ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 5, *pp.* 1645–1666

Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables

Markus Frölich

Center for Evaluation and Development, Mannheim, and University of Mannheim, Germany

and Martin Huber

University of Fribourg, Switzerland

[Received June 2015. Final revision February 2017]

Summary. The paper discusses the non-parametric 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 by using distinct instruments for the endogenous treatment and the endogenous mediator. We examine various settings to obtain non-parametric identification of (natural) direct and indirect as well as controlled direct effects for continuous and discrete mediators and continuous and discrete instruments. We also provide a simulation study and two empirical illustrations.

Keywords: Direct effect; Indirect effect; Instrument; Treatment effects

1. Introduction

A range of empirical studies focus on assessing the total effect of a treatment on an outcome of interest, such as the average treatment effect. However, in many applications, not only the average treatment effect 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 and the *indirect* effect that runs through an intermediate variable or so-called mediator. Early work on the evaluation of causal mechanisms or mediation analysis (see for instance Cochran (1957), Judd and Kenny (1981) and Baron and Kenny (1986)) typically relied on linear models. More recent research has focused on non-parametric and semiparametric identification, e.g. Pearl (2001), Robins (2003), Hong (2010), Imai *et al.* (2010), Tchetgen Tchetgen and Shpitser (2012) and Huber (2014). Most studies assume that the treatment and the mediator are exogenous given observed covariates.

In this paper, we analyse the non-parametric identification of causal mechanisms via instrumental variables (IVs) and permit both treatment and mediator endogeneity to be related to unobserved confounders. (This is an abridged version of Frölich and Huber (2014a).) We make use of two distinct IVs to control for either endogeneity problem. In our heterogeneous treatment effect model with a binary treatment, identification relies on particular monotonicity and exogeneity assumptions of the instruments, which might hold only conditionally given a set

Address for correspondence: Markus Frölich, Department of Economics, University of Mannheim, L7, 3–5, Mannheim 68131, Germany.
E-mail: froelich@uni-mannheim.de

 1369–7412/17/791645

of observed covariates. The methods proposed enable disentangling the so-called local average treatment effect (LATE) on the compliers into direct and indirect effects. As special cases, our results also cover the scenarios of a random treatment, which corresponds to a situation with perfect compliance, and of unconditional instrument validity, implying that one need not control for covariates. Our identification strategies consider various settings with either a continuous or a discrete mediator and a continuous or a discrete instrument for the mediator.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Model and parameters of interest

2.1. Direct and indirect effects in non-parametric model

We are interested in disentangling the total effect of a *binary* treatment D on an outcome Y into a direct effect and an indirect effect operating through some scalar mediator M. (Extensions to vector-valued mediators are possible but would require additional IVs.) Identification will be based on two instruments Z_1 and Z_2 for the endogenous variables D and M. We consider the following structural model consisting of a system of non-separable non-parametric equations:

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

where φ, ζ and χ are unknown functions. $\mathbf{1}(\cdot)$ is the indicator function which is equal to 1 if its argument is true and 0 otherwise. U, V and W comprise unobservables and may be arbitrarily associated, so that the treatment and the mediator are in general endogenous. X are other covariates. (Note that the X-variables are permitted to be correlated with the unobservables. In principle, (some of) the X-variables may even be causally affected by the treatment (post-treatment confounders), as long as the IV assumptions below are not violated. In principle, we could further permit different sets of X-variables in each of the equations, which would complicate the notation in the independence assumptions considerably, though.) Z_1 is the instrument for treatment D, which is henceforth denoted as the first instrument, whereas Z_2 denotes the instrument for mediator M, which is referred to as the second instrument hereafter.

In this paper, we assume that Z_1 is *binary*, which includes the special case of a binary randomization indicator in an experiment with imperfect compliance. Concerning the second instrument, we consider both discrete and continuous Z_2. Identification of the (total) LATE has been shown in Imbens and Angrist (1994) and Angrist *et al.* (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 (but could in principle run via further mediators other than M). Two endogeneity problems arise in this context. The first stems from the permitted association between W and U, even after conditioning on X, and is tackled by the first instrument Z_1. A second issue is that the mediator 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.

To ease our discussion we make use of the potential outcomes framework. Let Y^d and 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: $Y^{d,M^{d'}}$. In terms of our model, these parameters are defined for $d, d' \in \{0, 1\}$ as

$$M_i^d \equiv \zeta(d, Z_{2i}, X_i, V_i),$$
$$Y_i^{d,M^{d'}} \equiv \varphi(d, M_i^{d'}, X_i, U_i) = \varphi\{d, \zeta(d', Z_{2i}, X_i, V_i), X_i, U_i\}.$$

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

$$D_i(z_1) = \mathbf{1}\{\chi(z_1, X_i, W_i) \geqslant 0\}.$$

As discussed in Angrist *et al.* (1996), the population can be categorized into four subpopulations or types (denoted by T), according to the treatment behaviour as a function of the first instrument: the *always-takers* ($T_i = \text{at}$) take treatment irrespectively of Z_1, i.e. $D_i(0) = D_i(1) = 1$. The *never-takers* ($T_i = \text{nt}$) do not take treatment irrespectively of Z_1, i.e. $D_i(0) = D_i(1) = 0$. The *compliers* ($T_i = \text{co}$) take treatment only if Z_1 is 1, i.e. $D_i(0) = 0$ and $D_i(1) = 1$. Finally, the *defiers* ($T_i = \text{de}$) take treatment only if Z_1 is 0, i.e. $D_i(0) = 1$ and $D_i(1) = 0$. We shall assume that the last group has probability mass 0, 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 treatment model defined in expression (1) to $D = \mathbf{1}\{\chi(Z_1, Z_2, X, W) \geqslant 0\}$. This model is more general as it permits the second instrument to influence D also and bears some similarities with the idea of an 'included instrument' in D'Haultfoeuille *et al.* (2014). The main implication of this extension is that T_i is a function of Z_{2i}, X_i and W_i. Since all subsequent identification approaches make use of only 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.)

We now define the effects of interest: (natural) direct and indirect, as well as controlled direct effects among compliers. The total average effect among compliers corresponds to the LATE, which is also known as the complier average causal effect:

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

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

$$\theta(d) = E[Y^{1,M^d} - Y^{0,M^d} | T = \text{co}], \qquad \text{for } d \in \{0,1\}. \tag{2}$$

The *indirect* effect among compliers is the mean difference when exogenously shifting the mediator to its potential values with and without treatment, but keeping the treatment fixed at $D = d$:

$$\delta(d) = E[Y^{d,M^1} - Y^{d,M^0} | T = \text{co}], \qquad \text{for } d \in \{0,1\}. \tag{3}$$

(Because expressions (2) and (3) refer to the compliers alone, they are local versions of the natural or pure or total direct and indirect effects that were discussed in Robins and Greenland (1992), Robins (2003) and Pearl (2001) respectively. For convenience, we shall simply refer to them as direct and indirect effects in the subsequent discussion.)

The *controlled* direct effect is the mean difference when exogenously varying the treatment, but setting the mediator to a particular value, say m, rather than the potential mediator state:

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

i.e., contrary to the (natural) direct effect, which is the direct effect 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.

2.2. Relationship to the literature

Most IV approaches in the mediation literature use a single instrument and therefore cover less general problems than those analysed in this paper. Robins and Greenland (1992) and Geneletti (2007) considered an exogenous treatment and an endogenous mediator with a 'perfect' instrument that forces the mediator to take a particular (and desired) value. This is equally attractive as directly manipulating the mediator exogenously; see the discussion in Imai *et al.* (2013). Perfect instruments are, however, rare in applications. Ten Have *et al.* (2007) also assumed treatment exogeneity but exploited treatment–covariate interactions as instruments for the mediator while imposing the absence of treatment–mediator, mediator–covariate and treatment–covariate interactions in the outcome model, such that identification comes from structural restrictions. (The idea of creating instruments by interacting random-treatment assignment with covariates was also discussed in Gennetian *et al.* (2002).) See also Dunn and Bentall (2007), Albert (2008) and Small (2012) for related approaches. In contrast, no restrictions on interactions are imposed in our approach, where instruments reflect variables rather than functional form assumptions.

Imai *et al.* (2013) discussed non-parametric identification in experiments (again with an exogenous treatment) based on imperfect and discrete instruments for the mediator. One particular design identifies the indirect effect among individuals whose mediator reacts to the instrument ('mediator compliers'). (See their Section 4.2 on crossover encouragement designs or the corresponding discussion in Imai *et al.* (2011). Also Mattei and Mealli (2011) considered a random treatment and a binary instrument for the mediator to derive bounds on direct effects within principal strata defined on potential mediator states.) In contrast, our paper permits both treatment and mediator endogeneity. Secondly, our assumptions are sufficiently strong to identify the effects on all treatment compliers rather than the subgroup of mediator compliers (among treatment compliers). Under specific assumptions, this is even so for a binary mediator.

Joffe *et al.* (2008) assumed a single instrument that jointly affects the treatment and the mediator and discussed identification under particular structural restrictions. However, in a non-parametric framework, a single instrument for both endogeneity problems is generally not sufficient for identification. An exception is Yamamoto (2013), who considered identification based on an instrument for the treatment and a latent ignorability assumption that was similar to Frangakis and Rubin (1999) with respect to the mediator. This requires the mediator to be exogenous conditionally on treatment compliance (and observed covariates). The present work does not rely on this restriction, but on distinct instruments for the treatment and the mediator.

