Identification and Estimation of Causal Mediation Effects with Treatment Noncompliance*

Teppei Yamamoto†

First Draft: May 10, 2013
This Draft: December 20, 2013

Abstract

Treatment noncompliance, a common problem in program evaluation, poses a serious challenge to the identification of causal mechanisms via causal mediation analysis. This is because the mediated portion of an intention-to-treat (ITT) effect cannot be nonparametrically identified even when there is no unobserved confounding. In particular, the commonly-used naïve approach of ignoring the actual treatment status and applying the "mediation formula" to the assigned treatment, mediator and outcome leads to a biased estimate of the mediated ITT effect. This paper proposes an alternative approach. It is shown that the mediated ITT effects and the local average causal mediation effects (LACME) for compliers can be identified under a local sequential ignorability assumption as well as the standard instrumental variable assumptions. Bias in the naïve estimator is formally characterized. The proposed estimator is illustrated via a Monte Carlo simulation study and applied to data from a large-scale job training experiment. The proposed method, implemented in an open-source R package, enables researchers to investigate causal mechanisms by which the treatment affects the outcome of interest even when treatment noncompliance exists.

Key Words: Causal inference; Program evaluation; Instrumental variables; Encouragement design; Causal mechanisms; Natural direct and indirect effects.

*I am grateful to Luke Keele and Dustin Tingley for our ongoing collaboration which motivated this paper. I thank Alberto Abadie, Adam Glynn, Martin Huber, Kosuke Imai, Tyler VanderWeele, and participants at the 2013 Atlantic Causal Inference Conference for their helpful comments and suggestions. All errors are my own.

†Assistant Professor, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: teppei@mit.edu, URL: http://web.mit.edu/teppei/www

1 Introduction

Identification of causal mechanisms is an important goal in program evaluation. When researchers investigate the effectiveness of a public policy, they are often also interested in how and why the policy produces effects. Answering such questions enhances the understanding of the science behind the policy, enables policymakers to prescribe better alternatives, and often provides them with justifications for the continuation and future expansion of the program. Causal mediation analysis, a common framework for the statistical analysis of causal mechanisms in social and medical sciences, is becoming increasingly popular in program evaluation (e.g. Flores and Flores-Lagunes, 2009, 2010; Heckman et al., 2013; Huber, 2013). Based on the potential outcomes model of causal inference (Neyman, 1923; Rubin, 1974), causal mediation analysis formalizes the role of an intermediate variable ("mediator") which lies on the causal pathway between the treatment and the outcome (e.g. Pearl, 2001; Robins, 2003; Imai et al., 2010b) and uses various parametric and nonparametric estimation procedures for statistical inference (e.g. Imai et al., 2010a; Tchetgen Tchetgen and Shpitser, 2011).

In program evaluation, randomized experiments are widely viewed as a gold standard, for they allow researchers to obtain unbiased estimates of the average treatment effects without relying on untestable assumptions. Yet, in practice, many randomized experiments suffer from treatment noncompliance, where some individuals do not follow the treatments assigned by the researcher. When treatment noncompliance occurs, the treatments actually received by individuals are no longer independent of their unobserved characteristics, rendering the simple estimators of causal effects invalid. Fortunately, methodologists have developed various procedures to cope with the problem. A common approach is to focus on the intention-to-treat (ITT) effects of the assigned treatment on the outcome, regardless of whether the treatments are actually received. Another popular approach is to estimate the local average treatment effect (LATE) among the individuals who would actually comply with the treatment assignment using an instrumental variables technique (e.g. Imbens and Angrist, 1994; Angrist et al., 1996; Nagelkerke et al., 2000).

While treatment noncompliance is a challenging but surmountable problem for the analysis of policy

effectiveness, it poses a significantly harder problem for analyzing causal mechanisms behind a policy. Indeed, while the ITT effect itself can be nonparametrically identified even in the presence of treatment noncompliance as long as the treatment assignment is randomized (Angrist et al., 1996), its mediated portion cannot be identified even when the conditional ignorability of the mediator is additionally assumed. This result, although somewhat counterintuitive, can be obtained as a special case of the unidentifiability result by Robins (2003), who shows that neither the direct nor indirect effect of an intervention is identifiable whenever its effect on the mediator and outcome is intercepted by a post-treatment variable (Figure 1(a)). This implies that, when the randomized intervention is compromised by noncompliance, one cannot identify the mediated portion of the effect of the intervention without making additional unverifiable assumptions. This important result has never been noted in the literature in the context of treatment noncompliance, as far as I am aware.

The unidentifiability of the mediated ITT effect appears to pose a serious threat to the analysis of causal mechanisms in randomized experiments with treatment noncompliance, for it implies that any purported estimator for the mediated ITT effect is invalid without additional assumptions. In particular, the intuitively appealing approach of (1) treating the assigned treatment status as if it were the actual treatment and (2) applying a standard causal mediation analysis technique (e.g. Imai et al., 2010b) to decompose the effects of the assigned treatment does not generally produce a valid estimate of the mediated ITT effect. Unfortunately, some authors have taken the "naïve" approach of decomposing the ITT effect in analyzing empirical data with noncompliance. For example, Flores and Flores-Lagunes (2009) analyze the effect of random assignment to a job training program in the National Job Corps Study and attempt to "decompose the ITT" effect (p.24) on post-training earnings into the portions that are mediated and unmediated through the labor market experience lost during the study. However, because the effect of the assignment to the training program is almost surely transmitted through the actual participation in the program, the reported estimates are likely to be biased for the mediated and unmediated ITT effects.

In this paper, I propose an alternative approach to the analysis of causal mechanisms in the presence of treatment noncompliance. First, I show that one can still identify causal quantities relevant for the mech-

anism of interest once the standard instrumental variables assumptions are made (Angrist et al., 1996). Specifically, I prove that the local average causal mediation effects (LACME) and natural direct effects (LANDE) for compliers, as well as the mediated and unmediated ITT effects, can be nonparametrically identified under exclusion restrictions, monotonicity of the treatment, and a local sequential ignorability assumption for the compliers (Pearl, 2001; Imai et al., 2010b). Second, I formally characterize the bias in the naïve estimators for the mediated and unmediated ITT effects, LACME and LANDE under those assumptions. Third, I propose two estimators for these quantities of interest that can accommodate a wide variety of nonparametric and parametric models for the outcome, mediator and compliance indicator variables. The methods proposed in this paper have been implemented as part of `mediation`, an open-source R package publicly available on the Comprehensive R Archive Network (**?**).

The rest of the paper proceeds as follows. In Section 2, I introduce a randomized job training experiment with treatment noncompliance to motivate the methodology. In Section 3, I describe the proposed framework and present the main analytical results. In Section 4, I present results from a small-scale Monte Carlo experiment to analyze the small-sample properties of the proposed estimators. The results suggest that the proposed estimators have reasonable performance as long as the rate of noncompliance is not too high. In Section 5, I apply the proposed method to the empirical example in order to analyze whether the hours spent in job training mediate the causal effects of enrollment in the program on the probability of employment. Finally, Section 6 concludes with the discussion of directions for future research.

2 A Motivating Example

The National JTPA study is a large-scale randomized job training evaluation study conducted in the United States between the late 1980s and early 1990s (see Orr et al., 1996, for detailed information). The primary goal of the study was to measure the benefits and costs of the employment and training programs for economically disadvantaged adults and youths that were publicly funded on the basis of the Job Training Partnership Act of 1982 (JTPA). The study collected data on approximately 20,000 participants in the program at 16 study sites, which were systematically selected from the population of 649 JTPA service

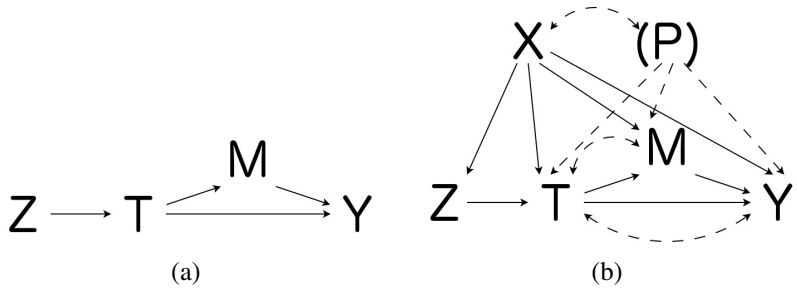

Figure 1: Causal DAGs Representing the Causal Structures Analyzed in This Paper. Figure 1(a) represents the nonparametric structural equations model under which the path-specific effect corresponding to the path $Z \to T \to M \to Y$ cannot be identified despite the lack of any unobserved confounding, as shown by Avin et al. (2005). The unidentifiability of the mediated ITT effects is a special case of this scenario where Z and T represent the assigned and actual treatments, respectively. Figure 1(b) represents the set of causal assumptions corresponding to those required for the proposed method in this paper. First, the missing direct arrow from Z to M or Y represents the exclusion restriction (Assumption 1). Second, the missing bidirected dashed arcs between Z and any of the other nodes imply that Z is conditionally ignorable given X (Assumption 3). Third, the missing bidirected dashed arc between M and Y implies the ignorability of the observed mediator conditional on T, X and P (Assumption 4). The remaining assumption (monotonicity, Assumption 2) is not directly represented in this diagram. The node P, indicating compliance types, is put in parentheses to indicate that it is not directly observable.

delivery areas so as to ensure a broad geographic coverage as well as diversity in the characteristics of participants. Within the study sites, approximately two thirds of the eligible participants were randomly assigned to the treatment condition, where they were offered to receive the job training services coordinated by JTPA administrators. However, not every participant in the treatment group chose to enroll in the JTPA program, and a small fraction of the participants in the control group managed to get enrolled. As shown in Table 1, the proportion of the actual enrollees is about two thirds in the treatment condition and a little below two percent in the control condition.

Once participants chose to enroll in the program, they typically received various types of training sessions (chosen based on their pre-treatment assignment characteristics) for the next 18 months. However, the actual amount of training received varied among the enrollees. In fact, approximately 45 percent of the participants are reported to have not received any job training services (except the initial enrollment session) even after enrollment. On the other hand, those who did not enroll were excluded from receiving JTPA-funded job training services for the duration of this period, although they were free to participate in other job training activities. The proportion of the non-enrolled participants who received some job training during the study period was about 29 percent. On average, the treatment group received significantly

greater number of hours of job training than the control group, as can be seen in Table 1.

