot f_{M|X,C}. \tag{29}$$

This is not possible for Theorem 2a, where one would need to condition on Z_2 and C, which would entail a degenerate distribution function. The weights in (28) may therefore also be expressed as

$$\Omega = \frac{1}{\Pi} \frac{f_{M|X,C}(m)}{f_{M|Z_1=1,X,C,T=co}(m)},$$

which provides the proper re-weighting approach to transform (27) into (23). Density expression (29) also links up with Assumption 6 in that $f_{M|X,C}(m)$ needs to be positive for almost all values of X and C (unless $E[D\,(Z_1 - \Pi)\,|M, X, C] = 0$).

3.2.2 Identification via instrument monotonicity

Instead of the control function approach, we consider an alternative identification strategy assuming monotonicity of the mediator in Z_2. We therefore drop Assumption 4 and impose Assumption 7 instead.

Assumption 7: Monotonicity of the mediator in the instrument

$\zeta(d, z_2, x, v)$ is strictly in z_2 for almost all d, x, v. We normalize ζ to be monotonically *increasing*.

Whether Assumption 4 or Assumption 7 is more plausible depends on the particular application. If the mediator is, for instance, household income and Z_2 represents unexpected windfall income (e.g. lottery wins, inheritances), assuming monotonicity with respect to Z_2 would appear very natural, whereas monotonicity with respect to V might be more debatable. Furthermore, Assumption 7 has the advantage of implying testable implications, whereas Assumption 4 is not testable. Finally, Assumption 7 permits the unobservable heterogeneity V to be multidimensional, whereas Assumption 4 requires it to be one-dimensional, see e.g. Kasy (2014) for a discussion.

With ζ strictly monotonic in z_2, the mediator equation $M = \zeta(D, Z_2, X, V)$ may be inverted

to obtain

$$Z_2 = \zeta^{-1}(D, M, X, V),$$

where ζ^{-1} is now the inverse function with respect to the second argument. Note that this is a different inverse function than in the previous section, where it referred to the fourth argument. To minimize the number of symbols, we, however, use the same notation here.

To see how Assumption 7 (along with several previous assumptions) entails identification, define the random variable Q as

$$Q \equiv \zeta^{-1}(1, m, X, V), \tag{30}$$

which is a stochastic function of the two random variables X and V. Hence, the distribution of Q is governed by the distributions of X and V. It follows that conditional on X, the only stochastic component in Q is V. We use this fact in the following expression for the mean potential outcome:

$$\begin{aligned} E\left[Y^{1,m}|T = co\right] &= \int \varphi(1, m, X, U) \cdot dF_{X,U,Q|T=co} \\ &= \int\int \varphi(1, m, X, U) \cdot dF_{U|Q,X,T=co} \cdot dF_{Q|X,T=co} \cdot dF_{X|T=co}. \end{aligned} \tag{31}$$

Let us now examine the terms in the second line of (31). Starting with $dF_{U|Q,X,T=co}$, as conditional on X the only stochastic element in Q is V and since $(U, V) \perp\!\!\!\perp (Z_1, Z_2)|X, T = co$, we have

$$dF_{U|Q,X,T=co} = dF_{U|Q,Z_1,Z_2,X,T=co}.$$

Now consider

$$\begin{aligned}
& F_{U|Q,Z_1,Z_2,X,T=co}(u,q,1,q,x) \\
= \;& \Pr\left(U \leq u | Q=q, Z_1=1, Z_2=q, X=x, T=co\right) \\
= \;& \Pr\left(U \leq u | \zeta^{-1}(1,m,X,V)=q, Z_1=1, Z_2=q, X=x, T=co\right) \\
= \;& \Pr\left(U \leq u | m=\zeta(1,q,X,V), Z_1=1, Z_2=q, X=x, T=co\right) \\
= \;& \Pr\left(U \leq u | m=\zeta(D,Z_2,X,V), Z_1=1, Z_2=q, X=x, T=co\right) \\
= \;& \Pr\left(U \leq u | m=M, Z_1=1, Z_2=q, X=x, T=co\right) \\
= \;& F_{U|M=m,Z_1=1,Z_2=q,X=x,T=co}(u).
\end{aligned}$$

Secondly, concerning $dF_{Q|X,T=co}$, note that conditional on X, the only stochastic component in Q is V. Because V is independent of Z_1, Z_2 conditional on X, so is Q. It follows that

$$dF_{Q|X,T=co} = dF_{Q|Z_1,Z_2,X,T=co}.$$

Now consider

$$\begin{aligned}
F_{Q|X,T=co}(q,x) &= F_{Q|Z_1,Z_2,X,T=co}(q,1,q,x) \\
&= \Pr\left(Q \leq q | Z_1=1, Z_2=q, X=x, T=co\right) \\
&= \Pr\left(\zeta^{-1}(1,m,X,V) \leq q | Z_1=1, Z_2=q, X=x, T=co\right) \\
&= \Pr\left(m \leq \zeta(1,q,X,V) | Z_1=1, Z_2=q, X=x, T=co\right) \\
&= \Pr\left(m \leq \zeta(D,Z_2,X,V) | Z_1=1, Z_2=q, X=x, T=co\right) \\
&= \Pr\left(m \leq M | Z_1=1, Z_2=q, X=x, T=co\right) \\
&= 1 - F_{M|Z_1=1,Z_2,X,T=co}\left(m,q,x\right).
\end{aligned}$$

Therefore, the density function is obtained by differentiation as

$$f_{Q|X,T=co}(q,x) = -\frac{\partial F_{M|Z_1=1,Z_2,X,T=co}(m,q,x)}{\partial q}. \tag{32}$$

Identification of the density functions requires that

$$Supp\left(Z_2|X,T=co\right) \supseteq Supp\left(Q|X,T=co\right).$$

That is, whenever Q has positive density, also Z_2 must have positive density such that Q is observable in that area of the support. In other words, sufficient variation in Z_2 is required to move M to take the value m for any individual. Put differently, for every x, v in the support of X, V, there exists a value z_2 in the support of Z_2 such that $\zeta^{-1}(1,m,x,v) = z_2$, which corresponds to Assumption 6.

Plugging the previous results into (31) yields

$$E\left[Y^{1,m}|T=co\right]$$

$$= \int\int \varphi(1,m,x,u) dF_{U|M=m,Z_1=1,Z_2=q,X=x,T=co}(u) \left(-\frac{\partial F_{M|Z_1=1,Z_2,X,T=co}(m,q,x)}{\partial q}\right) \cdot f_{X|T=co} dq dx$$

$$= \int \left(\int \varphi(D,M,X,U) dF_{U|M=m,Z_1=1,Z_2=z_2,X=x,T=co}\right) \left(-\frac{\partial F_{M|Z_1=1,Z_2,X,T=co}(m,z_2,x)}{\partial z_2}\right) f_{X|T=co} dz_2 dx$$

$$= \int E\left[Y|M=m,Z_1=1,Z_2=z_2,X=x,T=co\right] \left(-\frac{\partial F_{M|Z_1=1,Z_2,X,T=co}(m,z_2,x)}{\partial z_2}\right) f_{X|T=co}(x) dz_2 dx. \tag{33}$$

For making (33) operational, we need to identify $F_{M|Z_1,Z_2,X,T=co}$, which is derived in the appendix. With these preliminaries, the following identification result is obtained:

Theorem 3: Under Assumptions 1, 3, 6 and 7

$$E\left[Y^{1,m}|T=co\right]=\frac{1}{\Pr\left(T=co\right)}\int E\left[YD\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|Z_2,X,M=m\right]\cdot\Omega\cdot dF_{Z_2,X} \tag{34}$$

with weights

$$\begin{aligned}\Omega &= \omega(Z_2,X)=-\frac{\partial}{\partial z_2}\left(\frac{E\left[D\left(Z_1-\bar{\Pi}\right)|M\leq m,Z_2,X\right]}{E\left[D\left(Z_1-\bar{\Pi}\right)|Z_2,X\right]}F_{M|Z_2,X}\left(m\right)\right)\\ &\quad\times\frac{1}{f_{Z_2|X}}\frac{E\left[\frac{D}{\bar{\Pi}}\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|X\right]}{E\left[D\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|M=m,Z_2,X\right]}.\end{aligned}$$

Note that the expectation in the numerator conditions on $M\leq m$ whereas the expectation in the denominator conditions on $M=m$.

3.3 Natural effects with discrete mediator

The previous identification approaches are only applicable under a continuous mediator M. If M is discrete, it is not possible to point identify V, so that the methods relying on Assumption 4 cannot be used. Neither is Assumption 7 applicable. In the previous sections, identification was achieved by controlling for $f_{M^d|V,X,T=co}$ (via variation in Z_2), in particular by weighting with $\frac{f_{M^0|V,X,T=co}}{f_{M^1|V,X,T=co}}$. With M being discrete, observations need to by weighted by $\frac{\Pr\left(M^0|V,X,T=co\right)}{\Pr(M^1|V,X,T=co)}$. However, as V is no longer identified under a discrete M, we cannot estimate $\Pr\left(M^d|V,X,T=co\right)$ because V is unobserved. As an alternative, one may might find a weighting scheme that produces $\frac{\Pr\left(M^0|V,X,T=co\right)}{\Pr(M^1|V,X,T=co)}$ on average, via integration with respect to Z_2. The price to pay are somewhat stronger identifying assumptions. In the following we focus on the case where M is *binary*, which implies the following model:

$$\begin{aligned}Y &= \varphi(D,M,X,U), \\ M &= 1\left(\zeta(D,Z_2,X,V)\geq 0\right), \\ D &= 1\left(\chi(Z_1,X,W)\geq 0\right).\end{aligned} \tag{35}$$

In addition to Assumptions 1 to 3, identification requires strengthening the monotonicity condition:

Assumption 8: Monotonicity of mediator in the instrument and the unobservable

(i) V is a continuously distributed random variable with a cdf $F_{V|X=x,T=co}(v)$ that is strictly increasing in the support of V, for almost all values of x,
(ii) $\zeta(d, z_2, x, v)$ is strictly monotonic in z_2 and in v. We normalize $\zeta(d, z_2, x, v)$ to be monotonically *increasing* in z_2 and in v.

We thus assume monotonicity in two arguments (which is implicit also in parametric models such as probit and logit specifications). This implies that the values of z_2 can be ordered such that a model of type (35) exists. While monotonicity in v (which is not directly testable) is a fundamental assumption, monotonicity in z_2 (which implies testable restrictions on observed variables) is only needed for quantifying some conditional probabilities under the non-identifiability of V. The particular ordering of the values z_2 themselves is not important. I.e. it would suffice if a transformation of z_2 existed such that the transformed values of z_2 satisfied (35) with Assumption 8. For instance, conditional expectations of M should be increasing in z_2. If this was not the case, Z_2 could be transformed accordingly such that the condition was satisfied.

To develop the intuition for identification based on monotonicity of M in z_2 and v (due to the lack of identification of V), consider for a moment a simplified version of model (35), in which X is dropped (or kept implicit) for notational convenience. The mediator is then given by

$$M = 1\left(\zeta(D, Z_2, V) \geq 0\right).$$

Define ζ^{-1} to be the inverse function with respect to z_2. Since ζ is monotonically increasing in z_2, so is ζ^{-1}. Applying the inverse function on both sides, the mediator equation can be rewritten as

$$M = 1\left(Z_2 \geq \zeta^{-1}(D, 0, V)\right).$$

Furthermore, define $\xi_d(v) \equiv \xi(d,v) \equiv \zeta^{-1}(d,0,v)$. Note that while ζ^{-1} is strictly monotonically increasing in its second argument z_2, it is strictly *monotonically decreasing* in its third argument v, see Lemma 3 in the appendix. Therefore, also $\xi(d,v)$ is strictly *monotonically decreasing* in v. We may write

$$M = 1\left(\xi(D,V) \leq Z_2\ \right), \tag{36}$$

or alternatively,

$$M = 1\left(\xi_D(V) \leq Z_2\ \right),$$

where we use the notation $\xi_d(v) \equiv \xi(d,v)$, with d indexing the function. The latter is convenient because D only takes values 0 and 1.