Powdthavee *et al.* (2013), Burgess *et al.* (2015) and Jhun (2015) are among the few studies using two instruments, but they considered parametric models that do not permit treatment–mediator interaction effects and thus heterogeneity in direct and indirect effects. In contrast, our non-parametric results allow for heterogeneous effects across treatment states and observed covariates. Miquel (2002) and Blackwell (2015) considered a non-parametric framework with two binary instruments and showed the identification of controlled direct effects for subpopulations defined on compliance in either endogenous variable. (Note, however, that Blackwell (2015) did not allow for causal effects of one endogenous variable on another, so he considered a multiple treatment rather than a mediation framework.) In contrast, we identify both natural and controlled effects for all treatment compliers by imposing stronger assumptions on the second instrument than did Miquel (2002). (Our paper is also related to the literature on triangular systems, which mostly does not consider mediators or direct and indirect effects. An important exception is Jun *et al.* (2016), who assumed a model of the type (adapted to our notation and a binary treatment)

$$Y = \varphi\{\alpha(D, M, Z_3), U\},$$
$$M = \zeta(D, Z_2, V),$$
$$D = \mathbf{1}\{\chi(Z_1) \geqslant W\},$$

where α is an unknown real-valued function of d, m and z_3. They assumed a discrete M (and imposed some further structure on ζ), a discrete D and $(Z_1, Z_2, Z_3) \perp\!\!\!\perp (U, V, W)$ and $E[\varphi(\alpha, U)|V = v, W = w]$ is strictly monotonic in α for all v and w, along with some support conditions. A main distinction from our model is the single-index structure and the monotonicity of the outcome in the index. Furthermore, a special regressor Z_3 is required, which is excluded from the mediator and treatment equations and must be sufficiently powerful (at least partly) to offset the effects of D and M on α. In contrast, our outcome equation (1) is unrestricted. However, we rely on monotonicity restrictions with respect to M that are not required in Jun *et al.* (2016).)

3. Identifying direct and indirect effects

We shall focus on the identification of $E[Y^{1,M^0}|T = \text{co}]$ and $E[Y^{1,m}|T = \text{co}]$, whereas the further potential outcomes can be obtained in an analogous way to identify $\theta(d)$, $\delta(d)$ and $\gamma(m)$. (Furthermore, by replacing $Y^{1,m}$ with $\mathbf{1}(Y^{1,m} \leqslant a)$ in all expressions we obtain the cumulative distribution function $F_{Y^{1,m}|T=\text{co}}(a)$ that is required for quantile treatment effects or other inequality measures such as the Gini coefficient.)

3.1. Instrumental variables assumptions common to several identification approaches

Our first assumption imposes particular independence restrictions on the instruments conditionally on X and is (for ease of exposition) slightly stronger than needed for the various lemmas and theorems to follow. Letting the symbol '$\perp\!\!\!\perp$' denote statistical independence, we make the following assumptions.

Assumption 1 (IV independence).

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

Note that assumption 1 would be implied by the following, slightly stronger assumption:

$$(Z_1, Z_2) \perp\!\!\!\perp (U, V, W)|X. \tag{4}$$

The main difference is that assumption 1 permits some specific forms of dependence between Z_2 and W which are discussed in Appendix A by means of causal graphs, whereas assumption (4) does not. (As W determines T, allowing for dependence between Z_2 and W can be relevant in applications where Z_2 is not randomized but depends on D. Assumption 1 also allows for an association between Z_1 and W, as long as it vanishes when conditioning on Z_2. Condition (4) is not required for any results, but if it holds it implies that the probability of complying does not depend on Z_2. This is testable, as $\Pr(T = \text{co}|Z_2, X)$ is identified further below. It further implies that $Z_2 \perp\!\!\!\perp D|X, Z_1$. Hence, if both assumptions seem plausible, they may be used to test identification partially.)

For some (but not all) of our identification results we additionally require that the two instruments Z_1 and Z_2 are independent of each other conditionally on X.

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

Assumption 2 holds by construction in experiments if both instruments are independently randomized. If only Z_1 is randomized, it is also satisfied if Z_2 is assigned at the same time as or shortly before Z_1 because, in experiments, any prerandomization variable is independent of the randomization indicator Z_1. Even in observational studies, we may attain independence via a transformation of Z_2, even if Z_1 and Z_2 are not (conditionally) independent. (Suppose that Z_2 is continuously distributed with a *strictly* increasing cumulative distribution function. Define $\tilde{Z}_{2i} = \Phi^{-1}\{F_{Z_2|Z_1,X}(Z_{2i}, Z_{1i}, X_i)\}$, where Φ is the cumulative distribution function of the standard normal distribution and Φ^{-1} its quantile function. $\tilde{Z}_2|Z_1, X$ is standard normal with mean 0 and variance 1 and thus *independent* of Z_1 (and X). We can thus use $\tilde{Z}_2$ instead of Z_2 as second instrument throughout, which satisfies assumption 2. Hence, assumption 2 is a normalization rather than a substantial restriction. In practice, however, $F_{Z_2|Z_1,X}$ must be estimated, which probably makes effect estimation less reliable.)

In addition to independence, identification requires particular monotonicity assumptions. Assumption 3 imposes monotonicity of D in Z_1, which rules out defiers, and the existence of compliers; see also Imbens and Angrist (1994) and Angrist *et al.* (1996).

Assumption 3 (weak monotonicity of treatment choice).

$$\begin{aligned} &\Pr(T = \text{de}) = 0, \\ &\Pr(T = \text{co}) > 0. \end{aligned} \tag{5}$$

Assumptions 1 and 3 enable us to identify the fraction of compliers. To ease notational burden, we shall use the following symbols for the conditional instrument probabilities: $\Pi = \pi(X) = \Pr(Z_1 = 1|X)$ and $\bar{\Pi} = \bar{\pi}(Z_2, X) = \Pr(Z_1 = 1|Z_2, X)$. (In settings imposing also assumption 2, i.e. that $Z_1 \perp\!\!\!\perp Z_2|X$, we have that $\bar{\pi}(Z_2, X) = \pi(X)$ throughout.) Under assumptions 1 and 3, the probability mass of compliers is identified as

$$\Pr(T = \text{co}) = E\left[\frac{D}{\bar{\Pi}} \frac{Z_1 - \bar{\Pi}}{1 - \bar{\Pi}}\right]. \tag{6}$$

In the following sections we examine identification for various settings. In Sections 3.2 and 3.3 we consider a continuously distributed M and Z_2 and impose monotonicity of M in the unobservable V, which leads to a control function approach. For the controlled effect in Section 3.3 we may either invoke monotonicity in V or in Z_2 for identification, which even yields testable implications. In Sections 3.4 (natural effects) and 3.5 (controlled effects) we consider a binary M and continuous Z_2, which is more demanding in terms of identification. Finally, Section 3.6 assumes a continuous M for which only a discrete Z_2 is available.

3.2. Natural effects with continuous M and Z_2

We first consider the case of a continuous mediator M, and we exploit a control function approach that allows shifting D independently from movements in the mediator.

Assumption 4 (monotonicity of mediator (control function restriction)).

(a) V is a continuously distributed random variable with a cumulative distribution function $F_{V|X=x,T=\text{co}}(v)$ that is strictly increasing in the support of V, for almost all values of x.

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

Assumption 4 is quite strong, as it requires that V is scalar (or at least that the unobservables affecting M can be transformed into a scalar index V that satisfies the mediator equation in expression (1)) and continuously distributed with no values with zero densities in its support conditionally on X and $T = \text{co}$. This assumption is crucial for identifying effects among all (treatment) compliers. Invoking strict monotonicity of M in V enables pinning down the distribution function of V given X among compliers by means of the conditional distribution of M given D, Z_2 and X among compliers. For this, we define the control function $C_i = C(M_i, D_i, Z_{2i}, X_i)$, with

$$C(m, d, z_2, x) = \frac{E[(d + D - 1)\{Z_1 - \bar{\pi}(z_2, x)\} | M \leqslant m, Z_2 = z_2, X = x]}{E[D\{Z_1 - \bar{\pi}(z_2, x)\} | Z_2 = z_2, X = x]} F_{M|Z_2,X}(m, z_2, x). \quad (7)$$

Control function C identifies V_i, as shown in the following lemma.

Lemma 1. Under assumptions 1, 3 and 4 it follows that

(a) $C_i = F_{M|D,Z_2,X,T=\text{co}}(M_i, D_i, Z_{2i}, X_i) = F_{V|X=X_i,T=\text{co}}(V_i)$,
(b) $V_i = F^{-1}_{V|X=X_i,T=\text{co}}(C_i)$ and
(c) $M \perp\!\!\!\perp U | C, X, T = \text{co}$.

Part (a) of lemma 1 shows that the control function corresponds to the distribution function of V among compliers conditionally on X. Part (b) shows that C_i is a one-to-one mapping of V_i, i.e., conditionally on X and $T = \text{co}$, V is a one-to-one function of C, and V is thus identified. Therefore, conditioning on C or V is equivalent. Part (c) shows that by controlling for C (in addition to X) we can separate M from U in the outcome equation within the complier subpopulation.

