Identifying causal mechanisms in experiments (primarily) based on inverse probability weighting

Martin Huber

University of St. Gallen, Dept. of Economics

Abstract: This paper demonstrates the identification of causal mechanisms in experiments with a binary treatment, (primarily) based on inverse probability weighting. I.e., we consider the average indirect effect of the treatment, which operates through an intermediate variable (or mediator) that is situated on the causal path between the treatment and the outcome, as well as the (unmediated) direct effect. Even under random treatment assignment, subsequent selection into the mediator is generally non-random such that causal mechanisms are only identified when controlling for confounders of the mediator and the outcome. To tackle this issue, units are weighted by the inverse of their conditional treatment propensity given the mediator and observed confounders. We show that the form and applicability of weighting depend on whether the confounders are themselves influenced by the treatment or not. A simulation study gives the intuition for these results and an empirical application to the direct and indirect health effects (through employment) of the U.S. Job Corps program is also provided.

Keywords: causal mechanisms, mediation analysis, direct and indirect effects, experiment, inverse probability weighting.

JEL classification: C14, C21, I38.

I have benefited from comments by Guido Imbens, Michael Lechner, Teppei Yamamoto, and participants of the conference "Frontiers in the Analysis of Causal Mechanisms" (Harvard, March 2012). Financial support from the Swiss National Science Foundation grant PBSGP1_138770 is gratefully acknowledged. Address for correspondence: Martin Huber, SEW, University of St. Gallen, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland, martin.huber@unisg.ch.

1 Introduction

Randomized experiments, which in social sciences date at least back to Neyman (1923) and Fisher (1925, 1935), are a cornerstone of the evaluation of policy interventions and widely regarded to be the gold standard of causal inference, see for instance Cochran and Chambers (1965), Freedman (2006), and Rubin (2008). In a properly designed experiment, the average treatment effect (ATE) of some intervention simply corresponds to the difference in the expected values of the outcomes conditional on the presence and the absence of the intervention. However, in many economic problems, not only the (total) ATE appears relevant, but also the causal mechanisms through which it operates. In this case, one would like to disentangle the *direct* effect of the treatment on the outcome as well as the *indirect* ones that run through one or more intermediate variables, so-called mediators. E.g., when assessing the employment or earnings effects of an active labor market policy, researchers and policy makers may be interested to which extent the total impact comes from increased search effort, increased human capital, or other mediators that are themselves affected by the policy.

However, even in experiments, causal mechanisms are not easily identified. As discussed in Robins and Greenland (1992), random treatment assignment does not imply randomness of the mediator, which may be regarded as intermediate outcome. Therefore, the total effect cannot be disentangled by simply conditioning on a mediator that is itself affected by the treatment, because this generally introduces selection bias coming from confounders of the mediator and the outcome, see Rosenbaum (1984). For this reason, already the early work on mediation analysis of Judd and Kenny (1981) highlights the importance of controlling for such confounders. In the light of these contributions, it seems surprising that this issue has been ignored in so many applications in social sciences that claim to identify direct and indirect effects.

This paper demonstrates the identification of causal mechanisms under discrete or continuous mediators in experiments, mainly based on inverse probability weighting (IPW).[1] I.e., units are

[1]The idea of IPW goes back to Horvitz and Thompson (1952), who first proposed an estimator of the population mean in the presence of non-randomly missing data.

weighted by the inverse of their conditional propensity to be observed in a particular treatment state given the mediator and the observed covariates. Identification relies on the assumption that the mediator is conditionally exogenous given these variables, implying that the mediator and the potential outcomes for a particular mediator state are conditionally independent given the covariates and the treatment.[2] We also discuss that for the identification of indirect effects, the results depend on whether the covariates are themselves a function of the treatment. If the latter is the case, the identification of the "total" indirect effect, which also accounts for correlations between the covariates and the mediator, requires additional restrictions, see the proof in Avin, Shpitser, and Pearl (2005) and the discussion in Robins (2003) and Imai and Yamamoto (2011). In contrast, the "partial" indirect effect, which only considers the immediate link between the treatment and the mediator (and no "detour" via the covariates), is identified under weaker assumptions. We provide a simulation study that gives the intuition for these identification issues. Finally, we apply our methods to experimental data on Job Corps, an vocational/educational program for disadvantaged youths in the U.S. that also includes health care and health education. We disentangle its impact on general health into the indirect effect mediated by labor market success (finding employment) and a direct remainder component. The results are partly sensitive to whether the covariates are admitted to be affected by the treatment, which highlights the importance of carefully choosing the identifying assumptions.

Whereas the evaluation of direct and indirect effects, often referred to as mediation analysis, is still a comparably small field of research in economics, it is widespread in other social sciences such as epidemiology, political sciences, and psychology, see for instance MacKinnon (2008) for typical applications. Even though the idea is not entirely new, see Cochran (1957), it appears to be the paper of Baron and Kenny (1986) that triggered the popularity of mediation analysis. While most studies rely on inflexible linear specifications, more general identification under conditional exogeneity of the mediator has been considered by Pearl (2001), Robins (2003), Petersen, Sinisi, and van der Laan (2006), VanderWeele (2009), Imai, Keele, and Yamamoto (2010), Albert and

[2]This is similar in spirit, but yet different to observational evaluation studies of the (total) ATE, see for instance Imbens (2004) and Imbens and Wooldridge (2009), where conditional independence refers to the treatment rather than the mediator.

Nelson (2011), and Imai and Yamamoto (2011), among others. One of the rare studies in the field of economics is Flores and Flores-Lagunes (2009), who use the Job Corps experiment to evaluate the direct earnings effect of the intervention after controlling for the mediator "work experience". The issue is that participating in a training is likely to decrease work experience shortly after the program start compared to nonparticipation, a phenomenon known as "locking-in effect" due to a decreased job search effort during training participation. Assuming that the mediator is exogenous conditional on pre-treatment covariates, Flores and Flores-Lagunes (2009) estimate a positive effect on earnings based on a regression approach. As a further example, Simonsen and Skipper (2006) use a semiparametric identification strategy based on matching to assess the direct wage effect of motherhood in Denmark by controlling for several mediators through which motherhood may have an influence on wages. They find negative direct effects which vary little across different sectors.

Our paper makes four contributions to the literature on causal mechanisms in economics: Firstly, it derives identification results based on IPW by the treatment propensity score that are straightforward to implement by semiparametric (if the score ist estimated parametrically) or nonparametric estimation (if the score is estimated nonparametrically). Furthermore, if the mediator is exogenous conditional on pre-treatment covariates, our approach allows relaxing one functional form assumption imposed in Flores and Flores-Lagunes (2009) (their Assumption 3). It is also easier to implement than the nonparametric estimators of Imai, Keele, and Yamamoto (2010), which require estimating the conditional mean of the outcome and the conditional density of the mediator. Secondly and in contrast to Flores and Flores-Lagunes (2009) and Simonsen and Skipper (2006), we also discuss identification when mediator exogeneity only holds conditional on post-treatment covariates which are themselves a function of the treatment, such that pre-treatment variables do not fully capture the endogeneity. This appears realistic in most applications including Job Corps, where the treatment likely affects variables that potentially confound the mediator and the outcome, e.g., intermediate health shortly before the mediator. While direct effects are still identified by IPW in this set up after a modification of the initial

assumptions, the identification of indirect effects requires additional restrictions. We present a functional form restriction allowing us to do so, which, however, is less general than the entirely nonparametric identification under IPW. Thirdly, we show that IPW still identifies a partial indirect effect when keeping the confounders fixed, i.e., the part of the indirect effect which is not correlated with the confounders. Fourthly, as an empirical contribution, the present work appears to be the first which assesses the direct and indirect health effects of the Job Corps program.

The remainder of this paper is organized as follows. Section 2 defines the parameters of interest (the average direct and indirect effects) and discusses identification (mostly) based on IPW. Section 3 presents a simulation study which provides the intuition for the issues related to the identification of causal mechanisms in experiments. In Section 4, we apply our methods to the experimental study of the Job Corps program. Section 5 concludes.

2 Parameters of interest and identification

Suppose we are interested in the average treatment (ATE) effect of a binary treatment indicator D on some outcome variable Y. Furthermore, assume that we would like to disentangle the ATE into a direct component and an indirect effect operating through the mediator M which has bounded support and may be discrete or continuous. To define the parameters of interest, we use the potential outcome framework advocated by Rubin (1974) (among many others) and considered in the direct and indirect effects framework for instance by Rubin (2004), Ten Have, Joffe, Lynch, Brown, Maisto, and Beck (2007), and Albert (2008). Let $Y(d), M(d)$ denote the potential outcome and the potential mediator state under treatment $d \in \{0, 1\}$. For each unit only one of the two potential outcomes and mediator states, respectively, is observed, because the realized outcome and mediator values are $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$ and $M = D \cdot M(1) + (1 - D) \cdot M(0)$.

The ATE is defined by $\Delta = E[Y(1) - Y(0)]$. To disentangle this total effect into a direct and indirect (through M) causal channel, first note that the potential outcome can be rewritten as a

function of both the treatment and the intermediate variable M: $Y(d) = Y(d, M(d))$. It follows that the (average) direct effect is identified by

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

i.e., by exogenously varying the treatment but keeping the mediator fixed at its potential value for $D = d$. Equivalently, the (average) indirect effects is defined as

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

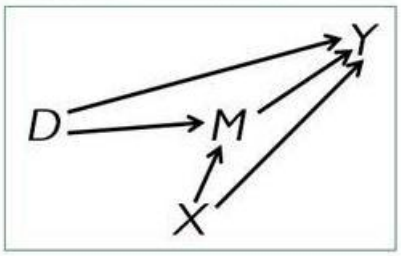
i.e., by exogenously shifting the mediator to its potential values under treatment and non-treatment but keeping the treatment fixed at $D = d$. Pearl (2001) named these parameters the natural direct and indirect effects, whereas Robins (2003) referred to them as the pure direct and indirect effects and Flores and Flores-Lagunes (2009) as net and mechanism average treatment effects, respectively. It is obvious that these 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 $Y(1, M(0))$ are never observed. Therefore, identification of direct and indirect effects hinges on the generation of exogenous variation (i) in the treatment and (ii) in the post-treatment mediator.