In a next step, we examine

$$\begin{aligned}
& \Pr\left(M = 0 | D = d, Z_2 = z_2, T = co\right) \\
= \ & \Pr\left(\xi(D,V) > Z_2\ | D = d, Z_2 = z_2, T = co\right) \\
= \ & \Pr\left(\xi(d,V) > z_2\ | Z_1 = d, Z_2 = z_2, T = co\right) \\
= \ & \Pr\left(\xi_d(V) > z_2\ | Z_1 = d, Z_2 = z_2, T = co\right) \\
= \ & \Pr\left(\xi_d(V) > z_2\ | T = co\right),
\end{aligned}$$

where we made use of the facts that $Z_1 = D$ for compliers and that $V \perp\!\!\!\perp (Z_1, Z_2) | X, T = co$ by Assumption 1. Since $\xi_d(v)$ is strictly *monotonically decreasing* in v, its inverse function ξ_d^{-1} exists and is also strictly *monotonically decreasing* in v such that the inverse function can be applied on both sides (where the inequality sign changes because ξ_d^{-1} is a decreasing function) to obtain

$$= \Pr\left(V \leq \xi_d^{-1}(z_2)\ \ | T = co\right) = F_{V|T=co}\left(\xi_d^{-1}(z_2)\right).$$

Now, we make conditioning on X explicit again and define the function $\mu_{d,x}(z_2)$ as

$$\mu_{d,x}(z_2) = \Pr\left(M = 0 | D = d, Z_2 = z_2, X = x, T = co\right). \tag{37}$$

Repeating the previous derivations yields

$$M = 1\left(\xi_{D,X}(V) \leq Z_2\right)$$

with

$$\mu_{d,x}(z_2) = F_{V|X=x,T=co}\left(\xi_{d,x}^{-1}(z_2)\right). \tag{38}$$

Further, $\mu_{d,x}(z_2)$ is identified by Assumptions 1 and 2 as

$$\begin{aligned}
\mu_{1,x}(z_2) &= \frac{E\left[(1-M)D\left(Z_1 - E\left[Z_1|X=x\right]\right)|Z_2 = z_2, X = x\right]}{E\left[D\left(Z_1 - E\left[Z_1|X=x\right]\right)|Z_2 = z_2, X = x\right]}, \\
\mu_{0,x}(z_2) &= \frac{E\left[(1-M)\left(1-D\right)\left(Z_1 - E\left[Z_1|X=x\right]\right)|Z_2 = z_2, X = x\right]}{E\left[\left(1-D\right)\left(Z_1 - E\left[Z_1|X=x\right]\right)|Z_2 = z_2, X = x\right]},
\end{aligned}$$

which can be shown by similar arguments as before.

Since $F_{V|X,T=co}$ is strictly increasing by Assumption 8 and $\xi_{d,x}$ is monotonic as discussed above, the relationship in (38) can be inverted. Let $\mu_{d,x}^{-1}$ denote the inverse function of $\mu_{d,x}(z_2)$, i.e. with respect to z_2. (38) implies that $\mu_{d,x}(z_2)$ and $\xi_{d,x}^{-1}(z_2)$ are both strictly monotonically *decreasing*. Using the shortcut notation $F_{V|x,co} \equiv F_{V|X=x,T=co}$ and denoting its inverse function by $F_{V|x,co}^{-1}$, we note the following relationships:

$$\begin{aligned}
\mu_{d,x}(z_2) &= F_{V|x,co}\left(\xi_{d,x}^{-1}(z_2)\right), \\
\xi_{d,x}^{-1}(z_2) &= F_{V|x,co}^{-1}\left(\mu_{d,x}(z_2)\right), \\
\xi_{d,x}(v) &= \mu_{d,x}^{-1}\left(F_{V|x,co}(v)\right).
\end{aligned}$$

Therefore, the model can be rewritten as

$$\begin{aligned} Y &= \varphi(D, M, X, U), \\ M &= 1\left(\mu_{D,X}^{-1}\left(F_{V|X,co}\left(V\right)\right) \leq Z_2\right), \\ D &= 1\left(\chi(Z_1, X, W) \geq 0\right), \end{aligned}$$

where the function $\mu_{D,X}^{-1}$ is identified and $\mu_{d,x}^{-1}(v)$ is strictly monotonically decreasing in v.

Theorem 4: Under Assumptions 1, 2, 3, 5 and 8

$$\begin{aligned} E\left[Y^{1,M^0}|T=co\right] &= E\left[YD\Omega\frac{Z_1-\Pi}{\Pi\left(1-\Pi\right)}\right]\frac{1}{\Pr\left(T=co\right)}, \\ E\left[Y^{1,M^1}|T=co\right] &= E\left[YD\frac{Z_1-\Pi}{\Pi\left(1-\Pi\right)}\right]\frac{1}{\Pr\left(T=co\right)}, \\ E\left[Y^{0,M^1}|T=co\right] &= E\left[Y(D-1)\bar{\Omega}\frac{Z_1-\Pi}{\Pi\left(1-\Pi\right)}\right]\frac{1}{\Pr\left(T=co\right)}, \\ E\left[Y^{0,M^0}|T=co\right] &= E\left[Y(D-1)\frac{Z_1-\Pi}{\Pi\left(1-\Pi\right)}\right]\frac{1}{\Pr\left(T=co\right)}, \end{aligned}$$

with the weights being functions of Z_2 and X,

$$\Omega = \frac{f_{Z_2|X,T=co}\left(\mu_{0,X}^{-1}\left(\mu_{1,X}(Z_2)\right)\right)}{f_{Z_2|X,T=co}\left(Z_2\right)} \cdot \frac{\mu_{1,X}'(Z_2)}{\mu_{0,X}'(\mu_{0,X}^{-1}(\mu_{1,X}(Z_2)))} \tag{39}$$

and

$$\bar{\Omega} = \frac{f_{Z_2|X,T=co}\left(\mu_{1,X}^{-1}\left(\mu_{0,X}(Z_2)\right)\right)}{f_{Z_2|X,T=co}\left(Z_2\right)} \cdot \frac{\mu_{0,X}'(Z_2)}{\mu_{1,X}'(\mu_{1,X}^{-1}(\mu_{0,X}(Z_2)))}, \tag{40}$$

where $\mu_{d,x}'(z_2) \equiv \frac{d\mu_{d,x}(z_2)}{dz_2}$ is the derivative with respect to z_2.

The weights Ω and $\bar{\Omega}$ are obtained by first estimating the functions $\mu_{d,x}(z_2)$ and the density of Z_2. The conditional density of Z_2 in the complier subpopulation is identified as

$$f_{Z_2|X,T=co}(z_2) = f_{Z_2|X}(z_2) \cdot \frac{E\left[D\left(Z_1-\Pi\right)|X, Z_2=z_2\right]}{E\left[D\left(Z_1-\Pi\right)|X\right]}.$$

Of course, if also Z_2 were known to be independent of the type, the previous expression would simplify to $f_{Z_2|X,T=co}(z_2) = f_{Z_2|X}(z_2)$.

The intuition for the identification results in Theorem 4 goes as follows. Consider the first line,

$$E\left[YD\Omega\frac{Z_1-\Pi}{\Pi(1-\Pi)}\right]\frac{1}{\Pr(T=co)}.$$

In the appendix it is shown that the expectation is zero within the subpopulation of always takers such that

$$= E\left[\frac{Y\Omega}{\Pi}|T=co,Z_1=1\right]\Pr(Z_1=1|T=co).$$

Inserting the model and using Bayes' theorem yields

$$\begin{aligned} &= \int \varphi(1,M,X,U)\Omega(Z_2,X)dF_{U,M,V,Z_2|X,T=co,Z_1=1}\cdot dF_{X|T=co} \\ &= \int \varphi(1,M,X,U)\Omega(Z_2,X)dF_{U,V,Z_2|X,T=co,Z_1=1}\cdot dF_{X|T=co}, \end{aligned}$$

where we used the fact that M is uniquely determined by Z_1, Z_2, X, V among compliers in the last equation. For M binary, we may split the integral into regions where M is 1 and 0, respectively:

$$\begin{aligned} = &\int 1(M=1)\cdot\varphi(1,M,X,U)\Omega(Z_2,X)dF_{U,V,Z_2|X,T=co,Z_1=1}dF_{X|T=co} \\ &+\int 1(M=0)\cdot\varphi(1,M,X,U)\Omega(Z_2,X)dF_{U,V,Z_2|X,T=co,Z_1=1}dF_{X|T=co} \end{aligned}$$

$$\begin{aligned} = &\int 1\left(\mu_{1,X}^{-1}\left(F_{V|X,co}(V)\right)\leq Z_2\right)\cdot\varphi(1,1,X,U)\Omega(Z_2,X)dF_{U,V,Z_2|X,T=co,Z_1=1}dF_{X|T=co} \\ &+\int 1\left(\mu_{1,X}^{-1}\left(F_{V|X,co}(V)\right)> Z_2\right)\cdot\varphi(1,0,X,U)\Omega(Z_2,X)dF_{U,V,Z_2|X,T=co,Z_1=1}dF_{X|T=co}. \end{aligned}$$

By Assumptions 1 and 2, we can write $dF_{U,V,Z_2|X,T=co,Z_1=1} = dF_{U,V|X,T=co}\cdot dF_{Z_2|X,T=co}$ such

that

$$
\begin{aligned}
= & \int 1\left(\mu_{1,X}^{-1}\left(F_{V|X,co}\left(V\right)\right) \leq Z_2\right) \cdot \varphi(1,1,X,U)\Omega(Z_2,X) dF_{U,V,X|T=co} \cdot dF_{Z_2|X,T=co} \\
& + \int 1\left(\mu_{1,X}^{-1}\left(F_{V|X,co}\left(V\right)\right) > Z_2\right) \cdot \varphi(1,0,X,U)\Omega(Z_2,X) dF_{U,V,X|T=co} \cdot dF_{Z_2|X,T=co}
\end{aligned}
$$

$$
\begin{aligned}
= & \int \varphi(1,1,X,U) dF_{U|V,X,T=co} \cdot \left\{ \underbrace{\int \Omega(Z_2,X) \cdot 1\left(\mu_{1,X}^{-1}\left(F_{V|X,co}\left(V\right)\right) \leq Z_2\right) dF_{Z_2|X,T=co}}_{=\Pr(M=1|V,X,T=co,Z_1=0)} \right\} \cdot dF_{V,X|T=co} \\
& + \int \varphi(1,0,X,U) dF_{U|V,X,T=co} \cdot \left\{ \underbrace{\int \Omega(Z_2,X) \cdot 1\left(\mu_{1,X}^{-1}\left(F_{V|X,co}\left(V\right)\right) > Z_2\right) dF_{Z_2|X,T=co}}_{=\Pr(M=0|V,X,T=co,Z_1=0)} \right\} \cdot dF_{V,X|T=co}.
\end{aligned}
$$

The appendix shows that the terms in curly brackets integrate to $\Pr\left(M|V,X,T=co,Z_1=0\right)$. Further using $M^0 \perp\!\!\!\perp Z_1|V,X,T=co$ (by Assumptions 1 and 2) gives

$$
\begin{aligned}
= & \int \varphi(1,1,X,U) dF_{U|V,X,T=co} \cdot \Pr\left(M^0=1|V,X,T=co\right) \cdot dF_{V,X|T=co} \\
& + \int \varphi(1,0,X,U) dF_{U|V,X,T=co} \cdot \Pr\left(M^0=0|V,X,T=co\right) \cdot dF_{V,X|T=co} \\
= & \int \varphi(1,M^0,X,U) \cdot dF_{U|V,X,T=co} \cdot dF_{M^0|V,X,T=co} \cdot dF_{V,X|T=co}.
\end{aligned}
$$

Using $M^0 \perp\!\!\!\perp U|V,X,T=co$ (by Assumption 1) gives

$$
= \int \varphi(1,M^0,X,U) \cdot dF_{M^0,U|V,X,T=co} \cdot dF_{V,X|T=co} = E\left[Y^{1,M^0}|T=co\right].
$$

We conclude this section by considering the example of a *single-index model* as an interesting special case that fits our framework:

$$
M = 1\left(\zeta(\alpha D + \beta Z_2 + \gamma X + V) \geq 0\right),
$$

where ζ represents an unknown monotonic function and α, β, γ denote unknown coefficients. For this particular model, one obtains after some calculations that the weights simplify to

$$\Omega = \frac{f_{Z_2|X,T=co}\left(Z_2 + \frac{\alpha}{\beta}\right)}{f_{Z_2|X,T=co}\left(Z_2\right)} \qquad \text{and} \qquad \bar{\Omega} = \frac{f_{Z_2|X,T=co}\left(Z_2 - \frac{\alpha}{\beta}\right)}{f_{Z_2|X,T=co}\left(Z_2\right)}.$$

Hence, the weights have a particularly simple form under the single-index model.

