UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

WORKING PAPERS SES

Direct and indirect effects under sample selection and outcome attrition

Martin Huber
and
Anna Solovyeva

N. 496
VIII.2018

Faculté des sciences économiques et sociales
Wirtschafts- und sozialwissenschaftliche Fakultät

Direct and indirect effects under sample selection and outcome attrition

Martin Huber and Anna Solovyeva

Department of Economics, University of Fribourg, Switzerland

Abstract: This paper considers the evaluation of direct and indirect treatment effects, also known as mediation analysis, when outcomes are only observed for a subpopulation due to sample selection or outcome attrition. For identification, we combine sequential conditional independence assumptions on the assignment of the treatment and the mediator, i.e. the variable through which the indirect effect operates, with either selection on observables/missing at random or instrumental variable assumptions on the outcome attrition process. We derive expressions for the effects of interest that are based on inverse probability weighting by specific treatment, mediator, and/or selection propensity scores. We also provide a brief simulation study and an empirical illustration based on U.S. Project STAR data that assesses the direct effect and indirect effect (via absenteeism) of smaller kindergarten classes on math test scores.

Keywords: Causal mechanisms, direct effects, indirect effects, causal channels, mediation analysis, causal pathways, sample selection, attrition, outcome nonresponse, inverse probability weighting, propensity score.

JEL classification: C21, I21.

Address for correspondence: Martin Huber, Anna Solovyeva, University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland; martin.huber@unifr.ch, anna.solovyeva@unifr.ch.

1 Introduction

Following the seminal papers of Judd and Kenny (1981), Baron and Kenny (1986), and Robins and Greenland (1992), the evaluation of direct and indirect effects, also known as mediation analysis, is widespread in social sciences, see for instance the applications in MacKinnon (2008). The aim is to disentangle the total causal effect of a treatment on an outcome of interest into an indirect component operating through one or several intermediate variables, i.e. mediators, as well as a direct component. As example, consider the effect of educational interventions on health, where part of the effect might be mediated by health behaviors, see Brunello, Fort, Schneeweis, and Winter-Ebmer (2016), or personality traits, see Conti, Heckman, and Pinto (2016). While earlier studies on mediation typically rely on tight linear models, the more recent literature considers more flexible and possibly nonlinear specifications. A large number of contributions assumes a 'sequential conditional independence' assumption, implying that the assignment of the treatment and the mediator is conditionally exogenous given observed covariates and given the treatment and the covariates, respectively. For examples, see Pearl (2001), Robins (2003), Petersen, Sinisi, and van der Laan (2006), van der Weele (2009), Flores and Flores-Lagunes (2009), Imai, Keele, and Yamamoto (2010), Hong (2010), Albert and Nelson (2011), Tchetgen Tchetgen and Shpitser (2012), Vansteelandt, Bekaert, and Lange (2012), Zheng and van der Laan (2012), and Huber (2014a), among many others.

Our main contribution is the extension of such mediation models to account for issues of outcome nonresponse and sample selection, implying that outcomes are only observed for a subset of the initial population or sample of interest. These problems frequently occur in empirical applications as, for instance, wage gap decompositions, where wages are only observed for those who select themselves into employment. In a range of studies evaluating total (rather than direct and indirect effects), sample selection is modelled by a so-called missing at random (MAR) restriction, which assumes conditional exogeneity of sample selection given observed variables, see for instance Rubin (1976), Little and Rubin (1987), Robins, Rotnitzky, and Zhao (1994), Robins, Rotnitzky, and Zhao (1995), Carroll, Ruppert, and Stefanski (1995), Shah, Laird, and Schoen-

feld (1997), Fitzgerald, Gottschalk, and Moffitt (1998), Abowd, Crepon, and Kramarz (2001), and Wooldridge (2002, 2007). In contrast, so-called sample selection or nonignorable nonresponse models permit sample selection to be related to unobservables. Unless strong parametric assumptions are imposed (see for instance Heckman (1976, 1979), Hausman and Wise (1979), and Little (1995)), identification requires an instrumental variable (IV) for sample selection (e.g. Das, Newey, and Vella (2003), Newey (2007), and Huber (2012, 2014b)).

In this paper, we combine the identification of average natural direct and indirect effects based on sequential conditional independence with specific MAR or IV assumptions about sample selection. We show under which conditions the parameters of interest in the total as well as the selected population (whose outcomes are actually observed) are identified by inverse probability weighting[1] (IPW) based on particular propensity scores for treatment and selection. Under MAR, effects in the total population are obtained through reweighing by the inverse of the selection propensity given observed characteristics. If selection is related to unobservables, we make use of a control function that can be regarded as a nonparametric version of the inverse Mill's ratio in Heckman-type selection models. Under specific conditions, reweighing observations by the inverse of the selection propensity given observed characteristics and the control function identifies the effects in the selected and the total population. To convey the intuition of our identification results, we provide a brief simulation study, in which the finite sample properties of semiparametric IPW estimation with probit-based propensity scores is investigated.

As an empirical illustration, we evaluate the average natural direct and indirect effects of Program STAR, an educational experiment in Tennessee, U.S., which randomly assigned children to small classes in kindergarten and primary school. The positive impact of STAR classes on academic achievement has been demonstrated for example in Krueger (1999), but less is known about the underlying causal mechanisms. We consider absenteeism in kindergarten as potential mediator of the overall effect. The outcome of interest is the score on a standardized math test in the first grade of primary school, which is unobserved for a non-negligible share of students in

[1]The idea of using inverse probability weighting to control for selection problems goes back to Horvitz and Thompson (1952).

the data due to attrition. We apply one of our proposed IPW-based estimators to account for outcome attrition and compare the results to several alternative mediation estimators that make no corrections for sample selection.

The remainder of this paper is organized as follows. Section 2 discusses the parameters of interest, the assumptions, and the nonparametric identification results based on inverse probability weighting. Section 3 outlines estimation based on the sample analogs of the identification results. Section 4 presents a simulation study. Section 5 provides an application to Project STAR data. Section 6 concludes.

2 Identification

2.1 Parameters of interest

We would like to disentangle the average treatment effect (ATE) of a binary treatment variable D on an outcome variable Y into a direct effect and an indirect effect operating through the mediator M, which has bounded support and may be a scalar or a vector and discrete and/or continuous. To define the effects of interest, we use the potential outcome framework, see Rubin (1974), which has been applied in the context of mediation analysis by Rubin (2004), Ten Have, Joffe, Lynch, Brown, Maisto, and Beck (2007), and Albert (2008), among others. $M(d), Y(d, M(d'))$ denote the potential mediator state as a function of the treatment and potential outcome as a function of the treatment and the potential mediator, respectively, under treatments $d, d' \in \{0, 1\}$. Only one potential outcome and mediator state, respectively, is observed for each unit, because the realized mediator and outcome values are $M = D \cdot M(1) + (1 - D) \cdot M(0)$ and $Y = D \cdot Y(1, M(1)) + (1 - D) \cdot Y(0, M(0))$.

The ATE is given by $\Delta = E[Y(1, M(1)) - Y(0, M(0))]$. To disentangle the latter, note that the (average) natural direct effect (using the denomination of Pearl (2001))[2] is identified

[2]Robins and Greenland (1992) and Robins (2003) refer to this parameter as the total or pure direct effect and Flores and Flores-Lagunes (2009) as net average treatment effect.

by exogenously varying the treatment but keeping the mediator fixed at its potential value for $D = d$:

$$\theta(d) \quad = \quad E[Y(1, M(d)) - Y(0, M(d))], \quad d \in \{0, 1\}, \tag{1}$$

Equivalently, by exogenously shifting the mediator to its potential values under treatment and non-treatment but keeping the treatment fixed at $D = d$, the (average) natural indirect effect[3] is obtained:

$$\delta(d) \quad = \quad E[Y(d, M(1)) - Y(d, M(0))], \quad d \in \{0, 1\}. \tag{2}$$

The ATE is the sum of the direct and indirect effects defined upon opposite treatment states:

$$\begin{aligned} \Delta &= E[Y(1, M(1)) - Y(0, M(0))] \\ &= E[Y(1, M(1)) - Y(0, M(1))] + E[Y(0, M(1)) - Y(0, M(0))] = \theta(1) + \delta(0) \\ &= E[Y(1, M(0)) - Y(0, M(0))] + E[Y(1, M(1)) - Y(1, M(0))] = \theta(0) + \delta(1). \end{aligned} \tag{3}$$

This follows from adding and subtracting $E[Y(0, M(1))]$ or $E[Y(1, M(0))]$, respectively. The notation $\theta(1), \theta(0)$ and $\delta(1), \delta(0)$ points to possible effect heterogeneity w.r.t. the potential treatment state, implying the presence of interaction effects between the treatment and the mediator. However, the effects cannot be identified without further assumptions, as either $Y(1, M(1))$ or $Y(0, M(0))$ is observed for any unit, whereas $Y(1, M(0))$ and $Y(0, M(1))$ are never observed.

In contrast to natural effects, which are functions of the potential mediators, the so-called controlled direct effect is obtained by setting the mediator to a predetermined value m, rather than $M(d)$:

$$\gamma(m) \quad = \quad E[Y(1, m) - Y(0, m)], \quad m \text{ in the support of } M. \tag{4}$$

[3]Robins and Greenland (1992) and Robins (2003) refer to this parameter as the total or pure indirect effect and Flores and Flores-Lagunes (2009) as mechanism average treatment effect.

Whether $\theta(d)$ or $\gamma(m)$ is of primary interest depends on the research question at hand. The controlled direct effect may provide policy guidance whenever mediators can be externally prescribed, as for instance in a sequence of active labor market programs assigned by a caseworker, where D and M denotes assignment of the first and second program, respectively. This allows analysing the direct effect of the first program under alternative combinations of program prescriptions. In contrast, the natural direct effect assesses the effectiveness of the first program given the status quo decision to participate in the second program in the light of participation or non-participation in the first program. We refer to Pearl (2001) for further discussion of what he calls the descriptive and prescriptive natures of natural and controlled effects.

Our identification results will make use of a vector of observed covariates, denoted by X, that may confound the causal relations between D and M, D and Y, and M and Y. A further complication in our evaluation framework is that Y is assumed to be observed for a subpopulation, i.e. conditional on $S = 1$, where S is a binary variable indicating whether Y is observed/selected, or not. We therefore also define the direct and indirect effects among the selected population:

$$\begin{aligned}\theta_{S=1}(d) &= E[Y(1, M(d)) - Y(0, M(d))|S = 1], \quad \delta_{S=1}(d) = E[Y(d, M(1)) - Y(d, M(0))|S = 1],\\ \gamma_{S=1}(m) &= E[Y(1, m) - Y(0, m)|S = 1].\end{aligned}$$

Empirical examples with partially observed outcomes include wage regressions, with S being an employment indicator, see for instance Gronau (1974), or the evaluation of the effects of policy interventions in education on test scores, with S being participation in the test, see Angrist, Bettinger, and Kremer (2006). Throughout our discussion, S is allowed to be a function of D, M, and X, i.e. $S = S(D, M, X)$. However, S must neither be affected by nor affect Y.[4] S is therefore not a mediator, as selection per se does not causally influence the outcome. An example

[4]See for instance Imai (2009) for an alternative set of restrictions, assuming that selection is related to the outcome but is independent of the treatment conditional on the outcome and other observable variables.

for such a set up in terms of nonparametric structural models is given by

$$Y = \phi(D, M, X, U), \quad S = \psi(D, M, X, V), \tag{5}$$

where U, V are unobserved characteristics and ϕ, ψ are general functions.[5]

2.2 Assumptions and identification results under MAR

This section presents identifying assumptions that formalize the sequential conditional independence of D and M as imposed by Imai, Keele, and Yamamoto (2010) and many others as well as a MAR restriction on Y that implies that S is related to observables.[6]

Assumption 1 (conditional independence of the treatment):

(a) $Y(d, m) \perp D | X = x$, (b) $M(d') \perp D | X = x$ for all $d, d' \in \{0, 1\}$ and m in the support of M.