Furthermore, note that the ATE is the sum of the direct and indirect effects defined upon opposite treatment states:

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

which follows from adding and subtracting $E[Y(0, M(1))]$ after the first and $E[Y(1, M(0))]$ after the third equality. Using this notation highlights the possibility of effect heterogeneity in causal mechanisms w.r.t. the treatment state, i.e., the presence of interaction effects between the treatment and the mediator. Unfortunately, this possibility has been ignored by the vast majority of studies relying on linear models for mediation analysis. Here and in other recent studies considering nonparametric identification (e.g., Imai, Keele, and Yamamoto (2010) and Imai and Yamamoto (2011)), the possibility of interaction effects between D and M is explicitly accounted for by using the potential outcome framework.

Figure 1: Causal paths under conditional exogeneity of the mediator



After having defined the parameters of interest, we now consider our identifying assumptions, maintaining an i.i.d. framework throughout the paper. We start with the framework of conditional mediator exogeneity given observed covariates (denoted by X) which are themselves *not* a function of D, see Figure 1 for a graphical illustration using a directed acyclic graph. The leading case for this framework is mediator exogeneity conditional on pre-treatment covariates (evaluated prior to treatment assignment), see Flores and Flores-Lagunes (2009) and Imai, Keele, and Yamamoto (2010). Further below we will consider another set of restrictions by which the mediator is assumed to be exogenous conditional on X which are partially themselves a function of D, and thus, post-treatment variables, see for instance Robins (2003) and Imai and Yamamoto (2011). While the latter case appears more realistic in applications, it also makes identification more difficult. In particular, it requires additional functional form assumptions for the identification of indirect effects which are not required under mediator exogeneity given pre-treatment X where identification is entirely non-parametric.

Our first assumption reflects the experimental context by requiring that actual treatment assignment is independent of any potential post-treatment variable, i.e., the potential mediator states and the potential outcomes. Furthermore, the treatment must be unrelated to X, which is naturally satisfied if X consists of pre-treatment covariates, because the latter do not determine treatment assignment in (successfully) randomized experiments.

Assumption 1 (random treatment assignment and no association with confounders): $\{Y(d', m), M(d), X\} \perp D$ for all $d', d \in \{0, 1\}$ and m in the support of M.

By Assumption 1, the treatment is independent of the observed characteristics and of any unobservable factors jointly affecting the treatment on the one hand and the mediator and/or the outcome on the other hand. This is closely related to Assumption 1 in Flores and Flores-Lagunes (2009), albeit the latter do not make the independence of X and D explicit. They nevertheless assume it implicitly by stating that X are pre-treatment covariates such that D and X are independent under random assignment. Note that our Assumption 1 also implies that $\{Y(d', m), M(d)\}$ is independent of D given X.

The second assumption imposes conditional independence (or exogeneity) of the mediator given the covariates and the treatment along with a common support restriction on the conditional treatment probability:

Assumption 2 (conditional independence of the mediator):
(a) $Y(d', m) \perp M | D = d, X = x$ for all $d', d \in \{0, 1\}$ and m, x in the support of M, X,
(b) $\Pr(D = d | M = m, X = x) > 0$ for all $d \in \{0, 1\}$ and m, x in the support of M, X.

I.e., conditional on D and X, the effect of the mediator on the outcome is assumed to be unconfounded. Note that Assumption 2(a) is equivalent to equation (5) in Imai, Keele, and Yamamoto (2010), which is part of their sequential ignorability assumption also considered in Tchetgen Tchetgen and Shpitser (2011b) (see also the closely related Theorem 2 of Pearl (2001)). Yet, our framework differs from theirs in that we impose Assumption 1 instead of

equation (4) in Imai, Keele, and Yamamoto (2010), saying that $\{Y(d', m), M(d)\} \perp D|X$. I.e., we consider the experimental context of random treatment assignment instead of conditional independence of the treatment given X. Flores and Flores-Lagunes (2009) impose a restriction similar to our Assumption 2(a) in their Assumption 2, albeit defined in terms of potential mediator states given X. Assumption 2(b) 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. Note that by Bayes' theorem, this equivalently 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: $f_{M|D,X}(m, d, x) > 0$). I.e., conditional on X, the mediator state must not be a deterministic function of the treatment, otherwise identification is infeasible due to the lack of comparable units in terms of the mediator across treatment states.

The evaluation of direct and indirect effects hinges on the identifiability of $E[Y(d, M(d))]$ and $E[Y(d, M(1-d))]$. The former is directly observed from the data because $E[Y(d, M(d))] = E[Y(d)] = E[Y|D = d] = E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d)}\right]$ by Assumption 1, where $I\{\cdot\}$ denotes the indicator function that is equal to one if its argument is true and zero otherwise. Concerning the latter, observe that by 2(a), it holds for all $d \in \{0, 1\}$ and m, x in the support of M, X that

$$E[Y(d, m)|D = d, X = x] = E[Y(d, m)|D = d, M = m, X = x] = E[Y|D = d, M = m, X = x].$$

Furthermore, $E[Y(d, m)|D = d, X = x] = E[Y(d, m)|X = x]$ by Assumption 1. I.e., the mean potential outcome under hypothetical treatment and mediator states d, m given $X = x$ is identified by the observed conditional mean outcome given $D = d, M = m, X = x$. By adequately averaging over the distributions of M and X, the unobserved expectation $E[Y(d, M(1-d))]$ is

identified. This implies the following identification result:

$$
\begin{aligned}
& E[Y(d, M(1-d))] \\
= & \int E[Y(d,m)|M=m, X=x] dF_{M(1-d),X}(m,x) \\
= & \int E[Y(d,m)|M=m, X=x] dF_{M,X|D=1-d}(m,x) \\
= & \int E[Y|D=d, M=m, X=x] \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d)} dF_{M,X}(m,x) \\
= & E\left[E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \middle| M=m, X=x \right] \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d)} \right] \\
= & E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d)} \right]. \qquad (4)
\end{aligned}
$$

The first equality follows from the law of iterated expectations and from replacing the outer expectation by an integral, the second from Assumption 1, the third from Assumption 2(a) and Bayes' theorem, the fourth from basic probability theory and from replacing the integral by an expectation, and the last from the law of iterated expectations. Therefore, the direct effect is identified by

$$\theta(d) = E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d)} \right] - E\left[\frac{Y \cdot I\{D=1-d\}}{\Pr(D=1-d|M,X)} \cdot \frac{\Pr(D=d|M,X)}{\Pr(D=d)} \right] \qquad (5)$$

and the indirect effect by

$$\delta(d) = E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d)} \right] - E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d)} \right]. \qquad (6)$$

After some simple algebra we obtain the numerically identical expressions given in Propositions 1 and 2:

Proposition 1:

Under Assumptions 1 and 2, the average direct effect is identified by

$$\theta(d) = E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X)} \right) \cdot \frac{\Pr(D=d|M,X)}{\Pr(D=d)} \right]. \qquad (7)$$

It is worth noting that this IPW-based expression corresponds to that obtained for the ATE on the treated (if $d = 1$) or non-treated (if $d = 0$) in the conceptually different framework of a conditionally independent treatment given observed covariates, see Hirano, Imbens, and Ridder (2003). There, conditioning on observed covariates in the propensity score controls for selection into the treatment, whereas here, conditioning on M, X controls for mediator endogeneity (whereas the treatment is random). Proposition 1 implies that observations are reweighted according to the distribution of $M(d)$ among observations with $D = d$. However, the latter is equivalent to the hypothetical distribution of $M(d)$ in the total population due to the random assignment of D.[3]

By (3), the indirect effect is simply the difference between the average and the direct effect defined by the opposite treatment state: $\delta(d) = \Delta - \theta(1-d)$. This implies that in the current set up, the identification of direct and indirect effects hinges on the same assumptions and that one cannot be identified without the other. Proposition 2 provides the representation of the indirect effect based on IPW, which is numerically identical to the difference $\Delta - \theta(1-d)$:

Proposition 2:

Under Assumptions 1 and 2, the average indirect effect is identified by

$$\delta(d) = E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \cdot \left(\frac{\Pr(D=1|M,X)}{\Pr(D=1)} - \frac{1-\Pr(D=1|M,X)}{1-\Pr(D=1)}\right)\right]. \tag{8}$$

From a practitioner's perspective, a nice feature of these identification results is that they are straightforward to implement. They only involve the (possibly parametric or nonparametric) estimation of a binary choice model for the propensity score which is then plugged into the sample analogs of Propositions 1 and 2. Alternatively, matching on the propensity score (see for instance Rosenbaum and Rubin (1983)) could also be used. In either case, no parametric restrictions

[3]It is worth noting that if the treatment was not randomly assigned but merely conditionally independent such that Assumption 1 was to be replaced by $\{Y(d', m), M(d)\} \perp D|X$ for all $d', d \in \{0, 1\}$ and m in the support of M, identification would require a slight modification of Proposition 1: $\Pr(D = d)$ would have to be substituted by $\Pr(D = d|X)$ to reweight observations according to the hypothetical distribution of $M(d)$ in the total population, which then differed from the treated population. An equivalent argument applies to Proposition 2, where $\Pr(D = 1)$ and $1 - \Pr(D = 1)$ would have to be replaced by $\Pr(D = 1|X)$ and $1 - \Pr(D = 1|X)$.