3.4 Controlled direct effects with discrete mediator

The identification of the controlled direct effect appears difficult, as the control function approach fails (due to the non-identifiability of V) and an identification strategy similar to the previous subsection 3.3 is not applicable. In the latter case, we only needed to re-weight the density function of Z_2 to switch from M^0 to M^1 and vice versa. For the controlled direct effect, however, all weights would need to be assigned to those portions of the density function of Z_2 such that it integrates to one irrespective of the value of V. This implies only using those values of Z_2 for which M attains a particular value m with probability one. Monotonicity in V is not useful here, therefore we only invoke monotonicity in Z_2.

Consider the model

$$\begin{aligned} Y &= \varphi(D, M, X, U), \\ M &= 1\left(\zeta(D, Z_2, X, V) \geq 0\right), \\ D &= 1\left(\chi(Z_1, X, W) \geq 0\right), \end{aligned} \tag{41}$$

and impose Assumption 7, assuming that ζ is monotonically increasing in z_2. Our object of interest is $E\left[Y^{1,m}|T = co\right]$ for $m \in \{0, 1\}$). Identification requires some support condition similar to Assumption 6, with the density function being replaced by a probability because M is now

assumed to be discrete. For the identification of $E\left[Y^{1,0}|T=co\right]$, it has to hold that

$$\Pr\left(M=0|X,V,Z_1=1,T=co\right)>0 \qquad a.s. \tag{42}$$

For identification of both $E\left[Y^{1,0}|T=co\right]$ and $E\left[Y^{0,0}|T=co\right]$ the following assumption must be satisfied:

Assumption 6'

$$\Pr\left(M=0|X,V,Z_1,T=co\right)>0 \qquad a.s.$$

In contrast, the identification of $E\left[Y^{1,1}|T=co\right]$ and $E\left[Y^{0,1}|T=co\right]$ would hinge on:

Assumption 6"

$$\Pr\left(M=0|X,V,Z_1,T=co\right)<1 \qquad a.s.$$

Note that we split the support condition into two parts (i.e. Assumption 6' and 6"), because it may occur in empirical applications that only one of them is satisfied so that only one of the controlled direct effects is identified.

To see how identification is achieved, consider expression (42), which is equivalent to requiring

$$\Pr\left(\zeta(1,Z_2,X,V)<0\ |X,V,Z_1=1,T=co\right)>0 \qquad a.s.$$

or

$$\Pr\left(Z_2<\zeta^{-1}(1,0,X,V)\ |X,V,Z_1=1,T=co\right)>0 \qquad a.s.,$$

where ζ^{-1} is the inverse function with respect to z_2. Hence, for almost every v and x there exists a value $\underline{z}_{v,x}\equiv\zeta^{-1}(1,0,x,v)$, such that M takes the value 0 for every $Z_2<\underline{z}_{v,x}$. By Assumption 6', these values of Z_2 have positive probability mass. Now, for a given x, consider the minimum

of these values $\underline{z}_{v,x}$ in the support of V:

$$\underline{z}_x \equiv \min_{v \in Supp(V|X=x)} \underline{z}_{v,x} \equiv \min_{v \in Supp(V|X=x)} \zeta^{-1}(1,0,x,v). \tag{43}$$

Consider a value $\tilde{z}$ which satisfies $\tilde{z} < \underline{z}_x$, for a given x. The previous considerations imply

$$\Pr\left(M=0|X=x, Z_2=\tilde{z}, Z_1=1, T=co\right)=1, \tag{44}$$

whily by Assumption 6', such values $\tilde{z} < \underline{z}_x$ exist with positive density. Hence, by only using those observations i with $Z_{2i} < \underline{z}_{X_i}$, where $\underline{z}_{X_i}$ is defined by (43) for x taking the value of the observed X_i, there is no endogeneity problem. On the other hand, for observations with $Z_{2i} \geq \underline{z}_{X_i}$, observing $M_i = 0$ implies a dependence between V_i and Z_{2i} which would lead to an improper weighting of U_i. This idea is exploited in the following theorem:

Theorem 5: Under Assumptions 1, 3, 6', 7

$$\begin{aligned}
E\left[Y^{1,0}|T=co\right] &= \frac{\Pr\left(Z_2<\underline{z}_X\right)}{\Pr\left(T=co\right)} E\left[YD\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}\cdot\Omega_{10}|Z_2<\underline{z}_X\right], \\
E\left[Y^{0,0}|T=co\right] &= \frac{\Pr\left(Z_2<\underline{z}_X\right)}{\Pr\left(T=co\right)} E\left[Y(D-1)\frac{Z_1-\bar{\Pi}}{\bar{\Pi}}\cdot\Omega_{00}|Z_2<\underline{z}_X\right],
\end{aligned}$$

with weights

$$\begin{aligned}
\Omega_{10} &= \omega_{10}(X) = \frac{E\left[\frac{D}{\bar{\Pi}}\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|X\right]}{E\left[1\left(Z_2<\underline{z}_X\right)D\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|X\right]}, \\
\Omega_{00} &= \omega_{00}(X) = \frac{E\left[\frac{D}{\bar{\Pi}}\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|X\right]}{E\left[1\left(Z_2<\underline{z}_X\right)(D-1)\frac{Z_1-\bar{\Pi}}{\bar{\Pi}}|X\right]},
\end{aligned}$$

where $\underline{z}_X$ is given by (43) for the random variable X.

Under Assumptions 1, 3, 6", 7

$$
\begin{aligned}
E\left[Y^{1,1}|T=co\right] &= \frac{\Pr\left(Z_2>\bar{\mathrm{z}}_X\right)}{\Pr\left(T=co\right)}E\left[YD\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}\cdot\Omega_{11}|Z_2>\bar{\mathrm{z}}_X\right],\\
E\left[Y^{0,1}|T=co\right] &= \frac{\Pr\left(Z_2>\bar{\mathrm{z}}_X\right)}{\Pr\left(T=co\right)}E\left[Y(D-1)\frac{Z_1-\bar{\Pi}}{\bar{\Pi}}\cdot\Omega_{01}|Z_2>\bar{\mathrm{z}}_X\right],
\end{aligned}
$$

with weights

$$
\begin{aligned}
\Omega_{11} &= \omega_{11}(X)=\frac{E\left[\frac{D}{\bar{\Pi}}\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|X\right]}{E\left[1\left(Z_2>\bar{\mathrm{z}}_X\right)D\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|X\right]},\\
\Omega_{01} &= \omega_{01}(X)=\frac{E\left[\frac{D}{\bar{\Pi}}\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|X\right]}{E\left[1\left(Z_2>\bar{\mathrm{z}}_X\right)(D-1)\frac{Z_1-\bar{\Pi}}{\bar{\Pi}}|X\right]},
\end{aligned}
$$

where $\bar{\mathrm{z}}_X$ is given by

$$
\bar{\mathrm{z}}_x \equiv \max_{v\in Supp(V|X=x)}\zeta^{-1}(1,0,x,v)
$$

for the random variable X.

Concerning the first result of Theorem 5, only the outcome information of observations satisfying $Z_{2i}<\underline{\mathrm{z}}_{X_i}$ is used. In practice, $\underline{\mathrm{z}}_x$ is unknown and needs to be estimated. According to (44), the largest value satisfying that the probability of observing $M=0$ is one (or very close to one) for all values of Z_2 below this threshold should be chosen for $\underline{\mathrm{z}}_x$. Note that the left hand side of (44), which is a conditional probability among compliers, is identified as

$$
\begin{aligned}
\Pr\left(M=0|X,Z_2,Z_1=1,T=co\right) &= \frac{E\left[(1-M)\,D\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|X,Z_2\right]}{\Pr\left(T=co,Z_1=1|X,Z_2\right)}\\
&= \frac{E\left[(1-M)\,D\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|X,Z_2\right]}{\bar{\Pi}E\left[\frac{D}{\bar{\Pi}}\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|X,Z_2\right]}\\
&= \frac{E\left[(1-M)\,D\left(Z_1-\bar{\Pi}\right)|X,Z_2\right]}{E\left[D\left(Z_1-\bar{\Pi}\right)|X,Z_2\right]},
\end{aligned}
$$

because $Z_1 \perp\!\!\!\perp T|X,Z_2$ and $V \perp\!\!\!\perp Z_1|X,Z_2,T=at$ and $\Pr\left(T=co|X,Z_2\right)=E\left[\frac{D}{\bar{\Pi}}\frac{Z_1-\bar{\Pi}}{1-\bar{\Pi}}|X,Z_2\right]$. With

this relationship, $\Pr(M = 0|X, Z_2, Z_1 = 1, T = co)$ is estimable and one may find those values of z_2 where this probability is one (or close to one) in order to estimate $\underline{z}_x$.

3.5 Natural effects with discrete instruments

In this section, we discuss identification when both Z_1 and Z_2 are *discrete*, and M is continuous. Note that the results obtained are also applicable when Z_2 is continuous, but rest on stronger assumptions than in the previous sections.

In many applications, discrete instruments for the mediator may be easier to find than continuous ones. However, if Z_2 is discrete, the previous identification approaches are infeasible. The key channel for identification in Section 3.1 was to (hypothetically) vary Z_1 in order to change treatment status D, while at the same time keeping M unchanged through a variation of Z_2 that undoes the effect of Z_1 on M. This generally requires a continuous Z_2 in order to keep M fixed when Z_1 switches, conditional on V and X. Otherwise if for instance Z_1 and Z_2 are both binary, the distribution of M conditional on V, X and Z_1 has only two mass-points, which are generally different for $Z_1 = 0$ and $Z_1 = 1$. To more clearly see the problem, reconsider the identification approach of Section 3.1:

$$\begin{aligned} E\left[Y^{1,M^0}|T = co\right] &= \int \varphi(1, M^0, X, U) \cdot dF_{M^0,X,U,V|T=co} \\ &= \int \varphi(1, M^0, X, U) \cdot dF_{M^0|X,U,V,T=co} \cdot dF_{X,U,V|T=co} \\ &= \int \varphi(1, M^0, X, U) \cdot dF_{M^0|X,V,T=co} \cdot dF_{X,U,V|T=co}, \end{aligned}$$

because $M^0 \perp\!\!\!\perp U|X, V, T = co$ by the control function assumption. We multiplied and divided by $dF_{M^1|X,V,T=co}$ to obtain

$$\begin{aligned} &= \int \varphi(1, M^0, X, U) \cdot dF_{X,U,V|T=co} \cdot dF_{M^0|X,V,T=co} \cdot \frac{dF_{M^1|X,V,T=co}}{dF_{M^1|X,V,T=co}} \\ &= \int \varphi(1, M, X, U) \cdot dF_{X,U,V|T=co} \cdot dF_{M|Z_1=0,X,V,T=co} \cdot \frac{dF_{M|Z_1=1,X,V,T=co}}{dF_{M|Z_1=1,X,V,T=co}} \end{aligned}$$

as long as $dF_{M|Z_1=1,X,V,T=co} > 0$ at every m for which $dF_{M|Z_1=0,X,V,T=co} > 0$

$$= \int \varphi(1, M, X, U) \cdot \underbrace{\frac{dF_{M|Z_1=0,X,V,T=co}}{dF_{M|Z_1=1,X,V,T=co}}}_{\Omega-Weights} \cdot dF_{M|Z_1=1,X,V,T=co} dF_{X,U,V|T=co}.$$

Identification thus required that

$$Supp\left(M|Z_1 = 0, X, V, T = co\right) \subseteq Supp\left(M|Z_1 = 1, X, V, T = co\right), \tag{45}$$

which corresponds to Assumption 5.