In this paper, I focus on the employment status of the study participants in Months 19 to 30 after randomization (i.e. the year after the program termination) as the outcome of interest. Among the 9,279 adult participants for whom data on employment and duration of training received are both available, there appears to be a marginally statistically significant ITT effect of the assignment to the JTPA treatment on the post-program 12-month employment (0.017, $p = 0.067$ for the two-sided alternative). The effect is greater among the 5,253 adult female participants (0.029, $p = 0.021$) while it is essentially zero for the 4,026 adult males (0.003, $p = 0.850$). Thus, it may be concluded that the assignment to JTPA programs caused a moderate increase in the probability of employment. Previous analyses of the JTPA data set (e.g. Abadie et al., 2002) generally reached similar conclusions about a range of labor market outcomes.

Given the increase in both the hours of job market training during the study period and the probability of post-program employment for the treatment group, one might hypothesize that the JTPA treatment caused the improvement of job prospect by increasing the amount of job training received, i.e. the treatment effect can be attributed to the causal mechanism represented by the hours of training. However, Flores and Flores-Lagunes (2009) raise the interesting possibility that the increased time spent in job training sessions may instead harm the participants by preventing them from spending that time on actual job search or job experience. The causal mediation effect of the hours in training may be zero or even negative under this alternative scenario, which would lead to very different policy implications. Thus, it is of scientific and practical interest to investigate whether the hours in job training mediate the causal effect of the JTPA treatment on post-program employment.

3 The Proposed Methodology

In this section, I describe the proposed methodology for the identification and estimation of the mediated and unmediated ITT effects, LACME and LANDE with treatment noncompliance. I first introduce the notation and assumptions required for the identification analysis. These assumptions are graphically summarized in Figure 1(b). Next, the main identification result is presented. I then formally characterize

	All Adults		Male		Female	
Assigned to Program (Z_i)	Yes	No	Yes	No	Yes	No
Actual Enrollment (T_i)	0.655	0.017	0.637	0.014	0.668	0.018
	(0.476)	(0.128)	(0.481)	(0.119)	(0.471)	(0.134)
Hours of Training Received (M_i)	318.9	164.3	267.0	131.0	358.9	189.6
	(588.0)	(452.6)	(559.5)	(399.5)	(606.1)	(487.7)
Employment in Months 19–30 (Y_i)	0.723	0.704	0.731	0.732	0.717	0.683
	(0.447)	(0.457)	(0.443)	(0.443)	(0.450)	(0.465)
Estimated Prob. of Compliance	0.638		0.622		0.650	
Number of Observations (N)	9279		4026		5253	

Table 1: Means and Standard Deviations (in Parentheses) of the Key Variables Used in the Analysis for Each Program Assignment and Gender Stratum.

the bias in the naïve estimator commonly used in the literature. Finally, I propose both parametric and nonparametric estimators for the mediated and unmediated ITT effects, LACME and LANDE.

3.1 Notation and Assumptions

Consider a simple random sample of size N obtained from a population of interest. Let Z_i represent the binary *treatment assignment* for unit i, i.e., $Z_i \in \{0, 1\}$ where $i = 1, ..., N$. For each unit, define the binary variables $T_i(0)$ and $T_i(1) \in \{0, 1\}$ indicating the *potential treatment* that would be *received* by unit i if the unit were assigned to the treatment (1) or control (0) condition, respectively. Then, the *observed treatment* status of unit i can be written as $T_i = T_i(Z_i) = Z_i T_i(1) + (1 - Z_i)T_i(0)$. I use X_i to denote the vector of observed pre-treatment assignment covariates with support $\mathcal{X}$. Implicit in the above notation is the stable unit treatment value assumption (SUTVA), which precludes interference between units and different versions of treatment assignment (Rubin, 1990). Throughout the rest of the paper, I also maintain the assumption that both treatment and control conditions have non-zero probability of assignment, i.e., $0 < \Pr(Z_i = 1 \mid X_i = x) < 1$ for any $x \in \mathcal{X}$.

Next, let the *potential mediators* and *potential outcomes* denoted by $M_i(z, t) \in \mathcal{M}$ and $Y_i(z, t, m) \in \mathcal{Y}$, respectively, where $t \in \{0, 1\}$, $z \in \{0, 1\}$, and $\mathcal{M}$ and $\mathcal{Y}$ represent the supports of the respective variables. The potential mediator represents the value of the mediator that would be realized when the treatment and treatment assignment took on the values t and z for unit i, and the potential outcome equals

the value of the outcome that would occur if the mediator, treatment and treatment assignment equaled m, t and z, respectively. The SUTVA is also assumed for the potential mediator and outcome variables.

Following Angrist et al. (1996), I maintain the following assumptions for the rest of the paper. First, I assume *exclusion restrictions* for both the mediator and outcome variables:

Assumption 1 (Exclusion Restrictions)

$$M_i(z,t) \;=\; M_i(z',t) \quad \textit{and} \quad Y_i(z,t,m) \;=\; Y_i(z',t,m), \quad \textit{for any } z,z',t \in \{0,1\}, \textit{ and } m \in \mathcal{M}.$$

This assumption states that the treatment assignment can only affect the mediator and outcome through its effect on the treatment actually received. Under Assumption 1, the potential mediator and outcome can be written more concisely as $M_i(t)$ and $Y_i(t,m)$, respectively.

Second, I assume the *monotonicity* of the treatment reception function:

Assumption 2 (Monotone Treatment Reception)

$$T_i(0) \;\leq\; T_i(1) \quad \textit{for all} \quad i = 1, ..., N.$$

This assumption states that there is no unit who would always take the treatment status opposite to the assigned treatment. Under Assumption 2, the units can be categorized into three *principal strata* (Frangakis and Rubin, 2002) based on their potential treatment values, which I denote by the trichotomous variable $P_i \in \{c, a, n\}$. Those units who would always follow the assigned treatment status (i.e. $T_i(0) = 0$ and $T_i(1) = 1$) are called *compliers* and denoted as $P_i = c$; those who would always take the treatment regardless of assignment (i.e. $T_i(1) = T_i(0) = 1$) are called *always-takers* ($P_i = a$); and those who would never take the treatment (i.e. $T_i(1) = T_i(0) = 0$) are *never-takers* ($P_i = n$). Below, I refer to these subpopulations as *compliance types*.

Third, I assume *local sequential ignorability*, consisting of the following two assumptions:

Assumption 3 (Ignorable Treatment Assignment)

$$\{Y_i(t,m), M_i(t'), T_i(z) : \ t,t',z \in \{0,1\}, m \in \mathcal{M}\} \ \perp\!\!\!\perp \ Z_i \ \mid \ X_i$$

Assumption 4 (Locally Ignorable Mediator among Compliers)

$$Y_i(t',m) \ \perp\!\!\!\perp \ M_i(t) \ \mid \ T_i = t, P_i = c, X_i, \quad \textit{for all } t,t' \in \{0,1\} \textit{ and } m \in \mathcal{M}.$$

Assumption 3 typically holds in randomized experiments where units are randomly assigned to the treatment conditions. The assumption may also be plausible in observational studies where Z_i represents an *instrumental variable* that can be regarded as exogenous after conditioning on observed confounders.

Assumption 4 requires several remarks. First, because the actual treatment always equals the assigned treatment among compliers, Assumption 4 can be equivalently stated by conditioning on the treatment assignment instead of the actual treatment, i.e, $Y_i(t',m) \perp\!\!\!\perp M_i(t) \mid Z_i = t, P_i = c, X_i$ for all $t,t' \in \{0,1\}$ and $m \in \mathcal{M}$. Second, Assumption 4 is closely related to the assumption Frangakis and Rubin (1999) called latent ignorability in the analysis of missing data, in that ignorability between two sets of potential response variables is assumed only after conditioning on a compliance type. That is, Assumption 4 states that, if the compliance types were to be directly observable, the analyst would accept the ignorability of the observed mediator among the compliers after also conditioning on the treatment and observed pre-treatment covariates. Third, Assumptions 3 and 4 together imply that the *sequential ignorability* assumption (Imai et al., 2010b) holds locally among compliers. As shown by Imai et al. (2010b), the (global) sequential ignorability assumption is sufficient to nonparametrically identify the average causal mediation effects when treatment compliance is perfect. In Section 3.3, I show that local sequential ignorability, which is weaker than global sequential ignorability, is sufficient for identifying the *local* average causal mediation effect for compliers (as defined shortly) when Assumptions 1–2 are also satisfied.

A causal diagram consistent with Assumptions 1–4 is given in Figure 1(b). The missing direct arrows from Z to M and Z to Y in the diagram reflect Assumption 1. Assumption 3 is represented by the

missing bidirected dashed arcs between Z and any of the other nodes in the diagram, while Assumption 4 is represented by the missing bidirected dashed arc between M and Y. Assumption 2 is not explicitly represented in the diagram, but it implicitly constrains the causal relationships underlying the arrows from Z to T and P to T.

3.2 Quantities of Interest

In causal mediation analysis, researchers are typically interested in decomposing the treatment effect on the outcome into two portions: the component of the effect that operates through the mediator and the component that does not. The former is often referred to as the causal mediation effect (Imai et al., 2010b), natural indirect effect (Pearl, 2001) or pure (total) indirect effect (Robins, 2003), while the latter is called the natural or pure (total) direct effect. The population averages of these quantities are called the average causal mediation effect (ACME) and the average natural direct effect (ANDE), respectively (Imai et al., 2010b), and the existing research has largely focused on the identification and inference for these quantities of interest (e.g. Robins and Greenland, 1992; Imai et al., 2010a). However, in the presence of treatment noncompliance, the nonparametric identification of these quantities requires assumptions that are unrealistically strong in most applied contexts, such as the ignorability of the actual treatment conditional on observed covariates.