Intuitively, the key idea of our identification approach is to vary Z_1 to affect D, while keeping M unchanged through a variation of Z_2 that undoes the effect of Z_1 on M. For this, we need to condition on V, which is replaced by its control function C:

$$\begin{aligned} E[Y^{1,M^0} | T = \text{co}] &= \int \varphi(1, M^0, X, U) \, \mathrm{d}F_{M^0,U|X,C,T=\text{co}} \, \mathrm{d}F_{X,C|T=\text{co}} \\ &= \int \varphi(1, M^0, X, U) \, \mathrm{d}F_{U|X,C,T=\text{co}} \, \mathrm{d}F_{M^0|X,C,T=\text{co}} \, \mathrm{d}F_{X,C|T=\text{co}}, \end{aligned}$$

where the last equation follows as M^0 is independent of U conditionally on X and C by assumption 1 and lemma 1. To identify the distribution of M^0, we require $M^0 \perp\!\!\!\perp Z_1 | X, C, T = \text{co}$, which is implied by $Z_2 \perp\!\!\!\perp Z_1 | X, V, T$. It follows that $F_{M^0|X,C,T=\text{co}} = F_{M|Z_1=0,X,C,T=\text{co}}$ and thus equals

$$\int \varphi(1, M, X, U) \, \mathrm{d}F_{U|X,C,T=\text{co}} \, \mathrm{d}F_{M|Z_1=0,X,C,T=\text{co}} \, \mathrm{d}F_{X,C|T=\text{co}}.$$

As $\mathrm{d}F_{M|Z_1=1,X,C,T=\text{co}}$ is identifiable (see the on-line appendix), we may multiply and divide by it to obtain

$$\int \{\varphi(1, M, X, U) \, \mathrm{d}F_{U|X,C,T=\text{co}}\} \, \omega(M, X, C) \, \mathrm{d}F_{M|Z_1=1,X,C,T=\text{co}} \, \mathrm{d}F_{X,C|T=\text{co}}. \quad (8)$$

The weighting function $\omega(M, X, C)$ corresponds to a ratio of conditional densities of M under $Z_1 = 0$ *versus* $Z_1 = 1$, i.e.

$$\omega(M, X, C) = \frac{\mathrm{d}F_{M|Z_1=0,X,C,T=\mathrm{co}}}{\mathrm{d}F_{M|Z_1=1,X,C,T=\mathrm{co}}},$$

which is equal to

$$1 - \frac{E[Z_1|M, C, X] - \pi(X)}{E[DZ_1|M, C, X] - E[D|M, C, X]\pi(X)};$$

see the on-line appendix. Using $U \perp\!\!\!\perp (M, Z_1)|X, C, T = \mathrm{co}$ by lemma 1

$$\int \left\{\varphi(1, M, X, U)\,\mathrm{d}F_{U|M,X,C,Z_1=1,T=\mathrm{co}}\right\}\omega(M, X, C)\,\mathrm{d}F_{M|X,C,Z_1=1,T=\mathrm{co}}\,\mathrm{d}F_{X,C|T=\mathrm{co}}$$
$$= E\left[YDZ_1\frac{\omega(M, X, C)}{\Pr(Z_1 = 1|X)}|T = \mathrm{co}\right]. \tag{9}$$

Formula (9) shows identification of the counterfactual outcome by reweighting among compliers. Yet, since the compliers are unknown, we require an expression for the entire population that is equal to 0 in the always- and never-taker populations. As shown in the on-line appendix,

$$E\left[\left\{\frac{YDZ_1}{\Pr(Z_1 = 1|X)} - \frac{YD(1 - Z_1)}{\Pr(Z_1 = 0|X)}\right\}\omega(M, X, C)\right] \tag{10}$$

satisfies this condition and equals equation (9) multiplied by $\Pr(T = \mathrm{co})$, which is identified by assumption 1. By estimating expression (10) and dividing by $\Pr(T = \mathrm{co})$, we obtain equation (9), which gives $E[Y^{1,M^0}|T = \mathrm{co}]$.

From equation (8), we can see the support condition that is required for identification. It must hold that $\mathrm{d}F_{M|Z_1=1,X,C,T=\mathrm{co}} > 0$ at every m where $\mathrm{d}F_{M|Z_1=0,X,C,T=\mathrm{co}} > 0$ or, in other words, that $\mathrm{supp}(M|Z_1 = 0, X, C, T = \mathrm{co}) \subseteq \mathrm{supp}(M|Z_1 = 1, X, C, T = \mathrm{co})$. Noting that, given $Z_1, X, T = \mathrm{co}$, variation in C comes from Z_2 alone, this implies that the second instrument must be both sufficiently rich and strong for relevant combinations of Z_1 and X among compliers to ensure common support. This may fail in many empirical applications when fully non-parametric specifications of C and $\omega(M, X, C)$ are used. (Imbens and Newey (2009), for instance, documented common support issues in their empirical application when using non-parametric control functions, albeit in a somewhat different methodological context.) Assuming parametric functions (permitting extrapolation) may alleviate such issues at the cost of imposing more structure. An alternative way of expressing common support is $\Pr(Z_1 = 1|M, C, X, T = \mathrm{co}) > 0$ almost surely. Because of the unique mapping between C and V (see lemma 1), this is equivalent to the following condition.

Assumption 5 (common support of M).

$$\Pr(Z_1 = 1|M, V, X, T = \mathrm{co}) > 0 \qquad \text{almost surely.} \tag{11}$$

Assumption 5 is equivalent to requiring that the weights $\omega(M, X, C)$ do not approach ∞. If in an empirical application some (estimated) weights are extremely large, this could indicate the violation of the support condition (at least in the sample at hand). One could then redefine the objects of interest on subsets of the support spaces of M, X and C for which common support holds.

Theorem 1. Under assumptions 1–5 the potential outcome is identified as

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

with weights

$$\Omega = \omega(M, C, X) = 1 - \frac{E[Z_1|M, C, X] - \pi(X)}{E[DZ_1|M, C, X] - E[D|M, C, X]\pi(X)}$$

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 $\Pr(T = \text{co})$ is identified by expression (6). The proof is provided in the on-line appendix.

In the special case that all individuals comply with their treatment assignment, $Z_1 = D$ and $\Pr(T = \text{co}) = 1$. In this case, Z_1 may be replaced with D everywhere in theorem 1, and it follows that $E[Y^{1,M^0}|T = \text{co}] = E[Y^{1,M^0}]$, as everyone is a complier if $\Pr(T = \text{co}) = 1$.

The potential outcome in theorem 1 can be estimated by replacing the expectation by a sample mean and plugging in estimates of Ω and Π. The estimator is of an inverse probability weighting type where the weights are products of various conditional means. One can apply the approach of Newey (1994) to show that the estimated potential outcome is $\sqrt{n}$ consistent and asymptotically normal (implying the validity of the bootstrap) under certain conditions. First, all terms in the denominator of theorem 1 must be strictly bounded away from zero and their respective estimators uniformly consistent. Furthermore, the bias terms of any non-parametric plug-in estimates (e.g. conditional density functions) must be sufficiently small. For kernel-based estimation, the structure of derivations for showing $\sqrt{n}$-consistency is similar to Frölich and Huber (2014b), implying that the plug-in estimates must have a rate of convergence that is faster than $n^{-1/4}$. Theorem 1 contains several conditional means where the highest dimensional non-parametric component conditions on $\dim(X) + 2$ (possibly continuous) covariates. Using a product kernel function that is compactly supported, bounded, Lipschitz, integrating to 1, and of order λ, it needs to hold that $nh^{2\lambda} \to 0$ and $nh^{\dim(X)+2}/\ln(n) \to \infty$. These conditions jointly require that $2\lambda > \dim(X) + 2$. This implies that, if X contains a single regressor, a second-order kernel can be used; otherwise the components in Ω must be estimated on the basis of higher order kernels. All conditional means in Ω must be $\lambda - 1$ times continuously differentiable with the $(\lambda - 1)$th derivative Hölder continuous. An asymptotically linear expression can then be derived similarly to Frölich and Huber (2014b).

3.3. Controlled direct effects with continuous M and Z_2

We consider the identification of the controlled direct effect for the mediator fixed at m. In contrast with the natural direct effect, knowledge of the distribution of M^d is not required, which allows assumption 2 to be dropped so that dependence between Z_2 and Z_1 is permitted. 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 imposes monotonicity in Z_2 instead of V. Before presenting the formal results, we provide some intuition for identification.

3.3.1. Control function approach

$E[Y^{1,m}|T = \text{co}]$ can also be expressed as $E[\varphi(1, m, X, U)|T = \text{co}]$. Note that

$$\begin{aligned} E[Y^{1,m}|T = \text{co}] &= \int \varphi(1, m, X, U)\, \mathrm{d}F_{U|X,C,T=\text{co}}\, \mathrm{d}F_{X,C|T=\text{co}} \\ &= \int \varphi(1, m, X, U)\, \mathrm{d}F_{U|M=m,Z_1=1,X,C,T=\text{co}}\, \mathrm{d}F_{X,C|T=\text{co}} \\ &= \int E[Y|M = m, Z_1 = 1, X, C, T = \text{co}]\, \mathrm{d}F_{X,C|T=\text{co}}, \end{aligned} \tag{12}$$

because $U \perp\!\!\!\perp (Z_1, Z_2)|X, V, T = \text{co}$. Finally, estimable expressions for $E[Y|M, Z_1, X, C, T = \text{co}]$ and $\mathrm{d}F_{X,C|T=\text{co}}$ based on observable variables can be obtained as outlined in the on-line appendix.