By Assumption 1, there are no unobservables jointly affecting the treatment, on the one hand, and the mediator and/or the outcome, on the other hand, conditional on X. In observational studies, the plausibility of this assumption crucially hinges on the richness of the data, while in experiments, it is satisfied if the treatment is randomized within strata defined by X or randomized independently of X.[7]

Assumption 2 (conditional independence of the mediator):

$Y(d, m) \perp M | D = d', X = x$ for all $d, d' \in \{0, 1\}$ and m, x in the support of M, X.

By Assumption 2, there are no unobservables jointly affecting the mediator and the outcome conditional on D and X. Assumption 2 only appears realistic if detailed information on possible confounders of the mediator-outcome relation is available in the data (even in experiments with random treatment assignment) and if post-treatment confounders of M and Y can be plausibly

[5] Note that $Y(d, M(d')) = \phi(d, M(d'), X, U)$, which means that fixing the treatment and the potential mediator yields the potential outcome.

[6] We implicitly also impose the Stable Unit Treatment Value Assumption (SUTVA, see Rubin, 1990), stating that the potential mediators and outcomes for any individual are stable in the sense that their values do not depend on the treatment allocations in the rest of the population.

[7] In the latter case, even the stronger condition $\{Y(d', m), M(d), X\} \perp D$ holds.

ruled out when controlling for D and X.[8]

Assumption 3 (conditional independence of selection):

$Y \perp S|D = d, M = m, X = x$ for all $d \in \{0, 1\}$ and m, x in the support of M, X.

By Assumption 3, there are no unobservables jointly affecting selection and the outcome conditional on D, M, X, such that outcomes are are missing at random (MAR) in the denomination of Rubin (1976). Put differently, selection is assumed to be selective w.r.t. observed characteristics only.

Assumption 4 (common support):

(a) $\Pr(D = d|M = m, X = x) > 0$ and (b) $\Pr(S = 1|D = d, M = m, X = x) > 0$ for all $d \in \{0, 1\}$ and m, x in the support of M, X.

Assumption 4(a) is a common support restriction requiring that the conditional probability to be treated given M, X, henceforth referred to as propensity score, is larger than zero in either treatment state. It follows that $\Pr(D = d|X = x) > 0$ must hold, too. By Bayes' theorem, Assumption 4(a) implies that $\Pr(M = m|D = d, X = x) > 0$, or in the case of M being continuous, that the conditional density of M given D, X is larger than zero. Conditional on X, M must not be deterministic in D, as otherwise identification fails due to the lack of comparable units in terms of the mediator across treatment states. Assumption 4(b) requires that for any combination of D, M, X, the probability to be observed is larger than zero. Otherwise, the outcome is not observed for some specific combinations of these variables implying yet another common support issue.

Figure 1 illustrates the causal framework underlying our assumptions by means of a causal graph, see for instance Pearl (1995), in which each arrow represents a potential causal effect. Further (unobserved) variables that only affect one of the variables explicitly displayed in the system are kept implicit. For instance, there may be unobservable variables U that affect the

[8]Several studies in the mediation literature discuss identification in the presence of post-treatment confounders of the mediator that may themselves be affected by the treatment. See for instance Robins and Richardson (2010), Albert and Nelson (2011), Tchetgen Tchetgen and VanderWeele (2014), Imai and Yamamoto (2011), and Huber (2014a).

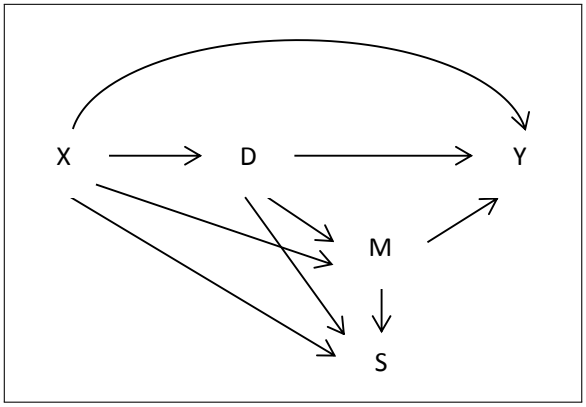

Figure 1: Causal framework under MAR

outcome, but do not influence D, M, or S; otherwise, there would be confounding. Under Assumptions 1 to 4, potential outcomes as well as direct and indirect effects in the total population are identified based on weighting by the inverse of the treatment and selection propensity scores.

Theorem 1:

(i) Under Assumptions 1, 2, 3, and 4, for $d \in \{0, 1\}$,

$$\begin{aligned} E[Y(d, M(1-d))] &= E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|M,X) \cdot \Pr(S=1|D,M,X)} \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)}\right], \\ E[Y(d, M(d))] &= E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|X) \cdot \Pr(S=1|D,M,X)}\right]. \end{aligned} \tag{6}$$

(ii) Under Assumptions 1(a), 2, 3, and 4, and M following a discrete distribution,

$$E[Y(d,m)] = E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\} \cdot S}{\Pr(D=d|X) \cdot \Pr(M=m|D,X) \cdot \Pr(S=1|D,M,X)}\right]. \tag{7}$$

Proof: See the appendix.

Using the results of Theorem 1, it can be easily shown that the direct and indirect effects are identified by

$$\begin{aligned}
\theta(d) &= E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X)}\right) \cdot \frac{\Pr(D=d|M,X) \cdot S}{\Pr(D=d|X) \cdot \Pr(S=1|D,M,X)}\right], \\
\delta(d) &= E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|M,X) \cdot \Pr(S=1|D,M,X)} \cdot \left(\frac{\Pr(D=1|M,X)}{\Pr(D=1|X)} - \frac{1-\Pr(D=1|M,X)}{1-\Pr(D=1|X)}\right)\right], \\
\gamma(m) &= E\left[\left(\frac{Y \cdot D}{\Pr(D=1|X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|X)}\right) \cdot \frac{I\{M=m\} \cdot S}{\Pr(M=m|D,X) \cdot \Pr(S=1|D,M,X)}\right].
\end{aligned}$$

These expressions are related to the IPW-based identification in Huber (2014a) for the case with no missing outcomes with the difference that here, multiplication by $S/\Pr(S=1|D,M,X)$ is included to account for sample selection. Furthermore, our results fit into the general framework of Wooldridge (2002), who considers the IPW-based M-estimation of missing data models. Finally, for the identification of $\gamma(m)$, Assumption 1 can be relaxed to Assumption 1(a) because (in contrast to $\theta(d), \delta(d)$) the distribution of the potential mediator $M(d)$ need not be identified.

2.3 Assumptions and identification results under selection related to unobservables

In the following discussion, we consider the case that selection is related to both observables and unobservables that are associated with the outcome. Assumptions 3 and 4 are therefore replaced. Rather, we assume that an instrumental variable for S is available to tackle sample selection.

Assumption 5 (Instrument for selection):

(a) There exists an instrument Z that may be a function of D, M, i.e. $Z = Z(D, M)$, is conditionally correlated with S, i.e. $E[Z \cdot S|D, M, X] \neq 0$, and satisfies (i) $Y(d, m, z) = Y(d, m)$ and (ii) $\{Y(d, m), M(d')\} \perp Z(d'', m')|X = x$ for all $d, d', d'' \in \{0, 1\}$ and z, m, m', x in the support of Z, M, X,

(b) $S = I\{V \leq \Pi(D, M, X, Z)\}$, where Π is a general function and V is a scalar (index of) unobservable(s) with a strictly monotonic cumulative distribution function conditional on X,

(c) $V \perp (D, M, Z)|X$.

Assumption 5 no longer imposes the independence of Y and S given observed characteristics. As the unobservable V in the selection equation is allowed to be associated with unobservables affecting the outcome, Assumptions 1 and 2 generally do not hold conditional on $S = 1$ due to the endogeneity of the post-treatment variable S. In fact, $S = 1$ implies that $\Pi(D, M, X, Z) > V$ such that conditional on X, the distribution of V generally differs across values of D, M. This entails a violation of the sequential conditional independence assumptions on D, M given $S = 1$ if potential outcome distributions differ across values of V. We therefore require an instrumental variable denoted by Z, which is allowed to be affected by D and M, but must not affect Y or be associated with unobservables affecting M or Y conditional on X, as invoked in (5a).[9] We apply a control function approach based on this instrument,[10] which requires further assumptions.

By the threshold crossing model postulated in 5(b), $\Pr(S = 1|D, M, X, Z) = \Pr(V \leq \Pi(D, M, X, Z)) = F_V(\Pi(D, M, X, Z))$, where $F_V(v)$ denotes the cumulative distribution function of V evaluated at v. We will henceforth use the notation $p(W) = \Pr(S = 1|D, M, X, Z)$ with $W = D, M, X, Z$ for the sake of brevity. Again by Assumption 5(b), the selection probability $p(W)$ increases strictly monotonically in Π conditional on X, such that there is a one-to-one correspondence between the distribution function F_V and specific values v given X. For X fixed, the identification of F_V by $p(W)$ is 'as good as good as' identifying V. By Assumption 5(c), V is independent of (D, M, Z) given X, implying that the distribution function of V given X is (nonparametrically) identified. Figure 2 illustrates the causal framework underlying Assumptions 1, 2, and 5 by means of a causal graph.

By comparing individuals with the same $p(W)$, we control for F_V and thus for the confounding associations of V with (i) D and $\{Y(d, m), M(d')\}$ and (ii) M and $Y(d, m)$ that occur conditional on $S = 1$. In other words, $p(W)$ serves as control function where the exogenous variation comes

[9] As an alternative set of IV restrictions in the context of selection, d'Haultfoeuille (2010) permits the instrument to be associated with the outcome, but assumes conditional independence of the instrument and selection given the outcome.

[10] Control function approaches have been applied in semi- and nonparametric sample selection models, e.g. Ahn and Powell (1993), Das, Newey, and Vella (2003), Newey (2007), and and Huber (2012, 2014b) as well as in nonparametric instrumental variable models, see for example Newey, Powell, and Vella (1999), Blundell and Powell (2004), and Imbens and Newey (2009).