In general, this support condition is not satisfied if Z_2 is discrete. The problem could be solved, however, if the satisfaction of IV validity would not hinge on conditioning on X, as the distribution of M conditional on V and Z_1 is usually continuous as long as at least one element in X is continuous. Variation in X may then be used to shift M to all points needed in the $Z_1 = 1$ and $Z_1 = 0$ populations. This, however, requires X to be exogenous, as discussed below. To be concise, it would suffice if only one element of X satisfied exogeneity. Under these conditions, the support assumption can be weakened to only satisfying $dF_{M|Z_1=1,V,T=co} > 0$ at every m where $dF_{M|Z_1=0,V,T=co} > 0$, i.e. without conditioning on X. Therefore, (45) is replaced by the weaker restriction

$$Supp\left(M|Z_1 = 0, V, T = co\right) \subseteq Supp\left(M|Z_1 = 1, V, T = co\right). \tag{46}$$

We can summarize this support condition in the following way, which admits identification of both Y^{1,M^0} and Y^{0,M^1}:

Assumption 9: Common support of M

$$0 < \Pr\left(Z_1 = 1|M, C, T = co\right) < 1 \qquad a.s.$$

A further requirement for the identification of the mean potential outcomes is that X is structurally separated from M. To this end, we assume that the outcome equation is additively separable in X, while the other equations are as unrestricted as in the previous discussion:

$$\begin{aligned} Y &= \varphi(D, M, U) + \psi(D, X), \\ M &= \zeta(D, Z_2, X, V), \\ D &= 1\left(\, \chi(Z_1, X, W) \geq 0 \,\right). \end{aligned} \tag{47}$$

Both Z_1 and Z_2 are discrete, so that X has to contain (at least) one continuous variable.

Finally, our conditional independence assumptions need to be strengthened to embrace exogeneity of X. That is, in *addition* to Assumptions 1 and 2, the following needs to hold:

Assumption 10: Exogeneity assumptions

$$\begin{aligned} X &\perp\!\!\!\perp Z_1, \\ X &\perp\!\!\!\perp (U, V) | T = co. \end{aligned}$$

Assumptions 1,2, and 10 jointly imply that

$$\begin{aligned} Z_1 &\perp\!\!\!\perp (Z_2, X, U, V, T), \\ (Z_1, Z_2, X) &\perp\!\!\!\perp (U, V) | T = co. \end{aligned}$$

While the first part of Assumption 10 is straightforward and easily testable, the second condition is more delicate as it requires independence of X in the subpopulation of compliers. Since the type is itself a function of X and W, conditioning on being a complier can introduce a dependency even if X were independent of U, V in the full population. The second part of Assumption 10 therefore appears more plausible if X does not affect D so that the type only depends on W.

Hence, one could extend the previous model so that only some covariates are exogenous and

structurally separated, while others are possibly endogenous as in previous sections. To this end, consider partitioning $X = (X_1, X_2)$, where the *exogenous* variables X_1 must contain (at least) one *continuous* variable, while the possibly endogenous variables X_2 are not restricted. We may then admit the model

$$\begin{aligned} Y &= \varphi(D, M, X_2, U) + \psi(D, X_1, X_2), \\ M &= \zeta(D, Z_2, X_1, X_2, V), \\ D &= 1\left(\chi(Z_1, X_1, X_2, W) \geq 0 \right), \end{aligned} \tag{48}$$

and replace Assumption 10 by

Assumption 10': Exogeneity assumption (modified)

$$\begin{aligned} X_1 &\perp\!\!\!\perp Z_1 | X_2, \\ X_1 &\perp\!\!\!\perp (U, V) | X_2, T = co. \end{aligned}$$

The second part of Assumption 10' would be more plausible if treatment choice was only affected by X_2, but not by X_1. (Also Assumption 9 would need to be extended to include X_2 in the conditioning set.) This extended model (48) nests the setups of Section 3.1 when X_1 is the empty set, as well as (47) when X_2 is the empty set. For ease of exposition, we focus on model (47) instead of (48) in this section, though.

We note that Lemma 1 also holds true with both instruments being discrete. By Lemma 1 we have $C_i = F_{V|X=X_i,T=co}(V_i)$ meaning that conditional on X, the control function C is a one-to-one mapping of V. By Assumption 10, $F_{V|X,T=co} = F_{V|T=co}$, which implies $C_i = F_{V|T=co}(V_i)$, so that C uniquely identifies V even *without* conditioning on X.

We subsequently discuss the identification of Y^{1,M^0} and note that the derivations for Y^{0,M^1}

follow by symmetry.[12] By (47), the mean potential outcome of interest corresponds to

$$E[Y^{1,M^0}|T = co] = E\left[\varphi(1, M^0, U) + \psi(1, X)|T = co\right].$$

Since the two functions φ and ψ may both contain an intercept, we need to (arbitrarily) fix one of these intercepts (as otherwise the two functions are not identified). We normalize the intercept of ψ such that for some value x_0 in the support of X

$$\psi(1, x_0) = 0.$$

Identification proceeds in three steps, requiring Assumptions 1, 2, 3, 4, 9, and 10. First, we identify the function $\psi(D, X)$. Second, we subtract the function $\psi(D, X)$ from Y, and then identify $E[\varphi(1, M^0, U)|T = co]$. Third, we identify $E[\psi(1, X)|T = co]$, and then combine the previous results for the identification of $E[Y^{1,M^0}|T = co]$.

Consider the following expression for some values m, c, x in the support of M, C, X:

$$\frac{E\left[Y D \frac{Z_1 - \Pi}{1 - \Pi}|M = m, C = c, X = x\right]}{\Pr\left(T = co, Z_1 = 1|M = m, C = c, X = x\right)}. \tag{49}$$

As shown in the appendix, this term is zero for the always takers, such that we obtain

$$\begin{aligned}
&= E\left[Y|T = co, Z_1 = 1, M = m, C = c, X = x\right] \\
&= E\left[\varphi(1, m, U) + \psi(1, x)|T = co, Z_1 = 1, M = m, C = c, X = x\right] \\
&= E\left[\varphi(1, m, U)|T = co, Z_1 = 1, M = m, C = c, X = x\right] + \psi(1, x).
\end{aligned}$$

Because $U \perp\!\!\!\perp M|C, Z_1, X, T = co$ due to $U \perp\!\!\!\perp Z_2|V, Z_1, X, T = co$ and $U \perp\!\!\!\perp Z_1|C, X, T = co$, and

[12] In applications, we could simply re-code D_i as $1 - D_i$ and Z_i as $1 - Z_i$, and then apply the formulae below to the re-coded data.

finally $U \perp\!\!\!\perp X | C, T = co$ by Assumption 10, it follows that

$$= E\left[\varphi(1, m, U) | T = co, C = c\right] + \psi(1, x)$$

$$= \chi(m, c) + \psi(1, x),$$

where $\chi(m, c) \equiv E\left[\varphi(1, m, U) | T = co, C = c\right]$ is an unknown function of m and c only.

We further note that the denominator of (49) can be simplified by using an auxiliary result of the proof of Theorem 1:

$$dF_{M,C|Z_1=1,X,T=co} = E\left[\frac{D}{\Pi}\frac{Z_1 - \Pi}{1 - \Pi} | M, C, X\right] \frac{dF_{M,C|X}}{\Pr\left(T = co | X\right)}.$$

Using Bayes' theorem and plugging in the previous expression, the denominator of (49) can therefore be written as

$$\begin{aligned}
\Pr\left(T = co, Z_1 = 1 | M, C, X\right) &= \frac{dF_{M,C|X,T=co,Z_1=1} \cdot \Pr\left(T = co, Z_1 = 1 | X\right)}{dF_{M,C|X}} \\
&= E\left[\frac{D}{\Pi}\frac{Z_1 - \Pi}{1 - \Pi} | M, C, X\right] \frac{\Pr\left(T = co, Z_1 = 1 | X\right)}{\Pr\left(T = co | X\right)} \\
&= E\left[D\frac{Z_1 - \Pi}{1 - \Pi} | M, C, X\right],
\end{aligned}$$

where the last result follows from $Z_1 \perp\!\!\!\perp T | X$. Combining all these results we obtain for some values m, c, x in the support of M, C, X,

$$\frac{E\left[Y D \left(Z_1 - \Pi\right) | M = m, C = c, X = x\right]}{E\left[D \left(Z_1 - \Pi\right) | M = m, C = c, X = x\right]} = \chi(m, c) + \psi(1, x). \tag{50}$$

Consider now two triplets (m, c, x_1) and (m, c, x_0) in the support of M, C, X. We obtain

$$\frac{E\left[Y D \left(Z_1 - \Pi\right) | M = m, C = c, X = x_1\right]}{E\left[D \left(Z_1 - \Pi\right) | M = m, C = c, X = x_1\right]} - \frac{E\left[Y D \left(Z_1 - \Pi\right) | M = m, C = c, X = x_0\right]}{E\left[D \left(Z_1 - \Pi\right) | M = m, C = c, X = x_0\right]} = \psi(1, x_1). \tag{51}$$

because $\psi(1, x_0)$ has been normalized to be zero. Similarly, consider two triplets (m', c', x_2) and

(m', c', x_1) in the support of M, C, X:

$$\frac{E\left[YD\left(Z_1-\Pi\right)|M=m', C=c', X=x_2\right]}{E\left[D\left(Z_1-\Pi\right)|M=m', C=c', X=x_2\right]}-\frac{E\left[YD\left(Z_1-\Pi\right)|M=m', C=c', X=x_1\right]}{E\left[D\left(Z_1-\Pi\right)|M=m', C=c', X=x_1\right]}=\psi(1,x_2)-\psi(1,x_1),$$

and as $\psi(1, x_1)$ is identified by (51), so is $\psi(1, x_2)$. We can thus identify $\psi(1, x_1)$, $\psi(1, x_2)$, and so forth. However, identification of the entire function $\psi(1, x)$ for all x in the support of X requires further conditions as we need to find triplets with identical m and c, but $x_2 \neq x_1$. Since M is a function of Z_2, X, V among compliers, it is only through variation of z_2 that identical values of m and c for different x may be obtained.

If ψ is a parametric function of, say, a k-dimensional parameter vector β, it generally suffices to identify $\psi(1, x) \equiv \psi_1(x; \beta)$ for k different values of x. In practice, one may for instance consider the following regression approach. Let $\hat{Y}_i$ be a (non-parametric) estimate of $E[YD\left(Z_1-\Pi\right)|M_i, C_i, X_i]/E[D\left(Z_1-\Pi\right)|M_i, C_i, X_i]$. We estimate the model

$$\hat{Y}_i = \chi(M_i, C_i) + \psi_1(X_i; \beta) + \epsilon_i, \tag{52}$$

with ϵ_i being the regression error, χ an unknown two-dimensional nonparametric function, and $\psi_1(x; \beta)$ a parametric function, using partially linear semiparametric regression.

On the other hand, if ψ is a non-parametric function, identification conditions are more difficult to characterize and require more than smoothness assumptions. If we do not want to impose any structure on ψ (apart from continuity), it would be required that for any x_1 in the support of X, there exists (at least) one value of m and c, respectively, in the supports of M and C which allows applying (51). This would be unproblematic if m did not appear on the right-hand-side of (50), because C and X are independent by Assumption 10. However, in the $T = co, Z_1 = 1$ subpopulation, M is a deterministic function of Z_2, X and C only, such that M, C, and X are closely related and the relationship is non-deterministic only because of Z_2.

Consider the most cumbersome case where both Z_1 and Z_2 are *binary*. M takes only two different values conditional on X and C and being in the $T = co, Z_1 = 1$ subpopulation. It is

difficult to find general identification conditions. One special case applies if the function $M = \zeta(D, Z_2, X, V)$ happens to be invertible in x, i.e. in its third argument. Since there is a one-to-one mapping between V and C (as established in Lemma 1), we can also find a function $\bar{\zeta}$ such that M can be expressed in terms of C instead of V, that is $M = \bar{\zeta}(D, Z_2, X, C)$. Let $\bar{\zeta}^{-1}_{(3)}$ denote the inverse function with respect to the third argument of $\bar{\zeta}$, such that $X = \bar{\zeta}^{-1}_{(3)}(D, Z_2, M, C)$.