For these derivations, we require the support condition $\text{supp}(X, C|T = \text{co}) \subseteq \text{supp}(X, C|M = m, Z_1 = 1, T = \text{co})$ or, equivalently, that the conditional mediator density $f_{M|X,C,Z_1=1,T=\text{co}}(m, x, c) > 0$ at every value x and c where $f_{X,C|T=\text{co}}(x, c)$ is positive. The one-to-one relationship between C and V implies that $f_{M|X,V,Z_1=1,T=\text{co}}(m, x, v) > 0$ whenever $f_{X,V|T=\text{co}}(x, v)$ is positive. In other words, $f_{M|X,V,Z_1=1,T=\text{co}}(m, X, V)$ must be positive almost everywhere.

Assumption 6 (common support).

$$f_{M|X,V,Z_1=1,T=\text{co}}(m) > 0 \qquad \text{almost surely.} \tag{13}$$

In terms of $M = \zeta(D, Z_2, X, V)$, this assumption requires that, for every x and v in the support of X and V between compliers, there is (at least) one z_2 with positive density such that $\zeta(1, z_2, x, v) = m$. As assumption 6 can be written as $f_{M|X,C,Z_1=1,T=\text{co}}(m) > 0$ almost surely, this is testable.

Theorem 2. Under assumptions 1, 3, 4 and 6

$$E[Y^{1,m}|T = \text{co}] = \frac{1}{\Pr(T = \text{co})} \int E\left[YD\frac{Z_1 - \bar{\Pi}}{1 - \bar{\Pi}}\Omega|X, M = m\right]\mathrm{d}F_X$$

with weights

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

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

3.3.2. Identification via monotonicity in Z_2

Instead of imposing assumption 4, we alternatively assume monotonicity of the mediator in Z_2.

Assumption 7 (monotonicity of the mediator in the instrument). $\zeta(d, z_2, x, v)$ is strictly monotonic in z_2 for almost all d, x and v. We normalize ζ to be *increasing*.

With ζ strictly monotonic in z_2, the 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. (This is a different inverse function from that 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 $Q \equiv \zeta^{-1}(1, m, X, V)$, which is a stochastic function of the two random variables X and V. Hence, Q is governed by the distributions of X and V so, conditionally on X, the only stochastic component in Q is V. We use this fact in the expression

$$E[Y^{1,m}|T = \text{co}] = \int\int \varphi(1, m, X, U)\,\mathrm{d}F_{U|Q,X,T=\text{co}}\,\mathrm{d}F_{Q|X,T=\text{co}}\,\mathrm{d}F_{X|T=\text{co}}. \tag{14}$$

As $(U, V) \perp\!\!\!\perp (Z_1, Z_2)|X, T = \text{co}$ we obtain $\mathrm{d}F_{U|Q,X,T=\text{co}} = \mathrm{d}F_{U|Q,Z_1,Z_2,X,T=\text{co}}$ and $\mathrm{d}F_{Q|X,T=\text{co}} = \mathrm{d}F_{Q|Z_1,Z_2,X,T=\text{co}}$. The functions on the right-hand side are equivalent to $F_{U|Q,Z_1,Z_2,X,T=\text{co}}(u, q, 1, q, x) = F_{U|M=m,Z_1=1,Z_2=q,X=x,T=\text{co}}(u)$ and $F_{Q|X,T=\text{co}}(q, x) = 1 - F_{M|Z_1=1,Z_2,X,T=\text{co}}(m, q, x)$; see the on-line appendix. Therefore, we obtain $f_{Q|X,T=\text{co}}(q, x) = -\partial F_{M|Z_1=1,Z_2,X,T=\text{co}}(m, q, x)/\partial q$.

Identification of the density functions requires that $\text{supp}(Z_2|X, T = \text{co}) \supseteq \text{supp}(Q|X, T = \text{co})$,

i.e., whenever Q has positive density, Z_2 also must have positive density such that Q is observable in that area of the support. Put differently, for every x and v in the support of X and V, there is 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 equation (14) yields

$$\begin{aligned} E[Y^{1,m}|T=\text{co}] &= \int\int \varphi(1,m,x,u)\,\mathrm{d}F_{U|M=m,Z_1=1,Z_2=q,X=x,T=\text{co}}(u) \\ &\quad \times \left\{ -\frac{\partial F_{M|Z_1=1,Z_2,X,T=\text{co}}(m,q,x)}{\partial q} \right\} f_{X|T=\text{co}}\,\mathrm{d}q\,\mathrm{d}x \\ &= \int E[Y|M=m,Z_1=1,Z_2=z_2,X=x,T=\text{co}] \\ &\quad \times \left\{ -\frac{\partial F_{M|Z_1=1,Z_2,X,T=\text{co}}(m,z_2,x)}{\partial z_2} \right\} f_{X|T=\text{co}}(x)\,\mathrm{d}z_2\,\mathrm{d}x. \end{aligned}$$

For making equation (15) operational, we need to identify $F_{M|Z_1,Z_2,X,T=\text{co}}$, which is derived in the on-line appendix.

Theorem 3. Under assumptions 1, 3, 6 and 7

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

with weights

$$\begin{aligned} \Omega = \omega(Z_2,X) = -\frac{\partial}{\partial z_2} \left\{ \frac{E[D(Z_1-\bar{\Pi})|M\leqslant m, Z_2, X]}{E[D(Z_1-\bar{\Pi})|Z_2,X]} F_{M|Z_2,X}(m) \right\} \\ \times \frac{1}{f_{Z_2|X}} \frac{E\left[\dfrac{D}{\bar{\Pi}} \dfrac{Z_1-\bar{\Pi}}{1-\bar{\Pi}} | X \right]}{E\left[D\dfrac{Z_1-\bar{\Pi}}{1-\bar{\Pi}} | M=m, Z_2, X \right]}. \end{aligned} \tag{15}$$

3.4. Natural effects with discrete M and continuous Z_2

In the previous sections, identification was achieved by controlling for $f_{M^d|V,X,T=\text{co}}$ (via variation in Z_2), in particular by weighting with $f_{M^0|V,X,T=\text{co}}/f_{M^1|V,X,T=\text{co}}$. With M being discrete, observations need to be weighted by $\Pr(M^0|V,X,T=\text{co})/\Pr(M^1|V,X,T=\text{co})$. However, V is no longer identified for a discrete M such that $\Pr(M^d|V,X,T=\text{co})$ is not either. An alternative is a weighting scheme that produces $\Pr(M^0|V,X,T=\text{co})/\Pr(M^1|V,X,T=\text{co})$ on average, via integration with respect to Z_2. The price to pay is stronger identifying assumptions. We focus on the case of a *binary* M, which implies the model

$$\left.\begin{aligned} Y &= \varphi(D,M,X,U), \\ M &= \mathbf{1}\{\zeta(D,Z_2,X,V)\geqslant 0\}, \\ D &= \mathbf{1}\{\chi(Z_1,X,W)\geqslant 0\}. \end{aligned}\right\} \tag{16}$$

In addition to assumptions 1–3, identification requires strengthening the monotonicity condition.

Assumption 8 (monotonicity of mediator in the instrument and the unobservable).

(a) V is a continuously distributed random variable with a cumulative distribution function $F_{V|X=x,T=\text{co}}(v)$ that is strictly increasing in the support of V, for almost all values of x,

(b) $\zeta(d, z_2, x, v)$ 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 in parametric models such as probit and logit), implying that the values of z_2 can be ordered such that a model of type (16) exists. (Although monotonicity in v (which is not directly testable) is a fundamental assumption, monotonicity in z_2 (which implies testable restrictions on observed variables) is needed only 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 expression (16) with assumption 8.) As we show in the on-line appendix, under assumptions 1, 2, 3 and 8, expression (16) can be rewritten as

$$\begin{aligned} Y &= \varphi(D, M, X, U), \\ M &= \mathbf{1}[\mu_{D,X}^{-1}\{F_{V|X,\mathrm{co}}(V)\} \leqslant Z_2], \\ D &= \mathbf{1}\{\chi(Z_1, X, W) \geqslant 0\}, \end{aligned}$$

where $\mu_{d,x}^{-1}(v)$ is strictly monotonically *decreasing* in v and is defined as the inverse function of

$$\mu_{d,x}(z_2) = \frac{E[(1-M)(Z_1 - E[Z_1|X=x])|D=d, Z_2=z_2, X=x]}{E[Z_1 - E[Z_1|X=x]|D=d, Z_2=z_2, X=x]}.$$

Theorem 4. Under assumptions 1, 2, 3, 5 and 8

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

with weights

$$\Omega = \frac{f_{Z_2|X,T=\mathrm{co}}[\mu_{0,X}^{-1}\{\mu_{1,X}(Z_2)\}]}{f_{Z_2|X,T=\mathrm{co}}(Z_2)}\frac{\mu'_{1,X}(Z_2)}{\mu'_{0,X}[\mu_{0,X}^{-1}\{\mu_{1,X}(Z_2)\}]},$$

where $\mu'_{d,x}(z_2) \equiv \mathrm{d}\mu_{d,x}(z_2)/\mathrm{d}z_2$. The proof is provided in the on-line appendix.