are imposed on the models of the outcome and the mediator such that arbitrary nonlinearities are allowed for. Only the related but independently developed method of Tchetgen Tchetgen (2012) using weighting by the inverse of the odds ratio relating the treatment and the mediator conditional on the covariates (which is to assume a logit model for the propensity score) appears to share these features. In contrast, the standard approach in the literature consists in estimating the ingredients of the following alternative representations of the parameters of interest, see for instance equations (8) and (26) in Pearl (2001) and Theorem 1 in Imai, Keele, and Yamamoto (2010):

$$\theta(d) = \int\int E[Y|D=1,M=m,X=x] - E[Y|D=0,M=m,X=x]dF_{M|D=d,X=x}(m)dF_X(x), \tag{9}$$

$$\delta(d) = \int\int E[Y|D=d,M=m,X=x]\{dF_{M|D=1,X=x}(m) - dF_{M|D=0,X=x}(m)\}dF_X(x). \tag{10}$$

This requires estimators for the conditional mean of Y given D, M, X and the conditional density of M given D, X. In the literature, parametric methods have been most commonly used, see for instance Pearl (2011) and VanderWeele (2009).[4] They, however, appear unattractive due to their severe functional form restrictions and the potentially difficult interpretability of direct and indirect effects under non-linear modeling (e.g., when both the outcome and the mediator are binary). Nonparametric estimation as recently proposed in Imai, Keele, and Yamamoto (2010) avoids these shortcomings, but might be cumbersome in empirical applications if X is high dimensional and/or M is continuous. In contrast, estimation based on Propositions 1 and 2 is less prone to such issues as it relies on a single propensity score model. Our IPW-based results are also more general than the regression approach of Flores and Flores-Lagunes (2009). The latter does not require the estimation of conditional density of the mediator, but imposes a functional form restriction (their Assumption 3) on the expected potential outcomes across potential mediator states (for the treatment fixed) which we need not invoke here. Only if the

[4]Furthermore, Tchetgen Tchetgen and Shpitser (2011a) and Zheng and van der Laan (2012) (among others) provide doubly robust parametric estimators based on conditional mean estimation of the outcome and conditional density estimation of both the treatment and the mediator.

mediator is exogenous conditional on post-treatment confounders which are themselves influenced by the treatment, we have to rely on a similar assumption, which is outlined further below.

As discussed before, it appears unlikely in most applications that conditioning on pre-treatment variables is sufficient to control for mediator endogeneity, given that the mediator is itself a post-treatment variable. Equivalent to the treatment evaluation literature, where potential confounders of the treatment are measured at or shortly before the treatment, potential confounders of the mediator should be controlled for just before the selection into the mediator takes place. Then, however, it appears likely that at least some of these covariates are also a function of the treatment, implying that they are themselves mediators that affect the mediator of interest. Therefore, Robins (2003) suspects that the set up relying on Assumptions 1 and 2 is of limited practical relevance. This most likely also applies to our application presented in Section 4, where we are interested in the effect of the Job Corps program on health. The mediator is employment and clearly, some potential confounders affecting both employability and health (such as the labor market state shortly prior to employment) are most likely a function of the treatment. Similar issues arguably arise in Flores and Flores-Lagunes (2009), who estimate the direct earnings effect of Job Corps that is not explained by differences in work experience induced by the program, e.g., due to locking-in effects (see for instance van Ours (2004)) during training participation. Even though the authors exclusively condition on pre-treatment covariates, it appears likely that the treatment changes motivation and search effort, which itself might affect both employment (and thus, work experience) and earnings.

Therefore, we also consider a framework in which D is permitted to have an effect on X, see also Robins (2003) and Imai and Yamamoto (2011). In this case, mediation analysis becomes more complicated and requires us to introduce additional notation by rewriting the mediator and outcome also as a function of X: $M(d) = M(d, X(d))$ and $Y(d, M(d)) = Y(d, M(d, X(d)), X(d))$, where $X(d)$ is the vector of potential values of X for $D = d$. Then, the total indirect effect is defined as

$$\delta^t(d) = E[Y(d, M(1, X(1)), X(d)) - Y(d, M(0, X(0)), X(d))]. \tag{11}$$

We refer to $\delta^t(d)$ as the total indirect effect because it comprises all effects via M which either come from D directly or "take a devious route" through X. I.e., this parameter accounts for the fact that M is affected by D both directly and indirectly through a change in X. In contrast, the partial indirect effect only identifies the effect through M directly coming from D, but not going through X:

$$\delta^p(d) = E[Y(d, M(1, X(d)), X(d)) - Y(d, M(0, X(d)), X(d))]. \tag{12}$$

I.e., $\delta^p(d)$ is the ceteris paribus indirect effect via the mediator when holding X constant at the level implied by d such that any channel through the covariates is shut down. Note that this is what a linear regression framework identifies as "indirect effect" when regressing Y on $(1, D, M, X)$ and multiplying the coefficient on M with the first stage effect of D on M. Obviously, this effect neglects any correlations between M and X. We therefore argue that the total indirect effect is the more interesting parameter,[5] but nevertheless discuss the identification of both parameters. However, it will be shown further below that $\delta^p(d)$ is more easily identified than $\delta^t(d)$.

The direct effect is defined as

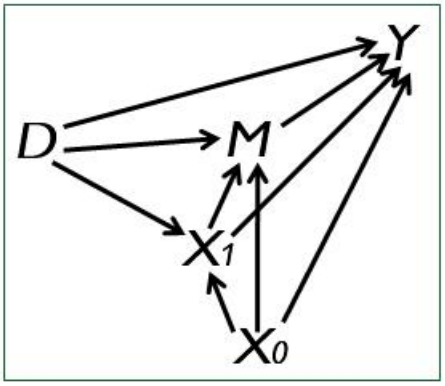
$$\theta(d) = E[Y(1, M(d, X(d)), X(d)) - Y(0, M(d, X(d)), X(d))], \tag{13}$$

i.e., it corresponds to the change in the mean potential outcome due to an exogenous change in the treatment, while keeping the mediator and the covariates fixed. Note that this definition differs from Imai and Yamamoto (2011), who consider the difference between the ATE and the total indirect effect to be the "direct" effect: $E[Y(d, M(d, X(d)), X(d)) - Y(1-d, M(d, X(d)), X(1-d))]$. However, this includes changes in the mean potential outcome which are due to a change in X which is not mediated by M. Here, we define the direct effect in a narrower sense that also excludes (inherently indirect) channels via X. For this reason, $\theta(d)$ and $\delta^t(d)$ or $\delta^p(d)$, respectively, do not add up to the ATE, as either $(E[Y(d, M(d, X(1)), X(1)) - Y(d, M(d, X(0)), X(0))])$ or

[5] Also Imai and Yamamoto (2011) focus on $\delta^t(d)$ as indirect effect and do not consider $\delta^p(d)$ at all.

$(E[Y(d, M(d, X(d)), X(1)) - Y(d, M(d, X(d)), X(0))])$ are not accounted for, respectively.

Figure 2: Causal paths with pre-treatment covariates (X_0) and post-treatment covariates (X_1)



The directed acyclic graph in Figure 2 displays a set up where the treatment affects the observed confounders of the mediator. To be specific, X is partitioned into X_0, the pre-treatment covariates, which are by randomization independent of D just as before, and X_1, which are a function of D and confound M. Identification requires that after conditioning on X_1, there are no unobserved confounders that jointly affect X_1 on the one hand and M and/or Y on the other hand. This is clearly more data demanding than the previous framework with the covariates not being a function of the treatment. A further identification issue arises from the fact that conditioning on post-treatment variables generally changes the distribution of pre-treatment variables across treatment states (which initially were balanced by randomization), which may confound the direct and indirect effects. For this reason, we also need to control for all pre-treatment covariates X_0 that are jointly related with the post-treatment confounders X_1 on the one hand and directly with the outcome or the mediator on the other hand in order to not break the randomization of D. We therefore replace Assumption 1 by Assumption 3:

Assumption 3 (random treatment assignment and conditional independence):

(a) $\{Y(d'', m, x'), M(d', x), X_0, X_1(d)\} \perp D$ for all $d'', d', d \in \{0, 1\}$ and m, x, x' in the support of M, X,

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

M, X.

Assumption 3(a) relaxes the restriction that D is independent of X in Assumption 1 to independence of the pre-treatment covariates X_0 only, while for the post-treatment covariates X_1, independence is only required to hold w.r.t. their potential values. As a side remark, the notation of $X(d)$ in the definitions of the effects in (11), (12), and (13) is accurate even in the presence of pre-treatment covariates and the partition of X into X_0 and X_1, because $X_0(d) = X_0$ under randomization. For this reason, '$X_0, X_1(d)$' in Assumption (3a) could be replaced by '$X(d)$' at the cost of not being explicit about the pre-treatment confounders. Assumption 3(b) is new and explicitly states that unconfoundedness of the treatment effects on the mediator and outcome must also hold when conditioning on X (i.e., both pre- and post-treatment covariates). This was trivially satisfied under Assumption 1 even if there existed unobserved confounders of X and M or Y, because X was independent of D.

As a further modification to our initial set up, we replace Assumption 2 by Assumption 4:

Assumption 4 (conditional independence of the potential mediator state):
(a) $Y(d'', m, X(d'))\perp M|D = d, X = x$ for all $d'', d', d \in \{0, 1\}$ and m, x in the support of M, X,
(b) $\Pr(D = d|M = m, X = x) > 0$ for all $d \in \{0, 1\}$ and m, x in the support of M, X.