In this paper, I instead investigate two related causal quantities that are relevant to the analysis of causal mechanisms. First, one might be interested in decomposing the ITT effect into the portion that can be attributed to the mediator of interest and the remaining portion that is transmitted through other unmodeled mediators. I call these quantities the *mediated ITT effect* and *unmediated ITT effect*, which are defined respectively as follows:

$$\text{Mediated ITT effect:} \quad \lambda(z) \equiv \mathbb{E}[Y_i(T_i(z), M_i(T_i(1))) - Y_i(T_i(z), M_i(T_i(0)))], \tag{1}$$

$$\text{Unmediated ITT effect:} \quad \mu(z) \equiv \mathbb{E}[Y_i(T_i(1), M_i(T_i(z))) - Y_i(T_i(0), M_i(T_i(z)))], \tag{2}$$

for $z \in \{0, 1\}$. Equation 1 represents the average hypothetical change in the outcome variable in response

to the change in the mediator from the value that would be realized when assigned to the treatment to the value that would occur when assigned to the control, while the actual treatment status is fixed at the value that would be realized under assignment to either the treatment ($z = 1$) or control ($z = 0$) for each i. On the other hand, equation 2 equals the average hypothetical change in the outcome when the actual treatment was changed from the value under assignment to the control to the value under assignment to the treatment, holding the value of the mediator constant at its natural value when assigned to either the treatment or control for each i. It is straightforward to show that these two effects sum up to the ITT effect as follows: $\mathbb{E}[Y_i(T_i(1), M_i(T_i(1))) - Y_i(T_i(0), M_i(T_i(0)))] = \lambda(1) + \mu(0) = \lambda(0) + \mu(1)$. Although the mediated and unmediated ITT effects have attractive substantive interpretations, these quantities are unfortunately not nonparametrically identifiable even under a strong set of ignorability conditions such as represented in Figure 1(a). However, once the standard instrumental assumptions are made (i.e. Assumptions 1 and 2), these effects become identifiable even under a substantially weaker set of ignorability assumptions such as Assumptions 3 and 4. A more detailed discussion of this issue can be found in Section 3.4.

Second, one might also be willing to focus on the similar decomposition of the treatment effect only among a specific subpopulation of units defined by their compliance status. Specifically, I call the conditional expectations of the causal mediation effects and natural direct effects among compliers the *local average causal mediation effect* (LACME) and *local average natural direct effect* (LANDE), respectively, defined as follows:

$$\text{LACME:} \qquad \delta(t) \equiv \mathbb{E}[Y_i(t, M_i(1)) - Y_i(t, M_i(0)) \mid P_i = c], \tag{3}$$

$$\text{LANDE:} \qquad \zeta(t) \equiv \mathbb{E}[Y_i(1, M_i(t)) - Y_i(0, M_i(t)) \mid P_i = c], \tag{4}$$

for $t \in \{0, 1\}$. Equation (3) represents the average hypothetical change in the outcome variable among compliers when the mediator was changed from the value that would be realized under the treatment condition to the value that would occur under the control condition while the treatment status itself is

held constant at t. On the other hand, equation (4) equals the average hypothetical change in the outcome among compliers when the treatment was changed from the control condition to the treatment condition while the mediator is held at the value that would realize under treatment condition t for each i. Thus, these quantities represent the portions of the average treatment effect that can and cannot be attributed to the mediator, respectively, among the subpopulation of compliers. Note that the sums of these two quantities equal the local average treatment effect (LATE; see Angrist et al., 1996) in the following way: $\mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0)) \mid P_i = c] = \delta(1) + \zeta(0) = \delta(0) + \zeta(1)$. Although these quantities are natural extensions of the ACME and ANDE to the case of imperfect treatment compliance, they have not heretofore been formally analyzed in the literature.

3.3 The Main Identification Results

The following theorem shows that the LACME and LANDE are nonparametrically identified under the assumptions considered in Section 3.1. (Here and for the rest of the paper, the mediator is assumed to be continuous unless otherwise stated; results for a discrete mediator can be obtained by replacing the appropriate integrals with summations over all values of $m \in \mathcal{M}$.)

Theorem 1 (Nonparametric Identification of LACME and LANDE) *Under Assumptions 1, 2 ,3 and 4, the LACME and LANDE are nonparametrically identified by the following expressions for $t \in \{0, 1\}$:*

$$\begin{aligned}
\delta(t) &= \int\int \frac{Q_{t|tx}G_{m|ttx}S_{mttx} - Q_{t|(1-t)x}G_{m|t(1-t)x}S_{mt(1-t)x}}{Q_{t|tx}G_{m|ttx} - Q_{t|(1-t)x}G_{m|t(1-t)x}} \\
&\qquad \times \frac{Q_{1|1x}G_{m|11x} - Q_{1|0x}G_{m|10x} - Q_{0|0x}G_{m|00x} + Q_{0|1x}G_{m|01x}}{Q_{1|1x} - Q_{1|0x}}\, dm\, dF(x), \qquad (5) \\
\zeta(t) &= \int\int \left\{ \frac{Q_{1|1x}G_{m|11x}S_{m11x} - Q_{1|0x}G_{m|10x}S_{m10x}}{Q_{1|1x}G_{m|11x} - Q_{1|0x}G_{m|10x}} - \frac{Q_{0|0x}G_{m|00x}S_{m00x} - Q_{0|1x}G_{m|01x}S_{m01x}}{Q_{0|0x}G_{m|00x} - Q_{0|1x}G_{m|01x}} \right\} \\
&\qquad \times \frac{Q_{t|tx}G_{m|ttx} - Q_{t|(1-t)x}G_{m|t(1-t)x}}{Q_{1|1x} - Q_{1|0x}}\, dm\, dF(x), \qquad (6)
\end{aligned}$$

where $S_{mtzx} = \mathbb{E}[Y_i \mid M_i = m, T_i = t, Z_i = z, X_i = x]$, $G_{m|tzx} = p(M_i = m \mid T_i = t, Z_i = z, X_i = x)$, $Q_{t|zx} = \Pr(T_i = t \mid Z_i = z, X_i = x)$, and $F(x)$ denotes the distribution function of X_i.

Theorem 1 shows that the LACME and LANDE can be expressed as functions of the following observed quantities: (1) the conditional expectation of the observed outcome given the observed mediator, actual treatment, treatment assignment and pre-treatment covariates (S_{mtzx}), (2) the conditional density (or probability mass) of the observed mediator given the actual and assigned treatments as well as the pre-treatment covariates ($G_{m|tzx}$), and (3) the conditional distribution of the actual treatment variable given the treatment assignment and pre-treatment covariates ($Q_{t|zx}$).

The intuition behind Theorem 1 is straightforward. First, because Assumptions 3 and 4 imply the sequential ignorability of Imai et al. (2010b) for compliers, the LACME and LANDE can be written in terms of the conditional mean and distribution of the observed outcome and mediator, respectively, given $P_i = c$ as well as T_i and X_i. Then, although the compliance type P_i cannot be directly observed and conditioned on, these latent conditional distributions can still be identified as mixtures of the observed conditional distributions given the actual and assigned treatment values (as well as the pre-treatment covariates) under Assumptions 1 and 2 (Abadie, 2003). A detailed proof can be found in Appendix A.1.

Theorem 1 calls for several additional remarks. First, note that S_{mtzx} and $G_{m|tzx}$ are in general not structural (i.e., causal) models for the potential outcomes and mediators. In fact, Assumption 1 implies that the true data generating process for $Y_i(t, m)$ and $M_i(t)$ cannot include the treatment assignment Z_i. The models for S_{mtzx} and $G_{m|tzx}$ are instead purely predictive models for the observed values of the outcome and mediator conditioning on the observed values of the variables in the conditioning sets. Various nonparametric and semiparametric estimators for conditional densities and moments have been proposed to make inference robust to model misspecification, and one can make use of such methods for the estimation of the LACME and LANDE as explained in Section 3.5.

Second, both expressions for the LACME and LANDE in Theorem 1 involve a division by $Q_{1|1x} - Q_{1|0x}$, which equals the conditional probability of being a complier under Assumptions 2 and 3. This implies that the problem of an "weak instrument" also applies for the estimation of these quantities (Bound et al., 1995). That is, the estimates of the LACME and LANDE based on Theorem 1 will be unstable and suffer from small-sample bias when the proportion of compliers is small. The simulation study reported

in Section 4 provides some evidence about the severity of this problem.

Based on Theorem 1, it can be shown that the mediated and unmediated ITT effects can also be nonparametrically identified under Assumptions 1–4. The following corollary summarizes the result.

Corollary 2 (Nonparametric Identification of mediated and unmediated ITT effects) *Under Assumptions 1, 2 ,3 and 4, the LACME and LANDE are nonparametrically identified for* $t \in \{0, 1\}$ *as follows:*

$$\begin{aligned}
\lambda(z) &= \int\int \frac{Q_{z|zx}G_{m|zzx}S_{mzzx} - Q_{z|(1-z)x}G_{m|z(1-z)x}S_{mz(1-z)x}}{Q_{z|zx}G_{m|zzx} - Q_{z|(1-z)x}G_{m|z(1-z)x}} \\
&\quad \times (Q_{1|1x}G_{m|11x} - Q_{1|0x}G_{m|10x} - Q_{0|0x}G_{m|00x} + Q_{0|1x}G_{m|01x})\, dm\, dF(x), \qquad (7) \\
\mu(z) &= \int\int \left\{ \frac{Q_{1|1x}G_{m|11x}S_{m11x} - Q_{1|0x}G_{m|10x}S_{m10x}}{Q_{1|1x}G_{m|11x} - Q_{1|0x}G_{m|10x}} - \frac{Q_{0|0x}G_{m|00x}S_{m00x} - Q_{0|1x}G_{m|01x}S_{m01x}}{Q_{0|0x}G_{m|00x} - Q_{0|1x}G_{m|01x}} \right\} \\
&\quad \times (Q_{z|zx}G_{m|zzx} - Q_{z|(1-z)x}G_{m|z(1-z)x})\, dm\, dF(x). \qquad (8)
\end{aligned}$$

A proof is given in Appendix A.2. It is important to note that the expressions in Corollary 2 do not equal the naïve formulas discussed in Section 3.4.

3.4 Bias in the Naïve Estimators

Without Assumptions 1 and 2, the mediated and unmediated ITT effects are nonparametrically unidentifiable even under a maximally strong set of ignorability assumptions. This result becomes apparent once these effects are regarded as particular types of the path-specific causal effects first noted in the seminal work by Robins (2003) and later discussed more generally by Avin et al. (2005). Consider the causal relationships among the assigned treatment (Z), actual treatment (T), mediator (M) and outcome of interest (Y) by a causal diagram in Figure 1(a). Here, the lack of any bidirected dashed arc in the diagram implies that there is no unobserved confounding among these variables (i.e., the causal model is Markovian). Robins (2003, p.5) shows that the average causal mediation effect of Z on Y with respect to M (i.e., the mediated ITT effect through M) cannot be nonparametrically identified even under the strong set of ignorability assumptions in Figure 1(a). The reason is that the counterfactual outcomes in equations (1) and (2) each nest two potential values of T that can never be jointly observed (i.e. $T_i(0)$

and $T_i(1)$) and the identification of such nested counterfactuals requires some assumptions about the joint distribution of those potential values (e.g. Albert and Nelson, 2011; Tchetgen Tchetgen and VanderWeele, 2014). More intuitively, the actual treatment acts as an observed post-treatment confounder between the mediator and outcome, which makes the mediation effects of the assigned treatment unidentifiable (Imai and Yamamoto, 2013).