The weights Ω are obtained by first estimating the functions $\mu_{d,x}(z_2)$ and the density of Z_2. (In a single-index model, $M = \mathbf{1}\{\zeta(\alpha D + \beta Z_2 + \gamma X + V) \geqslant 0\}$ where ζ represents an unknown monotonic function and α, β and γ denote unknown coefficients, the weights can be shown to simplify to $\Omega = f_{Z_2|X,T=\mathrm{co}}(Z_2 + \alpha/\beta)/f_{Z_2|X,T=\mathrm{co}}(Z_2)$.) The conditional density of Z_2 in the complier subpopulation is identified as

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

3.5. Controlled direct effects with discrete M and continuous Z_2

The identification of the controlled direct effect appears difficult, as the control function approach fails (because of the non-identifiability of V) and a strategy similar to that in Section 3.4 is not applicable. Identification requires that there are values of Z_2 for which M attains a particular value m with probability 1. This case is discussed in the on-line appendix.

3.6. Natural effects with continuous M and discrete Z_2

In this section, we discuss identification when both Z_1 and Z_2 are *discrete*, and M is continuous.

(The results are also applicable when Z_2 is continuous but rest on stronger assumptions than those in previous sections.) If Z_2 is discrete, common support as postulated in assumption 5 generally fails such that the approach of Section 3.2, which consisted of varying Z_1 to change D while keeping M unchanged through a variation of Z_2 that undoes the effect of Z_1 on M, is not applicable. However, identification is feasible if the IV validity does not hinge on conditioning on X, such that variation in X may be used to set M to appropriate values in the $Z_1 = 1$ and $Z_1 = 0$ populations. This requires X to be exogenous but allows replacing assumption 5 by the weaker assumption 9.

Assumption 9 (common support of M).

$$\Pr(Z_1 = 1 | M, C, T = \text{co}) > 0 \qquad \text{almost surely.}$$

A further requirement is that X is structurally separated from M. We assume that the outcome equation is additively separable in X, whereas the other equations remain unrestricted:

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

As both Z_1 and Z_2 are discrete, X must contain (at least) one continuous variable. Finally, our conditional independence assumptions need to be strengthened to embrace exogeneity of X.

Assumption 10 (exogeneity assumption).

$$X \perp\!\!\!\perp Z_1, \qquad X \perp\!\!\!\perp (U, V) | T = \text{co}.$$

(Assumptions 1, 2 and 10 jointly imply that $Z_1 \perp\!\!\!\perp (Z_2, X, U, V, T)$ and $(Z_1, Z_2, X) \perp\!\!\!\perp (U, V) | T = \text{co}$.)

Identification is outlined in the on-line appendix. For example, for ψ (which is required in theorem 5), it is shown that

$$\frac{E[YD(Z_1 - \Pi) | M = m, C = c, X = x]}{E[D(Z_1 - \Pi) | M = m, C = c, X = x]} = \Xi(m, c) + \psi(1, x), \tag{17}$$

where $\Xi(m, c) \equiv E[\varphi(1, m, U) | T = \text{co}, C = c]$ is an unknown function of m and c. 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. One may for instance estimate the model

$$\hat{Y}_i = \Xi(M_i, C_i) + \psi_1(X_i; \beta) + \epsilon_i, \tag{18}$$

using partially linear semiparametric regression, where $\hat{Y}_i$ is an estimate of the left-hand side of equation (17), Ξ an unknown non-parametric function, $\psi_1(x; \beta)$ a parametric function and ϵ_i the error. (Identifying conditions are more complicated for a non-parametric ψ. See lemma 2 (in the on-line appendix) for one possibility.)

Theorem 5. Under assumptions 1, 2, 3, 4, 9 and 10 and identification of $\psi(1, X)$

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

with weights

$$\Omega = \omega(M, C) = \frac{E[(D-1)\{Z_1 - \Pr(Z_1 = 1)\}|M, C]}{E[D\{Z_1 - \Pr(Z_1 = 1)\}|M, C]}.$$

4. Simulation study

The following simulation study provides some intuition for the results of theorems 1 and 5. The data-generating process when considering theorem 1 is given by

$$\begin{aligned} Y &= D + M + \beta DM + 0.5X + U, \\ M &= \alpha Z_2 + 0.5D + 0.5X + V, \\ D &= \mathbf{1}(\alpha Z_1 + 0.5X + W > 0.5\alpha), \\ Z_1 &= \mathbf{1}(0.5X + P > 0), \\ Z_2 &= 0.5X + Q, \\ (U, V, W) &\sim \mathcal{N}(\mu, \sigma), \end{aligned}$$

where $\mu = \mathbf{0}$ 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 Q are standard normal, independent of each other and of U, V and W.

M and D are endogenous because of the non-zero correlation of U, V and W. β gauges the interaction effect of D and M on Y, i.e. whether direct and indirect effects are heterogeneous across treatments, whereas α determines IV power of the binary Z_1 and continuous Z_2. We run 1000 simulations and set β either to 0 (no interaction) or 1 and consider $\alpha = 1, 2, 3$, entailing complier shares of 35%, 63% and 82% respectively. The sample sizes are $n = 1250$ and $n = 5000$.

We investigate several estimators of natural direct and indirect effects. The first approach is semiparametric and based on the sample analogues of theorem 1. For this, plug-in estimates of $\pi(X)$, $E[Z_1|M, C, X]$, $E[D|M, C, X]$ and $E[DZ_1|M, C, X]$ are obtained by probit regressions. Estimation of the control function C_i is based on expression (7), in which $\bar{\pi}(Z_2, X)$ is, however, replaced by $\pi(X)$, which is permitted because theorem 1 invokes assumption 2. We use ordinary least squares (OLS) to estimate $E[D\{Z_1 - \pi(X)\}|Z_2, X]$ and $E[(d + D - 1)\{Z_1 - \pi(X)\}|M \leqslant m, Z_2, X]$. Concerning the latter, regression on $(1, Z_2, X)$ is performed in the subset of observations satisfying $M \leqslant m$ in the data, with $m = M_i$ (i.e. the value of M for the ith observation in the data) if prediction is for observation i. This implies underidentification 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 40, implying that $m > M_i$ for the 39 observations with the lowest values of M. Finally, $F_{M|Z_2,X}(m, z_2, x)$, which enters expression (7), is obtained by non-parametric kernel estimation of conditional distributions by using the np package of Hayfield and Racine (2008) and the Silverman (1986) rule of thumb for bandwidth selection. We consider both untrimmed and trimmed versions of the estimators. Similarly to Huber *et al.* (2013) and Frölich and Huber (2014b), the trimmed versions discard observations that would obtain a relative weight that is larger than 5% in the estimation of some mean potential outcome.

Secondly, we examine multistep parametric IV estimation similarly to Powdthavee *et al.* (2013). In the first step, we run a probit regression of D on $(1, Z_1, X)$ to predict the treatment, which is denoted by $\tilde{D}$. Then, we linearly regress M on $(1, Z_2, \tilde{D}, X)$ to predict M, which is 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. condition (4)). 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 the latter regression and that on $\tilde{M}$ in the regression of Y. (If Y and M are linear in D and, thus, in its prediction, either linear or non-linear models might be used to predict D depending on its distribution. However, if M or Y were not linear in D (or Y not linear in M), an estimation strategy based on predicted residuals rather than predicted endogenous variables (see for instance Terza *et al.* (2008)), or a maximum likelihood approach would need to be chosen to avoid inconsistency.) In contrast with semiparametric estimation, this estimator does not allow interaction effects between M and D. Finally, we include a naive OLS approach neither considering confounding due to unobservables or X, nor interaction effects. The direct effect is given by the coefficient on D in a regression of Y on $(1, D, M)$; the indirect effect by that on M in the last regression times the coefficient on D when regressing M on $(1, D)$.

Table 1 presents the bias, standard deviation sd and root-mean-squared error RMSE of the various estimators of $\theta(d)$ and $\delta(d)$ (see expressions (2) and (3)) for $\beta = 0$ (no interactions) when varying the sample size and IV strength. Whereas OLS is severely biased, the correctly specified parametric IV estimators parIV are almost unbiased and competitive in terms of RMSE in any scenario. Semiparametric estimation without trimming, semIV, performs very poorly when $\alpha = 1$ and $n = 1250$. Trimming, semIVtr, improves the performance and entails a smaller bias than does OLS, but yet a substantially higher RMSE than parametric IV estimation and OLS. The competitiveness of semiparametric estimation increases in the IV strength and sample size, whereas the importance of trimming decreases (i.e. trimming has little effect in set-ups with larger α and n). Trimmed estimation based on theorem 1 dominates OLS when $\alpha \geqslant 2$ whereas, for $\alpha = 3$ and $n = 5000$, both the trimmed and the untrimmed versions perform almost as decently as the parametric IV estimator.

The situation changes with effect heterogeneity. Table 2 reports the results for $\beta = 1$ (effect heterogeneity). Biases are non-negligible for OLS and the (now misspecified) parametric IV estimator, but relatively small for the semiparametric IV methods when $\alpha \geqslant 2$. For $n = 5000$ and $\alpha \geqslant 2$, at least the trimmed estimators based on theorem 1 uniformly outperform the parametric IV method in terms of RMSE, implying that the gains in terms of reduction in bias outweigh the losses in efficiency.