Assumption 4(a) is equivalent to Assumption 2(a), but now accounts for our modified notation. The common support restriction 4(a) is exactly the same as before. When comparing our framework to other assumptions made in the literature, it turns out that Assumptions 3 and 4(a) are similar to FRCISTG (fully randomized causally interpretable structural tree graph) in Robins (2003) (see also Robins (1986)) and Assumption 2 in Imai and Yamamoto (2011), with two important differences. Firstly, here, the treatment is assumed to be randomly assigned, whereas Robins (2003) and Imai and Yamamoto (2011) consider the case of a conditionally independent treatment given observed pre-treatment characteristics. Secondly, Robins (2003) and Imai and

Yamamoto (2011) do not impose conditional independence of $Y(d'', m, X(d'))$ and $M(d)$ for possibly distinct d'', d', d as required by our Assumption 4(a), which therefore also holds for potential outcomes and potential mediator states defined on opposite treatments. They merely assume that $Y(d, m, X(d))$ and $M(d)$ are conditionally independent, i.e., when all parameters are defined on the same treatment. Robins and Richardson (2010) present a DGP in their Appendix B where indeed FRCISTG holds while our stronger restrictions are not satisfied. However, Robins (2003) argues that it seems hard to construct realistic scenarios where one set of assumptions holds while the other one does not.

We now consider the identification of the direct and indirect effects based on our modified assumptions. The obtainment of $\theta(d)$ hinges on the identifiability of $E[Y(1-d, M(d, X(d)), X(d))]$, which we show below:

$$\begin{aligned}
& E[Y(1-d, M(d, X(d)), X(d))] \\
= & \int E[Y(1-d, m, x)|M = m, X = x] dF_{M(d,X(d)),X(d)}(m, x) \\
= & \int E[Y(1-d, m, x)|M = m, X = x] dF_{M,X|D=d}(m, x) \\
= & \int E[Y|D = 1-d, M = m, X = x] \cdot \frac{\Pr(D = d|M, X)}{\Pr(D = d)} dF_{M,X}(m, x) \\
= & E\left[E\left[\frac{Y \cdot I\{D = 1-d\}}{\Pr(D = 1-d|M, X)}\middle| M = m, X = x\right] \cdot \frac{\Pr(D = d|M, X)}{\Pr(D = d)}\right] \\
= & E\left[\frac{Y \cdot I\{D = 1-d\}}{\Pr(D = 1-d|M, X)} \cdot \frac{\Pr(D = d|M, X)}{\Pr(D = d)}\right]. \qquad (14)
\end{aligned}$$

The first equality follows from the law of iterated expectations and from replacing the outer expectation by an integral, the second one from Assumption 3, the third from Assumption 4(a) and Bayes' theorem, the fourth from basic probability theory and from replacing integrals by expectations, and the last from the law of iterated expectations. From comparing (14) to (4) it becomes obvious that the identification result for the direct effect under Assumptions 3 and 4 is identical to that under Assumptions 1 and 2: $\theta(d) = E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d)}\right] - E\left[\frac{Y \cdot I\{D=1-d\}}{\Pr(D=1-d|M,X)} \cdot \frac{\Pr(D=d|M,X)}{\Pr(D=d)}\right]$. This gives rise to Proposition 3:

Proposition 3:

Under Assumptions 3 and 4, the average direct effect is identified equivalently to Proposition 1:

$$\theta(d) = E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X)}\right) \cdot \frac{\Pr(D=d|M,X)}{\Pr(D=d)}\right]. \tag{15}$$

We next consider the partial indirect effect (holding X fixed) which hinges on the identification of $Y(d, M(1-d, X(d)), X(d))$:

$$\begin{aligned}
& E[Y(d, M(1-d, X(d)), X(d))] \\
= & \int\int E[Y(d,m,x)|M=m, X=x] dF_{M(1-d,x)|X=x}(m) dF_{X(d)}(x) \\
= & \int\int E[Y(d,m,x)|M=m, X=x] dF_{M|D=1-d,X=x}(m) dF_{X|D=d}(x) \\
= & \int\int E[Y|D=d, M=m, X=x] \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)} dF_{M|X=x}(m) \frac{\Pr(D=1-d|X)}{\Pr(D=1-d)} dF_X(x) \\
= & E\left[E\left[E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)}\middle| M=m, X=x\right] \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)}\middle| X=x\right] \cdot \frac{\Pr(D=d|X)}{\Pr(D=d)}\right] \\
= & E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)} \cdot \frac{\Pr(D=d|X)}{\Pr(D=d)}\right].
\end{aligned} \tag{16}$$

The first equality follows from the law of iterated expectations and from replacing the outer expectations by integrals, the second one from Assumption 3, the third from Assumption 4(a) and Bayes' theorem, the fourth from basic probability theory and from replacing integrals by expectations, and the last from the law of iterated expectations. Therefore, the partial indirect effect is identified as outlined in Proposition 4:

Proposition 4:

Under Assumptions 3 and 4, the average partial indirect effect is identified by

$$\delta^p(d) = E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d)} - \frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)} \cdot \frac{\Pr(D=d|X)}{\Pr(D=d)}\right]. \tag{17}$$

The identification of the total indirect effect requires the knowledge of $E[Y(d, M(1-d, X(1-$

$d)), X(d))]$. Unfortunately, this is not feasible without further assumptions, see the proof in Avin, Shpitser, and Pearl (2005). As discussed in Robins and Greenland (1992) and Robins (2003), identification would require us in a first step to exogenously set D to $1-d$ to observe $M(1-d, X(1-d)) = M(1-d)$ and in a second step to exogenously set D to d and M to $M(1-d)$ in order to identify the distribution of $Y(d, M(1-d, X(1-d)), X(d))$. It is obviously impossible to do so because even in an experiment, we cannot manipulate the treatment state of any unit to be d and $1-d$ at the same time. Put differently, when randomly assigning units to $D = d$ and $D = 1-d$, identification requires us to adjust the distribution of M among the treated to that of the non-treated while at the same time keeping the distribution of X fixed, which is impossible if X and M are associated.

However, Robins (2003) shows that the total indirect effect is identified under an additional restriction, namely the absence of interaction effects between D and M. Formally, his assumption implies that the unit-level treatment effect (for any unit i) for the mediator fixed is constant across different values of the mediator:

$$Y_i(1, m, X_i(1)) - Y_i(0, m, X_i(0)) = Y_i(1, m', X_i(1)) - Y_i(0, m', X_i(0)) = B_i,$$

where B_i is an unit-level constant. Unfortunately, this assumption appears unattractive in empirical applications (see for instance the discussion in Imai, Tingley, and Yamamoto (2012), Section 3.1) and restricts the usefulness of nonparametric identification advocated in recent work. However, Imai and Yamamoto (2011) demonstrate that the assumption of no interaction effect can be relaxed to assuming a homogenous interaction effect:

$$Y_i(1, m, X_i(1)) - Y_i(0, m, X_i(0)) = Y_i(1, m', X_i(1)) - Y_i(0, m', X_i(0)) = B_i + Cm,$$

where C is constant for any m. I.e., the interaction between the treatment and the mediator varies homogenously for all observations.

Here, we propose an alternative (but still related) functional form restriction w.r.t. potential

mediator states across treatments that is comparable to that of Assumption 3 in Flores and Flores-Lagunes (2009) (who, however, use it in the set up where D does not affect X).

Assumption 5 (functional form restriction w.r.t. potential mediators):
For all m_d, x_d in the support of $M(d), X(d)$, write $E[Y(d, M(d, X(d)), X(d))|M(d, X(d)) = m_d, X(d) = x_d] = \mu_{d,x_d}(m_d)$, i.e., write the mean potential outcome for $D = d$ as a function of m_d. It is assumed that
(a) for all m_{1-d}, x_d in the support of $M(1-d), X(d)$, it holds that
$\mu_{d,x_d}(m_{1-d}) = E[Y(d, M(1-d, X(1-d)), X(d))|M(1-d, X(1-d)) = m_{1-d}, X(d) = x_d]$,
(b) $\mu_{d,x_d}(m_{1-d}) = \mu_{d,x_d}(E(m_{1-d}))$.

Assumption 5(a) states that one can predict $E[Y(d, M(1-d, X(1-d)), X(d))|M(1-d, X(1-d)) = m_{1-d}, X(d) = x_d]$ for any m_{1-d} and x_d based on the regression function μ_{d,x_d}. This implies that the interaction effect between D and M is the same for $M(1)$ and $M(0)$. Even though $M(1-d, X(1-d))$ is not known for units with $D = d$ (on which the identification of μ_{d,x_d} is based upon), we exploit the fact that $E[Y(d, M(1-d, X(1-d)), X(d))] = E[Y(d, E[M(1-d, X(1-d))], X(d))]$ by Assumption 5(b) and that $E[M(1-d, X(1-d))]$ is observed for $D = 1-d$. It has to be stressed that Assumption 5 is not innocuous. Firstly, Assumption 5(a) requires us to have a correctly specified model for the prediction across mediator states. Secondly, Assumption 5(b) further restricts μ, essentially to be linear in M such that predicting based on $E(m_{1-d})$ is asymptotically equivalent to the use of m_{1-d}. Given that these considerably stronger functional form assumptions are satisfied, $E[Y(d, M(1-d, X(1-d)), X(d))]$, which is required for the total indirect effect, is obtained

from the following result:

$$
\begin{aligned}
& E[Y(d, M(1-d, X(1-d)), X(d))] \\
= \; & E[Y(d, M(1-d, X(1-d)), X(d))|D=d] \\
= \; & E\left[E[Y(d,m,x)|D=d, M=M(1-d, X(1-d)), X=x]|D=d\right] \\
= \; & E[\mu_{d,x}(M(1-d, X(1-d)))|D=d] \\
= \; & E[\mu_{d,x}(E[M(1-d, X(1-d))])|D=d] \\
= \; & E[\mu_{d,x}(E[M|D=1-d])|D=d] \\
= \; & E[\mu_{d,x}(E[M|D=1-d])|D=d] \\
= \; & E\left[\frac{\mu_{d,x}(E[M|D=1-d]) \cdot I\{D=d\}}{\Pr(D=d)}\right]. \qquad (18)
\end{aligned}
$$

The first equality follows from Assumption 3, the second from the law of iterated expectations and Assumption 3, the third from Assumption 5(a), the fourth from Assumption 5(b), the fifth from Assumption 4(a), and the last from basic probability theory. It follows that the total indirect effect is identified by Proposition 5:

Proposition 5:

Under Assumptions 3, 4, and 5, the average total indirect effect is identified by

$$\delta^t(d) = E\left[\frac{\{Y - \mu_{d,x}(E[M|D=1-d])\} \cdot I\{D=d\}}{\Pr(D=d)}\right]. \qquad (19)$$

3 Simulations

This section presents a simulation study that provides some intuition for the identification results and the issues related to incorrectly imposing the wrong set of assumptions. For the ease of

exposition, we consider the following DGP based on linear equations:

$$
\begin{aligned}
Y &= 0.5D + M + \beta DM + X + \epsilon_1, && (20)\\
M &= 0.5D + 0.5X + \epsilon_2, && (21)\\
X &= \gamma D + \epsilon_3, && (22)\\
D &= I\{\epsilon_4 > 0\}, && (23)\\
&\text{with} \quad \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 \sim N(0,1), \text{ independently of each other.}
\end{aligned}
$$

(20) is the outcome equation, in which Y is a function of D, M, X and an unobserved term ϵ_1. β gauges the interaction effect between D and M such that $\beta = 0$ satisfies the assumption of no interaction discussed in Robins (2003). By (21), the mediator is a function of D, X and the unobservable ϵ_2. The parameter γ in (22) determines whether X is caused by D and which set of assumptions is valid. If $\gamma = 0$, Assumptions 1 and 2 are valid, otherwise identification has to be based on Assumptions 3 to 5, with μ_{d,x_d} simply corresponding to the prediction according to the linear outcome equation (20). By (23), the treatment is randomly assigned with a treatment probability of 0.5.

We now discuss the various effects and first consider $\gamma = 0$ such that Assumptions 1 and 2 hold. For $\beta = 0$, the average direct effect is homogenous and simply corresponds to the coefficient on D, i.e., $\theta(1) = \theta(0) = 0.5$, because there is no interaction between D and M. Likewise, the average indirect effect is $\delta(1) = \delta(0) = E[M(1)] - E[M(0)] = 0.5$, because by (21), $E[M(1)] = 0.5$ (as $E[X(1)] = E[X(0)] = 0$ by (22) and $\gamma = 0$) and $E[M(0)] = 0$. The ATE is the sum of both effects and thus equal to one. In the simulations, we also consider $\beta = 0.5$. This implies that the direct effect for $D = 1$ is $\theta(1) = E[0.5 + M + \beta M + X + \epsilon_1 | M = M(1)] - E[M + X + \epsilon_1 | M = M(1)] = E[0.5 + \beta M | M = M(1)] = 0.5 + \beta E[M(1)]$. Note that 0.5 is the direct effect not accounting for the interaction with M, whereas $\beta E[M(1)] = 0.25$ gives the interaction effect. Therefore, $\theta(1) = 0.75$. Likewise, $\theta(0) = 0.5 + \beta E[M(0)] = 0.5$. Concerning the indirect effects, $\delta(1) = E[0.5D + M(1) + \beta DM(1) + X + \epsilon_1 | D = 1] - E[0.5D + M(0) + \beta DM(0) + X + \epsilon_1 | D = 1] =$

$E[(1+\beta)\cdot(M(1)-M(0))] = 0.75$, whereas $\delta(0) = E[(M(1)-M(0))] = 0.5$. Again, $\theta(d)+\delta(1-d)$ yields the ATE, which is now equal to 1.25.

Secondly, we set $\gamma = 0.2$, such that Assumptions 1 and 2 are violated. This implies that $E[X(1)] = 0.2$, while $E[X(0)] = 0$. Therefore, the mean potential mediator states under treatment and non-treatment are, respectively, $E[M(1)] = E[M(1, X(1))] = 0.5 + 0.5 \cdot 0.2 = 0.6$ and $E[M(0)] = E[M(0, X(0))] = 0$. For $\beta = 0$, the average direct effect is again constant: $\theta(d) = E[0.5 + M + X + \epsilon_1 | M = M(d), X = X(d)] - E[M + X + \epsilon_1 | M = M(d, X(d)), X = X(d)] = 0.5 + E[M(d, X(d))] + E[X(d)] - E[M(d, X(d))] - E[X(d)] = 0.5$. The average total indirect effect is $\delta^t(d) = E[M(1, X(1))] + E[X(d)] - E[M(0, X(0))] - E[X(d)] = 0.6$. Note that the ATE is equal to 1.3, i.e., 0.2 higher than the sum of the total indirect effect and the direct effect, which corresponds to the change in the outcome due to the change in X which does not go through M. Concerning the average partial indirect effect, $E[M(1, X(0))] = 0.5 + 0.5 \cdot E[X(0)] = 0.5$ and $E[M(0, X(1))] = 0.5 \cdot E[X(1)] = 0.1$. Therefore, $\delta^p(1) = E[M(1, X(1))] + E[X(1)] - E[M(0, X(1))] - E[X(1)] = 0.6 - 0.1 = 0.5$, whereas $\delta^p(0) = E[M(1, X(0))] + E[X(0)] - E[M(0, X(0))] - E[X(0)] = 0.5 - 0 = 0.5$. Finally, we consider the case that $\gamma = 0.2$ and $\beta = 0.5$. Now, $\theta(1) = 0.5 + \beta E[M(1, X(1))] = 0.8$ and $\theta(0) = 0.5$, $\delta^t(1) = E[(1 + \beta) \cdot (M(1, X(1)) - M(0, X(0)))] = 1.5 \cdot (0.6 - 0) = 0.9$, and $\delta^t(0) = E[(M(1, X(1)) - M(0, X(0)))] = 0.6$ and $\delta^p(1) = 1.5 \cdot (E[M(1, X(1))] - E[M(0, X(1))]) = 0.75$, $\delta^p(0) = E[M(1, X(0))] - E[M(0, X(0))] = 0.5$. Again, the sum of the direct effect and the total indirect effect is 0.2 lower than the ATE (1.4). Table 1 summarizes the direct and indirect effects for the various scenarios.

Table 1: True direct and indirect effects for various scenarios

	γ=0				γ=0.2			
	β=0		β=0.5		β=0		β=0.5	
Effect	D=1	D=0	D=1	D=0	D=1	D=0	D=1	D=0
θ	0.5	0.5	0.75	0.5	0.5	0.5	0.8	0.5
δ	0.5	0.5	0.75	0.5	-	-	-	-
δ^t	-	-	-	-	0.6	0.6	0.9	0.6
δ^p	-	-	-	-	0.5	0.5	0.75	0.5

We run 5000 Monte Carlo simulations with 2000 observations and estimate the models by the sample analogs of either Propositions 1 and 2 (for $\gamma = 0$) or Propositions 3 to 5 (for $\gamma = 0.2$), respectively. For the estimation of the propensity score $\Pr(D = 1|M, X)$ a probit specification is used. In addition, we also consider OLS regression, which estimates the direct effect as the coefficient on D in a regression of Y on $(1, D, M, X)$. The indirect effect corresponds to the coefficient on M in the latter regression multiplied with the coefficient on D in a regression of M on $(1, D, X)$.[6] Obviously, this approach omits interactions between D and M, an issue often encountered in empirical work, see Section 4.1 of Imai, Keele, Tingley, and Yamamoto (2010) for a revision and discussion of the shortcomings of the standard linear framework. Finally, we also include a naive OLS estimator, where conditioning on X in the aforementioned regressions is omitted. I.e., this approach is naive in the sense that it does not control for the confounder of the mediator.

Table 2: Bias, variance, and MSE of various estimators under Assumptions 1 and 2 ($\gamma = 0$)

	β=0						β=0.5					
	D=1			D=0			D=1			D=0		
Est.	bias	var	MSE	bias	var	MSE	bias	var	MSE	bias	var	MSE
$\hat{\theta}_{\text{IPW}}$	0.000	0.003	0.003	0.001	0.003	0.003	-0.000	0.003	0.003	0.001	0.004	0.004
$\hat{\theta}_{\text{OLS}}$	0.001	0.002	0.002	0.001	0.002	0.002	-0.124	0.002	0.018	0.126	0.002	0.018
$\hat{\theta}_{\text{naive}}$	-0.199	0.004	0.043	-0.199	0.004	0.043	-0.324	0.004	0.109	-0.074	0.004	0.009
$\hat{\delta}_{\text{IPW}}$	-0.002	0.008	0.008	-0.001	0.008	0.008	-0.003	0.012	0.012	-0.001	0.008	0.008
$\hat{\delta}_{\text{OLS}}$	-0.001	0.002	0.002	-0.001	0.002	0.002	-0.127	0.003	0.019	0.123	0.003	0.018
$\hat{\delta}_{\text{naive}}$	0.198	0.005	0.044	0.198	0.005	0.044	0.072	0.007	0.012	0.322	0.007	0.111

Table 2 presents the bias, variance, and mean squared error (MSE) of the various estimators for $\gamma = 0$. We see that the IPW-based methods are close to being unbiased and that their MSEs are moderate in any scenario. For β=0, also the OLS estimators work well. Their MSEs are even somewhat lower than those of the IPW estimators due to their generally smaller variance related to tighter functional form restrictions. However, non-negligible biases arise for β=0.5, because the OLS estimators do not account for interactions between D and M. The naive estimators are

[6]For $\gamma = 0$, conditioning on X is actually not required in the regression of M. As D and X are independent under Assumptions 1 and 2, (not) conditioning on X does not matter in terms of identification, but may affect efficiency.

biased in any scenario, as they omit the confounder X. The results for $\gamma = 0.2$ are displayed in Table 3. Concerning the direct effects, we observe a similar pattern of the estimators' properties as before. Taking a look at the total indirect effect, it, however, becomes obvious that OLS is no longer consistent even if β=0, because it does not account for correlations between X and M. This highlights that under Assumptions 3 and 4, identification is not obtained by the standard OLS approach heavily used in the mediation literature if the interest lies in the total indirect effect. However, for β=0, OLS still identifies the partial indirect effect, i.e., the ceteris paribus impact going through M for X fixed. For β=0.5, OLS is no longer consistent, whereas the naive estimator is severely biased in any scenario.