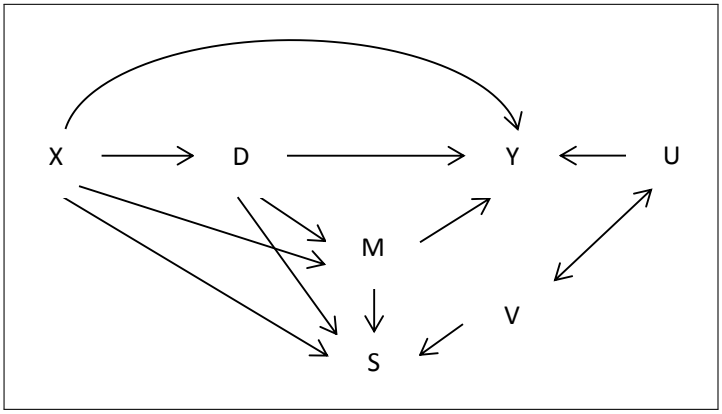

Figure 2: Causal framework under selection on unobservables

from Z. More concisely, it follows from our assumptions for any bounded function g that

$$E\left[g(Y(d,m))|D,M,X,p(W),S=1\right]=E\left[g(Y(d,m))|D,M,X,F_V,S=1\right]$$
$$=E\left[g(Y(d,m))|D,X,F_V,S=1\right]=E\left[g(Y(d,m))|X,F_V,S=1\right].$$

The first equality follows from $p(W)=F_V$ under Assumption 5, the second from the fact that when controlling for F_V, conditioning on $S=1$ does not result in an association between $Y(d,m)$ and M given D,X such that $Y(d,m)\perp M|D,X,p(W),S=1$ holds by Assumptions 2 and 5. The third equality follows from the fact that when controlling for F_V, conditioning on $S=1$ does not result in an association between $Y(d,m)$ and D given X such that $Y(d,m)\perp D|X,p(W),S=1$ holds by Assumptions 1 and 5. Similarly,

$$E\left[g(M(d))|D,X,p(W),S=1\right]=E\left[g(M(d))|D,X,F_V,S=1\right]=E\left[g(M(d))|X,F_V,S=1\right]$$

follows from the fact that when controlling for F_V, conditioning on $S=1$ does not result in an association between $M(d)$ and D given X such that $M(d)\perp D|X,p(W),S=1$ holds by Assumptions 1 and 5. These results will be useful in the proofs of Theorems 2 and 3, see Appendix A.2.

Furthermore, identification requires the following common support assumption, which is similar to Assumption 4(a), but in contrast to the latter also includes $p(W)$ as a conditioning

variable.

Assumption 6 (common support):

$\Pr(D = d|M = m, X = x, p(W) = p(w), S = 1) > 0$ for all $d \in \{0, 1\}$ and m, x, z in the support of M, X, Z.

By Bayes' theorem, Assumption 6 implies that the conditional density of $p(W) = p(w)$ given $D, M, X, S = 1$ is larger than zero. This means that in fully nonparametric contexts, the instrument Z must in general be continuous and strong enough to importantly shift the selection probability $p(W)$ conditional on D, M, X in the selected population. Assumptions 1, 2, 5, and 6 are sufficient for the identification of mean potential outcomes as well as direct and indirect effects in the selected population.

Theorem 2:

(i) Under Assumptions 1, 2, 5, and 6 for $d \in \{0, 1\}$,

$$\begin{aligned} E[Y(d, M(1-d))|S=1] &= E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X,p(W))} \cdot \frac{\Pr(D=1-d|M,X,p(W))}{\Pr(D=1-d|X,p(W))}\middle| S=1\right], \\ E[Y(d, M(d))|S=1] &= E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|X,p(W))}\middle| S=1\right]. \end{aligned} \tag{8}$$

(ii) Under Assumptions 1(a), 2, 5, and 6, and M following a discrete distribution,

$$E[Y(d, m)|S=1] = E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\}}{\Pr(D=d|X,p(W)) \cdot \Pr(M=m|D,X,p(W))}\middle| S=1\right]. \tag{9}$$

Proof: See the appendix.

Therefore, the direct and indirect effects are identified by

$$\begin{aligned} \theta_{S=1}(d) &= E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,X,p(W))} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X,p(W))}\right) \cdot \frac{\Pr(D=d|M,X,p(W))}{\Pr(D=d|X,p(W))}\middle| S=1\right], \\ \delta_{S=1}(d) &= E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X,p(W))} \cdot \left(\frac{\Pr(D=1|M,X,p(W))}{\Pr(D=1|X,p(W))} - \frac{1-\Pr(D=1|M,X,p(W))}{1-\Pr(D=1|X,p(W))}\right)\middle| S=1\right], \\ \gamma_{S=1}(m) &= E\left[\left(\frac{Y \cdot D}{\Pr(D=1|X,p(W))} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|X,p(W))}\right) \cdot \frac{I\{M=m\}}{\Pr(M=m|D,X,p(W))}\middle| S=1\right]. \end{aligned}$$

In nonparametric models that allow for general forms of effect heterogeneity related to unobservables, direct and indirect effects can generally only be identified among the selected

population. The reason is that effects among selected observations cannot be extrapolated to the non-selected population if the effects of D and M interact with unobservables that are distributed differently across $S = 1, 0$. The identification of effects in the total population therefore requires additional assumptions. In Assumption 7 below, we impose homogeneity in the direct and indirect effects across selected and non-selected populations conditional on X, V. A sufficient condition for effect homogeneity is the separability of observed and unobserved components in the outcome variable, i.e. $Y = \eta(D, M, X) + \nu(U)$, where η, ν are general functions and U is a scalar or vector of unobservables. Furthermore, common support as postulated in Assumption 6 needs to be strengthened to hold in the entire population. In addition, the selection probability $p(w)$ must be larger than zero for any w in the support of W; otherwise, outcomes are not observed for some values of D, M, X. Assumption 8 formalizes these common support restrictions.

Assumption 7 (conditional effect homogeneity):

$E[Y(1,m) - Y(0,m)|X = x, V = v, S = 1] = E[Y(1,m) - Y(0,m)|X = x, V = v]$ and $E[Y(d, M(1)) - Y(d, M(0))|X = x, V = v, S = 1] = E[Y(d, M(1)) - Y(d, M(0))|X = x, V = v]$, for all $d \in \{0, 1\}$ and m, x, v in the support of M, X, V.

Assumption 8 (common support):

(a) $\Pr(D = d|M = m, X = x, p(W) = p(w)) > 0$ and (b) $p(w) > 0$ for all $d \in \{0, 1\}$ and m, x, z in the support of M, X, Z.

While the mean potential outcomes in the total population remain unknown even under Assumptions 7 and 8, the effects of interest are nevertheless identified by the separability of U.

Theorem 3:

(i) Under Assumptions 1, 2, 5, 6, 7, and 8 for $d \in \{0, 1\}$,

$$\theta(d) = E\left[\left(\frac{Y \cdot D}{\Pr(D = 1|M, X, p(W))} - \frac{Y \cdot (1 - D)}{1 - \Pr(D = 1|M, X, p(W))}\right) \cdot \frac{\Pr(D = d|M, X, p(W)) \cdot S}{\Pr(D = d|X, p(W)) \cdot p(W)}\right] \quad (10)$$

$$\delta(d) = E\left[\frac{Y \cdot I\{D = d\} \cdot S}{\Pr(D = d|M, X, p(W)) \cdot p(W)} \cdot \left(\frac{\Pr(D = 1|M, X, p(W))}{\Pr(D = 1|X, p(W))} - \frac{1 - \Pr(D = 1|M, X, p(W))}{1 - \Pr(D = 1|X, p(W))}\right)\right].$$

(ii) Under Assumptions 1(a), 2, 5, 6, 7, and 8, and M following a discrete distribution,

$$\gamma(m) = E\left[\left(\frac{Y\cdot D}{\Pr(D=1|X,p(W))} - \frac{Y\cdot(1-D)}{1-\Pr(D=1|X,p(W))}\right)\cdot\frac{I\{M=m\}\cdot S}{\Pr(M=m|D,X,p(W))\cdot p(W)}\right]. \tag{11}$$

Proof: See the appendix.

2.4 Extensions to further populations, parameters, and variable distributions

This section briefly sketches how the identification results can be extended to further populations of interest, policy-relevant parameters, and richer distributions of the treatment and/or the mediator. First and in analogy to the concept of weighted treatment effects in Hirano, Imbens, and Ridder (2003), direct and indirect effects can be identified for particular target populations by reweighing observations according to the distribution of X in the target population. To this end, we define $\omega(X)$ to be a well-behaved weighting function depending on X. Including $\frac{\omega(X)}{E[\omega(X)]}$ in the expectation operators presented in the theorems above yields the parameters of interest for the target population. As an important example, consider $\omega(X) = \Pr(D=1|X)$. For some well-behaved function $f(Y,D,M,S,X,Z)$ of the observed data,

$$\begin{aligned} &E\left[\tfrac{\omega(X)}{E[\omega(X)]}\cdot f(Y,D,M,S,X,Z)\right] = E\left[\tfrac{\Pr(D=1|X)}{\Pr(D=1)}\cdot f(Y,D,M,S,X,Z)\right] \\ &= E\left[\tfrac{\Pr(D=1|X)}{\Pr(D=1)}\cdot f(Y,D,M,S,X,Z)\right] = E\left[f(Y,D,M,S,X,Z)|D=1\right], \end{aligned} \tag{12}$$

i.e. the expected value of that function among the treated is identified. Likewise, defining $\omega(X) = 1-\Pr(D=1|X)$ gives the expected value among the non-treated. Any of the expressions in the expectation operators of the theorems may serve as $f(Y,D,M,S,X,Z)$ in (12).[11]

[11] For instance, the weighted versions of the parameters identified in Theorem 1 correspond to

$$\begin{aligned} E_\omega[Y(d,M(1-d))] &= E\left[\frac{\omega(X)}{E[\omega(X)]}\cdot\frac{Y\cdot I\{D=d\}\cdot S}{\Pr(D=d|M,X)\cdot\Pr(S=1|D,M,X)}\cdot\frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)}\right], \\ E_\omega[Y(d,M(d))] &= E\left[\frac{\omega(X)}{E[\omega(X)]}\cdot\frac{Y\cdot I\{D=d\}\cdot S}{\Pr(D=d|X)\cdot\Pr(S=1|D,M,X)}\right], \\ E_\omega[Y(d,m)] &= E\left[\frac{\omega(X)}{E[\omega(X)]}\cdot\frac{Y\cdot I\{D=d\}\cdot I\{M=m\}\cdot S}{\Pr(D=d|X)\cdot\Pr(M=m|D,X)\cdot\Pr(S=1|D,M,X)}\right]. \end{aligned}$$

Second, the identification results may be extended to well-behaved functions of Y, rather than Y itself. For instance, replacing Y by $I\{Y \leq a\}$, the indicator function that Y is not larger than some value a, everywhere in the theorems permits identifying distributional features or effects. The inversion of potential outcome distribution functions allows identifying quantile treatment effects.