Despite this result, some authors estimate the mediated and unmediated ITT effects by simply "ignoring" the actual treatment that is potentially confounded, as they would (appropriately) do for the estimation of the total ITT effect. That is, the following "naïve" formulas are often used to estimate the mediated and unmediated ITT effects:

$$\begin{aligned}
\lambda^*(z) &= \int\int \mathbb{E}[Y_i \mid M_i = m, Z_i = z, X_i = x] \\
&\qquad \left\{dF_{M_i|Z_i=1,X_i=x}(m) - dF_{M_i|Z_i=0,X_i=x}(m)\right\} dF_{X_i}(x), \qquad (9) \\
\mu^*(z) &= \int\int \{\mathbb{E}[Y_i \mid M_i = m, Z_i = 1, X_i = x] - \mathbb{E}[Y_i \mid M_i = m, Z_i = 0, X_i = x]\} \\
&\qquad dF_{M_i|Z_i=z,X_i=x}(m) dF_{X_i}(x), \qquad (10)
\end{aligned}$$

for $z \in \{0, 1\}$. Equations (9) and (10) are identical to the so-called "mediation formulas" for the average causal mediation effect and natural direct effect (Pearl, 2001; Imai et al., 2010b), except that the assigned treatment (Z_i) is used in place of the actual treatment (T_i). For example, Flores and Flores-Lagunes (2009) use regression models for $\mathbb{E}[Y_i \mid M_i, Z_i, X_i]$ and $F_{M_i|Z_i,X_i}(m)$ and estimate the mediated and unmediated ITT effects based on the formulas equivalent to equations (9) and (10) (p.16).

Unfortunately, these formulas turn out to produce biased estimates even under Assumptions 1–4. The following theorem formally characterizes the bias in the naïve estimators based on $\lambda^*(z)$ and $\mu^*(z)$.

Theorem 3 (Bias in the Naïve Estimators) *Under Assumptions 1, 2 ,3 and 4, the bias in the naïve estimators given in equations 9 and 10 for the mediated and unmediated ITT effects can respectively be*

written for $z \in \{0, 1\}$ *as follows,*

$$\begin{aligned}
&\lambda^*(z) - \lambda(z) \\
&= (-1)^z \int\int \left\{ \frac{S_{mzzx} G_{m|zzx} Q_{z|zx} - S_{mz(1-z)x} G_{m|z(1-z)x} Q_{z|(1-z)x}}{G_{m|zzx} Q_{z|zx} - G_{m|z(1-z)x} Q_{z|(1-z)x}} (G_{m|(1-z)zx} Q_{(1-z)|zx} + G_{m|z(1-z)x} Q_{z|(1-z)x}) \right. \\
&\qquad \left. - S_{m10x} G_{m|10x} Q_{1|0x} - S_{m01x} G_{m|01x} Q_{0|1x} \right\} \left(1 - \frac{p(M_i = m \mid Z_i = 1 - z)}{p(M_i = m \mid Z_i = z)} \right) \, dm \, dF(x), \qquad (11)
\end{aligned}$$

$$\mu^*(z) - \mu(z) \;=\; \lambda(1 - z) - \lambda^*(1 - z). \qquad (12)$$

A proof is provided in Appendix A.3. Theorem 3 makes it clear that the naïve estimates of the mediated and unmediated ITT effects are generally biased and the direction of the bias can be either upward or downward. The theorem also implies that the naïve estimator will be unbiased if (1) treatment compliance is perfect, such that $Q_{z|(1-z)x} = 0$ for $z \in \{0, 1\}$ and all x, or (2) the ITT effect of treatment assignment on the mediator is zero, such that $p(M_i \mid Z_i = 1) = p(M_i \mid Z_i = 0)$. Finally, the theorem confirms the previously known result that the (total) ITT effect is nonparametrically identified under Assumptions 1–4, since the bias in the sum of the naïve estimates for the mediated and unmediated ITT effects is shown to be zero, i.e., $\lambda^*(z) + \mu^*(1 - z) = \lambda(z) + \mu(1 - z)$ for $z \in \{0, 1\}$.

3.5 Estimation and Inference

Theorem 1 shows that the mediated and unmediated ITT effects, LACME and LANDE can be written in terms of the joint population distribution of Y_i, M_i, T_i, Z_i, and X_i. In practice, these population quantities must be estimated from a random sample. Here, I consider two estimation strategies for the LACME and LANDE. The estimators for the mediated and unmediated ITT effects can be found analogously.

First, for a discrete mediator and pre-treatment covariates, one can consistently estimate the quantities in Theorem 1 by replacing $Q_{t|zx}$, $G_{m|tzx}$ and S_{mtzx} with their sample analogues. That is, the following estimators are consistent for $\delta(t)$ and $\zeta(t)$ respectively:

Estimator 1 (Fully Nonparametric Estimator)

$$\widehat{\delta}(t) = \sum_{x\in\mathcal{X}}\sum_{m\in\mathcal{M}} \frac{\widehat{P}_{mt|tx}\widehat{S}_{mttx} - \widehat{P}_{mt|(1-t)x}\widehat{S}_{mt(1-t)x}}{\widehat{P}_{mt|tx} - \widehat{P}_{mt|(1-t)x}} \times \frac{\widehat{P}_{m1|1x} - \widehat{P}_{m1|0x} - \widehat{P}_{m0|0x} + \widehat{P}_{m0|1x}}{\widehat{Q}_{1|1x} - \widehat{Q}_{1|0x}}, \quad (13)$$

$$\widehat{\zeta}(t) = \sum_{x\in\mathcal{X}}\sum_{m\in\mathcal{M}} \left\{ \frac{\widehat{P}_{m1|1x}\widehat{S}_{m11x} - \widehat{P}_{m1|0x}\widehat{S}_{m10x}}{\widehat{P}_{m1|1x} - \widehat{P}_{m1|0x}} - \frac{\widehat{P}_{m0|0x}\widehat{S}_{m00x} - \widehat{P}_{m0|1x}\widehat{S}_{m01x}}{\widehat{P}_{m0|0x} - \widehat{P}_{m0|1x}} \right\} \times \frac{\widehat{P}_{mt|tx} - \widehat{P}_{mt|(1-t)x}}{\widehat{Q}_{1|1x} - \widehat{Q}_{1|0x}}, \quad (14)$$

where $\widehat{S}_{mtzx} = (\sum_{i=1}^{N} Y_i \cdot \mathbf{1}\{M_i = m\} \cdot \mathbf{1}\{T_i = t\} \cdot \mathbf{1}\{Z_i = z\} \cdot \mathbf{1}\{X_i = x\})/(\sum_{i=1}^{N} \mathbf{1}\{M_i = m\} \cdot \mathbf{1}\{T_i = t\} \cdot \mathbf{1}\{Z_i = z\} \cdot \mathbf{1}\{X_i = x\})$, $\widehat{P}_{mt|zx} = (\sum_{i=1}^{N} \mathbf{1}\{M_i = m\} \cdot \mathbf{1}\{T_i = t\} \cdot \mathbf{1}\{Z_i = z\} \cdot \mathbf{1}\{X_i = x\})(\sum_{i=1}^{N} \mathbf{1}\{Z_i = z\} \cdot \mathbf{1}\{X_i = x\})$, $\widehat{Q}_{t|zx} = (\sum_{i=1}^{N} \mathbf{1}\{T_i = t\} \cdot \mathbf{1}\{Z_i = z\} \cdot \mathbf{1}\{X_i = x\})(\sum_{i=1}^{N} \mathbf{1}\{Z_i = z\} \cdot \mathbf{1}\{X_i = x\})$, *and* $\mathbf{1}\{\cdot\}$ *represents the indicator function.*

Estimator 1 is fully nonparametric and does not rely on any assumption other than Assumptions 1–4. In practice, however, this estimator will be difficult to use when either the mediator or the pre-treatment covariates take on many levels and/or X_i consists of many variables relative to the sample size, because each of the sample averages in the expression requires subclassification into the strata defined by the levels of X_i (as well as the levels of M_i for $\widehat{S}_{mtzx}$).

The second estimator is based on a more flexible approach where S_{mtzx}, $G_{m|tzx}$ and $Q_{t|zx}$ are estimated with regression models. This estimator can be expressed as follows.

Estimator 2 (Regression-based Estimator)

$$\widetilde{\delta}(t) = \frac{1}{N}\sum_{i=1}^{N} \int \frac{\widetilde{Q}_{t|tX_i}\widetilde{G}_{m|ttX_i}\widetilde{S}_{mttX_i} - \widetilde{Q}_{t|(1-t)X_i}\widetilde{G}_{m|t(1-t)X_i}\widetilde{S}_{mt(1-t)X_i}}{\widetilde{Q}_{t|tX_i}\widetilde{G}_{m|ttX_i} - \widetilde{Q}_{t|(1-t)X_i}\widetilde{G}_{m|t(1-t)X_i}} \times \frac{\widetilde{Q}_{1|1X_i}\widetilde{G}_{m|11X_i} - \widetilde{Q}_{1|0X_i}\widetilde{G}_{m|10X_i} - \widetilde{Q}_{0|0X_i}\widetilde{G}_{m|00X_i} + \widetilde{Q}_{0|1X_i}\widetilde{G}_{m|01X_i}}{\widetilde{Q}_{1|1X_i} - \widetilde{Q}_{1|0X_i}} \, dm, \quad (15)$$

$$\widetilde{\zeta}(t) = \frac{1}{N}\sum_{i=1}^{N} \int \left\{ \frac{\widetilde{Q}_{1|1X_i}\widetilde{G}_{m|11X_i}\widetilde{S}_{m11X_i} - \widetilde{Q}_{1|0X_i}\widetilde{G}_{m|10X_i}\widetilde{S}_{m10X_i}}{\widetilde{Q}_{1|1X_i}\widetilde{G}_{m|11X_i} - \widetilde{Q}_{1|0X_i}\widetilde{G}_{m|10X_i}} - \frac{\widetilde{Q}_{0|0X_i}\widetilde{G}_{m|00X_i}\widetilde{S}_{m00X_i} - \widetilde{Q}_{0|1X_i}\widetilde{G}_{m|01X_i}\widetilde{S}_{m01X_i}}{\widetilde{Q}_{0|0X_i}\widetilde{G}_{m|00X_i} - \widetilde{Q}_{0|1X_i}\widetilde{G}_{m|01X_i}} \right\}$$

$$\times \frac{\widetilde{Q}_{t|tX_i}\widetilde{G}_{m|ttX_i} - \widetilde{Q}_{t|(1-t)X_i}\widetilde{G}_{m|t(1-t)X_i}}{\widetilde{Q}_{1|1X_i} - \widetilde{Q}_{1|0X_i}}\, dm, \qquad (16)$$

where $\widetilde{S}_{mtzx}$, $\widetilde{G}_{m|tzx}$ and $\widetilde{Q}_{t|zx}$ are regression-based sample estimates of S_{mtzx}, $G_{m|tzx}$ and $Q_{t|zx}$, respectively.