Finally, we consider estimation based on theorem 5, when Z_1 is independent of the covariates and Z_2 is discrete. For this, we change the specifications of Z_1 and Z_2 of the data-generating process:

$$Z_1 = \mathbf{1}(P > 0),$$
$$Z_2 = \text{round}(0.5X + Q) \qquad Q \sim \mathcal{U}(-2, 2),$$

where $\mathcal{U}$ stands for the uniform distribution and 'round' rounds its argument to the next integer such that Z_2 is discrete. Whenever possible, the same first step estimators as for estimation based on theorem 1 are used in the procedure based on theorem 5, whereas the estimate of the numerator and denominator of the left-hand side of equation (17) as well as of equation (18) is based on OLS.

Table 3 reports the results for $\beta = 1$ and $\alpha = 1, 2$, which qualitatively match those in Table 2: semiparametric methods become more competitive as the sample size and IV strength increase (and trimming is important in scenarios with small α and n). All in all, the simulation results

Table 1. Bias, standard deviation and RMSE for $\beta = 0$†

Method	$\theta(1)$			$\theta(0)$			$\delta(1)$			$\delta(0)$		
	Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
$\alpha=1$, $n=1250$												
semIV	3.542	126.618	126.668	−16.598	531.847	532.106	16.580	531.857	532.115	−3.559	126.613	126.663
semIVtr	0.087	0.994	0.998	0.112	0.745	0.753	−0.128	0.744	0.755	−0.103	1.001	1.006
parIV	−0.006	0.178	0.178	−0.006	0.178	0.178	−0.010	0.245	0.245	−0.010	0.245	0.245
OLS	0.529	0.060	0.533	0.529	0.060	0.533	1.941	0.120	1.945	1.941	0.120	1.945
$\alpha=1$, $n=5000$												
semIV	−0.160	7.526	7.528	0.826	11.808	11.836	−0.834	11.812	11.842	0.152	7.522	7.524
semIVtr	−0.043	0.931	0.932	0.082	0.914	0.918	−0.090	0.905	0.909	0.035	0.940	0.940
parIV	−0.004	0.080	0.081	−0.004	0.080	0.081	−0.001	0.121	0.121	−0.001	0.121	0.121
OLS	0.528	0.029	0.529	0.528	0.029	0.529	1.940	0.057	1.941	1.940	0.057	1.941
$\alpha=2$, $n=1250$												
semIV	−0.061	0.922	0.924	0.032	0.715	0.716	−0.040	0.718	0.719	0.053	0.949	0.951
semIVtr	−0.022	0.289	0.290	0.006	0.226	0.226	−0.014	0.259	0.260	0.014	0.366	0.366
parIV	−0.002	0.098	0.098	−0.002	0.098	0.098	−0.006	0.215	0.215	−0.006	0.215	0.215
OLS	0.541	0.061	0.545	0.541	0.061	0.545	2.006	0.165	2.013	2.006	0.165	2.013
$\alpha=2$, $n=5000$												
semIV	0.042	1.011	1.012	0.032	0.120	0.124	−0.035	0.143	0.148	−0.046	1.018	1.019
semIVtr	−0.012	0.178	0.178	0.032	0.110	0.114	−0.036	0.135	0.139	0.008	0.207	0.207
parIV	−0.002	0.044	0.044	−0.002	0.044	0.044	−0.000	0.107	0.107	−0.000	0.107	0.107
OLS	0.539	0.029	0.540	0.539	0.029	0.540	2.001	0.082	2.003	2.001	0.082	2.003
$\alpha=3$, $n=1250$												
semIV	−0.008	0.225	0.225	−0.024	0.188	0.190	0.018	0.273	0.274	0.002	0.333	0.333
semIVtr	−0.008	0.225	0.225	−0.017	0.151	0.152	0.010	0.240	0.240	0.002	0.333	0.333
parIV	−0.001	0.076	0.076	−0.001	0.076	0.076	−0.005	0.236	0.236	−0.005	0.236	0.236
OLS	0.421	0.062	0.425	0.421	0.062	0.425	2.080	0.223	2.092	2.080	0.223	2.092
$\alpha=3$, $n=5000$												
semIV	−0.005	0.068	0.068	0.003	0.063	0.063	−0.005	0.118	0.118	0.002	0.129	0.129
semIVtr	−0.005	0.068	0.068	0.003	0.063	0.063	−0.005	0.118	0.118	0.002	0.129	0.129
parIV	−0.002	0.034	0.034	−0.002	0.034	0.034	−0.000	0.118	0.118	−0.000	0.118	0.118
OLS	0.419	0.029	0.420	0.419	0.029	0.420	2.074	0.112	2.077	2.074	0.112	2.077

†Results are based on 1000 simulations: semIV, semiparametric IV estimation based on theorem 1 without trimming; semIVtr, semiparametric IV estimation based on theorem 1 with trimming; parIV, parametric IV estimation. The true effects under $\beta=0$ are $\theta(1)=\theta(0)=1$ and $\delta(1)=\delta(0)=0.5$. The complier share is 35% for $\alpha=1$, 63% for $\alpha=2$ and 82% for $\alpha=3$. Conditionally on X and Z_1, Z_2 explains 29%, 52% and 62% of the total variation of Y (i.e. the total sum of squares) in a linear regression, respectively for $\alpha=1, 2, 3$.

Table 2. Bias, standard deviation and RMSE for $\beta = 1$†

Method	$\theta(1)$			$\theta(0)$			$\delta(1)$			$\delta(0)$		
	Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
$\alpha = 1$, $n = 1250$												
semIV	3.544	126.625	126.674	−11.367	370.675	370.850	11.353	370.690	370.864	−3.559	126.613	126.663
semIVtr	0.090	1.034	1.038	0.256	1.044	1.075	−0.270	1.089	1.122	−0.103	1.001	1.006
parIV	−0.254	0.231	0.343	0.246	0.231	0.337	−0.266	0.367	0.453	0.234	0.367	0.435
OLS	0.277	0.086	0.290	0.777	0.086	0.782	2.415	0.170	2.421	2.915	0.170	2.920
$\alpha = 1$, $n = 5000$												
semIV	−0.162	7.528	7.530	0.737	16.835	16.851	−0.747	16.845	16.861	0.152	7.522	7.524
semIVtr	−0.045	0.939	0.940	0.195	1.232	1.247	−0.205	1.237	1.254	0.035	0.940	0.940
parIV	−0.254	0.107	0.276	0.246	0.107	0.268	−0.252	0.182	0.311	0.248	0.182	0.308
OLS	0.276	0.043	0.280	0.776	0.043	0.777	2.413	0.083	2.415	2.913	0.083	2.915
$\alpha = 2$, $n = 1250$												
semIV	−0.062	0.948	0.950	0.018	0.872	0.873	−0.027	0.917	0.918	0.053	0.949	0.951
semIVtr	−0.023	0.346	0.347	0.008	0.379	0.379	−0.017	0.491	0.491	0.014	0.366	0.366
parIV	−0.254	0.160	0.301	0.246	0.160	0.293	−0.259	0.323	0.413	0.241	0.323	0.403
OLS	0.287	0.104	0.306	0.787	0.104	0.794	2.610	0.239	2.620	3.110	0.239	3.119
$\alpha = 2$, $n = 5000$												
semIV	0.041	1.015	1.016	0.059	0.171	0.181	−0.064	0.241	0.250	−0.046	1.018	1.019
semIVtr	−0.013	0.198	0.198	0.058	0.164	0.174	−0.063	0.238	0.246	0.008	0.207	0.207
parIV	−0.253	0.076	0.264	0.247	0.076	0.258	−0.250	0.161	0.298	0.250	0.161	0.297
OLS	0.288	0.052	0.293	0.788	0.052	0.790	2.602	0.121	2.605	3.102	0.121	3.105
$\alpha = 3$, $n = 1250$												
semIV	−0.012	0.301	0.302	−0.055	0.388	0.392	0.045	0.547	0.549	0.002	0.333	0.333
semIVtr	−0.012	0.301	0.301	−0.039	0.304	0.306	0.029	0.477	0.478	0.002	0.333	0.333
parIV	−0.253	0.164	0.302	0.247	0.164	0.296	−0.258	0.353	0.438	0.242	0.353	0.428
OLS	0.168	0.132	0.214	0.668	0.132	0.681	2.762	0.327	2.781	3.262	0.327	3.278
$\alpha = 3$, $n = 5000$												
semIV	−0.003	0.126	0.126	0.014	0.123	0.123	−0.015	0.234	0.234	0.002	0.129	0.129
semIVtr	−0.003	0.126	0.126	0.014	0.123	0.123	−0.015	0.234	0.234	0.002	0.129	0.129
parIV	−0.251	0.075	0.262	0.249	0.075	0.260	−0.250	0.177	0.306	0.250	0.177	0.306
OLS	0.168	0.064	0.180	0.668	0.064	0.671	2.754	0.166	2.759	3.254	0.166	3.259

†Results are based on 1000 simulations: semIV, semiparametric IV estimation based on theorem 1 without trimming; semIVtr, semiparametric IV estimation based on theorem 1 with trimming; parIV, parametric IV estimation. The true effects under $\beta = 1$ are $\theta(1) = 1.5$, $\theta(0) = 1$, $\delta(1) = 1$ and $\delta(0) = 0.5$. The complier share is 35% for $\alpha = 1$, 63% for $\alpha = 2$ and 82% for $\alpha = 3$. Conditionally on X and Z_1, Z_2 explains 29%, 52% and 62% of the total variation of Y (i.e. the total sum of squares) in a linear regression, respectively for $\alpha = 1, 2, 3$.