Table 3: Bias, variance, and MSE of various estimators under Assumptions 3 to 5 ($\gamma = 0.2$)

	β=0						β=0.5					
	D=1			D=0			D=1			D=0		
Est.	bias	var	MSE	bias	var	MSE	bias	var	MSE	bias	var	MSE
$\hat{\theta}_{\text{IPW}}$	-0.002	0.004	0.004	-0.003	0.003	0.003	-0.002	0.004	0.004	-0.003	0.005	0.005
$\hat{\theta}_{\text{OLS}}$	-0.001	0.002	0.002	-0.001	0.002	0.002	-0.151	0.002	0.025	0.149	0.002	0.024
$\hat{\theta}_{\text{naive}}$	-0.041	0.004	0.006	-0.041	0.004	0.006	-0.191	0.004	0.040	0.109	0.004	0.016
$\hat{\delta}^t_{\text{IPW}}$	-0.001	0.003	0.003	-0.001	0.003	0.003	-0.001	0.006	0.006	-0.001	0.003	0.003
$\hat{\delta}^t_{\text{OLS}}$	-0.101	0.002	0.012	-0.101	0.002	0.012	-0.276	0.003	0.079	0.024	0.003	0.004
$\hat{\delta}^t_{\text{naive}}$	0.239	0.005	0.062	0.239	0.005	0.062	0.089	0.007	0.015	0.389	0.007	0.159
$\hat{\delta}^p_{\text{IPW}}$	0.000	0.003	0.003	-0.000	0.003	0.003	0.000	0.006	0.006	-0.000	0.003	0.003
$\hat{\delta}^p_{\text{OLS}}$	-0.001	0.002	0.002	-0.001	0.002	0.002	-0.126	0.003	0.019	0.124	0.003	0.019
$\hat{\delta}^p_{\text{naive}}$	0.339	0.005	0.120	0.339	0.005	0.120	0.239	0.007	0.064	0.489	0.007	0.246

Finally, we investigate the properties of the IPW estimator of the indirect effect based on Proposition 2 which is valid for $\gamma = 0$, when in fact $\gamma = 0.2$. Table 4 reveals that by omitting the link between D and X, $\hat{\delta}_{\text{IPW}}$ neither identifies the total, nor the partial indirect effect, no matter whether β is zero or 0.5. The biases are substantial and comparable to the naive estimator (see Table 3). This suggests that invoking the wrong assumptions when controlling for mediator endogeneity may be equally harmful as ignoring the endogeneity problem altogether, urging researchers to carefully think about which set assumptions appears most plausible in their application at hand.

Table 4: Bias, variance, and MSE of various estimators under Assumptions 3 to 5 ($\gamma = 0.2$)

	β=0						β=0.5					
	D=1			D=0			D=1			D=0		
Est.	bias	var	MSE	bias	var	MSE	bias	var	MSE	bias	var	MSE
$\hat{\delta}_{\text{IPW}}$ for δ^t	0.201	0.008	0.049	0.200	0.008	0.048	0.201	0.013	0.054	0.200	0.008	0.048
$\hat{\delta}_{\text{IPW}}$ for δ^p	0.301	0.008	0.099	0.300	0.008	0.098	0.351	0.013	0.137	0.300	0.008	0.098

4 Application

We apply the estimators resulting from Propositions 1 to 5 to a welfare policy experiment with a binary treatment assignment (D) which was conducted in the mid-1990s to assess the publicly funded U.S. Job Corps program.[7] The program, which is currently administered by more than 120 local Job Corps centers throughout the U.S., targets young individuals (aged 16-24 years) that have a legal residence in the U.S. and come from a low-income household. It provides participants with approximately 1200 hours of vocational training and education as well as with housing and board over an average duration of 8 months. Participants also receive health education as well as health and dental care. Schochet, Burghardt, and Glazerman (2001) and Schochet, Burghardt, and McConnell (2008) discuss in detail the experimental design[8] and the main results, i.e., the ATEs on a broad range of outcomes. Their findings suggest that Job Corps increases educational attainment, reduces criminal activity, and increases employment and earnings (at least for some years after the program).

Flores and Flores-Lagunes (2009) appear to be the first to assess the causal mechanisms of the program and find a positive direct effect on earnings after controlling for the mediator work experience which they assume to be conditionally exogenous given pre-treatment variables. Alternatively, Flores and Flores-Lagunes (2010) suggest a partial identification approach[9] that

[7]Note that the compliance with the treatment assignment was not perfect. According to Schochet, Burghardt, and McConnell (2008) only 73 % of eligible individuals actually enrolled at Job Corps centers. Here, we abstract from this issue and consider the assignment as treatment variable. Strictly speaking, we therefore consider (direct and indirect) "intention to treat" effects rather than treatment effects.

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

[9]Partial identification of economic parameters in general goes back to Manski (1989, 1994) and Robins (1989).

does not require to control for mediator endogeneity, at the cost of sacrificing point identification. They bound the indirect effects of Job Corps on employment and earnings which are mediated by the achievement of a GED, high school degree, or vocational degree as well as the direct effects.

In contrast to these studies which are concerned with labor market outcomes, we focus on the program's effects on general health. To be precise, we consider a binary health indicator (Y) evaluated 2.5 years after randomization, which is equal to one if self-assessed general health is stated to be very good and zero otherwise. In this context, employment appears to be an interesting mediator, as it is affected by Job Corps and may itself have an impact on health. In line with this idea, Huber, Lechner, and Wunsch (2011) find that entering employment increases self-assessed mental health when investigating a sample of German welfare recipients. Furthermore, several studies in medicine and social sciences conclude that there is a negative association between unemployment and health, see for instance the surveys by Jin, Shah, and Svoboda (1997), Björklund and Eriksson (1998), and Mathers and Schofield (1998). We therefore disentangle the total health effect into a direct and an indirect component that is due to a change in the likelihood to work. If there existed a positive total effect which, however, only operated through employment, this would imply that health care and health education were less decisive for general health than the human capital related interventions of Job Corps which affect employability. In this context, the analysis of causal mechanisms may help to assess the usefulness of different components of a program in place.

We define employment in the first half of the second year after randomization (i.e., half way between the treatment assignment and the measurement of the outcome) as our mediator of interest (M). I.e., $M = 1$ in case of any kind of employment and $M = 0$ otherwise. We argue that the covariates to be controlled for should include potential confounders that are measured shortly before the mediator, as they may change over time, in particular as a function of the treatment. In contrast to Flores and Flores-Lagunes (2009), we therefor do not exclusively rely on pre-treatment covariates, but also use variables that were measured in the year after treatment assignment, just before the assessment of the mediator. Nevertheless, we also condition on a rich

set of pre-treatment variables, not only to control for mediator endogeneity, but also to control for confounding of the treatment effect that may be induced by conditioning on post-treatment variables only, see the discussion of Assumption (3b) in Section 2.

The empirical literature, see for instance Mulatu and Schooler (2002) and Llena-Nozal, Lindeboom, and Portrait (2004) among many others, suggests that socio-economic factors such as education, age, and income are strongly correlated with health while they also determine an individual's employment perspectives. As discussed in Huber, Lechner, and Wunsch (2011), similar arguments are likely to hold for the labor market history. E.g., previous jobs might have a positive or negative effect on health depending on an individual's level of stress, willingness/reluctance to work, or physical strain. Furthermore, as acknowledged in Böckerman and Ilmakunnas (2009), it appears important to condition on initial (in our case: pre-mediator) health, which allows controlling for time-constant unobservable confounders. In the data, we do not only observe initial health, but also health behavior prior to the mediator period such as alcohol and drug abuse.

We analyze the direct and indirect effects of the program separately by gender, in order to account for potential effect heterogeneity. We restrict the initial data set (14,327 youths with completed baseline survey prior to the treatment assignment) to the 4,352 females and 5,673 males for which the post-treatment variables M and Y are observed in the follow-up survey after 2.5 years. Table 5 presents descriptive evidence that the selection into the mediator is indeed selective for females and males in our evaluation sample. Individuals entering employment 1 to 1.5 years after randomization are on average slightly older, (in the case of females) more educated, (in the case of males) less often arrested, more likely to be white, less likely to receive on public housing, transfer payments, and food stamps, and living in smaller households at assignment. Interestingly, the association with household income is non-monotonic, whereas the number of kids is (as expected) negatively associated with female employment and positively with male employment. Concerning the labor market history, we see a strong positive correlation between previous employment and the mediator and a negative association of the latter with being in a training activity in the year before the mediator assessment, pointing to locking-in