This estimator requires that the analyst postulate regression models for S_{mtzx}, $G_{m|tzx}$ and $Q_{t|zx}$. Then, the estimated conditional densities and expectations will be calculated for the specified levels of the treatment assignment, treatment and mediator variables, with the pre-treatment covariates fixed at the observed values for each observation. Finally, the consistent estimates of the LACME and LANDE will be calculated via equations (15) and (16), respectively. When the mediator is a continuous variable, this final step often requires numerical integration over $\mathcal{M}$ at each observed value of X_i, which may be computationally difficult depending on the assumed mediator model. Computation is substantially less demanding when the mediator takes on discrete levels. Estimator 2 is highly general and accommodates a broad range of statistical models, including semiparametric and nonparametric regressions.

In principle, the Delta method can be employed to obtain asymptotic variances of Estimators 1 and 2. This approach however involves tedious algebra and therefore is not further pursued here. Instead, uncertainty estimates for these estimators can be calculated via simulation-based approaches, including the quasi-Bayesian Monte Carlo and the nonparametric bootstrap. In Section 4, I investigate the performance of the confidence intervals based on the nonparametric bootstrap under various conditions via a Monte Carlo experiment.

4 Simulation Study

In this section, I conduct a Monte Carlo experiment to investigate the finite-sample performance of the proposed estimators for the LACME.

4.1 Setup

I use a simple population data generating process where the outcome is continuous, the mediator is binary, and there is no pre-treatment confounders. The potential outcomes and mediators are generated from the following structural models:

$$\begin{aligned} \Pr(M_i(t) = 1) &= \text{logit}^{-1}\left(\alpha_{Mi} + \beta_{Mi} t\right), \\ Y_i(t, m) &= \alpha_{Yi} + \beta_{Yi} t + \gamma_i m + \kappa_i t m + \varepsilon_i, \end{aligned}$$

where $\alpha_{Mi} \sim \mathcal{N}(-\mathbf{1}\{P_i = c\} + \mathbf{1}\{P_i = a\} - 0.5 \cdot \mathbf{1}\{P_i = n\}, 1)$, $\beta_{Mi} \sim \mathcal{N}(\mathbf{1}\{P_i = c\} + 0.2 \cdot \mathbf{1}\{P_i = a\} + 0.6 \cdot \mathbf{1}\{P_i = n\}, 1)$, $\alpha_{Yi} \sim \mathcal{N}(\mathbf{1}\{P_i = c\} - 2 \cdot \mathbf{1}\{P_i = a\} + 0.5 \cdot \mathbf{1}\{P_i = n\}, 1)$, $\beta_{Yi} \sim \mathcal{N}(\mathbf{1}\{P_i = c\} + 0.4 \cdot \mathbf{1}\{P_i = a\} + 0.2 \cdot \mathbf{1}\{P_i = n\}, 1)$, $\gamma_i \sim \mathcal{N}(\mathbf{1}\{P_i = c\} - 0.4 \cdot \mathbf{1}\{P_i = a\} + 0.3 \cdot \mathbf{1}\{P_i = n\}, 1)$, $\kappa_i \sim \mathcal{N}(\mathbf{1}\{P_i = c\} + 0.8 \cdot \mathbf{1}\{P_i = n\}, 1)$, and $\varepsilon_i \sim \mathcal{N}(0, 1)$. This data generating process guarantees that observed mediators are distributed independently from the potential outcomes conditional on the treatment status and compliance type, but not without conditioning of the compliance type, i.e., $Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid T_i = t, P_i = p$ for $p \in \{c, a, n\}$ (implying Assumption 4) but $Y_i(t', m) \not\perp\!\!\!\perp M_i(t) \mid T_i = t$. Thus, the data generating process does not satisfy the global sequential ignorability of Imai et al. (2010b) or NPSEM of Pearl (2001).

The compliance types are generated from the categorical distribution $\Pr(P_i = c) = \pi$ and $\Pr(P_i = a) = \Pr(P_i = n) = (1 - \pi)/2$. Importantly, I set the proportion of compliers π to three different values (0.8, 0.4 and 0.2, corresponding to minor, moderate and severe noncompliance, respectively) in order to assess how the performance of the proposed estimators will change according to the severity of treatment noncompliance. As it turns out, the finite-sample performance appears to depend heavily on this parameter. Under this population model, the true values of the LACME under the treatment condition is $\delta(1) = 0.396$ regardless of the value of π.

Using this data generating process, I analyze the performance of the proposed estimator across the sample sizes (N) of 200, 1000 and 2000. For each sample size, I generate the treatment assignment

indicator Z_i via a complete randomization such that $\sum_{i=1}^{N} Z_i/N = 0.5$. I then construct the estimate of $\delta(1)$ along with the 95% nonparametric bootstrap percentile interval based on 2000 resamples. Note that, under the current data generating process, Estimators 1 and 2 produce identical estimates of the quantities of interest if the regression models for Estimator 2 are specified correctly (a normal linear regression for S_{mtz} and a logit model for $G_{m|tz}$) and the model parameters are estimated via maximum likelihood.

For the sake of illustration, I also calculate the naïve estimate of the LACME for each sample size and compliance probability by first estimating the mediated ITT effect using equation (9) and dividing the resulting quantity by the estimated proportion of compliers. As discussed in Section 3.4, this intuitively appealing approach produces a biased estimate of the LACME even though the mediated ITT is identified under the current data generating process. Based on Theorem 3, bias in this naïve estimate is expected to equal -0.134, -0.368 and -0.467 when π is set to 0.8, 0.4 and 0.2, respectively.

4.2 Results

Table 2 presents the results of this experiment after 2000 simulations. Overall, the performance quickly improves for the proposed estimator (top panel) as sample size becomes large regardless of the proportion of compliers in the population data generating process. The absolute level of performance, however, depends on compliance probability. When most individuals comply with the assigned treatment ($\pi = 0.8$, left side of the table), the estimator essentially has zero bias even with a sample size as small as 200. The bootstrap confidence intervals also have approximately correct coverage for all sample sizes. However, when only two fifths of the individuals are compliers ($\pi = 0.4$, middle of the table), the estimator performs rather poorly with the sample size of 200. The bias is large, and the root mean squared error is greater than that of the naïve estimator under the same condition. The estimator, though, starts to perform expectedly once the sample size is increased to 1000, and bias effectively disappears when the sample size reaches 2000. The confidence intervals also have approximately correct coverage for all sample sizes. Finally, the estimator becomes highly unstable when compliance probability is very low ($\pi = 0.2$, right side of the table). When the sample size equals 200, the point estimates suffer from extremely large bias and

		80% Compliance			40% Compliance			20% Compliance		
Estimator	Sample Size	Bias	RMSE	95% CI Coverage	Bias	RMSE	95% CI Coverage	Bias	RMSE	95% CI Coverage
Proposed	200	-0.026	0.211	0.938	-0.317	1.705	0.967	—	—	0.999
$\hat{\delta}(1)$	1000	-0.003	0.090	0.950	-0.030	0.211	0.943	-0.149	1.761	0.944
	2000	-0.002	0.063	0.946	-0.020	0.139	0.941	-0.074	0.367	0.940
Naïve	200	-0.147	0.227	0.873	-0.370	0.403	0.576	—	—	0.889
$\hat{\delta}^*(1)$	1000	-0.133	0.153	0.492	-0.367	0.371	0.009	-0.461	0.470	0.023
	2000	-0.136	0.146	0.340	-0.367	0.369	0.000	-0.463	0.467	0.000

Table 2: Finite-Sample Performance of the Proposed Estimator. The table shows the results of a Monte Carlo experiment with varying sample sizes and complier proportions. The results are based on 2000 Monte Carlo iterations. The rows correspond to different sample sizes as shown. Results for the causal parameters other than $\delta(1)$ are similar and omitted for the sake of space. For each of the three blocks corresponding to different population proportions of compliers, the columns represent (from left to right): estimated biases of the nonparametric estimator (Estimator 1), its root mean squared errors (RMSEs), and the coverage probabilities of the 95% percentile intervals based on 2000 nonparametric bootstrap resamples. The true values of the parameters are given in text in Section 4. The results indicate that the performance of the proposed estimator ($\hat{\delta}(1)$, top panel) quickly improves as sample size becomes large, although the absolute level of performance varies across compliance probabilities. The naïve estimator ($\hat{\delta}^*(1)$, bottom panel) is biased regardless of the sample size.

root mean squared errors due to a handful of simulation draws that produce estimates close to positive or negative infinity. The confidence intervals also tend to be too wide because of the problem with outlying simulation draws. However, the performance does improve as sample size becomes large. In contrast to these results, the naïve estimator (bottom panel) always produces biased estimates regardless of sample size, and the magnitude of biases increases as compliance probability becomes smaller. The bias matches the theoretically expected value for each compliance probability based on Theorem 3. The coverage probabilities of the confidence intervals also become increasingly poor as sample size becomes larger.

Based on these results, it may be concluded that the proposed estimator performs well as long as compliance probability is adequate and sample size is not too small. The estimator shows somewhat extreme behavior when either of these conditions is not met. Although this result may raise concerns, it is in fact an example of the well-documented problem common to many techniques based on instrumental variables (Bound et al., 1995). Methods for detecting and possibly correcting for the finite-sample problems caused by such "weak instruments" have been proposed (e.g. Stock et al., 2002).