Third, our framework can be adapted to allow for multiple or multivalued (rather than binary) treatments. If D is multivalued discrete, the derived expressions may be applied under minor adjustments. For instance, for any $d \neq d'$ in the discrete support of D, the expression for potential outcomes in Theorem 1 becomes

$$E[Y(d, M(d'))] = E\left[\frac{Y \cdot I\{D = d\} \cdot S}{\Pr(D = d|M, X) \cdot \Pr(S = 1|D, M, X)} \cdot \frac{\Pr(D = d'|M, X)}{\Pr(D = d'|X)}\right]$$

under appropriate common support conditions. If D is continuous, any indicator functions for treatment values, which are only appropriate in the presence of mass points, need to be replaced by kernel functions, while treatment propensity scores need to be substituted by conditional density functions. In analogy to Hsu, Huber, Lee, and Pipoz (2018), who consider mediation analysis with continuous treatments in the absence of sample selection, the expression for potential outcomes in Theorem 1 becomes

$$\begin{aligned} E[Y(d, M(d'))] &= \lim_{h \to 0} E\left[\frac{Y \cdot \omega(D; d, h) \cdot S}{E[\omega(D; d, h)|M, X] \cdot \Pr(S = 1|D, M, X)}\right. \\ &\times \left.\frac{E[\omega(D; d', h)|M, X]}{E[\omega(D; d', h)|X]}\right]. \end{aligned}$$

The weighting function $\omega(D; d) = K\left((D - d)/h\right)/h$, with K being a symmetric second order kernel function assigning more weight to observations closer to d and h being a bandwidth operator. For h going to zero, i.e. $\lim_{h \to 0}$, $E[\omega(D; d', h)|X]$ and $E[\omega(D; d', h)|M, X]$ correspond to the conditional densities of D given X and given M, X, respectively, also known as generalized propensity scores. We refer to Hsu, Huber, Lee, and Pipoz (2018) for more discussion on direct

.

and indirect effects of continuous treatments and how estimation may proceed based on generalized propensity scores. We also note that in the context of controlled direct effects, such kernel methods not only allow for a continuous treatment, but (contrarily to our theorems) also for a continuous mediator.

3 Estimation

The parameters of interest can be estimated using the normalized versions of the sample analogs of the IPW-based identification results in Section 2. This implies that the weights of the observations used for the computation of mean potential outcomes add up to unity, as advocated in Imbens (2004) and Busso, DiNardo, and McCrary (2009). For instance, the normalized sample analogs of the results in Theorem 1, part (i) are given by

$$\begin{aligned}
\hat{\mu}_{1,M(0)} &= \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \cdot D_i \cdot S_i}{\hat{p}(M_i, X_i) \cdot \hat{\pi}(D_i, M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} \frac{D_i \cdot S_i}{\hat{p}(M_i, X_i) \cdot \hat{\pi}(D_i, M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)}, \\
\hat{\mu}_{0,M(1)} &= \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \cdot (1-D_i) \cdot S_i}{(1-\hat{p}(M_i, X_i)) \cdot \hat{\pi}(D_i, M_i, X_i)} \frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} \frac{(1-D_i) \cdot S_i}{(1-\hat{p}(M_i, X_i)) \cdot \hat{\pi}(D_i, M_i, X_i)} \frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)}, \\
\hat{\mu}_{1,M(1)} &= \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \cdot D_i \cdot S_i}{\hat{p}(X_i) \cdot \hat{\pi}(D_i, M_i, X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} \frac{D_i \cdot S_i}{\hat{p}(X_i) \cdot \hat{\pi}(D_i, M_i, X_i)}, \\
\hat{\mu}_{0,M(0)} &= \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \cdot (1-D_i) \cdot S_i}{(1-\hat{p}(X_i)) \cdot \hat{\pi}(D_i, M_i, X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} \frac{(1-D_i) \cdot S_i}{(1-\hat{p}(X_i)) \cdot \hat{\pi}(D_i, M_i, X_i)}.
\end{aligned}$$

i indexes observations in an i.i.d. sample of size n and $\hat{\mu}_{d,M(d')}$ is an estimate of $\mu_{d,M(d')} = E[Y(d, M(d'))]$ with $d, d' \in \{1, 0\}$. $\hat{p}(M_i, X_i)$, $\hat{p}(X_i)$ are estimates of the treatment propensity scores $\Pr(D = 1|M_i, X_i)$, $\Pr(D = 1|X_i)$, respectively, while $\hat{\pi}(D_i, M_i, X_i)$ is an estimate of the selection propensity score $\Pr(S = 1|D, M, X)$. Direct and indirect effect estimates are obtained by $\hat{\theta}(d) = \hat{\mu}_{1,M(d)} - \hat{\mu}_{0,M(d)}$ and $\hat{\delta}(d) = \hat{\mu}_{d,M(1)} - \hat{\mu}_{d,M(0)}$.

When propensity scores are estimated parametrically, e.g. based on probit models as in the simulations and application below, then $\hat{\mu}_{d,M(d')}, \hat{\theta}(d), \hat{\delta}(d)$ satisfy the sequential GMM framework discussed in Newey (1984), with propensity score estimation representing the first step and parameter estimation the second step. This approach is $\sqrt{n}$-consistent and asymptotically nor-

mal under standard regularity conditions. When the propensity scores are estimated nonparametrically, $\sqrt{n}$-consistency and asymptotic normality can be obtained if the first step estimators satisfy particular regularity conditions. See Hsu, Huber, and Lai (2018), who consider series logit estimation of the propensity scores, however, for the case without sample selection. Furthermore, the bootstrap is consistent for inference as the proposed IPW estimators are smooth and asymptotically normal.

4 Simulation study

This section provides a brief simulation study, in which we investigate the finite sample properties of estimation of natural direct and indirect effects based on the sample analogs of Theorems 1 to 3. To this end, the following data generating process is considered:

$$\begin{aligned} Y &= 0.5D + M + 0.5DM + X - \alpha DU + U, \quad Y \text{ is observed if } S = 1, \\ S &= I\{0.5D - 0.5M + 0.25X + Z + V > 0\}, \\ M &= 0.5D + 0.5X + W, \quad D = I\{0.5X + Q > 0\}, \quad Z = 0.25X - 0.25M + R, \\ X, U, V, W, Q, R &\sim \mathcal{N}(0,1), \text{ independently of each other.} \end{aligned}$$

The outcome Y is a linear function of the observed variables D, M, X and an unobserved term U, and is only observed if the selection indicator S – which depends on D, M, X, an instrument Z, and an unobservable V – is equal to one. α gauges the interaction of D and U in the outcome equation. For $\alpha \neq 0$, the treatment effect is heterogeneous in U such that Assumption 7 is violated. W and R denote the unobservables in the linearly modelled mediator M and instrument Z, respectively. Any unobservable as well as the observed covariate X are standard normally distributed independent of each other. In this framework, the assumptions underlying Theorem 1 are satisfied.

We run 5,000 Monte Carlo simulations with sample sizes $n = 1000, 4000$ and consider estimation of the natural direct and indirect effects in the total population ($\theta(d)$, $\delta(d)$) based on three different estimators: (i) normalized IPW as suggested in Huber (2014a) among the selected ('IPW w. $S = 1$') that controls for X but ignores selection bias, (ii) normalized IPW based on Theorem 1 assuming MAR ('IPW MAR'), and (iii) normalized IPW based on Theorem 3 ('IPW IV'). We estimate the treatment and selection propensity scores by probit and apply a trimming rule that discards observations with $\hat{p}(M, X)$ smaller than 0.05 or larger than 0.95 or with $\hat{\pi}(D, M, X)$ smaller than 0.05 to prevent exploding weights due to small denominators. Trimming hardly affects IPW estimator (i), but reduces the variance of estimation based on Theorems 1 and 3 in several cases.

Table 1: Simulations under selection on observables, total population

	$\hat{\theta}(1)$			$\hat{\theta}(0)$			$\hat{\delta}(1)$			$\hat{\delta}(0)$		
	bias	std	rmse	bias	std	rmse	bias	std	rmse	bias	std	rmse
	$\alpha = 0.25$, $n = 1000$											
IPW w. $S = 1$	-0.16	0.14	0.21	-0.17	0.16	0.23	-0.01	0.15	0.15	-0.02	0.11	0.12
IPW MAR	0.03	0.28	0.28	0.01	0.20	0.20	-0.03	0.13	0.14	-0.05	0.14	0.15
IPW IV	-0.01	0.30	0.30	-0.02	0.31	0.31	-0.02	0.18	0.18	-0.03	0.15	0.15
	$\alpha = 0.25$, $n = 4000$											
IPW w. $S = 1$	-0.16	0.07	0.18	-0.17	0.08	0.19	0.00	0.08	0.08	-0.01	0.06	0.06
IPW MAR	0.01	0.15	0.15	0.01	0.10	0.10	-0.02	0.07	0.07	-0.03	0.08	0.09
IPW IV	-0.01	0.15	0.15	-0.02	0.16	0.16	-0.01	0.09	0.09	-0.02	0.08	0.08

Note: 'std' and 'rmse' report the standard deviation and root mean squared error, respectively.

Table 1 reports the simulations results under $\alpha = 0.25$,[12] namely the bias, standard deviation (std), and the root mean squared error (RMSE) of the various estimators for the natural direct and indirect effects in the total population. Ignoring selection (IPW w. $S = 1$) yields biased estimates of the direct effects under either sample size, while biases are generally small for estimation based on Theorem 1. Interestingly, the latter result also holds for estimation related to Theorem 3, where the selection process accounts for the same observed factors as under the correct MAR assumption, plus the control function. Even though including the control function is not required for consistency, it does not jeopardize identification either, even if Assumption 7 requiring $\alpha = 0$

[12]Results are very similar when $\alpha = 0$ and therefore omitted.

is not satisfied,[13] as reflected in the low biases. However, accounting for this unnecessary variable entails an increase of the standard deviation in some cases. In general, the estimators based on Theorems 1 and 3 are (due to the estimation of the sample selection propensity score) less precise than IPW without selection correction in the selected sample. The proposed methods become relatively more competitive in terms of the RMSE as the sample size increases and gains in bias reduction become relatively more important compared to losses in precision.

As a modification to our initial setup, we introduce a correlation between U and V, which implies that the assumptions underlying Theorem 1 no longer hold, while those of Theorem 2 are satisfied and those of Theorem 3 are satisfied when $\alpha = 0$:

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma), \text{ where } \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

Table 2: Simulations with selection on unobservables, total population

	$\hat{\theta}(1)$			$\hat{\theta}(0)$			$\hat{\delta}(1)$			$\hat{\delta}(0)$		
	bias	std	rmse	bias	std	rmse	bias	std	rmse	bias	std	rmse
	$\alpha = 0$, $n = 1000$											
IPW w. $S = 1$	-0.28	0.13	0.31	-0.27	0.16	0.32	0.07	0.16	0.18	0.07	0.12	0.14
IPW MAR (Theorem 1)	-0.09	0.30	0.31	-0.11	0.21	0.24	0.06	0.14	0.15	0.04	0.15	0.16
IPW IV (Theorem 3)	0.02	0.32	0.32	-0.01	0.31	0.31	-0.02	0.18	0.18	-0.05	0.16	0.16
	$\alpha = 0$, $n = 4000$											
IPW w. $S = 1$	-0.28	0.07	0.29	-0.28	0.08	0.29	0.08	0.08	0.12	0.09	0.06	0.11
IPW MAR (Theorem 1)	-0.11	0.16	0.20	-0.11	0.10	0.15	0.06	0.07	0.09	0.06	0.09	0.11
IPW IV (Theorem 3)	0.01	0.17	0.17	-0.01	0.16	0.16	-0.02	0.09	0.09	-0.04	0.08	0.09
	$\alpha = 0.25$, $n = 1000$											
IPW w. $S = 1$	-0.37	0.13	0.39	-0.35	0.15	0.38	0.05	0.16	0.16	0.07	0.12	0.14
IPW MAR (Theorem 1)	-0.20	0.30	0.36	-0.20	0.21	0.28	0.03	0.14	0.14	0.04	0.15	0.16
IPW IV (Theorem 3)	-0.14	0.32	0.34	-0.16	0.31	0.35	-0.02	0.18	0.18	-0.05	0.16	0.16
	$\alpha = 0.25$, $n = 4000$											
IPW w. $S = 1$	-0.38	0.07	0.38	-0.36	0.08	0.36	0.06	0.08	0.10	0.09	0.06	0.11
IPW MAR (Theorem 1)	-0.22	0.16	0.27	-0.20	0.10	0.22	0.04	0.07	0.08	0.06	0.09	0.11
IPW IV (Theorem 3)	-0.14	0.16	0.22	-0.16	0.16	0.23	-0.01	0.09	0.09	-0.04	0.08	0.09

Note: 'std' and 'rmse' report the standard deviation and root mean squared error, respectively.