Consider some value c in the support of C. For x_0 and c given, two different values of m can be observed, depending on whether Z_2 takes the value 0 or 1. (Note that we are always in the $T = co, Z_1 = 1$ subpopulation.) Suppose $z_2 = 1$. Then, there exists a different value x_1 which in combination with $z_2 = 0$ would deliver the same value of m, i.e. $\bar{\zeta}(1, z_2 = 0, x_1, c) = \bar{\zeta}(1, z_2 = 1, x_0, c)$. This value x_1 is given by $x_1 = \bar{\zeta}^{-1}_{(3)}(d = 1, z_2 = 0, \bar{\zeta}(d = 1, z_2 = 1, x_0, c), c)$. Similarly, for the same x_0 and c, one could observe $Z_2 = 0$. This would give us a value, say, $\bar{m}$. Now, there exists a different value x_2 which in combination with $z_2 = 1$ would deliver the same value $\bar{m}$. This value x_2 is given by $x_2 = \bar{\zeta}^{-1}_{(3)}(d = 1, z_2 = 1, \bar{\zeta}(d = 1, z_2 = 0, x_0, c), c)$. These two values x_1 and x_2 are observable with positive density if $0 < \Pr(Z_2 = 1|C, X) < 1$ so that both values of Z_2 actually occur for given c and x.[13] Hence, for *each* value of c there exist two values x_1 and x_2 at which $\psi(1, x)$ is identified.[14] Identification of $\psi(1, \bar{x})$ at a general $\bar{x}$ requires the existence of a c in the support of C, such that either of the two mappings delivers $\bar{x}$. This is formally stated in the following lemma.

Lemma 2: For Z_1 and Z_2 being binary random variables, the function $\psi(1, x)$ is identified at x if there is a value c in the support of C such that either

$$x = \bar{\zeta}^{-1}_{(3)}(1, 1, \bar{\zeta}(1, 0, x_0, c), c)$$

or

$$x = \bar{\zeta}^{-1}_{(3)}(1, 0, \bar{\zeta}(1, 1, x_0, c), c),$$

[13] Obviously, they are only identified in the support of X among compliers, but we anyhow do not require the function $\psi(1, x)$ to be identified outside this support.

[14] This result applies to a Z_2 having only *two* mass-points. If Z_2 was discrete with k mass-points, we could identify $k \cdot (k - 1)$ values for each c.

under Assumptions 1, 2, 3, 4, 10 and the following assumptions:

i) $\zeta(D, Z_2, X, V)$ **is invertible in its third argument,**

ii) $0 < \Pr(Z_2 = 1|C, X) < 1$.

In the following, we suppose that function $\psi(1, x)$ is identified, without specifying exactly how (e.g. via being a parametric function, Lemma 2, or some alternative support and structural assumptions). Based on $\psi(1, x)$, $E[\varphi(1, M^0, U)|T = co]$ can be identified, too. Define the weights

$$\Omega = \omega(M, C) = \frac{dF_{M,C|Z_1=0,T=co}}{dF_{M,C|Z_1=1,T=co}}$$

and consider the following expression

$$E\left[(Y - \psi(1, X)) \cdot \Omega \cdot D\frac{Z_1 - \Pi}{1 - \Pi}\right], \tag{53}$$

which, by inserting the outcome equation, is equivalent to

$$= E\left[\varphi(D, M, U) \cdot \Omega \cdot D\frac{Z_1 - \Pi}{1 - \Pi}\right].$$

As we show in the appendix this is corresponds to

$$= E\left[\varphi(1, M^0, U)|T = co\right] \cdot \Pr\left(T = co, Z_1 = 1\right).$$

By Assumptions 1, 2 and 10, it also holds $Z_1 \perp\!\!\!\perp T$, implying $\Pr\left(T = co, Z_1 = 1\right) = \Pr\left(T = co\right)\Pr\left(Z_1 = 1\right)$ such that we obtain

$$E\left[\varphi(1, M^0, U)|T = co\right] = \frac{E\left[(Y - \psi(1, X)) \cdot \Omega \cdot D\frac{Z_1 - \Pi}{1 - \Pi}\right]}{\Pr\left(T = co\right)\Pr\left(Z_1 = 1\right)}.$$

Finally, we need to identify

$$E[\psi(1, X)|T = co].$$

One can show that by using $Z_1 \perp\!\!\!\perp T|X$ one gets

$$\frac{1}{\Pr(T=co)} E\left[\psi(1,X)\cdot\frac{D}{\Pi}\frac{Z_1-\Pi}{1-\Pi}\right] = \int \psi(1,X) dF_{X|T=co} = E[\psi(1,X)|T=co].$$

The proof is straightforward via iterated expectations with respect to the type and X.

We combine all these results in the following theorem, which also makes use of $\Pr(Z_1=1|X) = \Pr(Z_1=1)$ by Assumption 10.

Theorem 6: Under Assumptions 1, 2, 3, 4, 9 and 10 and assuming $\psi(1,X)$ to be identifiable in the support of X, we obtain:

$$E\left[Y^{1,M^0}|T=co\right] = \frac{E\left[\{Y\cdot\Omega+(1-\Omega)\cdot\psi(1,X)\}\cdot D\cdot(Z_1-\Pr(Z_1=1))\right]}{\Pr(T=co)\Pr(Z_1=1)\Pr(Z_1=0)}$$

with

$$\begin{aligned}\Omega &= \omega(M,C) = \frac{E\left[(D-1)\{Z_1-\Pr(Z_1=1)\}|M,C\right]}{E\left[D\{Z_1-\Pr(Z_1=1)\}|M,C\right]} \\ &= 1-\frac{E[Z_1|M,C]-E[Z_1]}{E[DZ_1|M,C]-E[D|M,C]\cdot E[Z_1]}.\end{aligned}$$

4 Simulation study

This section presents a brief simulation study that provides some intuition for the identification results underlying Theorem 1 and the issues related to model misspecification. The data gener-

ating process (DGP) consists of the following models for the outcome, mediator, and treatment:

$$\begin{aligned} Y &= D + M + \beta DM + 0.5X + U, \\ M &= 2Z_2 + 0.5D + 0.5X + V, \\ D &= 1\left(2Z_1 + 0.5X + W > 1 \right), \\ Z_1 &= 1\left(0.5X + P > 0 \right), \\ Z_2 &= 0.5X + Q, \end{aligned}$$

$$(U, V, W) \sim N(\mu, \sigma), \text{ where } \mu = \mathbf{0} \text{ and } \sigma = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix},$$

and X, P and W are standard normal, independently of each other and of U, V, W.

M and D are endogenous due to the non-zero correlation of the error terms U, V, and W. The first instrument Z_1 is binary and has a strong impact on D, shifting the treatment probability by roughly 68 percentage-points. This is comparable to compliance rates in many field experiments, see for instance Abadie, Angrist, and Imbens (2002) and our application to the Job Corps experiment further below. Z_2 is continuously distributed and affects the endogenous mediator M. β gauges the interaction effect of D and M on Y, i.e., whether direct and indirect effects are heterogenous across treatment states. In the simulations, $\beta = 0$ (no interaction) and $\beta = 1$ are considered, for sample sizes $n = 2500$ and 5000.

We investigate the performance of three different approaches for the estimation of natural direct and indirect effects. The first one is semiparametric and based on the sample analogs of Theorem 1. The conditional probabilities, namely $\pi(x)$, $E[Z_1|M, C, X]$ and $E[D|M, C, X]$ and $E[DZ_1|M, C, X]$ are estimated by probit regression. Before we can proceed so, we need estimates of the control function C_i, which we obtain via Lemma 1. Therefore, we need estimates of $E[D(Z_1 - \pi(x))|Z_2, X]$ and of $E[(d + D - 1) \cdot (Z_1 - \pi(x))|M \leq m, Z_2 = z_2, X = x]$, which we estimate by OLS. For the latter conditional expectation, linear regression on $(1, Z_2, X)$ is

performed with the subset of observations satisfying $M \leq m$ in the data, with $m = M_i$ (that is, the value of M for the i^{th} observation in the data) if the conditional expectation is predicted for observation i. This obviously implies underidentification and sparse data problems for the lowest value(s) of M_i. We therefore set m such that the number of observations in the linear regression is never below 10, implying $m > M_i$ for the 9 observations with the lowest values of M. Even though this approach biases estimates at the lower bound of the empirical support of M (i.e. for 9 observations), their relative importance in the estimation of direct and indirect effects vanishes as the sample size increases. Finally for Lemma 1 we need $F_{M|Z_2,X}(m, z_2, x)$, which we estimate by nonparametric kernel estimation of conditional distribution functions using the np package of Hayfield and Racine (2008). The kernel bandwidths are selected based on the Silverman (1986) rule of thumb.

We consider both untrimmed and trimmed versions of the estimators of direct and indirect effects. Trimming prevents that influential observations receive arbitrarily large weights in IPW, which may cause an explosion in the variance of the estimator. Similarly to Huber, Lechner, and Wunsch (2013) and Frölich and Huber (2014), observations that would obtain a relative weight larger than 5% in the estimation of each of the mean potential outcomes of Theorem 1 are therefore discarded.

Secondly, we examine multi-step parametric IV estimation as applied in Powdthavee, Lekfuangfu, and Wooden (2013), see their equations (3) to (5). In a first step, we run a probit regression of D on $(1, Z_1, X)$ to predict the treatment, denoted by $\tilde{D}$. Then, we linearly regress M on $(1, Z_2, \tilde{D}, X)$ to predict M, denoted by $\tilde{M}$. As these predictions are based on variations in the instruments unrelated to (U, V, W) given X, they are exogenous (if we impose the additional assumption that W is independent of Z_2, i.e. assumption (7)). Therefore, the estimated direct effect corresponds to the coefficient on $\tilde{D}$ in an OLS regression of Y on $(1, \tilde{D}, \tilde{M}, X)$. Finally, we linearly regress M on $(1, \tilde{D}, X)$ and estimate the indirect effect as the product of the coefficient on $\tilde{D}$ in this last regression and the coefficient on $\tilde{M}$ in the regression of Y. Note that in contrast to non- or semiparametric estimation, this parametric IV estimator does not allow for

interaction effects between M and D.

Finally, the last estimator considered consists of a naive OLS approach that does neither control for endogeneity due to the unobservables or X, nor allow for interaction effects, see Baron and Kenny (1986). The direct effect is estimated by the coefficient on D in an OLS regression of Y on $(1, D, M)$, the indirect effect by the coefficient on M in the last regression times the coefficient on D in an OLS regression of M on $(1, D)$.

Table 1 presents the bias, standard deviation (sd), and root mean squared error (RMSE) of the various estimators of the direct and indirect effects $\theta(d)$ and $\delta(d)$, defined in (4) and (5), based on 1000 simulations. Considering the upper two panels with $\beta = 0$ (no treatment-mediator interactions), naive OLS is severely biased despite the substantial share of compliers. In contrast, the correctly specified parametric IV estimators are close to being unbiased and (due to their tighter specification) more efficient than the semiparametric IV methods. However, also the latter perform satisfactorily in terms of bias and RMSE in the larger sample. Trimming reduces the standard deviation and the RMSE in the small sample, and has no effect in the larger sample where extreme weights apparently do not occur any more.

The situation changes, though, once we permit for treatment effect heterogeneity: For $\beta = 1$, biases are large for OLS and also for parametric IV (even for $n = 5000$), but small for the semiparametric methods. Nevertheless, for $n = 2500$ the latter (even with trimming) do not always outperform parametric IV in terms of RMSE because of their lower precision. With the larger sample size, however, the gains in terms of bias reduction outweigh the losses in efficiency so that the RMSE of semiparametric IV is always smaller than that of parametric IV. Hence, the semiparametric IV estimators dominate all others in terms of RMSE. We therefore conclude that semiparametric estimation can be preferable to parametric methods in samples with several thousand observations, as in many recently conducted field experiments and the applications presented in the next section.