Table 3. Bias, standard deviation and RMSE for $\beta = 1$†

Method	$\theta(1)$			$\theta(0)$			$\delta(1)$			$\delta(0)$		
	Bias	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*	*Bias*	*SD*	*RMSE*
$\alpha = 1, n = 1250$												
semIV	−0.455	17.346	17.352	0.068	0.450	0.455	−0.112	0.641	0.651	0.411	17.360	17.364
semIVtr	−0.153	2.181	2.186	0.064	0.446	0.451	−0.108	0.644	0.653	0.109	2.206	2.208
parIV	−0.256	0.250	0.358	0.244	0.250	0.349	−0.288	0.483	0.562	0.212	0.483	0.527
OLS	0.457	0.093	0.466	0.957	0.093	0.961	1.944	0.211	1.956	2.444	0.211	2.454
$\alpha = 1, n = 5000$												
semIV	−0.021	0.179	0.180	0.083	0.152	0.173	−0.090	0.293	0.306	0.014	0.222	0.223
semIVtr	−0.021	0.179	0.180	0.083	0.152	0.173	−0.090	0.293	0.306	0.014	0.222	0.223
parIV	−0.253	0.122	0.281	0.247	0.122	0.276	−0.257	0.241	0.352	0.243	0.241	0.343
OLS	0.454	0.048	0.457	0.954	0.048	0.956	1.945	0.100	1.948	2.445	0.100	2.447
$\alpha = 2, n = 1250$												
semIV	0.007	0.282	0.282	0.020	0.270	0.271	−0.056	0.632	0.634	−0.042	0.335	0.338
semIVtr	0.007	0.282	0.282	0.020	0.270	0.271	−0.056	0.632	0.634	−0.042	0.335	0.338
parIV	−0.254	0.206	0.327	0.246	0.206	0.320	−0.280	0.471	0.548	0.220	0.471	0.520
OLS	0.353	0.131	0.377	0.853	0.131	0.863	1.636	0.346	1.673	2.136	0.346	2.164
$\alpha = 2, n = 5000$												
semIV	0.025	0.136	0.138	0.037	0.134	0.140	−0.043	0.323	0.326	−0.031	0.166	0.168
semIVtr	0.025	0.136	0.138	0.037	0.134	0.140	−0.043	0.323	0.326	−0.031	0.166	0.168
parIV	−0.250	0.102	0.271	0.250	0.102	0.270	−0.256	0.239	0.350	0.244	0.239	0.341
OLS	0.352	0.066	0.358	0.852	0.066	0.855	1.647	0.171	1.656	2.147	0.171	2.154

†Results are based on 1000 simulations: semIV, semiparametric IV estimation based on theorem 5 without trimming; semIVtr, semiparametric IV estimation based on theorem 5 with trimming; parIV, parametric IV estimation. The true effects under $\beta = 1$ are $\theta(1) = 1.5$, $\theta(0) = 1$, $\delta(1) = 1$ and $\delta(0) = 0.5$. The complier share is 35% for $\alpha = 1$ and 63% for $\alpha = 2$. Conditionally on X and Z_1, Z_2 explains 56% and 63% of the total variation of Y (i.e. the total sum of squares) in a linear regression, respectively for $\alpha = 1, 2$.

suggest that semiparametric estimation can be preferable to fully parametric methods under sufficiently strong instruments and in sufficiently large samples with several 1000 observations.

5. Empirical illustrations

5.1. Empirical illustration based on theorem 1

Our first application is based on theorem 1 and data from the British Household Panel Survey. We aim at assessing the effect of education on the outcome 'social functioning', which reflects the (mental and physical) ability to participate in social life. We distinguish the indirect effect via *income* from the direct effect. The treatment is a binary indicator D which is 1 if an individual has obtained more than lower secondary education according to the international standard classification of education of the United Nations Educational, Scientific and Cultural Organization. D is instrumented by an increase in the UK minimum school leaving age in 1971 from 15 to 16 years, affecting all cohorts born in 1956 or later, Z_1. The change in law induced some 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 in Oreopoulos (2006) and Brunello *et al.* (2013).) To disentangle the effect of education into a direct and an indirect component driven by income, annual individual income (in British

Table 4. Direct and indirect effects on social functioning, British Household Panel Survey application, cohorts 1945–1965

Parameter	*LATE* Δ	*Results for semiparametric estimation*				*Results for parametric estimation*	
		Direct $\hat{\theta}(1)$	*Direct* $\hat{\theta}(0)$	*Indirect* $\hat{\delta}(1)$	*Indirect* $\hat{\delta}(0)$	*Direct* $\hat{\theta}_{\text{para}}$	*Indirect* $\hat{\delta}_{\text{para}}$
Estimate	3.272	3.934	3.472	−0.199	−0.661	3.397	−0.029
Standard error	1.090	11.516	20.142	20.222	11.404	1.165	0.303
p-value	0.003	0.733	0.863	0.992	0.954	0.004	0.925

pounds) serves as mediator M, which 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).) As covariates X we include gender and a dummy for Scotland. (In Frölich and Huber (2014a), we applied the methods of Huber and Mellace (2015) and Kitagawa (2015) to test the IV validity of Z_1 and Z_2. The p-values of all test statistics turned out to be insignificant. The same holds for the instruments of our second application presented below.)

Our empirical illustration is based on the four waves 5, 6, 8 and 9 of the British Household Panel Survey (which started in 1991 with 10 300 individuals), which were conducted in 1995, 1996, 1998 and 1999 respectively. The covariates X are measured in 1995 and educational attainment D is measured in 1996. In wave 8 annual income M and windfall profits Z_2 are measured. Finally, in wave 9 the social functioning index Y is measured. We restrict the sample to observations born between (and including) 1944 and 1967, i.e. at most 12 years before or after the beginning of 1956, the year of the first cohort that was affected by the 1971 schooling reform. We refer to the on-line appendix for descriptive statistics on our evaluation sample, which contains $n = 3428$ observations.

Table 4 presents the (total) LATE as well as the direct and indirect effects by using semiparametric and parametric IV methods along with bootstrap standard errors (based on 999 bootstrap draws) and p-values (based on the t-statistic). The LATE is estimated by weighting based on the (parametric) instrument propensity score; see Frölich (2007) and Tan (2006). The semiparametric estimators of the direct and indirect effects $\hat{\theta}(1)$, $\hat{\theta}(0)$, $\hat{\delta}(1)$ and $\hat{\delta}(0)$ based on theorem 1 (and theorem 5 further below) are identical to those in the simulations; see Section 4. The final two columns provide the results for the parametric IV estimators $\hat{\theta}_{\text{para}}$ and $\hat{\delta}_{\text{para}}$ that were also considered in the simulations. The results show a positive effect of education on social functioning: the LATE amounts to roughly 3 points and is significant at the 1% level. Whereas the semiparametric indirect effects are close to 0, the direct effects are similar in magnitude to the total effect (although rather imprecise). They are similar in size to the parametric estimates, where the direct effect is highly significant and the indirect effect is close to 0. We therefore conclude that education affects social functioning mostly through mechanisms other than income.

5.2. *Empirical illustration based on theorem 5*

To illustrate estimation based on theorem 5, we consider a welfare policy experiment that was conducted in the 1990s to assess the US job corps programme; see Schochet *et al.* (2001, 2008).

Table 5. Effect of the job corps programmes on earnings ($n = 4603$)

Parameter	*LATE* Δ	*Results for semiparametric IV estimation*				*Results for parametric IV estimation*	
		Direct $\hat{\theta}(1)$	*Direct* $\hat{\theta}(0)$	*Indirect* $\hat{\delta}(1)$	*Indirect* $\hat{\delta}(0)$	*Direct* $\hat{\theta}_{\text{para}}$	*Indirect* $\hat{\delta}_{\text{para}}$
Estimate	12.797	−6.855	−1.322	14.119	19.651	−0.824	13.188
Standard error	6.325	50.205	3.519	6.030	50.343	3.405	6.572
p-value	0.043	0.891	0.707	0.019	0.696	0.809	0.045

The programme targets young individuals (aged 16–24 years) from low income households, providing them with vocational training and education, housing, board and health services. The treatment D is enrolment in the programme in the first or second year after randomization, which is instrumented by randomized treatment assignment, Z_1. The mediator M reflects hours worked per week in the third year after randomization; the outcome Y is weekly earnings in that year. As is common in labour economics, the numbers of children in the household who are younger than 6 and younger than 15 years serve as (discrete) instruments Z_2 for M, and only the female sample is considered in our analysis. Furthermore, we control for several covariates X that potentially confound Z_2: education, race, age, labour market state and school attendance before randomization and dependence on 'Aid to families with dependent children' or food stamps.