[13] Note that in spite of $\alpha = 0.25$, estimation based on (the incorrect) Theorem 3 is consistent because the distribution of U is not associated with S conditional on D, M, X.

Table 2 reports the results for the estimation of natural effects in the total population under $\alpha = 0$ and 0.25 using the same methods as before. Non-negligible biases occur not only when ignoring sample selection ('IPW w. $S = 1$'), but also when selection is assumed to be related to observables only (IPW MAR). When $\alpha = 0$, estimation based on Theorem 3 (IPW IV) is close to being unbiased and dominates the other methods in terms of RMSE under the larger sample size ($n = 4000$). When $\alpha = 0.25$, however, also the latter approach is biased due to the violation of Assumption 7. Therefore, Table 3 considers the estimation of natural effects among the selected population only ($\theta_{S=1}(d)$, $\delta_{S=1}(1)$) in the presence of the D-U-interaction effect. We investigate the performance of estimation based on Theorem 2 ('IPW IV w. $S = 1$'), as well as of IPW among the selected ignoring selection. While the latter approach is biased, the former is close to being unbiased, but less precise. Under the larger sample size, our approach dominates both in terms of unbiasedness and RMSE.[14]

Table 3: Simulations with selection on unobservables, selected population ($S = 1$)

	$\hat{\theta}_{S=1}(1)$			$\hat{\theta}_{S=1}(0)$			$\hat{\delta}_{S=1}(1)$			$\hat{\delta}_{S=1}(0)$		
	bias	std	rmse	bias	std	rmse	bias	std	rmse	bias	std	rmse
	$\alpha = 0.25$, $n = 1000$											
IPW w. $S = 1$	-0.11	0.13	0.17	-0.09	0.15	0.17	0.05	0.16	0.16	0.07	0.12	0.14
IPW IV w. $S = 1$ (Theorem 2)	0.00	0.21	0.21	-0.03	0.23	0.23	0.02	0.17	0.17	-0.01	0.12	0.12
	$\alpha = 0.25$, $n = 4000$											
IPW w. $S = 1$	-0.12	0.07	0.14	-0.10	0.08	0.12	0.06	0.08	0.10	0.09	0.06	0.11
IPW IV w. $S = 1$ (Theorem 2)	0.01	0.10	0.10	-0.02	0.11	0.12	0.03	0.08	0.08	-0.00	0.06	0.06

Note: 'std' and 'rmse' report the standard deviation and root mean squared error, respectively.

5 Empirical Application

This section illustrates the evaluation of direct and indirect treatment effects in the presence of sample selection using data from Project STAR (Student-Teacher Achievement Ratio), an educational experiment conducted from 1985 to 1989 in Tennessee, USA. In the experiment, a cohort of students entering kindergarten and their teachers were randomly assigned within their school to one of three class types: small (13 – 17 students), regular (22 – 26 students), or

[14]Results are very similar when setting $\alpha = 0$ and therefore omitted.

regular with an additional teacher's aid. Students were supposed to remain in the assigned class type through third grade, returning to regular classes afterwards. The goal of Project STAR was investigating the impact of class size on academic achievement measured by standardized and curriculum-based tests in mathematics, reading, and basic study skills. Numerous studies found positive effects of reduced class size on academic performance both short- (Folger and Breda (1989), Finn and Achilles (1990), Krueger (1999)), mid- (Finn, Fulton, Zaharias, and Nye (1989)), long-term (Nye, Hedges, and Konstantopoulos (2001), Krueger and Whitmore (2001)), and even on later-life outcomes (Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011)). While benefits of small class size are well documented, the causal mechanisms underlying the effect are less well-understood. Finn and Achilles (1990) argue that the impact is likely driven by classroom processes related to higher teacher morale and satisfaction translated to students, increased teacher-student interactions and time for individual attention, and student involvement in learning activities.

We investigate whether the effect of reduced class size on academic performance is mediated by the number of days absent from school. There might be several explanations for why class size affects days of absence. A smaller concentration of children in a classroom may be related to reduced transmission of infectious diseases and hence absenteeism.[15] Increased student involvement and closer teacher-student relationships in smaller classes may represent further channels making children and their parents more engaged and less likely to miss classes. As for the link between school absence and academic performance, a number of studies demonstrated a negative association between the two, see for instance Gershenson, Jacknowitz, and Brannegan (2017), Gottfried (2009), and Morrissey, Hutchison, and Winsler (2014).

We compare results using the IPW MAR estimator ('IPW MAR' in Table 5) based on Theorem 1 (relying on Assumptions 1 through 4) in Section 2 to three previously considered

[15] Odongo, Wakhungu, and Stanley (2017) find a positive correlation between school size and communicable disease prevalence rates in Kenya. We are, however, not aware of any such study considering class (rather than school) size.

mediation estimators that ignore sample selection:[16] (i) a linear mediation estimator allowing for treatment-mediator interactions but neither accounting for observed pre-treatment confounders, nor selection, which is numerically equivalent to the decomposition of Blinder (1973) and Oaxaca (1973) ('Lin w. $S = 1$, no X');[17] (ii) a semiparametric IPW-based analog of the linear mediation estimator not accounting for confounding also considered in Huber (2015) ('IPW w. $S = 1$, no X'); and (iii) the IPW estimator suggested in Huber (2014a) that incorporates observed pre-treatment covariates X but ignores sample selection when estimating the effect for the total population ('IPW w. $S = 1$'). We apply the same trimming rule as in the simulations presented in Section 4, which discards observations with treatment propensity scores $\hat{p}(M, X)$ smaller than 0.05 or larger than 0.95 or with $\hat{\pi}(D, M, X)$ smaller than 0.05. However, no observations are dropped for any IPW method as such extreme propensity scores do not occur in our sample.

The treatment (D) is a binary indicator which is one if a child entering kindergarten was enrolled in a small class and zero otherwise.[18] The outcome (Y) is the first grade score in the Stanford Achievement Test (SAT) in mathematics. For IPW MAR estimation, a selection indicator S for missing outcomes is generated and all observations in our evaluation sample are preserved, such that effects are estimated for the entire population. In the case of the remaining three estimators, the evaluation is based on the data with non-missing Y, such that estimation relies on the selected sample only. The mediator (M) is the number of days a child was absent during the kindergarten year. Observed covariates (X) consist of child's race, gender, year of birth, and free lunch status as a proxy for socio-economic status. They are controlled for in the 'IPW w. $S = 1$' and 'IPW MAR' estimators. Even if these variables are initially balanced due to the random assignment of D, they might confound M and Y, implying that they are imbalanced when conditioning on the mediator for estimating direct and indirect effects.[19]

[16] We do not consider IPW IV estimation based on Theorems 2 and 3, as our data do not contain credible instruments.

[17] See Huber (2015) on the equivalence of conventional wage gap decompositions and a simple mediation model.

[18] Following Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011), we consider regular class size with and without additional teaching aid to be one treatment.

[19] For example, Ready (2010) reports a stronger negative impact of absenteeism on early literacy outcomes for students with lower socioeconomic status, which implies that socioeconomic status and absenteeism interact in explaining the outcome. If socioeconomic status in addition affects absenteeism, it is a confounder of the association between absenteeism and the literacy outcomes.

We restrict the initial sample of 11,601 children to 6,325 observations who were part of Project STAR in kindergarten such that their treatment status was observed.[20] About 30% of participants in the kindergarten year were randomized into small classes. Table 4 presents summary statistics for the variables included in our empirical illustration for individuals without any missing values in the covariates. It shows a positive and statistically significant association between reduced class size and the average score in the standardized math test. Furthermore, children in small classes are, on average, about 0.7 days less absent and this difference is significant at the 5% level. There are no statistically significant differences in students' gender, race,[21] and free lunch status across treatment states due to treatment randomization. The sample is not perfectly balanced in terms of students' years of birth: children born in 1978 and 1980 are less likely to be in small classes (differences are statistically significant at the 1 and 10% levels, respectively), while those born in 1979 are more likely to be in small classes (significant at the 5% level). There is substantial attrition: math SAT scores in the first grade are observed for only 70% of program participants in the kindergarten year. The number of missing values in other key variables is much smaller. In the estimations, observations with missing values in M or X are dropped, which concerns all in all 83 cases, or about 1% of the sample.

Table 5 provides point estimates ('est.'), cluster-robust standard errors ('s.e.') based on blockbootstrapping the effects 1,999 times, and p-values for the total treatment effect, as well as natural direct and indirect effects under treatment and non-treatment ($\hat{\theta}(1)$, $\hat{\theta}(0)$, $\hat{\delta}(1)$, $\hat{\delta}(0)$) for the four estimators. The total average effect of small class assignment is very similar across all methods and highly statistically significant, amounting to an increase of almost 10 points. Furthermore, we find that if anything, the contribution of the indirect effects due to reduced days of absence is positive, but rather modest, ranging 0.18 to 0.99 points across different methods and treatment states. The IPW MAR estimator yields the largest indirect effects (amounting to 3 – 11% of the total effect), and the indirect effect on the non-treated group is statistically

[20] 5,276 students joined the program in subsequent years. About 2,200 entered the experiment in the first grade, 1,600 in the second and 1,200 in the third grade.

[21] Less than 1% of students in the sample are Asian, Hispanic, Native American or other race. In our analysis, they are included in one group with black students.