Table 5: Descriptives

	Females				*Males*			
Variable	$M=1$	$M=0$	diff	p-val	$M=1$	$M=0$	diff	p-val
socio-economic factors								
age at assignment	18.750	18.377	0.373	0.000	18.506	17.880	0.627	0.000
years of education at ass.	10.523	10.047	0.476	0.000	10.242	10.116	0.126	0.397
in school in yr. before ass.	0.629	0.635	-0.006	0.662	0.644	0.696	-0.052	0.000
number of kids after 1st yr	0.496	0.671	-0.175	0.000	0.129	0.092	0.037	0.001
ethnicity: black*	0.509	0.587	-0.078	0.000	0.415	0.529	-0.114	0.000
ethnicity: white*	0.247	0.152	0.094	0.000	0.354	0.223	0.131	0.000
household size at ass.	4.503	4.825	-0.323	0.000	4.351	4.492	-0.141	0.011
low household income in yr before ass.**	0.313	0.382	-0.069	0.000	0.227	0.288	-0.062	0.000
high household income in yr before ass.**	0.329	0.393	-0.063	0.000	0.366	0.405	-0.040	0.003
number of times in welfare before ass.	2.079	2.326	-0.247	0.000	1.915	2.134	-0.218	0.000
food stamps in yr before ass.	0.506	0.593	-0.086	0.000	0.337	0.423	-0.086	0.000
public assistance in yr before ass.	0.251	0.283	-0.032	0.020	0.240	0.262	-0.022	0.067
in public housing 1 yr after ass.	0.164	0.231	-0.067	0.000	0.118	0.171	-0.053	0.000
transfer payments in 1st yr after ass.	0.507	0.621	-0.114	0.000	0.276	0.376	-0.099	0.000
ever arrested before ass.	0.160	0.172	-0.013	0.276	0.317	0.329	-0.012	0.357
number of arrests in 1st yr after ass.	0.071	0.056	0.015	0.123	0.242	0.340	-0.098	0.000
pre-mediator labor market state								
ever worked before ass.	0.132	0.183	-0.051	0.000	0.123	0.155	-0.032	0.001
worked in yr before ass.	0.707	0.480	0.227	0.000	0.728	0.526	0.202	0.000
worked in 1st yr after ass.	0.812	0.389	0.423	0.000	0.828	0.424	0.404	0.000
worked in months 9-12 after ass.	0.720	0.200	0.520	0.000	0.743	0.225	0.518	0.000
worked fulltime in months 9-12	0.384	0.009	0.375	0.000	0.394	0.014	0.381	0.000
in training in yr. before ass.	0.017	0.017	-0.000	0.903	0.019	0.019	0.000	0.911
vocational training in months 9-12	0.208	0.257	-0.048	0.000	0.186	0.231	-0.045	0.000
academic training in months 9-12	0.323	0.411	-0.089	0.000	0.317	0.430	-0.112	0.000
pre-mediator health (behavior)								
health at ass. (1=very good, 4=bad)	1.721	1.762	-0.041	0.074	1.619	1.648	-0.029	0.143
health after 1 year (1=very good, 4=bad)	1.847	1.868	-0.020	0.387	1.721	1.747	-0.026	0.213
phys./emot. problems at ass.	0.055	0.056	-0.001	0.852	0.045	0.049	-0.004	0.468
phys./emot. problems 1 yr after ass.	0.157	0.140	0.017	0.115	0.126	0.114	0.012	0.200
alcohol abuse before ass.	0.555	0.460	0.094	0.000	0.654	0.553	0.101	0.000
alcohol abuse 1 yr after ass.	0.239	0.159	0.080	0.000	0.366	0.245	0.121	0.000
illegal drugs before ass.	0.004	0.004	0.000	0.998	0.005	0.006	-0.001	0.664
illegal drugs 1 yr after ass.	0.010	0.008	0.002	0.413	0.020	0.013	0.007	0.053

Note: *: Baseline category is neither black, nor white. **: Baseline category is intermediate household income.

effects. In contrast, pre-mediator health is not strongly correlated with the mediator employment. Both the differences in general health (evaluated on a scale) and the incidence of physical or emotional problems (dummy variable) are insignificant. Maybe surprisingly, alcohol abuse is higher among the working than among the non-working, while differences in illegal drug use are mostly insignificant.

We control for all of these potential confounders in the estimation of the propensity score $\Pr(D=1|M,X)$ by a flexible probit specification (separately for females and males).[10] We test the latter using the nonparametric specification test for propensity score models proposed by

[10]The distributions of the propensity scores across treatment states and gender are provided in the appendix.

Shaikh, Simonsen, Vytlacil, and Yildiz (2009) which is based on kernel density estimation of the score along with an application of Bayes' theorem.[11] The p-values are 0.657 and 0.431 for the models used in the female and male samples, respectively, when choosing the bandwidth for kernel density estimation (using the Gaussian kernel) according to the Silverman (1986) rule of thumb. The non-rejection of the models is insensitive to using twice or half this bandwidth.

To estimate the direct and indirect effects, we use normalized versions of the sample analogs of the IPW-based identification results in Propositions 1 to 4 such that the weights of the observations in either treatment state add up to unity, as advocated in Imbens (2004) and Busso, DiNardo, and McCrary (2009b).[12] E.g., the normalized estimators of the direct effects under treatment and non-treatment are given by

$$\begin{aligned}\hat{\theta}(1) &= \frac{\sum Y_i \cdot D_i}{\sum D_i} - \frac{\sum Y_i \cdot (1-D_i) \cdot \hat{p}(M_i,X_i)/(1-\hat{p}(M_i,X_i))}{\sum (1-D_i) \cdot \hat{p}(M_i,X_i)/(1-\hat{p}(M_i,X_i))},\\ \hat{\theta}(0) &= \frac{\sum Y_i \cdot D_i \cdot (1-\hat{p}(M_i,X_i))/\hat{p}(M_i,X_i)}{\sum D_i \cdot (1-\hat{p}(M_i,X_i))/\hat{p}(M_i,X_i)} - \frac{\sum Y_i \cdot (1-D_i)}{\sum 1-D_i},\end{aligned}$$

where i is the index of the observations in the i.i.d. sample and $\hat{p}(M_i, X_i)$ denotes the estimate of the propensity score $\Pr(D = 1|M_i, X_i)$. Our semiparametric IPW methods (into which the propensity scores enter parametrically) can be expressed as sequential GMM estimators[13] where the propensity score estimation represents the first step and the effect estimation the second step, see Newey (1984). It follows from his results that our methods are $\sqrt{n}$-consistent under standard regularity conditions. Note that $\sqrt{n}$-consistency might be obtained even if a nonparametric estimator is used for the propensity score that satisfies particular regularity conditions, see Hirano, Imbens, and Ridder (2003) and Li, Racine, and Wooldridge (2009) for estimation based on series and kernel regression, respectively. Furthermore, IPW estimators are sufficiently smooth such

[11] Shaikh, Simonsen, Vytlacil, and Yildiz (2009) show that $f_{\Pr(D=1|M,X)|D=1}(\rho|D=1) = \frac{\Pr(D=0)}{\Pr(D=1)} \frac{\rho}{1-\rho} f_{\Pr(D=1|M,X)|D=0}(\rho|D=0) \quad \forall\, \rho \in (0,1)$, with $f_{\Pr(D=1|M,X)|D=d}(\cdot|D=d)$ being the pdf of $\Pr(D=1|M,X)$ conditional on $D=d$, is a testable implication of a correctly specified propensity score.

[12] However, we do not use any propensity score trimming (see for instance Busso, DiNardo, and McCrary (2009a), Crump, Hotz, Imbens, and Mitnik (2009), and Huber, Lechner, and Wunsch (2010)) in the estimation, because propensity scores close to the boundaries 0 and 1 do not occur in our application.

[13] See Hansen (1982) and Newey and McFadden (1994) for the assumptions underlying the standard GMM framework.

that the bootstrap is consistent for inference. We therefore estimate the standard errors by 1999 bootstrap draws. Concerning the estimation of the total indirect effects based on Proposition 5, $\mu_{d,x_d}(m)$ is specified as a linear model that includes all covariates entering the propensity score and again, the bootstrap is used for inference.

Table 6: Effects on the incidence of very good general health after 2.5 years

	$\hat{\Delta}$	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}^t(1)$	$\hat{\delta}^t(0)$	$\hat{\delta}^p(1)$	$\hat{\delta}^p(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$
					Females				
Effect	0.028	0.031	0.023	-0.000	0.000	-0.000	0.000	0.005	-0.003
S.E.	0.014	0.018	0.015	0.001	0.001	0.000	0.001	0.006	0.011
p-value	0.045	0.078	0.123	0.737	0.783	0.784	0.891	0.405	0.813
					Males				
Effect	0.022	-0.003	0.003	-0.000	-0.000	0.000	0.000	0.019	0.025
S.E.	0.013	0.019	0.014	0.000	0.000	0.000	0.000	0.007	0.013
p-value	0.099	0.861	0.845	0.584	0.782	0.540	0.677	0.004	0.060

Note: Standard errors (S.E.) are estimated based on 1999 bootstrap draws.

Table 6 presents the estimated effects on females and males. The second column gives the ATE, i.e., the mean difference between treated and non.treated outcomes, along with the standard error (s.e.) and the p-value. Taking a look at the females, the estimate suggests that Job Corps increases the incidence of a very good general health state by 2.8 % points.[14] The direct effect under treatment (column 3) is borderline significant and somewhat larger than that under non-treatment (column 4), which is, however, insignificant at the 10 % level. At least for the treated, the program appears to have a sizeable effect that is not mediated by employment. In contrast, all indirect effects are close to zero and insignificant under either set of assumptions, such that employment does not seem to mediate the effectiveness of the program in any important way. For the males, the ATE amounts to 2.2 % points and is borderline significant.[15] In contrast to the females, however, we do not find any sizeable direct effects, which points to effect heterogeneity w.r.t. gender. An interesting picture arises when looking at the indirect effects. While the partial and total indirect effects based on Assumptions 3 to 5 are all zero, estimation based on Assumptions 1 and 2 leads to conflicting results. In fact, $\hat{\delta}(1)$ and $\hat{\delta}(0)$ are significantly positive (at the 1 and 10 % levels, respectively) and economically non-negligible. This again demonstrates

[14]The mean outcome is 0.343 among the treated and 0.315 among the non-treated such that the ATE amounts to roughly 8 to 9 % of the mean outcomes.

[15]The mean outcomes are 0.432 under treatment and 0.410 under non-treatment.

the importance of carefully considering the choice of the set of identifying assumptions.

To check the sensitivity of our results to potential attrition bias due to restricting our sample to individuals with observed post-treatment variables, we consider the response behavior in the follow-up period to be a function of the observed variables D, X. This corresponds to the missing at random assumption of Rubin (1976).[16] The latter allows correcting for attrition bias by weighting observations in the estimation by $R/\Pr(R = 1|X, D)$ with R being the binary response indicator, see for instance Wooldridge (2002, 2007). We estimate the response propensity $\Pr(R = 1|X, D)$ using a probit model and find that controlling for attrition substantially decreases the precision of the estimates, but does not overthrow our results. We therefore conclude that for our sample of disadvantaged youths in the U.S., the health effects mediated by employment appear to be negligible. In contrast, our estimates point to a considerable direct effect of the program on the subjective health state of females, at least among the treated.