Table 1: Bias, standard deviation, and RMSE of various estimators

	$\theta(1)$			$\theta(0)$			$\delta(1)$			$\delta(0)$		
estimator	bias	sd	RMSE	bias	sd	RMSE	bias	sd	RMSE	bias	sd	RMSE
n=2500, $\beta = 0$												
semipara IV	-0.050	0.416	0.419	0.023	0.169	0.171	-0.028	0.193	0.195	0.045	0.438	0.440
semip IV trim	-0.033	0.280	0.282	0.026	0.145	0.148	-0.031	0.172	0.175	0.028	0.313	0.314
parametric IV	-0.000	0.069	0.069	-0.000	0.069	0.069	-0.004	0.148	0.148	-0.004	0.148	0.148
OLS	0.540	0.044	0.542	0.540	0.044	0.542	2.001	0.118	2.004	2.001	0.118	2.004
n=5000, $\beta = 0$												
semipara IV	-0.019	0.196	0.197	0.036	0.089	0.095	-0.038	0.114	0.120	0.017	0.227	0.227
semi IV trim	-0.019	0.196	0.197	0.036	0.089	0.095	-0.038	0.114	0.120	0.016	0.226	0.227
parametric IV	-0.002	0.047	0.047	-0.002	0.047	0.047	-0.000	0.107	0.107	-0.000	0.107	0.107
OLS	0.539	0.030	0.540	0.539	0.030	0.540	2.006	0.082	2.007	2.006	0.082	2.007
n=2500, $\beta = 1$												
semipara IV	-0.052	0.441	0.444	0.030	0.284	0.286	-0.037	0.373	0.375	0.045	0.438	0.440
semi IV trim	-0.035	0.318	0.320	0.037	0.226	0.229	-0.045	0.325	0.328	0.028	0.313	0.314
parametric IV	-0.250	0.116	0.276	0.250	0.116	0.276	-0.255	0.222	0.338	0.245	0.222	0.330
OLS	0.290	0.077	0.300	0.790	0.077	0.794	2.601	0.173	2.607	3.101	0.173	3.106
n=5000, $\beta = 1$												
semipara IV	-0.019	0.218	0.219	0.061	0.143	0.155	-0.062	0.222	0.231	0.017	0.227	0.227
semi IV trim	-0.018	0.218	0.219	0.061	0.143	0.155	-0.062	0.222	0.231	0.016	0.226	0.227
parametric IV	-0.253	0.081	0.265	0.247	0.081	0.260	-0.250	0.161	0.298	0.250	0.161	0.297
OLS	0.288	0.055	0.293	0.788	0.055	0.790	2.609	0.122	2.612	3.109	0.122	3.111

Note: Results are based on 1000 simulations. The true effects under $\beta = 0$ are $\theta(1) = \theta(0) = 1$, $\delta(1) = \delta(0) = 0.5$. Under $\beta = 1$, the true effects are $\theta(1) = 1.5$, $\theta(0) = 1$, $\delta(1) = 1$, $\delta(0) = 0.5$.

5 Applications

This section presents two empirical applications to illustrate Theorem 1 (continuous Z_2) and Theorem 6 (discrete Z_2).

5.1 Effects of education with continuous mediator income

Our first application is based on data from the British Household Panel Survey (BHPS), which includes information at the household and individual levels for a representative sample of the population of the United Kingdom. We aim at assessing the direct effect and the indirect effect (via income) of *education* on health. The outcome variable is measured as the (mental and physical) capability to participate in social life, more precisely as the ability to interact in the normal or usual way in society, which we will refer to as 'social functioning'. That is, we investigate

whether any effect of education on social functioning is driven by a change in income or by ('direct') causal mechanisms other than income.

We investigate the effects of a binary schooling indicator (D), which is one if an individual has obtained more than lower secondary education according to the International Standard Classification of Education (ISCED) of the UNESCO and zero otherwise. D is instrumented by a change in the UK compulsory school leaving age in 1971 (Z_1). In that year, the minimum age at which one could leave school was increased from 15 to 16 years, which affected all cohorts born in 1956 or later. The law change Z_1 should induce some individuals (compliers) to increase schooling, but is arguably not directly associated with social functioning (Y), which is measured on a scale from 0 (worst) to 9 (best). Changes in schooling laws have also been used as instruments for instance in Oreopoulos (2006), Spasojevic (2010), and Brunello, Fabbri, and Fort (2013). To disentangle the effect of education into a direct and an indirect component driven by income changes, we consider *annual individual income* (in GBP) as mediator (M). The latter is instrumented by windfall income (Z_2), the sum of four arguably exogenous income sources: accident claims, redundancy payments, lottery wins, and other lump sum payments. Similar exogenous variations in income were also exploited in Lindahl (2005) and Gardner and Oswald (2007), among others. We assume the IV assumptions underlying Theorem 1 to hold conditional on the covariate gender (X) and present several tests further below.

The BHPS started in 1991 with 10,300 individuals drawn from 250 areas in Great Britain and interviews the sample participants annually. The panel was again enlarged several times. In 1999 (wave 9) for instance, additional samples of 1500 households in each of Scotland and Wales were added. Our empirical illustration is based on four waves, namely 5, 6, 8, and 9, which were conducted in 1995, 1996, 1998, and 1999, respectively. As we exploit the panel structure of the data, we do not make use of the individuals added in wave 9 (or later). Wave 5 is used for measuring X, wave 6 for educational attainment D, wave 8 for M (annual income 1998) and Z_2 (windfall profit 1998), and wave 9 for Y, the social functioning index in 1999. The stock sample in wave 5 consists of 9,249 observations. 1,834 of the latter are not observed in at least one of the

other 3 waves, so that the balanced panel includes 7,415 individuals. Furthermore, we restrict the sample to observations born between (and including) 1946 and 1965, i.e. in a 10 years window around 1956, the year of the first cohort affected by the 1971 schooling reform. This criterion is satisfied by 3,009 individuals. Finally, all observations with missing values in X, D, M, Y, Z_1 or Z_2 are dropped, entailing a final evaluation sample of $n = 2,886$ observations. Table 2 provides descriptive statistics (means and standard deviations) of X, Y, Z_1 and Z_2, separately by the treatment state and the values of the mediator.

Table 2: Descriptive statistics

	$D=1$		$D=0$		$M>12,000$		$M\leq 12,000$	
	mean	std.dev	mean	std.dev	mean	std.dev	mean	std.dev
Gender (X) (binary)	0.541	0.498	0.575	0.495	0.346	0.476	0.765	0.424
Social functioning (Y)	8.149	1.741	7.825	2.095	8.280	1.596	7.859	2.029
School leaving age 16 years (Z_1) (binary)	0.559	0.497	0.475	0.500	0.546	0.498	0.534	0.499
Windfall income in 1000 GBP (Z_2)	0.476	3.756	0.469	5.291	0.674	4.616	0.262	3.572
# of observations	2,242		644		1,488		1,398	

Table 3 presents the total, direct, and indirect treatment effects using various semiparametric and parametric methods along with bootstrap standard errors (based on 1999 bootstrap draws) and p-values (based on the t-statistic). The total LATE on the compliers is estimated by IPW using the instrument propensity score as outlined in Frölich (2007) and Tan (2006). The semiparametric estimators of the direct and indirect effects $\hat{\theta}(1)$, $\hat{\theta}(0)$, $\hat{\delta}(1)$, $\hat{\delta}(0)$ based on the sample analogs of Theorem 1 are identical to those used in Section 4. As in the simulations, we also considered trimmed versions of the estimators. Since the point estimates were unaffected and the bootstrap p-values very similar, we do not report those results in the table. The final two columns provide the results for the parametric IV estimators $\hat{\theta}_{\text{para}}$ and $\hat{\delta}_{\text{para}}$, as described in Section 4. The results suggest that the total effect of education on social functioning is positive. The LATE is roughly 3 points and significant at the 10% level. When looking at the point estimates of the direct and indirect effects, this effect appears to be mainly driven by the direct channel, while the impact of the indirect income mechanism is generally much closer to zero. Despite their non-negligible magnitude, the semiparametric direct effects are, however, very imprecise and far from

being statistically significant at any conventional level. On the other hand, their sizes by and large match with the parametric direct effect, which is significant at the 10% level. We therefore conclude that education appears to affect social functioning mostly through mechanisms other than income.

Table 3: BHSP application, cohorts 1945-65 ($n = 2886$)

		semiparametric estimation				parametric	
	LATE	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$	$\hat{\theta}_{\text{para}}$	$\hat{\delta}_{\text{para}}$
estimate	3.094	3.747	4.118	-1.025	-0.653	3.072	0.071
s.e.	1.728	16.762	35.483	35.284	16.762	1.786	0.452
p-value	0.073	0.823	0.908	0.977	0.969	0.085	0.876

Note: Standard errors (s.e.) are based on 1999 bootstrap replications.

As a robustness check, we estimate the effects in a second sample in which we increase the time window around 1956 (the birth year of the first cohort affected by the schooling reform) for cohorts considered in the analysis to 15 years. This implies that all individuals born between (and including) 1940 and 1970 are part of the evaluation sample ($n = 4107$). The idea is to investigate the sensitivity of the estimates with respect to the cohorts included, as a too large time window could lead to confounding of the schooling law instrument by cohort or age effects. That is, cohort and age, which are deterministically related to the instrument, could also directly affect social functioning. Albeit our check is not a formal test for IV validity, implausibly large differences in the effects under different time windows would nevertheless cast doubts on the usefulness of the instrument. Table 4 shows that the LATE and the parametric direct effect are quite similar to the previous results and with the larger sample size now highly significant (at the 0.1% level).[15] The semiparametric direct and indirect effects, on the other hand, are again very imprecise.

We subsequently discuss several methods for the (partial) testability of the IV assumptions in the evaluation sample used in Table 3. Firstly, the independence of Z_1 and Z_2 given X implied by Assumptions 1 and 2 can be easily tested. Linearly regressing windfall income on the schooling reform instrument separately among females and males yields p-values of 0.344 and

[15] In contrast, using a time window of just 5 years entails an explosion of the variances of all estimators. The results are not reported but available upon request.

Table 4: BHSP application, cohorts 1940-70 ($n = 4107$)

		semiparametric estimation				parametric	
	LATE	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$	$\hat{\theta}_{\text{para}}$	$\hat{\delta}_{\text{para}}$
estimate	2.750	2.960	0.689	2.061	-0.210	2.784	-0.005
s.e.	0.668	2.857	11.853	11.803	2.776	0.701	0.206
p-value	0.000	0.300	0.954	0.861	0.940	0.000	0.980

Note: Standard errors (s.e.) are based on 1999 bootstrap replications.

0.351, respectively, so that independence cannot be rejected at conventional levels of significance. Secondly, we use the methods suggested in Kitagawa (2008) and Huber and Mellace (2014) to jointly test whether (i) Z_1 is independent of U given X, which is a further implication of Assumptions 1 and 2, and (ii) D is monotonic in Z_1 as required in Assumption 3.