Table 4: Mean covariate values by treatment status

Variable	Total	$d=0$	$d=1$	Difference	p-value
Student's gender: male	0.51	0.51	0.51	0.00	0.96
	[0.50]	[0.50]	[0.50]	(0.01)	
Student's race: white	0.67	0.67	0.68	0.01	0.42
	[0.47]	[0.47]	[0.47]	(0.02)	
Free lunch	0.48	0.49	0.47	-0.02	0.25
	[0.50]	[0.50]	[0.50]	(0.02)	
Born 1978	0.01	0.01	0.00	-0.01	0.00
	[0.08]	[0.09]	[0.05]	(0.00)	
Born 1979	0.23	0.22	0.25	0.03	0.04
	[0.42]	[0.42]	[0.43]	(0.01)	
Born 1980	0.76	0.77	0.74	-0.02	0.09
	[0.43]	[0.42]	[0.44]	(0.01)	
Born 1981	0.00	0.00	0.00	0.00	0.87
	[0.03]	[0.03]	[0.03]	(0.00)	
Kindergarten days absent	10.51	10.72	10.01	-0.71	0.02
	[9.76]	[9.95]	[9.29]	(0.31)	
Math SAT grade 1	534.54	531.52	541.25	9.73	0.00
	[43.83]	[42.92]	[45.10]	(2.14)	

Note: Standard deviations are in squared brackets. Cluster-robust standard errors are in parentheses.

significant at the 10% level. It is thus the direct effects, which are highly statistically significant for any method, that mostly drive the total effect. IPW MAR yields direct effect estimates of 8.52 points under treatment and 7.75 points under non-treatment, which is slightly smaller than those of the other estimators exploiting the subsample with non-missing outcomes only (ranging from 9.01 to 9.55 points under treatment and from 8.77 to 9.55 points under non-treatment). We therefore conclude that causal mechanisms not observed in the data (possibly including teacher motivation and individual teacher-student interaction) and entering the direct effect are much more important than absenteeism for explaining the effect of small kindergarten classes on math performance.

Table 5: Effects of small class size in kindergarten on the math SAT in grade 1

	Total effect			$\hat{\theta}(1)$			$\hat{\theta}(0)$			$\hat{\delta}(1)$			$\hat{\delta}(0)$		
	est.	s.e.	p-val	est.	s.e.	p-val	est.	s.e.	p-val	est.	s.e.	p-val	est.	s.e.	p-val
IPW MAR	8.74	2.37	0.00	8.52	2.36	0.00	7.75	2.70	0.00	0.99	0.79	0.21	0.23	0.13	0.09
Lin w. $S=1$, no X	9.73	2.16	0.00	9.46	2.17	0.00	9.55	2.15	0.00	0.27	0.18	0.12	0.18	0.13	0.16
IPW w. $S=1$, no X	9.73	2.16	0.00	9.55	2.15	0.00	9.43	2.18	0.00	0.30	0.21	0.16	0.18	0.13	0.15
IPW w. $S=1$	9.20	2.14	0.00	9.01	2.14	0.00	8.77	2.19	0.00	0.43	0.32	0.18	0.19	0.14	0.18

Note: Cluster-robust standard errors ('s.e.') and p-values ('p-val') for the point estimates ('est.') are obtained by bootstrapping the latter 1,999 times.

6 Conclusion

In this paper, we proposed an approach for disentangling a total causal effect into a direct component and a indirect effect operating through a mediator in the presence of outcome attrition or sample selection. To this end, we combined sequential conditional independence assumptions about the assignment of the treatment and the mediator with either selection on observables/missing at random or instrumental variable assumptions on the outcome attrition process. We demonstrated the identification of the parameters of interest based on inverse probability weighting by specific treatment, mediator, and/or selection propensity scores and outlined estimation based on the sample analogs of these results. We also provided a brief simulation study and an empirical illustration based on the Project STAR experiment in the U.S. to evaluate the direct and indirect effects of small classes in kindergarten on math test scores in first grade.

References

Abowd, J., B. Crepon, and F. Kramarz (2001): "Moment Estimation With Attrition: An Application to Economic Models," *Journal of the American Statistical Association*, 96, 1223–1230.

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.

Angrist, J., E. Bettinger, and M. Kremer (2006): "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia," *American Economic Review*, 96, 847–862.

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.

Blinder, A. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 8, 436–455.

Blundell, R. W., and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *The Review of Economic Studies*, 71, 655–679.

Brunello, G., M. Fort, N. Schneeweis, and R. Winter-Ebmer (2016): "The Causal Effect of Education on Health: What is the Role of Health Behaviors?," *Health Economics*, 25, 314–336.

Busso, M., J. DiNardo, and J. McCrary (2009): "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," *IZA Discussion Paper No. 3998*.

Carroll, R., D. Ruppert, and L. Stefanski (1995): *Measurement Error in Nonlinear Models*. Chapman and Hall, London.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star," *The Quarterly Journal of Economics*, 126, 1593–1660.

Conti, G., J. J. Heckman, and R. Pinto (2016): "The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviour," *The Economic Journal*, 126, F28–F65.

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. (2010): "A new instrumental method for dealing with endogenous selection," *Journal of Econometrics*, 154, 1–15.

Finn, J. D., and C. M. Achilles (1990): "Answers and Questions about Class Size: A Statewide Experiment," *American Educational Research Journal*, 27, 557–577.

Finn, J. D., D. Fulton, J. Zaharias, and B. A. Nye (1989): "Carry-Over Effects of Small Classes," *Peabody Journal of Education*, 67, 75–84.

Fitzgerald, J., P. Gottschalk, and R. Moffitt (1998): "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics," *Journal of Human Resources*, 33, 251–299.

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*.

Folger, J., and C. Breda (1989): "Evidence from Project STAR about Class Size and Student Achievement," *Peabody Journal of Education*, 67, 17–33.

Gershenson, S., A. Jacknowitz, and A. Brannegan (2017): "Are Student Absences Worth the Worry in U.S. Primary Schools?," *Education Finance and Policy*, 12, 137–165.

Gottfried, M. A. (2009): "Excused Versus Unexcused: How Student Absences in Elementary School Affect Academic Achievement," *Educational Evaluation and Policy Analysis*, 31, 392–415.

Gronau, R. (1974): "Wage comparisons-a selectivity bias," *Journal of Political Economy*, 82, 1119–1143.

Hausman, J., and D. Wise (1979): "Attrition Bias In Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, 47, 455–473.

Heckman, J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

Heckman, J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

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

Hong, G. (2010): "Ratio of mediator probability weighting for estimating natural direct and indirect effects," in *Proceedings of the American Statistical Association, Biometrics Section*, p. 24012415. Alexandria, VA: American Statistical Association.

Horvitz, D., and D. Thompson (1952): "A Generalization of Sampling without Replacement from a Finite Population," *Journal of American Statistical Association*, 47, 663–685.

Hsu, Y., M. Huber, and T. Lai (2018): "Nonparametric estimation of natural direct and indirect effects based on inverse probability weighting," *forthcoming in the Journal of Econometric Methods*.

Hsu, Y., M. Huber, Y. Lee, and L. Pipoz (2018): "Direct and indirect effects of continuous treatments based on generalized propensity score weighting," *SES Working papers 495, University of Fribourg*.

Huber, M. (2012): "Identification of average treatment effects in social experiments under alternative forms of attrition," *Journal of Educational and Behavioral Statistics*, 37, 443–474.

——— (2014a): "Identifying causal mechanisms (primarily) based on inverse probability weighting," *Journal of Applied Econometrics*, 29, 920–943.

——— (2014b): "Treatment evaluation in the presence of sample selection," *Econometric Reviews*, 33, 869–905.

——— (2015): "Causal pitfalls in the decomposition of wage gaps," *Journal of Business and Economic Statistics*, 33, 179–191.

Imai, K. (2009): "Statistical analysis of randomized experiments with non-ignorable missing binary outcomes: an application to a voting experiment," *Journal of the Royal Statistical Society Series C*, 58, 83–104.

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 5171.

Imai, K., and T. Yamamoto (2011): "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments," *unpulished manuscript*.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

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

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

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

Krueger, A. B., and D. M. Whitmore (2001): "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR," *The Economic Journal*, 111, 1–28.

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

Little, R. J. A. (1995): "Modeling the Drop-Out Mechanism in Repeated-Measures Studies," *Journal of the American Statistical Association*, 90, 1112–1121.

MacKinnon, D. P. (2008): *Introduction to Statistical Mediation Analysis.* Taylor and Francis, New York.

Morrissey, T. W., L. Hutchison, and A. Winsler (2014): "Family Income, School Attendance, and Academic Achievement in Elementary School," *Developmental Psychology*, 50, 741–753.

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

Newey, W. K. (1984): "A method of moments interpretation of sequential estimators," *Economics Letters*, 14, 201–206.

——— (2007): "Nonparametric continuous/discrete choice models," *International Economic Review*, 48, 1429–1439.

Nye, B., L. V. Hedges, and S. Konstantopoulos (2001): "The Long-Term Effects of Small Classes in Early Grades: Lasting Benefits in Mathematics Achievement at Grade 9," *The Journal of Experimental Education*, 69, 245–257.

Oaxaca, R. (1973): "Male-Female Wage Differences in Urban Labour Markets," *International Economic Review*, 14, 693–709.

Odongo, D. O., W. J. Wakhungu, and O. Stanley (2017): "Causes of variability in prevalence rates of communicable diseases among secondary school Students in Kisumu County, Kenya," *Journal of Public Health*, 25, 161–166.

Pearl, J. (1995): "Causal Diagrams for Empirical Research," *Biometrika*, 82, 669–710 (with discussion).

——— (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.

Ready, D. D. (2010): "Socioeconomic Disadvantage, School Attendance, and Early Cognitive Development: The Differential Effects of School Exposure," *Sociology of Education*, 83, 271–286.

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.

Robins, J. M., and T. Richardson (2010): "Alternative graphical causal models and the identification of direct effects," in *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, ed. by P. Shrout, K. Keyes, and K. Omstein. Oxford University Press.

Robins, J. M., A. Rotnitzky, and L. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors Are not Always Observed," *Journal of the American Statistical Association*, 90, 846–866.

——— (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of American Statistical Association*, 90, 106–121.

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

——— (1976): "Inference and Missing Data," *Biometrika*, 63, 581–592.

——— (1990): "Formal Modes of Statistical Inference For Causal Effects," *Journal of Statistical Planning and Inference*, 25, 279–292.

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

Shah, A., N. Laird, and D. Schoenfeld (1997): "A Random-Effects Model for Multiple Characteristics With Possibly Missing Data," *Journal of the American Statistical Association*, 92, 775–779.

Tchetgen Tchetgen, E. J., and I. Shpitser (2012): "Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis," *The Annals of Statistics*, 40, 1816–1845.

Tchetgen Tchetgen, E. J., and T. J. VanderWeele (2014): "On Identification of Natural Direct Effects when a Confounder of the Mediator is Directly Affected by Exposure," *Epidemiology*, 25, 282–291.

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.

Vansteelandt, S., M. Bekaert, and T. Lange (2012): "Imputation Strategies for the Estimation of Natural Direct and Indirect Effects," *Epidemiologic Methods*, 1, 129–158.

Wooldridge, J. (2002): "Inverse Probability Weigthed M-Estimators for Sample Selection, Attrition and Stratification," *Portuguese Economic Journal*, 1, 141–162.

——— (2007): "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 141, 1281–1301.

Zheng, W., and M. J. van der Laan (2012): "Targeted Maximum Likelihood Estimation of Natural Direct Effects," *The International Journal of Biostatistics*, 8, 1–40.