The IV assumptions underlying Z_1 appear plausible in our empirical context. As it is randomly assigned, it is per design unrelated to unobservables affecting the treatment, mediator or outcome as postulated in assumption 1 or to X as postulated in assumption 10. Furthermore, it seems credible that mere assignment does not directly affect the wage outcome such that the exclusion restriction holds and that D is weakly monotonic in Z_1 (assumption 3). Finally, regressing Z_1 on Z_2 and X yields statistically insignificant coefficients and therefore does not point to a violation of assumption 2. The IV validity of Z_2 is arguably more disputable. The presence of small children certainly is not a purely random event, and we aim to mitigate this by controlling for background characteristics X. Assumption 4 is satisfied if hours worked increase strictly monotonically in an unobserved index that reflects the unobserved eagerness to work. This residual 'eagerness to work' must be unrelated to the control variables X. Basically, we need to assume that any unobservables affecting hours of work (such as ability and motivation) can be split into a part that is related to X (such as the average ability that is associated with people with a certain amount of education) and residual unobservables that are independent of education. This assumption was not needed in Section 5.1 where identification was based on theorem 1.

Our sample consists of all female job corps applicants without missing values in Z_1, Z_2, D, M, Y and X. The on-line appendix provides descriptive statistics about the 4603 observations. Table 5 presents the effect estimates. The LATE amounts to roughly \$13 and is significant at the 5% level. Both $\hat{\delta}_{\text{para}}$ and $\hat{\delta}(1)$ are of a similar magnitude to that of the LATE and significant (in contrast with $\hat{\delta}(0)$, which is quite imprecise), whereas the direct effects are closer to 0 and never significantly different from 0. Our results therefore suggest that the job corps programme mainly affects earnings indirectly through increasing hours worked, rather than hourly wages.

Acknowledgements

We have benefitted from comments by Kosuke Imai and Teppei Yamamoto and seminar participants in Zurich (Eidgenössiche Technische Hochschule research seminar) and Laax (em-

pirical labour seminar). We are grateful to Thomas Aeschbacher for his excellent research assistance. The first author acknowledges financial support from the Research Center *Sonderforschungsbereich* 884 'Political economy of reforms' project B5, funded by the German Research Foundation. The second author acknowledges financial support from Swiss National Science Foundation grant PBSGP1 138770.

References

Albert, J. M. (2008) Mediation analysis via potential outcomes models. *Statist. Med.*, **27**, 1282–1304.
Angrist, J., Imbens, G. and Rubin, D. (1996) Identification of causal effects using instrumental variables (with discussion). *J. Am. Statist. Ass.*, **91**, 444–472.
Baron, R. M. and Kenny, D. A. (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Persnlty Socl Psychol.*, **51**, 1173–1182.
Blackwell, M. (2015) Identification and estimation of joint treatment effects with instrumental variables. *Working Paper*. Department of Government, Harvard University, Cambridge.
Brunello, G., Fabbri, D. and Fort, M. (2013) The causal effect of education on body mass: evidence from Europe. *J. Lab. Econ.*, **31**, 195–223.
Burgess, S., Daniel, R. M., Butterworth, A. S. and Thompson, S. G. (2015) Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *Int. J. Epidem.*, **44**, 484–495.
Cochran, W. G. (1957) Analysis of covariance: its nature and uses. *Biometrics*, **13**, 261–281.
D'Haultfoeuille, X., Hoderlein, S. and Sasaki, Y. (2014) Included instruments. *Discussion Paper*. Boston College, Chestnut Hill.
Dunn, G., and Bentall, R. (2007) Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). *Statist. Med.*, **26**, 4719–4745.
Frangakis, C. and Rubin, D. (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. *J. Econmetr.*, **139**, 35–75.
Frölich, M. and Huber, M. (2014a) Direct and indirect treatment effects: causal chains and mediation analysis with instrumental variables. *Discussion Paper 8280*. Institute for the Study of Labor, Bonn.
Frölich, M. and Huber, M. (2014b) Treatment evaluation with multiple outcome periods under endogeneity and attrition. *J. Am. Statist. Ass.*, **109**, 1697–1711.
Gardner, J. and Oswald, A. J. (2007) Money and mental wellbeing: a longitudinal study of medium-sized lottery wins. *J. Hlth Econ.*, **26**, 49–60.
Geneletti, S. (2007) Identifying direct and indirect effects in a non-counterfactual framework. *J. R. Statist. Soc.* B, **69**, 199–215.
Gennetian, L., Bos, J. and Morris, P. (2002) Using instrumental variables to learn more from social policy experiments. *Working Paper on Research Methodology*. MDRC, New York.
Hayfield, T. and Racine, J. (2008) Nonparametric econometrics: the np package. *J. Statist. Softwr.*, **27**, 1–32.
Hong, G. (2010) Ratio of mediator probability weighting for estimating natural direct and indirect effects. *Proc. Biometr. Sect. Am. Statist. Ass.*, 2401–2415.
Huber, M. (2014) Identifying causal mechanisms (primarily) based on inverse probability weighting. *J. Appl. Econmetr.*, **29**, 920–943.
Huber, M., Lechner, M. and Wunsch, C. (2013) The performance of estimators based on the propensity score. *J. Econmetr.*, **175**, 1–21.
Huber, M. and Mellace, G. (2015) Testing instrument validity for LATE identification based on inequality moment constraints. *Rev. Econ. Statist.*, **97**, 398–411.
Imai, K., Keele, L., Tingley, D. and Yamamoto, T. (2011) Unpacking the black box: learning about causal mechanisms from experimental and observational studies. *Polit. Sci. Rev.*, **105**, 765–789.
Imai, K., Keele, L. and Yamamoto, T. (2010) Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.*, **25**, 51–71.
Imai, K., Tingley, D. and Yamamoto, T. (2013) Experimental designs for identifying causal mechanisms (with discussion). *J. R. Statist. Soc.* A, **176**, 5–51.
Imbens, G. W. and Angrist, J. (1994) Identification and estimation of local average treatment effects. *Econometrica*, **62**, 467–475.
Imbens, G. W. and Newey, W. K. (2009) Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, **77**, 1481–1512.
Jhun, M. A. (2015) Epidemiologic approaches to understanding mechanisms of cardiovascular diseases: genes, environment, and DNA methylation. *Dissertation*. University of Michigan, Ann Arbor.
Joffe, M. M., Small, D., Have, T. T., Brunelli, S. and Feldman, H. I. (2008) Extended instrumental variables estimation for overall effects. *Int. J. Biostatist.*, **4**.

Judd, C. M. and Kenny, D. A. (1981) Process analysis: estimating mediation in treatment evaluations. *Evaln Rev.*, **5**, 602–619.
Jun, S., Pinkse, J., Xu, H. and Yildiz, N. (2016) Multiple discrete endogenous variables in weakly-separable triangular models. *Econometrics*, **4**, 1–21.
Kitagawa, T. (2015) A test for instrument validity. *Econometrica*, **83**, 2043–2063.
Lindahl, M. (2005) Estimating the effect of income on health and mortality using lottery prizes as an exogenous source of variation in income. *J. Hum. Resour.*, **40**, 144–168.
Mattei, A. and Mealli, F. (2011) Augmented designs to assess principal strata direct effects. *J. R. Statist. Soc.* B, **73**, 729–752.
Miquel, R. (2002) Identification of dynamic treatment effects by instrumental variables. *Discussion Paper 2002-11*. Department of Economics, University of St Gallen, St Gallen.
Newey, W. K. (1994) The asymptotic variance of semiparametric estimators. *Econometrica*, **62**, 1349–1382.
Oreopoulos, P. (2006) Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *Am. Econ. Rev.*, **96**, 152–175.
Pearl, J. (2001) Direct and indirect effects. In *Proc. 17th Conf. Uncertainty in Artificial Intelligence*, pp. 411–420. San Francisco: Morgan Kaufmann.
Powdthavee, N., Lekfuangfu, W. N. and Wooden, M. (2013) The marginal income effect of education on happiness: estimating the direct and indirect effects of compulsory schooling on well-Being in Australia. *Discussion Paper 7365*. Institute for the Study of Labor, Bonn.
Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems* (eds P. Green, N. Hjort and S. Richardson), pp. 70–81. Oxford: Oxford University Press.
Robins, J. M. and Greenland, S. (1992) Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143–155.
Schochet, P. Z., Burghardt, J. and Glazerman, S. (2001) National Job Corps Study: the impacts of Job Corps on participants' employment and related outcomes. *Report*. Mathematica Policy Research, Inc., Washington DC.
Schochet, P. Z., Burghardt, J. and McConnell, S. (2008) Does Job Corps Work?: Impact findings from the National Job Corps Study. *Am. Econ. Rev.*, **98**, 1864–1886.
Silverman, B. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
Small, D. S. (2012) Mediation analysis without sequential ignorability: using baseline covariates interacted with random assignment as instrumental variables. *J. Statist. Res.*, **46**, 91–103.
Tan, Z. (2006) Regression and weighting methods for causal inference using instrumental variables. *J. Am. Statist. Ass.*, **101**, 1607–1618.
Tchetgen Tchetgen, E. J. and Shpitser, I. (2012) Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Ann. Statist.*, **40**, 1816–1845.
Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A. and Beck, A. T. (2007) Causal mediation analyses with rank preserving models. *Biometrics*, **63**, 926–934.
Terza, J. V., Basu, A. and Rathouz, P. J. (2008) Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *J. Hlth Econ.*, **27**, 531–543.
Yamamoto, T. (2013) Identification and estimation of causal mediation effects with treatment noncompliance. *Manuscript*. Department of Political Science, Massachusetts Institute of Technology, Cambridge.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Appendix'.