Using the approach of Huber and Mellace (2014) and keeping conditioning on X implicit, we test the following constraints which need to be satisfied under independence and monotonicity:

$$E(Y|D=1,Z_1=1,Y\leq y_q) \leq E(Y|D=1,Z_1=0) \leq E(Y|D=1,Z_1=1,Y\geq y_{1-q})$$
$$E(Y|D=0,Z_1=0,Y\leq y_r) \leq E(Y|D=0,Z_1=1) \leq E(Y|D=0,Z_1=0,Y\geq y_{1-r}) \quad (54)$$

where (under our identifying assumptions) $q = \frac{\Pr(D=1|Z_1=0)}{\Pr(D=1|Z_1=1)}$ is the share of always treated in the population with $D = 1$, $Z_1 = 1$, and $r = \frac{\Pr(D=0|Z_1=1)}{\Pr(D=0|Z_1=0)}$ is the share of never treated in the population with $D = 0$, $Z_1 = 0$. As discussed in Huber and Mellace (2014) in more detail, the intuition of the test is that $E(Y|D = 1, Z_1 = 0)$ and $E(Y|D = 0, Z_1 = 1)$ point identify the mean potential outcomes (as a function of D) of the always and never treated under $D = 1, 0$, respectively. At the same time, the mean potential outcomes of the latter groups can be bounded in the (mixed) populations with $D = 1$, $Z_1 = 1$ and $D = 0$, $Z_1 = 0$, respectively, which also contain compliers. One can therefore test whether the points lie within the bounds as postulated in (54). We do so by applying the minimum p-value-based test of Chen and Szroeter (2012) for multiple inequality constraints. The distribution of the test statistic is estimated by bootstrapping (1999 bootstrap draws) and relies on pre-estimating which inequality constraints are (close to being)

violated to increase testing power in finite samples.[16]

Alternatively, Kitagawa (2008) suggests testing the following constraints on the joint probabilities of Y and M given Z_1 (again, conditioning on X is implicit):

$$\begin{aligned} \Pr(Y \in A, D = 1 | Z_1 = 1) &\geq \Pr(Y \in A, D = 1 | Z_1 = 0) \\ \Pr(Y \in A, D = 0 | Z_1 = 0) &\geq \Pr(Y \in A, D = 0 | Z_1 = 1) \end{aligned} \tag{55}$$

where A denotes a subset of the support of Y. The intuition for the first constraint is that under $D = 1$, the joint probability of having a particular outcome ($Y \in A$) and belonging to the always treated or compliers ($Z_1 = 1$) must not be larger than the corresponding joint probability of always treated alone ($Z_1 = 0$). An equivalent argument holds for the never treated in the second constraint. Note that this must hold for any A, so that depending on the definition of the subsets, multiple constraints can be tested. Kitagawa (2008) proposes a test that makes use of a two sample Kolmogorov-Smirnov-type statistic on the supremum of $\Pr(Y \in A, D = 1|Z_1 = 0) - \Pr(Y \in A, D = 1|Z_1 = 1)$ and $\Pr(Y \in A, D = 0|Z_1 = 1) - \Pr(Y \in A, D = 0|Z_1 = 0)$, respectively, across all subsets A. The distribution of the test statistic is estimated by a bootstrap method (or more concisely, by permutation) similar to Abadie (2002). We implement the test using 1999 bootstrap draws with 5, 10, and 20 subsets A, respectively, based on equi-quantile grids over the distribution of Y.

Table 5: P-values of IV validity tests for Z_1

	social functioning (Y)				annual income (M)			
	HM14	K08(5)	K08(10)	K08(20)	HM14	K08(5)	K08(10)	K08(20)
female sample	0.338	0.561	0.554	0.550	0.988	0.812	0.969	0.849
male sample	0.989	0.900	0.966	0.834	0.977	0.861	0.876	0.790

Note: Testing is based on 1999 bootstrap replications.

The left panel of Table 5 reports the p-values of the tests, when testing the schooling law instrument in the male and female samples. HM14 refers to the mean-based tests of Huber and

[16] See for instance Andrews and Soares (2010) and Chen and Szroeter (2012) for a more detailed discussion of moment selection based on pre-estimating which constraints are (almost) violated.

Mellace (2014) and K08(s) refers to the Kitagawa (2008) probability-based tests with s subsets, for $s = 5$, 10, and 20. We do not find evidence against the IV validity of Z_1 at conventional levels of significance. As a word of caution, however, note that even asymptotically these tests cannot find all possible violations of the instrument assumptions, because the outcomes of the always/never treated in the mixed populations with compliers are only partially identified. We apply the same methods to test a further implication of Assumptions 1 and 2 (again jointly with monotonicity of D in Z_1), namely that Z_1 is independent of V given X. This can be done by replacing outcome Y by the mediator M in any of the expressions in (54) and (55). Again, IV validity is not rejected, see the right panel of Table 5.

Finally, note that by Assumptions 1 and 2, Z_2 is independent of U given (Z_1, W, X) and thus, given (D, X) (because D is a deterministic function of Z_1, W, X). Under the additional assumption that income M is monotonic in windfall income Z_2 (which appears innocuous), we can therefor use the Huber and Mellace (2014) and Kitagawa (2008) methods to partially test IV validity of windfall income. As the currently available tests only apply to binary instruments and endogenous variables, we, however, need to dichotomize Z_2 and M. Let $\tilde{Z}_2$ and $\tilde{M}$ denote indicators for windfall income larger than zero and income larger than 7,000 GBP, respectively.[17] We test IV validity of windfall income by replacing D by $\tilde{M}$ and Z_1 by $\tilde{Z}_2$ in (54) and (55) and apply the methods in subsamples defined upon X and D. The p-values are again larger than any conventional significance level, see Table 6. We conclude that the various tests do not raise concerns about the validity of the IV assumptions on statistical grounds.

Table 6: P-values of IV validity tests for $\tilde{Z}_2$

	$D = 1$ (more than lower secondary education)				$D = 0$ (lower secondary education or less)			
	HM14	K08(5)	K08(10)	K08(20)	HM14	K08(5)	K08(10)	K08(20)
female sample	0.858	0.750	0.774	0.777	0.813	0.621	0.642	0.613
male sample	0.776	0.232	0.223	0.212	0.971	0.622	0.609	0.641

Note: Testing is based on 1999 bootstrap replications.

[17] We investigated several other cut-off values for the dichotomizations, which did not affect the IV tests in any important way.

5.2 Effects of Job Corps with discrete instrument

In many applications, only discrete instruments are available. To illustrate the use of Theorem 6 in Section 3.5, we in our second application consider a welfare policy experiment conducted in the mid-1990s to assess the publicly funded U.S. Job Corps program, which targets young individuals (aged 16-24 years) who legally reside in the U.S. and come from a low-income household. Participants are provided with approximately 1200 hours of vocational training and education, housing, board, and health services over an average duration of 8 months. Schochet, Burghardt, and Glazerman (2001) and Schochet, Burghardt, and McConnell (2008) discuss in detail the experimental design[18] and the main effects, which suggest that Job Corps increases educational attainment, employment, and earnings, and reduces criminal activity (at least for some years after the program). Several studies have investigated the causal mechanisms through which the program operates based on different identifying assumptions. Flores and Flores-Lagunes (2009) aim at assessing the direct effect on earnings after controlling for the mediator work experience. Assuming the latter to be conditionally exogenous given pre-treatment covariates only, they find a positive direct effect. In contrast, Huber (2013) considers mediator exogeneity conditional on both pre- and post-treatment covariates and estimates the program's direct and indirect (via employment) health effects. Finally, Flores and Flores-Lagunes (2010) invoke considerably weaker assumptions than exogeneity and derive upper and lower bounds for the direct and indirect effects on employment and earnings which are mediated by the achievement of a GED, high school degree, or vocational degree as well as the direct effects. Their approach allows for mediator endogeneity at the price of sacrificing point identification.

We complement these studies by assessing the causal mechanisms of the Job Corps program based on our IV approach. We aim at disentangling the program's earnings effect among female compliers into the indirect effect due to switching from no or part time employment into full time employment and the direct remainder effect. That is, our research question is whether the earnings

[18] In particular, Schochet, Burghardt, and Glazerman (2001) report that the randomization of the program was successful: Of 94 observed pre-treatment covariates, only 5 where statistically significantly different across treatment groups at the 5 % level, which is what one would expect by chance.

effect is indirectly generated by an increased labor force attachment or whether other channels like an increase in human capital play a role, too.[19] The treatment variable D is enrolment in Jobs Corps in the first year after randomization, which is instrumented by the randomized treatment assignment indicator (Z_1). The mediator M is the number of hours worked per week in the third year after randomization, while the outcome Y is weekly earnings in that year. The challenging part is to find a plausible instrument for M. As it is common in the empirical labor literature, we use the number of children in the household who are younger than 6 and younger than 15 two years after random assignment as (discrete) instruments for M. For this reason, we only consider the female sample, whose labor market state is more likely to respond to children in the household. Furthermore, we aim at controlling for potential confounders of the arguably disputable instrument Z_2 by conditioning on a range of pre-assignment characteristics X that are associated with the number of children and also likely affect M and Y: Education, race, age, labor market state and school attendance prior to randomization, and dependence on AFDC or foodstamps (as proxy for socio-economic background). Table 7 reports the OLS coefficients of the number of children under 6 on these variables.

Table 7: OLS regression of the number of children under 6 on X

	estimate	s.e.	t-value	p-value
high school degree at randomization	-0.230	0.034	-6.825	0.000
at least some college at randomization	-0.199	0.075	-2.649	0.008
black	0.179	0.031	5.859	0.000
Hispanic	0.132	0.038	3.439	0.001
age	0.357	0.099	3.598	0.000
age^2	-0.009	0.003	-3.440	0.001
was in school in year before randomization	-0.101	0.030	-3.396	0.001
had a job in year before randomization	-0.119	0.027	-4.376	0.000
AFDC before randomization	0.226	0.033	6.957	0.000
food stamps before randomization	0.137	0.032	4.254	0.000
constant	-2.897	0.959	-3.021	0.003
R^2	0.074			

Our evaluation sample consists of all female Job Corps applicants without missing values in Z_1, Z_2, D, M, Y, X, which gives 4,603 observations. Table 8 provides descriptive statistics

[19] In contrast, the *controlled* direct effect does not appear interesting in this example, because labor supply can typically not be enforced 'from outside', but is chosen by the individuals. It therefore seems irrelevant to assess the direct effect of the program if every one was forced to work full time or part time.

(means and standard deviations) of X, Y, Z_1, and Z_2, separately by the treatment state and the values of the mediator, respectively. The means of several pre-treatment characteristics differ importantly across D and M. Education, age, and going to school in the year before randomization have statistically significant correlations (at the 5% level) with both the treatment and the mediator, whereas ethnicity, having a job in year before randomization, and the receipt of food stamps/AFDC are significantly correlated with the mediator only.

Table 8: Descriptive statistics

	$D=1$		$D=0$		$M=1$		$M=0$	
	mean	std.dev	mean	std.dev	mean	std.dev	mean	std.dev
high school degree at rand. (binary)	0.226	0.418	0.253	0.435	0.324	0.468	0.214	0.410
at least some college at rand. (binary)	0.032	0.176	0.034	0.182	0.056	0.230	0.026	0.158
black (binary)	0.541	0.498	0.535	0.499	0.489	0.500	0.553	0.497
Hispanic (binary)	0.195	0.396	0.180	0.384	0.189	0.391	0.186	0.389
age	18.489	2.180	18.666	2.168	19.070	2.221	18.427	2.136
in school in year before rand. (binary)	0.661	0.473	0.626	0.484	0.607	0.489	0.653	0.476
had a job in year before rand. (binary)	0.615	0.487	0.631	0.483	0.754	0.431	0.581	0.493
AFDC before randomization	0.413	0.492	0.431	0.495	0.373	0.484	0.439	0.496
food stamps before randomization	0.538	0.499	0.559	0.497	0.497	0.500	0.567	0.496
weekly earnings 3^{rd} year after rand. (Y)	143.3	134.4	134.9	143.2	312.4	134.7	81.6	81.6
assignment to Job Corps (Z_1) (binary)	0.992	0.088	0.361	0.481	0.668	0.471	0.638	0.481
# of kids < 6 in 3^{rd} year after rand. (Z_2)	0.765	0.898	0.772	0.900	0.617	0.819	0.819	0.918
# kids < 15 in 3^{rd} year after rand. (Z_2)	1.140	1.248	1.163	1.267	0.907	1.117	1.233	1.292
# of observations	2,074		2,529		1,139		3,464	

The total, direct, and indirect effects are given in Table 9 along with bootstrap standard errors and p-values. Concerning semiparametric estimation based on Theorem 6, the first step estimates (e.g. propensity scores and conditional densities) are computed in the same way as for the estimators based on Theorem 1 whenever applicable (see Section 4 for details). Making use of (50), we first need an estimate of $\frac{E[YD(Z_1-\Pi)|M,C,X]}{E[D(Z_1-\Pi)|M,C,X]}$. The numerator and denominator are obtained from separate linear regressions of $YD(Z_1-\Pi)$ and $D(Z_1-\Pi)$, respectively, on $(1, M, \hat{C}, X)$, where $\hat{C}$ is an estimate of C. Then, the estimated $\frac{E[YD(Z_1-\Pi)|M,C,X]}{E[D(Z_1-\Pi)|M,C,X]}$ is itself linearly regressed on $(1, M, \hat{C}, X)$ to predict $\psi_1(X;\beta)$ in equation (52) by $X'\hat{\beta}$, where $\hat{\beta}$ are the coefficient estimates on X (excluding the constant). The parametric IV estimators are the same as in the simulations and first application.