5 Conclusion

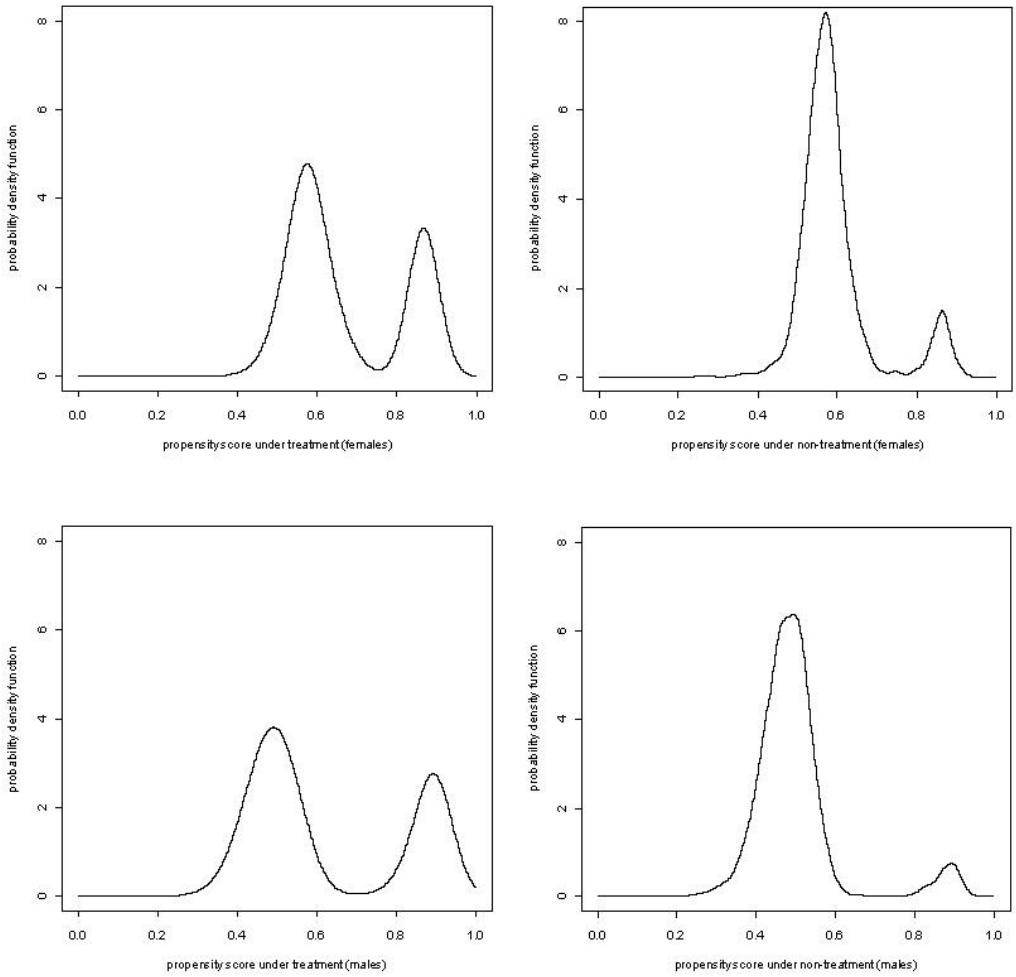
This paper has demonstrated how to identify causal mechanisms (or direct and indirect effects) in randomized experiments with a binary treatment (mainly) based on inverse probability weighting (IPW) using the treatment propensity score. Identification relies on the assumption of conditional exogeneity of the mediator, i.e., of the intermediate variable of interest through which the indirect effect operates, given a set of observed covariates and the treatment. We have discussed two sets of assumptions: Mediator exogeneity (i) given covariates which are not influenced by the treatment (with the leading case being pre-treatment variables) and (ii) given covariates which are themselves a function of the treatment. It has been shown that direct effects can be straightforwardly identified in either case, whereas the identification of indirect effects becomes more cumbersome in the latter case, which, however, appears more realistic in empirical applications. The identification issues for either set of assumptions have been demonstrated in a sim-

[16] For a discussion of alternative forms of missingness in experiments and remedies based on weighting, see Huber (2012).

ulation study. Finally, we have provided an application to the experimental evaluation study of the Job Corps program. As the results are partly sensitive to the choice of the set of conditional exogeneity assumptions, the importance of carefully considering the plausibility of the imposed identifying restrictions in the analysis of causal mechanisms cannot be overemphasized.

A Appendix: Propensity score distributions

Figure 3: Distribution of the propensity scores across treatment states and gender



Note: Kernel density estimation is based on the Gaussian kernel and the Silverman (1986) rule of thumb for bandwidth selection.

References

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.

Avin, C., I. Shpitser, and J. Pearl (2005): "Identifiability of path-specific effects," in *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pp. 357–363, Edinburgh, UK.

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.

Björklund, A., and T. Eriksson (1998): "Unemployment and mental health: evidence from research in the Nordic countries," *Scandinavian Journal of Social Welfare*, 7, 219–235.

Böckerman, P., and P. Ilmakunnas (2009): "Unemployment and self-assessed health: evidence from panel data," *Health Economics*, 18, 161–179.

Busso, M., J. DiNardo, and J. McCrary (2009a): "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," *unpublished manuscript*.

——— (2009b): "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," *IZA Discussion Paper No. 3998*.

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

Cochran, W. G., and S. P. Chambers (1965): "The planning of observational studies of human populations," *Journal of the Royal Statistical Society Series A*, 128, 234–265.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009): "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96, 187–199.

Fisher, R. (1925): *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

——— (1935): *The Design of Experiments*. Oliver and Boyd, Edinburgh.

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

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

Freedman, D. (2006): "Statistical Models for Causation: What Inferential Leverage Do They Provide," *Evaluation Review*, 30, 691–713.

Hansen, L. (1982): "Large Sample Properties of Generalized Method of Moment Estimators," *Econometrica*, 50, 1029–1054.

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

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

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

Huber, M., M. Lechner, and C. Wunsch (2010): "How to control for many covariates? Reliable estimators based on the propensity score," *IZA Discussion Paper no. 5268*.

——— (2011): "Does leaving welfare improve health? Evidence for Germany," *Health Economics*, 20, 484–504.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto (2010): "Unpacking the Black Box: Learning about Causal Mechanisms from Experimental and Observational Studies," *mimeo*.

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71.

Imai, K., D. Tingley, and T. Yamamoto (2012): "Experimental Designs for Identifying Causal Mechanisms," *forthcoming in the Journal of the Royal Statistical Society, Series A*.

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 J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Jin, R. L., C. P. Shah, and T. J. Svoboda (1997): "The impact of unemployment on health: a review of the evidence," *Journal of Public Health Policy*, 18, 275–301.

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

Li, Q., J. Racine, and J. Wooldridge (2009): "Efficient Estimation of Average Treatment Effects With Mixed Categorical and Continuous Data," *Journal of Business and Economics Statistics*, 27, 206–223.

Llena-Nozal, A., M. Lindeboom, and F. Portrait (2004): "The effect of work on mental health: does occupation matter?," *Health Economics*, 13, 1045–1062.

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

Manski, C. F. (1989): "Anatomy of the selection problem," *The Journal of Human Resources*, 24, 343–360.

——— (1994): "The selection problem," in *Advances in Econometrics: Sixth World Congress*, ed. by C. Sims., pp. 143–170. Cambridge University Press.

Mathers, C. D., and D. J. Schofield (1998): "The health consequences of unemployment: the evidence," *The Medical Journal of Australia*, 168, 178–182.

Mulatu, S., and C. Schooler (2002): "Causal Connections between Socio-Economic Status and Health: Reciprocal Effects and Mediating Mechanisms," *Journal of Health and Social Behavior*, 43, 22–41.

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

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. Engle, and D. McFadden. Elsevier, Amsterdam.

Neyman, J. (1923): "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles.," *Statistical Science*, Reprint, 5, 463–480.

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

——— (2011): "The Causal Mediation Formula - A practitioner guide to the assessment of causal pathways," *Technical report R-379.*

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

Robins, J. M. (1986): "A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect," *Mathematical Modelling*, 7, 1393–1512.

——— (1989): "The Analysis of Randomized and Non-RandomizedAIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS*, ed. by L. Sechrest, H. Freeman, and A. Mulley, pp. 113–159. U.S. Public Health Service, Washington, DC.

——— (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.

Rosenbaum, P. (1984): "The consequences of adjustment for a concomitant variable that has been affected by the treatment," *Journal of Royal Statistical Society, Series A*, 147, 656–666.

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

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.

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

——— (2008): "For objective causal inference, design trumps analysis," *The Annals of Applied Statistics*, 2, 808–840.

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

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

Shaikh, A. M., M. Simonsen, E. J. Vytlacil, and N. Yildiz (2009): "A specification test for the propensity score using its distribution conditional on participation," *Journal of Econometrics*, 151, 33–46.

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

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

Tchetgen Tchetgen, E. J. (2012): "Inverse Odds Ratio-Weighted Estimation for Causal Mediation Analysis," *mimeo*.

Tchetgen Tchetgen, E. J., and I. Shpitser (2011a): "Semiparametric Estimation of Models for Natural Direct and Indirect Effects," *Harvard University Biostatistics Working Paper 129*.

——— (2011b): "Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis," *Technical report, Harvard University School of Public Health*.

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 Ours, J. C. (2004): "The locking-in effect of subsidized jobs," *Journal of Comparative Economics*, 32, 37–55.

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

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, Article 3.