A Appendix

A.1 Proof of Theorem 1

$$
\begin{aligned}
& E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|M,X)\cdot\Pr(S=1|D,M,X)} \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)}\right] \qquad (A.1)\\
= & \underset{X}{E}\left[\underset{M|X=x}{E}\left[E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|M,X)\cdot\Pr(S=1|D,M,X)}\Big|M=m,X=x\right]\cdot\frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)}\right]\right]\\
= & \underset{X}{E}\left[\underset{M|X=x}{E}\left[E\left[\frac{Y \cdot S}{\Pr(S=1|D,M,X)}\Big|D=d,M=m,X=x\right]\cdot\frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)}\right]\right]\\
= & \underset{X}{E}\left[\underset{M|X=x}{E}\left[E\left[Y|D=d,M=m,X=x,S=1\right]\cdot\frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)}\right]\right]\\
= & \underset{X}{E}\left[\underset{M|D=1-d,X=x}{E}\left[E\left[Y|D=d,M=m,X=x,S=1\right]\right]\right]\\
= & \underset{X}{E}\left[\underset{M|D=1-d,X=x}{E}\left[E\left[Y|D=d,M=m,X=x\right]\right]\right]\\
= & \underset{X}{E}\left[\underset{M|D=1-d,X=x}{E}\left[E\left[Y(d,m)|D=d,M=m,X=x\right]\right]\right]\\
= & \underset{X}{E}\left[\underset{M|D=1-d,X=x}{E}\left[E\left[Y(d,m)|D=d,X=x\right]\right]\right]\\
= & \underset{X}{E}\left[\underset{M(1-d)|X=x}{E}\left[E\left[Y(d,m)|D=1-d,X=x\right]\right]\right]\\
= & \underset{X}{E}\left[\underset{M(1-d)|X=x}{E}\left[E\left[Y(d,m)|D=1-d,M(1-d)=m,X=x\right]\right]\right]\\
= & \underset{X}{E}\left[\underset{M(1-d)|X=x}{E}\left[E\left[Y(d,m)|M(1-d)=m,X=x\right]\right]\right]\\
= & \underset{X}{E}\left[\underset{M(1-d)|X=x}{E}\left[E\left[Y(d,M(1-d))|X=x\right]\right]\right] = E[Y(d,M(1-d))].
\end{aligned}
$$

Note that $\underset{A|B=b}{E}[C]$ denotes the expectation of C taken over the distribution of A conditional on $B=b$. The first equality follows from the law of iterated expectations, the second and third from basic probability theory, the fourth from Bayes' theorem, the fifth from Assumption 3, the sixth from the observational rule (implying for instance that Y given $D=d$ and $M=m$ is $Y(d,m)$), the seventh from Assumption 2, the eighth from Assumption 1, the ninth from Assumption 2, the tenth from Assumption 1, which implies that $Y(d,m) \perp D | M(1-d)=m, X=x$, and the last from the law of iterated expectations.

$$
\begin{aligned}
& E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|X)\cdot\Pr(S=1|D,M,X)}\right] \qquad (A.2)\\
= & \underset{X}{E}\left[E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|X)\cdot\Pr(S=1|D,M,X)}\Big|X=x\right]\right]\\
= & \underset{X}{E}\left[E\left[\frac{Y \cdot S}{\Pr(S=1|D,M,X)}\Big|D=d,X=x\right]\right]\\
= & \underset{X}{E}\left[\underset{M|D=d,X=x}{E}\left[\frac{E[Y\cdot S|D=d,M=m,X=x]}{\Pr(S=1|D,M,X)}\Big|D=d,X=x\right]\right]\\
= & \underset{X}{E}\left[\underset{M|D=d,X=x}{E}\left[E[Y|D=d,M=m,X=x,S=1]|D=d,X=x\right]\right]\\
= & \underset{X}{E}\left[\underset{M|D=d,X=x}{E}\left[E[Y|D=d,M=m,X=x]|D=d,X=x\right]\right]\\
= & \underset{X}{E}\left[E\left[Y|D=d,X=x\right]\right]\\
= & \underset{X}{E}\left[E\left[Y(d,M(d))|D=d,X=x\right]\right]\\
= & \underset{X}{E}\left[E\left[Y(d,M(d))|X=x\right]\right] = E[Y(d,M(d))].
\end{aligned}
$$

The first, third, sixth, and ninth equalities follow from the law of iterated expectations, the second and fourth from basic probability theory, the fifth from Assumption 3, the seventh from the observational rule, and the eighth from Assumption 1.

$$
\begin{aligned}
& E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\} \cdot S}{\Pr(D=d|X) \cdot \Pr(M=m|D,X) \cdot \Pr(S=1|D,M,X)}\right] \qquad \text{(A.3)}\\
= \; & \underset{X}{E}\left[E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\} \cdot S}{\Pr(D=d|X) \cdot \Pr(M=m|D,X) \cdot \Pr(S=1|D,M,X)}\Big|X=x\right]\right]\\
= \; & \underset{X}{E}\left[E\left[Y|D=d,M=m,X=x,S=1\right]\right]\\
= \; & \underset{X}{E}\left[E\left[Y|D=d,M=m,X=x\right]\right]\\
= \; & \underset{X}{E}\left[E\left[Y(d,m)|D=d,M=m,X=x\right]\right]\\
= \; & \underset{X}{E}\left[E\left[Y(d,m)|D=d,X=x\right]\right]\\
= \; & \underset{X}{E}\left[E\left[Y(d,m)|X=x\right]\right] = E[Y(d,m)]
\end{aligned}
$$

The first and seventh equalities follow from the law of iterated expectations, the second from basic probability theory, the third from Assumption 3, the fourth from the observational rule, the fifth from Assumption 2, and the sixth from Assumption 1.

A.2 Proof of Theorem 2

$$
\begin{aligned}
& E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X,p(W))} \cdot \frac{\Pr(D=1-d|M,X,p(W))}{\Pr(D=1-d|X,p(W))}\Big|S=1\right] \qquad \text{(A.4)}\\
= \; & \underset{X,p(W)|S=1}{E}\left[\underset{M|X=x,p(W)=p(w),S=1}{E}\left[E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X,p(W))}\Big|M=m,X=x,p(W)=p(w),S=1\right]\right.\right.\\
\times \; & \left.\left.\frac{\Pr(D=1-d|M,X,p(W)}{\Pr(D=1-d|X,p(W))}\right]\right]\\
= \; & \underset{X,p(W)|S=1}{E}\left[\underset{M|X=x,p(W)=p(w),S=1}{E}\left[E\left[Y|D=d,M=m,X=x,p(W)=p(w),S=1\right]\right.\right.\\
\times \; & \left.\left.\frac{\Pr(D=1-d|M,X,p(W))}{\Pr(D=1-d|X,p(W))}\right]\right]\\
= \; & \underset{X,p(W)|S=1}{E}\left[\underset{M|D=1-d,X=x,p(W)=p(w),S=1}{E}\left[E\left[Y(d,m)|D=d,M=m,X=x,p(W)=p(w),S=1\right]\right]\right]\\
= \; & \underset{X,p(W)|S=1}{E}\left[\underset{M|D=1-d,X=x,p(W)=p(w),S=1}{E}\left[E\left[Y(d,m)|D=d,X=x,p(W)=p(w),S=1\right]\right]\right]\\
= \; & \underset{X,p(W)|S=1}{E}\left[\underset{M(1-d)|X=x,p(W)=p(w),S=1}{E}\left[E\left[Y(d,m)|D=1-d,X=x,p(W)=p(w),S=1\right]\right]\right]\\
= \; & \underset{X,p(W)|S=1}{E}\left[\underset{M(1-d)|X=x,p(W)=p(w),S=1}{E}\left[E\left[Y(d,m)|D=1-d,M(1-d)=m,X=x,p(W)=p(w),S=1\right]\right]\right]\\
= \; & \underset{X,p(W)|S=1}{E}\left[\underset{M(1-d)|X=x,p(W)=p(w),S=1}{E}\left[E\left[Y(d,m)|M(1-d)=m,X=x,p(W)=p(w),S=1\right]\right]\right]\\
= \; & \underset{X,p(W)|S=1}{E}\left[\underset{M(1-d)|X=x,p(W)=p(w),S=1}{E}\left[E\left[Y(d,M(1-d))|X=x,p(W)=p(w),S=1\right]\right]\right]\\
= \; & E[Y(d,M(1-d))|S=1].
\end{aligned}
$$

The first equality follows from the law of iterated expectations, the second from basic probability theory, the third from Bayes' theorem and the observational rule, the fourth from Assumptions 2 and 5 (which imply $Y(d,m) \perp M|D=d', X=x, p(W)=p(w), S=1$), the fifth from Assumptions 1 and 5 (which imply $\{Y(d,m), M(1-d)\} \perp D|X=x, p(W)=p(w), S=1$), the sixth from Assumptions 2 and 5, the seventh from

Assumptions 1 and 5 (which imply $Y(d,m) \perp D | M(1-d) = m, X = x, p(W) = p(w), S = 1$), and the last from the law of iterated expectations.

$$
\begin{aligned}
& E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|X,p(W))}\Big|S=1\right] \qquad (A.5)\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|X,p(W))}\Big|X=x,p(W)=p(w),S=1\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y|D=d,X=x,p(W)=p(w),S=1\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y(d,M(d))|D=d,X=x,p(W)=p(w),S=1\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y(d,M(d))|X=x,p(W)=p(w),S=1\right]\right] = E[Y(d,M(d))|S=1].
\end{aligned}
$$

The first and last equalities follow from the law of iterated expectations, the second from basic probability theory, the third from the observational rule, and the fourth from Assumptions 1 and 5 (which imply $Y(d,m) \perp D | X = x, p(W) = p(w), S = 1$).

$$
\begin{aligned}
& E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\}}{\Pr(D=d|X,p(W)) \cdot \Pr(M=m|D,X,p(W))}\Big|S=1\right] \qquad (A.6)\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\}}{\Pr(D=d|X,p(W)) \cdot \Pr(M=m|D,X,p(W))}\Big|X=x,p(W)=p(w),S=1\right]\Big|S=1\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y|D=d,M=m,X=x,p(W)=p(w),S=1\right]\Big|S=1\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y(d,m)|D=d,M=m,X=x,p(W)=p(w),S=1\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y(d,m)|D=d,X=x,p(W)=p(w),S=1\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y(d,m)|X=x,p(W)=p(w),S=1\right]\right] = E[Y(d,m)|S=1]
\end{aligned}
$$

The first and sixth equalities follow from the law of iterated expectations, the second from basic probability theory, the third from the observational rule, the fourth from Assumptions 2 and 5 (which imply $Y(d,m) \perp M | D = d, X = x, p(W) = p(w), S = 1$), and the fifth from Assumptions 1 and 5 (which imply $Y(d,m) \perp D | X = x, p(W) = p(w), S = 1$).