The results in Table 9 point to a total earnings effect of the program among compliers of roughly 13 USD, which is significant at the 5% level. The total effect seems to be driven mainly by the *indirect* effect. The indirect effect is of a similar magnitude as the total effect, whereas the direct effect is closer to zero. As before, the semiparametric estimates are substantially more noisy than the parametric ones, but point to a similar overall picture. The parametric indirect effect $\hat{\delta}_{\text{para}}$ is significant at the 10% level and the semiparametric indirect effect $\hat{\delta}(1)$ is significant at the 5% level. The direct effects are much smaller in magnitude and never significantly different from zero. Although we cannot draw very strong conclusions, it seems that Job Corps mainly affects labor force attachment through increasing the number of hours worked (indirect channel), whereas the hourly wages themselves do not appear to be much affected.

Table 9: Job Corps application (n=4,603)

		semiparametric estimation				parametric	
	LATE	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$	$\hat{\theta}_{\text{para}}$	$\hat{\delta}_{\text{para}}$
estimate	12.797	-6.855	-1.322	14.119	19.651	-0.824	13.188
s.e.	6.446	16.718	3.787	6.214	17.602	3.540	6.780
p-value	0.047	0.682	0.727	0.023	0.264	0.816	0.052

Note: Standard errors (s.e.) are based on 1999 bootstrap replications. P-values are based on the quantiles

We again briefly discuss testing of the IV assumptions. Firstly, the independence of Z_1 and Z_2 is not rejected at conventional levels when linearly regressing the number of children under 6 or 15 on Z_1 and X. Furthermore, we apply the Kitagawa (2008) and Huber and Mellace (2014) methods to Z_2. To this end, the instrument is dichotomized: $\tilde{Z}_2$ is one if no children under 6 or 15, respectively, are present in the household and zero otherwise. We jointly test whether (i) Z_2 and U are independent and (ii) fulltime employment M monotonically increases in $\tilde{Z}_2$. Here, the tests are performed conditional on D only, rather than X, as the currently available test procedures are not suitable for conditioning on many covariates. Even without controlling for X, all p-values exceed conventional significance levels, see Table 10. This is in line with Huber and Mellace (2013), who test the validity of the number of children instruments for female labor supply in several empirical data sets and found no statistical evidence for its violation either. Finally, we also apply the tests to the randomization indicator Z_1, an a priori rather undisputable

instrument. As expected, the tests yield very large p-values and the results are therefore omitted. We conclude that also in this application, the various checks on Z_2 and Z_1 do not refute the IV assumptions.

Table 10: P-values of IV validity tests of the number of children instrument

	$D = 1$ (Job Corps participation)				$D = 0$ (non-participation)			
	HM14	K08(5)	K08(10)	K08(20)	HM14	K08(5)	K08(10)	K08(20)
$\tilde{Z}_2 = 1$ (No of children under $6 = 0$)	0.933	0.844	0.936	0.782	0.394	0.807	0.633	0.853
$\tilde{Z}_2 = 1$ (No of children under $15 = 0$)	0.977	0.863	0.897	0.704	0.981	1.000	0.995	0.979

Note: Testing is based on 1999 bootstrap replications.

6 Conclusion

Contrary to much of the literature on causal mechanisms relying on conditional exogeneity assumptions, this paper has demonstrated the nonparametric identification of (local) average direct and indirect effects based on (distinct) instruments for the endogenous treatment and the endogenous mediator. Tackling both treatment and mediator endogeneity based on conditionally valid instruments (given observed covariates), we identified natural direct and indirect as well as controlled direct effects on the subpopulation of compliers, whose treatment reacts on the corresponding instrument. In the special case of full compliance, the direct and indirect effects on the total population are obtained. To consider a range of relevant cases, we discussed various approaches that differed in terms of assumptions about the distributions of the mediator and its instrument, monotonicity (of the mediator in its arguments), and support conditions. For further intuition and illustration, a brief simulation study and two applications were also provided.

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.

Ahn, H., and J. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29.

Albert, J. M. (2008): "Mediation analysis via potential outcomes models," *Statistics in Medicine*, 27, 1282–1304.

Albert, J. M., and S. Nelson (2011): "Generalized causal mediation analysis," *Biometrics*, 67, 1028–1038.

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

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

Antman, F. (2011): "International Migration and Gender Discrimination among Children Left Behind," *American Economic Review*, 101(3), 645–49.

Bargain, O., and D. Boutin (2014): "Remittances and Child Labour in Africa: Evidence from Burkina Faso," *IZA Discussion Paper*, 8007.

Baron, R. M., and D. A. Kenny (1986): "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology*, 51, 1173–1182.

Binzel, C., and R. Assaad (2011): "Egyptian men working abroad: Labour supply responses by the women left behind," *Labour Economics*, 18, Supplement 1, 98–114.

Blundell, R., and J. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics*, ed. by L. H. M. Dewatripont, and S. Turnovsky, pp. 312–357. Cambridge University Press, Cambridge.

Brunello, G., D. Fabbri, and M. Fort (2013): "The Causal Effect of Education on Body Mass: Evidence from Europe," *Journal of Labor Economics*, 31, 195–223.

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*, ed. by L. Christofides, E. Grant, and R. Swidinsky, pp. 201–222. University of Toronto Press, Toronto.

Chen, L.-Y., and J. Szroeter (2012): "Testing Multiple Inequality Hypotheses: A Smoothed Indicator Approach," *CeMMAP working paper 16/12*.

Cochran, W. G. (1957): "Analysis of Covariance: Its Nature and Uses," *Biometrics*, 13, 261–281.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

D'Haultfoeuille, X., S. Hoderlein, and Y. Sasaki (2014): "Included Instruments," *Discussion Paper, Boston College*.

Flores, C. A., and A. Flores-Lagunes (2009): "Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness," *IZA DP No. 4237*.

——— (2010): "Nonparametric Partial Identification of Causal Net and Mechanism Average Treatment Effects," *mimeo, University of Florida*.

Frandsen, B., M. Frölich, and B. Melly (2012): "Quantile treatment effects in the regression discontinuity design," *Journal of Econometrics*, 168 (2), 382–395.

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.

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," *forthcoming in the Journal of the American Statistical Association.*

Frölich, M., and B. Melly (2013): "Unconditional quantile treatment effects under endogeneity," *Journal of Business and Economic Statistics (JBES)*, 31:3, 346–357.

Gardner, J., and A. J. Oswald (2007): "Money and mental wellbeing: A longitudinal study of medium-sized lottery wins," *Journal of Health Economics*, 26, 49–60.

Geneletti, S. (2007): "Identifying direct and indirect effects in a non-counterfactual framework," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 199–215.

Hayfield, T., and J. Racine (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27, 1–32.

Heckman, J., R. Pinto, and P. Savelyev (2013): "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, 103, 2052–2086.

Huber, M. (2013): "Identifying causal mechanisms (primarily) based on inverse probability weighting," *forthcoming in the Journal of Applied Econometrics.*

Huber, M., M. Lechner, and C. Wunsch (2013): "The performance of estimators based on the propensity score," *Journal of Econometrics*, 175, 1–21.

Huber, M., and G. Mellace (2013): "Testing exclusion restrictions and additive separability in sample selection models," *forthcoming in Empirical Economics.*

——— (2014): "Testing instrument validity for LATE identification based on inequality moment constraints," *forthcoming in the Review of Economics and Statistics*.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto (2011): "Unpacking the Black Box: Learning about Causal Mechanisms from Experimental and Observational Studies," *Political Science Review*, 105, 765–789.

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51Ű71.

Imai, K., D. Tingley, and T. Yamamoto (2013): "Experimental Designs for Identifying Causal Mechanisms," *Journal of the Royal Statistical Society, Series A*, 176, 5–51.

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

Imbens, G. W., and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512.

Joffe, M. M., D. Small, T. T. Have, S. Brunelli, and H. I. Feldman (2008): "Extended Instrumental Variables Estimation for Overall Effects," *The International Journal of Biostatistics*, 4.

Judd, C. M., and D. A. Kenny (1981): "Process Analysis: Estimating Mediation in Treatment Evaluations," *Evaluation Review*, 5, 602–619.

Kane, T. J., and C. E. Rouse (1995): "Labor-Market Returns to Two- and Four-Year College," *The American Economic Review*, 85, 600–614.

Kasy, M. (2014): "Instrumental variables with unrestricted heterogeneity and continuous treatment," *forthcoming in Review of Economic Studies*.

Kitagawa, T. (2008): "A Bootstrap Test for Instrument Validity in Heterogeneous Treatment Effect Models," *mimeo*.

Krueger, A. B. (1999): "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114, 497–532.

Lindahl, M. (2005): "Estimating the Effect of Income on Health and Mortality Using Lottery Prizes as an Exogenous Source of Variation in Income," *The Journal of Human Resources*, 40, 144–168.

Mattei, A., and F. Mealli (2011): "Augmented Designs to Assess Principal Strata Direct Effects," *Journal of Royal Statistical Society Series B*, 73, 729–752.

Mealli, F., and D. B. Rubin (2003): "Assumptions allowing the estimation of direct causal effects," *Journal of Econometrics*, 112, 79–87.

Mu, R., and D. van de Walle (2011): "Left behind to farm? Women's labor re-allocation in rural China," *Labour Economics*, 18, Supplement 1, 83–97.

Newey, W., J. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

Oreopoulos, P. (2006): "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter," *The American Economic Review*, 96, 152–175.

Pearl, J. (1995): "Causal Diagrams for Empirical Research," *Biometrika*, 82, 669–688.

Pearl, J. (2001): "Direct and indirect effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, San Francisco. Morgan Kaufman.

Petersen, M. L., S. E. Sinisi, and M. J. van der Laan (2006): "Estimation of Direct Causal Effects," *Epidemiology*, 17, 276–284.

Powdthavee, N., W. N. Lekfuangfu, and M. Wooden (2013): "The Marginal Income Effect of Education on Happiness: Estimating the Direct and Indirect Effects of Compulsory Schooling on Well-Being in Australia," *IZA Discussion Paper No. 7365*.

Robins, J. M. (2003): "Semantics of causal DAG models and the identification of direct and indirect effects," in *In Highly Structured Stochastic Systems*, ed. by P. Green, N. Hjort, and S. Richardson, pp. 70–81, Oxford. Oxford University Press.

Robins, J. M., and S. Greenland (1992): "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 3, 143–155.

Rose, H., and J. Betts (2004): "The effect of high school courses on earnings," *Review of Economics and Statistics*, 86, 497–513.

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

——— (2004): "Direct and Indirect Causal Effects via Potential Outcomes," *Scandinavian Journal of Statistics*, 31, 161–170.

Schochet, P. Z., J. Burghardt, and S. Glazerman (2001): "National Job Corps Study: The Impacts of Job Corps on ParticipantsŠ Employment and Related Outcomes," *Report (Washington, DC: Mathematica Policy Research, Inc.)*.

Schochet, P. Z., J. Burghardt, and S. McConnell (2008): "Does Job Corps Work? Impact Findings from the National Job Corps Study," *The American Economic Review*, 98, 1864–1886.

Silverman, B. (1986): *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Simonsen, M., and L. Skipper (2006): "The Costs of Motherhood: An Analysis Using Matching Estimators," *Journal of Applied Econometrics*, 21, 919–934.

Spasojevic, J. (2010): "Effects of Education on Adult Health in Sweden: Results from a Natural Experiment," in *Current Issues in Health Economics*, ed. by R. T. Daniel Slottje, vol. 290 of *Contributions to Economic Analysis*, pp. 179–199. Emerald Group Publishing Limited, Current Issues in Health Economics.

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

Ten Have, T. R., M. M. Joffe, K. G. Lynch, G. K. Brown, S. A. Maisto, and A. T. Beck (2007): "Causal mediation analyses with rank preserving models," *Biometrics*, 63, 926–934.

van der Weele, T. J. (2009): "Marginal Structural Models for the Estimation of Direct and Indirect Effects," *Epidemiology*, 20, 18–26.

Yamamoto, T. (2013): "Identification and Estimation of Causal Mediation Effects with Treatment Noncompliance," *unpublished manuscript, MIT Department of Political Science*.