5 Empirical Analysis

In this section, I apply the proposed method to the National JTPA study data introduced in Section 2.

5.1 Data

As discussed in Section 2, the interest here lies in whether the causal effects of assignment to the JTPA treatment (Z_i) and actual enrollment to the JTPA program (T_i) on post-program 12-month employment status (Y_i) are mediated by the hours spent in job training (M_i). Following the prominent study by Abadie et al. (2002) who analyzed the JTPA data, I estimate the causal parameters of interest for both gender subgroups as well as for the entire sample of adults. The estimated compliance probability was 0.638 for the entire sample, which falls between the "high" and "moderate" compliance rates investigated in the simulation study in Section 4. These numbers are similar for each gender subgroup (see Table 1). With the sample size of more than a few thousands, it may be safe to conclude that the proposed estimators for the mediated and unmediated ITT effects, LACME and LANDE are likely to have a reasonable performance.

The mediator is a non-negative continuous variable and equals zero for many participants in both the treatment and control groups. The outcome variable is a binary indicator which takes on the value of one for the participants who were employed at any point within the 12-month follow-up period. In addition to these variables, I include a set of pre-treatment covariates (X_i) in the analysis. I follow the choice of Abadie et al. (2002) and use the same set of variables used in their paper (age, education, race, marriage status, prior employment, recommended service type, earnings data source, and AFDC benefit receipt for female participants; descriptive statistics can be found in their article). Unlike the standard analysis of the LATE and ITT effects in randomized experiments with noncompliance, where pre-treatment covariates are not required for identification, these variables play a crucial role in the analysis of causal mechanisms because the plausibility of Assumption 4 depends on the choice of the covariates.

5.2 Assumptions and Estimation

I now use Estimator 2 to estimate the causal parameters of interest. As explained in Section 3, these estimators are consistent under Assumptions 1 to 4 as well as the assumption that the predictive models for the treatment, mediator and outcome are all correctly specified. In the current context, Assumption 1 implies that receiving an offer to enroll in a JTPA program must affect both the hours spent in job training and the post-program employment only through actual enrollment. This assumption is generally regarded as plausible by experts and in fact maintained in previous studies (e.g. Orr et al., 1996; Abadie et al., 2002), although Orr et al. (1996) note the possibility of minor violation. Assumption 2 implies that there must not be any participant who would enroll in a JTPA program only when assigned to the control condition. Because the experimental protocol only allowed one-sided noncompliance in principle, this assumption is guaranteed to hold almost exactly, although as noted above, the existence of a small fraction of the control group who took the treatment leaves the possibility of slight violation.

The remaining two identification assumptions are both ignorability conditions. Assumption 3 is guaranteed to be true by the successful randomization of the treatment assignment. In contrast, Assumption 4 is probably the most controversial assumption, for it is neither implied by the experimental design nor empirically verifiable (Robins and Richardson, 2010). It is, however, important to note that Assumption 4 represents a slight relaxation of the standard sequential ignorability (or NPSEM) assumption that is typically made in the previous literature on causal mediation (e.g. Pearl, 2001; Imai et al., 2010b; Tchetgen Tchetgen and Shpitser, 2011; Albert and Nelson, 2011), since it requires the conditional ignorability of the mediator only among compliers. Indeed, this additional conditioning on the unobserved compliance type appears to make the assumption substantially more plausible in the current application, where the potential mediators (hours participants would choose to spend in training if enrolled or not enrolled) are likely to be strongly correlated with the potential treatment (program enrollment decision given treatment assignment) due to unobserved common causes such as participants' willingness and motivation to receive training.

Given the types of the mediator and outcome variables, I use the following parametric models to estimate the conditional density and expectation of these variables ($\widetilde{G}_{m|tzx}$ and $\widetilde{S}_{mtzx}$):

$$\begin{aligned}
\Pr(M_i = 0 \mid T_i, Z_i, X_i) &= \Phi\left(\alpha_1 + \beta_1 T_i + \gamma_1 Z_i + \kappa_1 T_i Z_i + \xi_1^\top X_i\right), && (17)\\
\log(M_i) \mid M_i > 0, T_i, Z_i, X_i &\sim \mathcal{N}\left(\alpha_2 + \beta_2 T_i + \gamma_2 Z_i + \kappa_2 T_i Z_i + m_2^\top X_i,\ \sigma^2\right), && (18)\\
\Pr(Y_i = 1 \mid M_i, T_i, Z_i, X_i) &= \text{logit}^{-1}\big(\alpha_3 + \beta_3 T_i + \gamma_3 Z_i + \kappa_3 T_i Z_i \\
&\qquad + \eta M_i + \phi T_i M_i + \psi Z_i M_i + \nu T_i Z_i M_i + \xi_3^\top X_i\big), && (19)
\end{aligned}$$

where $\Phi(\cdot)$ denotes the distribution function of a standard normal random variable. Together, equations (17) and (18) imply a two-part probit-lognormal regression model for the observed mediator, and I use the two-step estimation approach implemented in the `mhurdle` package in R (Carlevaro et al., 2012) to estimate the model coefficients and obtain $\widetilde{G}_{m|tzx}$. The logit model for the observed outcome in equation (19) is estimated via maximum likelihood implemented in the `glm` function in R to obtain $\widetilde{S}_{mtzx}$. Finally, the estimated conditional probability of treatment ($\widetilde{Q}_{t|zx}$) is calculated via the standard approach of regressing the observed treatment indicator (T_i) on the treatment assignment indicator (Z_i) and the pre-treatment covariates (X_i) via the `lm` function in R. Once the point estimates for the mediated and unmediated ITT effects, LACME and LANDE are obtained based on these estimates, uncertainty estimates are calculated via the nonparametric bootstrap.

5.3 Results

Figure 2 presents the results of the proposed analysis for the entire sample of adult participants (left panels) as well as for the two gender groups (center and right panels). The top three panels show the LACME, LANDE and LATE, while the bottom three panels present the mediated, unmediated and total ITT effects. In each panel, the point estimates based on the proposed method (solid circles) are shown along with the 90% and 95% confidence intervals (thick and thin horizontal bars, respectively) based on percentiles of 5000 nonparametric bootstrap resamples. For the sake of comparison, I also calculate the "naïve"

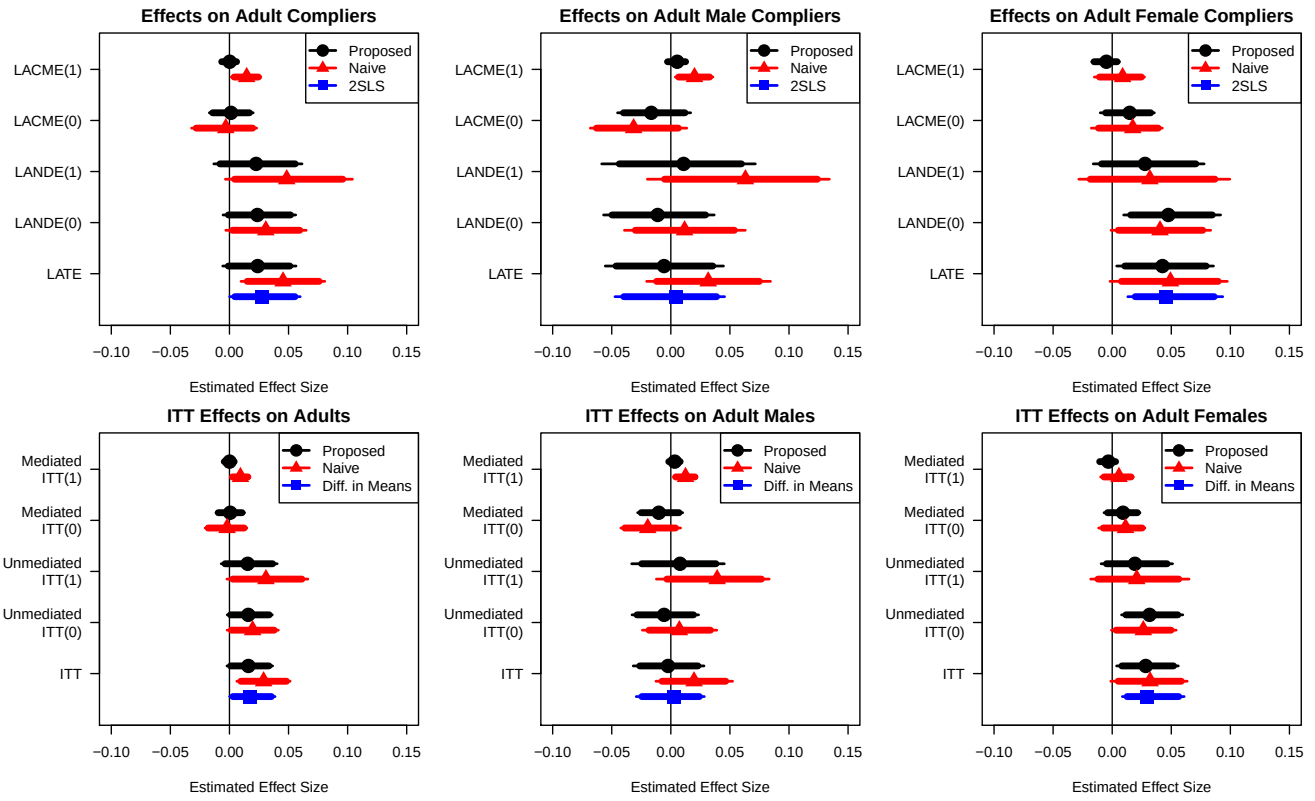

Figure 2: Estimated Causal Effects of JTPA Enrollment on Employment Mediated and Unmediated by the Actual Hours Spent in Job Training. In each plot, the solid circles represent the quantities of interest estimated based on Estimator 2 and the models given in Section 5.2, while the solid triangles represent the "naïve" estimates described in Section 5.3. The solid squares represent the benchmark nonparametric estimates of the LATE and the ITT effects. The thick and thin horizontal bars represent the 90% and 95% confidence intervals, respectively, based on the appropriate percentiles of 5000 nonparametric bootstrap resamples.

estimates of these quantities based on the strategy of applying Imai et al.'s (2010b) estimator to the ITT effects (using two-part probit-lognormal and logit models for the mediator and outcome, respectively) and dividing the results by the estimated probability of compliance (solid triangles). As discussed in Section 3, the latter estimates are generally invalid, although previous studies in program evaluation often used this strategy (e.g. Flores and Flores-Lagunes, 2009).