A.3 Proof of Theorem 3

$$
\begin{aligned}
& E\left[\left(\frac{Y\cdot D}{\Pr(D=1|M,X,p(W))}-\frac{Y\cdot(1-D)}{1-\Pr(D=1|M,X,p(W))}\right)\cdot\frac{\Pr(D=d|M,X,p(W))\cdot S}{\Pr(D=d|X,p(W))\cdot p(W)}\right] \qquad (A.7)\\
= & \underset{X,p(W)}{E}\left[\underset{M|X=x,p(W)=p(w)}{E}\left[E\left[\frac{Y\cdot D\cdot S}{\Pr(D=1|M,X,p(W))\cdot p(W)}\right.\right.\right.\\
- & \left.\left.\left.\frac{Y\cdot(1-D)\cdot S}{1-\Pr(D=1|M,X,p(W))\cdot p(W)}\Big|M=m,X=x,p(W)=p(w)\right]\cdot\frac{\Pr(D=d|M,X,p(W))}{\Pr(D=d|X,p(W))}\right]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M|X=x,p(W)=p(w)}{E}\left[E\left[\frac{Y\cdot S}{p(W)}\Big|D=1,M=m,X=x,p(W)=p(w)\right]\right.\right.\\
- & \left.\left.E\left[\frac{Y\cdot S}{p(W)}\Big|D=0,M=m,X=x,p(W)=p(w)\right]\cdot\frac{\Pr(D=d|M,X,p(W))}{\Pr(D=d|X,p(W))}\right]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M|X=x,p(W)=p(w)}{E}[E[Y|D=1,M=m,X=x,p(W)=p(w),S=1]\right.\\
- & \left.E[Y|D=0,M=m,X=x,p(W)=p(w),S=1]\cdot\frac{\Pr(D=d|M,X,p(W))}{\Pr(D=d|X,p(W))}\right]\Big]\\
= & \underset{X,p(W)}{E}\left[\underset{M|D=d,X=x,p(W)=p(w)}{E}[E[Y(1,m)|D=1,M=m,X=x,p(W)=p(w),S=1]\right.\\
- & \left.E[Y(0,m)|D=0,M=m,X=x,p(W)=p(w),S=1]\right]\Big]\\
= & \underset{X,p(W)}{E}\left[\underset{M|D=d,X=x,p(W)=p(w)}{E}[E[Y(1,m)|D=1,X=x,p(W)=p(w),S=1]\right.\\
- & \left.E[Y(0,m)|D=0,X=x,p(W)=p(w),S=1]\right]\Big]\\
= & \underset{X,p(W)}{E}\left[\underset{M(d)|X=x,p(W)=p(w)}{E}[E[Y(1,m)-Y(0,m)|X=x,p(W)=p(w),S=1]]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M(d)|X=x,p(W)=p(w)}{E}[E[Y(1,m)-Y(0,m)|X=x,p(W)=p(w)]]\right]=\theta(d)
\end{aligned}
$$

The first and last equalities follow from the law of iterated expectations, the second from basic probability theory, the third from basic probability theory and the fact that $\Pr(S=1|D,M,X,p(W))=\Pr(S=1|D,M,X,Z)=p(W)$ (as $p(W)$ is a deterministic function of Z conditional on D,M,X), the fourth from Bayes' theorem and the observational rule, the fifth from Assumptions 2 and 5 (which imply $Y(d,m)\perp M|D=d', X=x,p(W)=p(w),S=1$), the sixth from Assumptions 1 and 5 (which imply $\{Y(d,m),M(d')\}\perp D|X=x,p(W)=p(w),S=1$), and the seventh from Assumption 7 by acknowledging that $p(W)=F_V$.

$$
\begin{aligned}
& E\left[\frac{Y\cdot I\{D=d\}\cdot S}{\Pr(D=d|M,X,p(W))\cdot p(W)}\cdot\left(\frac{\Pr(D=1|M,X,p(W))}{\Pr(D=1|X,p(W))}-\frac{1-\Pr(D=1|M,X,p(W))}{1-\Pr(D=1|X,p(W))}\right)\right] \qquad (A.8)\\
= & \underset{X,p(W)}{E}\left[\underset{M|X=x,p(W)=p(w)}{E}\left[E\left[\frac{Y\cdot I\{D=d\}\cdot S}{\Pr(D=d|M,X,p(W))\cdot p(W)}\Big|M=m,X=x,p(W)=p(w)\right]\right.\right.\\
\times & \left.\left.\left(\frac{\Pr(D=1|M,X,p(W))}{\Pr(D=1|X,p(W))}-\frac{1-\Pr(D=1|M,X,p(W))}{1-\Pr(D=1|X,p(W))}\right)\right]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M|X=x,p(W)=p(w)}{E}\left[E[Y|D=d,M=m,X=x,p(W)=p(w),S=1]\cdot\left(\frac{\Pr(D=1|M,X,p(W))}{\Pr(D=1|X,p(W))}-\frac{1-\Pr(D=1|M,X,p(W))}{1-\Pr(D=1|X,p(W))}\right)\right]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M|D=1,X=x,p(W)=p(w)}{E}[E[Y(d,m)|D=d,M=m,X=x,p(W)=p(w),S=1]]\right.\\
- & \left.\underset{M|D=0,X=x,p(W)=p(w)}{E}[E[Y(d,m)|D=d,M=m,X=x,p(W)=p(w),S=1]]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M|D=1,X=x,p(W)=p(w)}{E}[E[Y(d,m)|D=d,X=x,p(W)=p(w),S=1]]\right.\\
- & \left.\underset{M|D=0,X=x,p(W)=p(w)}{E}[E[Y(d,m)|D=d,X=x,p(W)=p(w),S=1]]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M(1)|X=x,p(W)=p(w)}{E}[E[Y(d,m)|X=x,p(W)=p(w),S=1]]-\underset{M(0)|X=x,p(W)=p(w)}{E}[E[Y(d,m)|X=x,p(W)=p(w),S=1]]\right]\\
= & \underset{X,p(W)}{E}[E[Y(d,M(1))-Y(d,M(0))|X=x,p(W)=p(w)]]=\delta(d)
\end{aligned}
$$

The first and last equalities follow from the law of iterated expectations, the second from basic probability theory and the fact that $\Pr(S = 1|D, M, X, p(W)) = \Pr(S = 1|D, M, X, Z) = p(W)$, the third from Bayes' theorem and the observational rule, the fourth from Assumptions 2 and 5 (which imply $Y(d, m) \perp M|D = d', X = x, p(W) = p(w), S = 1$), the fifth from Assumptions 1 and 5 (which imply $\{Y(d, m), M(d')\} \perp D|X = x, p(W) = p(w), S = 1$), and the sixth from Assumption 7 by acknowledging that $p(W) = F_V$.

$$
\begin{aligned}
& E\left[\left(\frac{Y \cdot D}{\Pr(D = 1|X, p(W))} - \frac{Y \cdot (1 - D)}{1 - \Pr(D = 1|X, p(W))}\right) \cdot \frac{I\{M = m\} \cdot S}{\Pr(M = m|D, X, p(W)) \cdot p(W)}\right] \qquad \text{(A.9)} \\
= & \underset{X, p(W)}{E}\left[E\left[\left(\frac{Y \cdot D}{\Pr(D = 1|X, p(W))} - \frac{Y \cdot (1 - D)}{1 - \Pr(D = 1|X, p(W))}\right) \cdot \frac{I\{M = m\} \cdot S}{\Pr(M = m|D, X, p(W)) \cdot p(W)} \Big| X = x, p(W) = p(w)\right]\right] \\
= & \underset{X, p(W)}{E}[E[Y|D = 1, M = m, X = x, p(W) = p(w), S = 1] - E[Y|D = 0, M = m, X = x, p(W) = p(w), S = 1]] \\
= & \underset{X, p(W)}{E}[E[Y(1, m)|D = 1, M = m, X = x, p(W) = p(w), S = 1] - E[Y(0, m)|D = 0, M = m, X = x, p(W) = p(w), S = 1]] \\
= & \underset{X, p(W)}{E}[E[Y(1, m)|D = 1, X = x, p(W) = p(w), S = 1] - E[Y(0, m)|D = 0, X = x, p(W) = p(w), S = 1]] \\
= & \underset{X, p(W)}{E}[E[Y(1, m) - Y(0, m)|X = x, p(W) = p(w), S = 1]] \\
= & \underset{X, p(W)}{E}[E[Y(1, m) - Y(0, m)|X = x, p(W) = p(w)]] = \gamma(m)
\end{aligned}
$$

The first and last equalities follow from the law of iterated expectations, the second from basic probability theory and the fact that $\Pr(S = 1|D, M, X, p(W)) = \Pr(S = 1|D, M, X, Z) = p(W)$, the third from the observational rule, the fourth from Assumptions 2 and 5 (which imply $Y(d, m) \perp M|D = d, X = x, p(W) = p(w), S = 1$), the fifth from Assumptions 1 and 5 (which imply $Y(d, m) \perp D|X = x, p(W) = p(w), S = 1$), and the sixth from Assumption 7 by acknowledging that $p(W) = F_V$.

Working Paper 496 august 2018

Authors

Martin HUBER
University of Fribourg, Department of Economics, Chair of Applied Econometrics - Evaluation of Public Policies, Bd. de Pérolles 90, 1700 Fribourg, Switzerland.
Phone: +41 26 300 8274; Email: martin.huber@unifr.ch; Website: http://www.unifr.ch/appecon/en/team/martin-huber

Anna SOLOVYEVA
University of Fribourg, Department of Economics, Chair of Applied Econometrics - Evaluation of Public Policies, Bd. de Pérolles 90, 1700 Fribourg, Switzerland.
Phone: +41 26 300 8255; Email: anna.solovyeva@unifr.ch; Website: http://www.unifr.ch/appecon/en

Abstract

This paper considers the evaluation of direct and indirect treatment effects, also known as mediation analysis, when outcomes are only observed for a subpopulation due to sample selection or outcome attrition. For identification, we combine sequential conditional independence assumptions on the assignment of the treatment and the mediator, i.e. the variable through which the indirect effect operates, with either selection on observables/missing at random or instrumental variable assumptions on the outcome attrition process. We derive expressions for the effects of interest that are based on inverse probability weighting by specific treatment, mediator, and/or selection propensity scores. We also provide a brief simulation study and an empirical illustration based on U.S. Project STAR data that assesses the direct effect and indirect effect (via absenteeism) of smaller kindergarten classes on math test scores.

Citation proposal

Martin Huber, Anna Solovyeva. 2018. «Direct and indirect effects under sample selection and outcome attrition». Working Papers SES 496, Faculty of Economics and Social Sciences, University of Fribourg (Switzerland)

Jel Classification

C21, I21.

Keywords

Causal mechanisms, direct effects, indirect effects, causal channels, mediation analysis, causal pathways, sample selection, attrition, outcome nonresponse, inverse probability weighting, propensity score.

Working Papers SES collection

Last published

489 Frölich M., Huber M.: Including covariates in the regression discontinuity design; 2017
490 Eugster N., Isakov D.: Founding family ownership, stock market returns, and agency problems; 2017
491 Eugster N.: Family Firms and Financial Analyst Activity; 2017
492 Andresen M.E., Huber M.: Instrument-based estimation with binarized treatments: Issues and tests for the exclusion restriction; 2018
493 Bodory H., Huber M.: The causalweight package for causal inference in R; 2018
494 Huber M., Imhof D.: Machine Learning with Screens for Detecting Bid-Rigging Cartels; 2018
495 Hsu Y.-C., Huber M., Lee Y.-Y.: Direct and indirect effects of continuous treatments based on generalized propensity score weighting; 2018

Catalogue and download links

http://www.unifr.ch/ses/wp
http://doc.rero.ch/collection/WORKING_PAPERS_SES

Publisher

Université de Fribourg, Suisse, Faculté des sciences économiques et sociales
Universität Freiburg, Schweiz, Wirtschafts- und sozialwissenschaftliche Fakultät
University of Fribourg, Switzerland, Faculty of Economics and Social Sciences

Bd de Pérolles 90, CH-1700 Fribourg
Tél.: +41 (0) 26 300 82 00
decanat-ses@unifr.ch www.unifr.ch/ses