We begin with the discussion of LATE, the total effect of JTPA enrollment on the probability of employment after the study period among the compliers. The total LATE is estimated to be 0.024 for the adult participants, with the 90% and 95% confidence intervals of $[-0.001, 0.051]$ and $[-0.006, 0.056]$, respectively. The estimates are somewhat larger for females (0.043, $[0.011, 0.079]$ and $[0.004, 0.086]$) and nearly zero for males (-0.006, $[-0.046, 0.035]$, and $[-0.055, 0.045]$). These estimates indicate that

enrollment in JTPA training programs among the compliers had a positive effect on the probability of employment that barely misses conventional thresholds for statistical significance, and the effects were stronger and statistically significant among the female enrollees. Of note, these point estimates are nearly identical to the estimates of the LATE based on two-stage least squares (2SLS; solid squares) for both all adult compliers and each gender subgroup, although the confidence intervals are slightly wider. Given the known robustness properties of the 2SLS estimator (Imbens and Angrist, 1994; Angrist et al., 1996), this finding provides some evidence that the proposed method produces reliable estimates.

Based on the proposed method, the estimated LACME for adult compliers equals almost zero, regardless of whether the baseline treatment status is set to the treatment (0.000, with the 90% and 95% confidence intervals of $[-0.007, 0.006]$ and $[-0.008, 0.008]$, respectively) or to the control (0.001, $[-0.014, 0.017]$, and $[-0.017, 0.020]$). The LANDE, on the other hand, is estimated to be positive (0.023 for the treatment and 0.024 for the control), although the 90% confidence intervals slightly overlap with zero ($[-0.008, 0.056]$ for the treatment and $[-0.001, 0.052]$ for the control). These estimates suggest that the mechanism underlying the causal effect of JTPA enrollment on post-program employment does not on average involve the number of hours spent in job training. The results for female compliers are less clear-cut, but follow the same general pattern. That is, the estimated LACME is close to zero and statistically insignificant (-0.005, $[-0.015, 0.004]$ and $[-0.017, 0.006]$ for the treatment baseline) while the corresponding LANDE is positive and significant (0.047, $[0.016, 0.085]$ and $[0.010, 0.092]$ for the control). Finally, the estimated LACME and LANDE for male compliers are all close to zero, which is not surprising given that the total LATE is also nearly zero for this subgroup.

In contrast, the naïve estimator provides a substantively different picture of the causal mechanism behind the effect of JTPA enrollment. The naïve estimate of the LACME for the treatment baseline is positive (0.015) and statistically significant at the 0.05 level, with the 90% and 95% confidence intervals of $[0.004, 0.025]$ and $[0.002, 0.027]$, respectively. Thus, based on the naïve estimator, one might falsely conclude that the effect of JTPA enrollment on employment can be partially attributed to the increased time spent in job training. This result also holds for male compliers, for whom the naïve estimate of

LACME equals 0.022 and the 95% confidence interval does not overlap with zero ($[0.004, 0.036]$). Indeed, even the estimates of the total LATE appear to be substantially biased when they are constructed from the naïve estimator, especially for male compliers. The results for the mediated and unmediated ITT effects turn out to be qualitatively identical to those for the LACME and LANDE for all subgroups.

Overall, the results based on the proposed estimator suggest that the causal effect of JTPA enrollment on post-program employment was on average not mediated through the number of hours the participants spent receiving job training among the compliers. Several interpretations of this finding are possible. First, as discussed in Section 2, spending longer hours in job training programs may have been counter-productive for some participants who could otherwise have spent that time on actually searching for jobs or working. Second, it may be the experience involved in the process of JTPA enrollment — turning up at the study site, signing the forms, receiving basic instructions and consultation, etc. — that had a positive impact on the participants' employment prospect (perhaps by motivating them to put more efforts to find jobs) and the actual contents of training programs only had negligible impact. Finally, it is of course a possibility that the estimated effects are misrepresentation of the unobserved true effects due to problems in sampling process, mismeasurement, or the violation of identification or modeling assumptions. In particular, Assumption 4 is a strong assumption and must be evaluated with care. In summary, further investigation of causal mechanisms underlying job training programs is called for in future research.

6 Conclusion

Over the past couple of decades, statisticians and econometricians have made major advances in the analysis of randomized experiments with treatment noncompliance. Instrumental variables methods are now part of applied researchers' routine toolkit in medical, biological, social and behavioral sciences. However, implications of treatment noncompliance have not been well understood in the context of causal mediation analysis. Although noncompliance often occurs in program evaluation where researchers are interested in the identification of causal pathways in addition to that of average causal effects, the topic has attracted little attention to date.

This paper contributes to the literature on causal mediation by shedding light on this rather neglected, but nonetheless important topic. I show that one can nonparametrically identify the mediated and unmediated ITT effects, LACME and LANDE for compliers under a set of assumptions slightly weaker than the standard instrumental variables assumptions and the sequential ignorability assumption. Based on the identification result, I propose several estimators including a fully nonparametric estimator and a regression-based estimator. Through a Monte Carlo simulation study, I show that the proposed estimator performs well under a reasonable sample size and moderate degree of noncompliance, although the estimator suffers from the "weak instrument" problem which also plagues other instrumental variables methods. Finally, I apply the proposed method to a major job training evaluation study and find evidence that the effect of enrollment in the program is not mediated by the actual hours spent receiving job training, while the intuitively plausible "naïve" method that has been used in previous studies might lead to a different substantive conclusion.

This paper leaves several important questions for future research. Among others, it has been assumed throughout the paper that, in addition to the standard instrumental variables assumptions (Assumptions 1–3), the observed mediator is statistically independent of all the potential outcomes among compliers after conditioning on the treatment status and pre-treatment covariates (Assumption 4). Although this assumption is slightly weaker than the standard assumptions employed in the literature, it is still a strong assumption that is inherently unverifiable from observed data. A natural future direction of the research would therefore be the development of sensitivity analysis for this assumption. Other possible topics include the adaptation of the existing instrumental variables techniques that have desirable robustness properties for the estimation of the LACME and related quantities.

A Mathematical Appendix

A.1 Proof of Theorem 1

First, note that Assumptions 3 and 4 imply that the sequential ignorability assumption of Imai et al. (2010b) holds after conditioning on $P_i = c$ because

$$\Pr(T_i = Z_i \mid P_i = c, X_i = x) \;=\; 1 \tag{20}$$

for any i and $x \in \mathcal{X}$. This implies that Theorem 1 of Imai et al. (2010b) also holds after conditioning on $P_i = c$. That is, we have:

$$\begin{aligned}
& \mathbb{E}(Y_i(t, M_i(t')) \mid P_i = c, X_i = x) \\
= & \int \mathbb{E}(Y_i(t,m) \mid M_i(t') = m, P_i = c, Z_i = T_i = t', X_i = x)\, p(M_i(t') = m \mid P_i = c, X_i = x) dm \\
= & \int \mathbb{E}(Y_i(t,m) \mid Z_i = T_i = t', P_i = c, X_i = x)\, p(M_i(t') = m \mid P_i = c, X_i = x) dm \\
= & \int \mathbb{E}(Y_i(t,m) \mid Z_i = T_i = t, P_i = c, X_i = x)\, p(M_i(t') = m \mid Z_i = T_i = t', P_i = c, X_i = x) dm \\
= & \int \mathbb{E}(Y_i(t,m) \mid M_i(t) = m, P_i = c, Z_i = T_i = t, X_i = x)\, p(M_i(t') = m \mid Z_i = T_i = t', P_i = c, X_i = x) dm \\
= & \int \mathbb{E}(Y_i \mid M_i = m, Z_i = T_i = t, P_i = c, X_i = x)\, p(M_i = m \mid Z_i = T_i = t', P_i = c, X_i = x) dm
\end{aligned} \tag{21}$$

for any t and $t' \in \{0, 1\}$, where the first and third equalities follow from Assumption 3 and equation (20) and the second and fourth equalities follow from Assumption 4.

Next, note that under Assumptions 2 and 3, we have

$$Q_{1|1x} \;=\; \Pr(T_i(1) = 1 \mid X_i = x) \;=\; \Pr(P_i = c \mid X_i = x) + \Pr(P_i = a \mid X_i = x), \tag{22}$$

$$Q_{1|0x} \;=\; \Pr(T_i(0) = 1 \mid X_i = x) \;=\; \Pr(P_i = a \mid X_i = x), \tag{23}$$

implying that $\Pr(P_i = c \mid X_i = x) = Q_{1|1x} - Q_{1|0x}$ and $\Pr(P_i = n \mid X_i = x) = Q_{0|1x}$ for any $x \in \mathcal{X}$.

Similarly by Assumption 3, we have

$$\begin{aligned} G_{m|11x} &= p(M_i(1) = m \mid T_i(1) = 1, X_i = x) \\ &= \sum_{t\in\{0,1\}} p(M_i(1) = m \mid T_i(1) = 1, T_i(0) = t, X_i = x) \Pr(T_i(0) = t \mid T_i(1) = 1, X_i = x) \\ &= p(M_i(1) = m \mid P_i = c, X_i = x)\frac{Q_{1|1x} - Q_{1|0x}}{Q_{1|1x}} + p(M_i(1) = m \mid P_i = a, X_i = x)\frac{Q_{1|0x}}{Q_{1|1x}}, \end{aligned} \tag{24}$$

where the second equality follows from the law of total probability and the last equality from equations (22) and (23). By the same logic, we have $G_{m|01x} = p(M_i(0) = m \mid P_i = n, X_i = x)$, $G_{m|10x} = p(M_i(1) = m \mid P_i = a, X_i = x)$ and $G_{m|00x} = p(M_i(0) = m \mid P_i = c, X_i = x)\frac{Q_{1|1x}-Q_{1|0x}}{Q_{0|0x}} + p(M_i(0) = m \mid P_i = n, X_i = x)\frac{Q_{0|1x}}{Q_{0|0x}}$. Using the above results, we can now show that the second component of equation (21) (i.e., the conditional probability density or mass function of M_i given $P_i = c$, $T_i = Z_i$ and $X_i = x$) can be expressed in terms of observed quantities. First, we have

$$\begin{aligned} p(M_i = m \mid Z_i = T_i = t, P_i = c, X_i = x) &= p(M_i(t) = m \mid P_i = c, X_i = x) \\ &= \frac{Q_{t|tx}G_{m|ttx} - Q_{t|(1-t)x}G_{m|t(1-t)x}}{Q_{1|1x} - Q_{1|0x}}, \end{aligned} \tag{25}$$

for $t \in \{0, 1\}$. Next, note that

$$\begin{aligned} S_{m11x} &= \mathbb{E}[Y_i(1, m) \mid M_i(1) = m, T_i(1) = 1, Z_i = 1, X_i = x] \\ &= \mathbb{E}[Y_i(1, m) \mid M_i(1) = m, T_i(1) = 1, X_i = x] \\ &= \sum_{t\in\{0,1\}} \mathbb{E}[Y_i(1, m) \mid M_i(1) = m, T_i(1) = 1, T_i(0) = t, X_i = x] \\ &\qquad \times \Pr(T_i(0) = t \mid M_i(1) = m, T_i(1) = 1, X_i = x) \\ &= \sum_{t\in\{0,1\}} \mathbb{E}[Y_i(1, m) \mid M_i(1) = m, T_i(1) = 1, T_i(0) = t, X_i = x] \\ &\quad \times \frac{p(M_i(1) = m \mid T_i(1) = 1, T_i(0) = t, X_i = x) \Pr(T_i(0) = t \mid T_i(1) = 1, X_i = x)}{\sum_{t'\in\{0,1\}} p(M_i(1) = m \mid T_i(1) = 1, T_i(0) = t', X_i = x) \Pr(T_i(0) = t' \mid T_i(1) = 1, X_i = x)} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{p\in\{c,a\}} \mathbb{E}[Y_i(1,m) \mid M_i(1)=m, P_i=p, X_i=x] \\
&\qquad \times \frac{p(M_i(1)=m \mid P_i=p, X_i=x)\Pr(P_i=p \mid X_i=x)}{\sum_{p'\in\{c,a\}} p(M_i(1)=m \mid P_i=p', X_i=x)\Pr(P_i=p' \mid X_i=x)} \\
&= \mathbb{E}[Y_i(1,m) \mid M_i(1)=m, P_i=c, X_i=x]\frac{Q_{1|1x}G_{m|11x} - Q_{1|0x}G_{m|10x}}{Q_{1|1x}G_{m|11x}} \\
&\qquad + \mathbb{E}[Y_i(1,m) \mid M_i(1)=m, P_i=a, X_i=x]\frac{Q_{1|0x}G_{m|10x}}{Q_{1|1x}G_{m|11x}}, \qquad (26)
\end{aligned}
$$

where the second equality follows from Assumption 3, the third equality is due to the law of total expectation, the fourth equality is because of Bayes' rule, the fifth equality is the result of the definition of the compliance types and the last equality uses equation (25). Using the same logic, we can show that $S_{m01x} = \mathbb{E}[Y_i(0,m) \mid M_i(0)=m, P_i=n, X_i=x]$, $S_{m10x} = \mathbb{E}[Y_i(1,m) \mid M_i(1)=m, P_i=a, X_i=x]$ and $S_{m00x} = \mathbb{E}[Y_i(0,m) \mid M_i(0)=m, P_i=c, X_i=x]\frac{Q_{0|0x}G_{m|00x}-Q_{0|1x}G_{m|01x}}{Q_{0|0x}G_{m|00x}} + \mathbb{E}[Y_i(0,m) \mid M_i(0)=m, P_i=n, X_i=x]\frac{Q_{0|1x}G_{m|01x}}{Q_{0|0x}G_{m|00x}}$. These equations imply that the first component of equation (21) (i.e., the conditional expectation of Y_i given $P_i=c$, $T_i=Z_i$ and $X_i=x$) can also be expressed in terms of observed quantities as follows,

$$
\begin{aligned}
\mathbb{E}[Y_i \mid M_i=m, Z_i=T_i=t, P_i=c, X_i=x] &= \mathbb{E}[Y_i(t,m) \mid M_i(t)=m, P_i=c, X_i=x] \\
&= \frac{Q_{t|tx}G_{m|ttx}S_{mttx} - Q_{t|(1-t)x}G_{m|t(1-t)x}S_{mt(1-t)x}}{Q_{t|tx}G_{m|ttx} - Q_{t|(1-t)x}G_{m|t(1-t)x}} \qquad (27)
\end{aligned}
$$

for $t \in \{0,1\}$. Substituting the above results into equation (21) and taking the differences between the combinations of the resulting quantities corresponding to $\delta(1)$, $\delta(0)$, $\zeta(1)$ and $\zeta(0)$ defined in Section 3.1 yields the expressions in Theorem 1. ∎

A.2 Proof of Corollary 2

First, let $\lambda(z,x) = \mathbb{E}[Y_i(T_i(z), M_i(T_i(1))) - Y_i(T_i(z), M_i(T_i(0))) \mid X_i=x]$ and $\delta(t,x) = \mathbb{E}[Y_i(t, M_i(1)) - Y_i(t, M_i(0)) \mid P_i=c, X_i=x]$. Then Assumptions 1 and 2 imply that $\lambda(z,x) = \sum_{p\in\{c,a,n\}} \mathbb{E}[Y_i(T_i(z), M_i(T_i(1))) - Y_i(T_i(z), M_i(T_i(0))) \mid P_i=p, X_i=x]\Pr(P_i=p \mid X_i=x) = \delta(z,x)\Pr(P_i=c \mid$

$X_i = x)$. Noting that $\lambda(z) = \int \lambda(z,x) dF_X(x)$ and that $Pr(P_i = c \mid X_i = x) = Q_{1|1x} - Q_{1|0x}$ under Assumptions 2 and 3, it is immediate that equation (5) implies equation (7). By an analogous argument equation (6) can be shown to imply equation (8), completing the proof. ∎

A.3 Proof of Theorem 3

We begin with the case of $z = 1$. Under Assumptions 1–3 and equation (9), we have

$$
\begin{aligned}
&\lambda^*(1) - \mathbb{E}[Y_i(T_i(1), M_i(T_i(1)))] \\
&= -\iint \mathbb{E}[Y_i \mid M_i = m, Z_i = 1, X_i = x] p(M_i = m \mid Z_i = 0, X_i = x) dm dF_{X_i}(x) \\
&= -\iint \left\{ \begin{array}{l} \mathbb{E}[Y_i \mid M_i = m, Z_i = 1, X_i = x, P_i = c] \Pr(P_i = c \mid M_i = m, Z_i = 1, X_i = x) \\ +\mathbb{E}[Y_i \mid M_i = m, Z_i = 1, X_i = x, P_i = a] \Pr(P_i = a \mid M_i = m, Z_i = 1, X_i = x) \\ +\mathbb{E}[Y_i \mid M_i = m, Z_i = 1, X_i = x, P_i = n] \Pr(P_i = m \mid M_i = m, Z_i = 1, X_i = x) \end{array} \right\} \\
&\qquad\qquad \times p(M_i = m \mid Z_i = 0, X_i = x) dm dF_{X_i}(x) \\
&= -\iint \left\{ \begin{array}{l} \mathbb{E}[Y_i(1,m) \mid M_i(1) = m, X_i = x, P_i = c] p(M_i(1) = m \mid P_i = c, X_i = x) \Pr(P_i = c \mid X_i = x) \\ +\mathbb{E}[Y_i(1,m) \mid M_i(1) = m, X_i = x, P_i = a] p(M_i(1) = m \mid P_i = a, X_i = x) \Pr(P_i = a \mid X_i = x) \\ +\mathbb{E}[Y_i(0,m) \mid M_i(0) = m, X_i = x, P_i = n] p(M_i(0) = m \mid P_i = n, X_i = x) \Pr(P_i = n \mid X_i = x) \end{array} \right\} \\
&\qquad\qquad \times \frac{p(M_i(T_i(1)) = m \mid X_i = x)}{p(M_i(T_i(0)) = m \mid X_i = x)} dm dF_{X_i}(x),
\end{aligned}
$$

where the first equality follows from Assumption 3, the second equality follows from the law of total expectation and Assumption 2 and the third equality follows from Bayes' rule, Assumption 3, and the definition of the compliance types. Similarly, for $\lambda(1)$ in equation (1) we have

$$
\begin{aligned}
&\lambda(1) - \mathbb{E}[Y_i(T_i(1), M_i(T_i(1)))] \\
&= -\iint \mathbb{E}[Y_i(T_i(1), m) \mid M_i(T_i(0)) = m, X_i = x] p(M_i(T_i(0)) = m \mid X_i = x) dm dF_{X_i}(x) \\
&= -\iint \left\{ \begin{array}{l} \mathbb{E}[Y_i(1,m) \mid M_i(0) = m, X_i = x, P_i = c] p(M_i(0) = m \mid P_i = c, X_i = x) \Pr(P_i = c \mid X_i = x) \\ +\mathbb{E}[Y_i(1,m) \mid M_i(1) = m, X_i = x, P_i = a] p(M_i(1) = m \mid P_i = a, X_i = x) \Pr(P_i = a \mid X_i = x) \\ +\mathbb{E}[Y_i(0,m) \mid M_i(0) = m, X_i = x, P_i = n] p(M_i(0) = m \mid P_i = n, X_i = x) \Pr(P_i = n \mid X_i = x) \end{array} \right\}
\end{aligned}
$$

$$dmdF_{X_i}(x).$$

Now, note that Assumptions 3 and 4 implies $\mathbb{E}[Y_i(1,m) \mid M_i(t)=m, X_i=x, P_i=c] = \mathbb{E}[Y_i(1,m) \mid M_i(t)=m, T_i=Z_i=t, X_i=x, P_i=c] = \mathbb{E}[Y_i(1,m) \mid X_i=x, P_i=c]$ for $t \in \{0,1\}$. Therefore, the bias can be expressed as,

$$\begin{aligned}
&\lambda^*(1) - \lambda(1) \\
&= \iint \left\{ \begin{array}{l}
\mathbb{E}[Y_i(1,m) \mid X_i=x, P_i=c]\big\{p(M_i(0)=m \mid X_i=x, P_i=c) \\
\qquad\qquad -p(M_i(1)=m \mid X_i=x, P_i=c)\phi(m,x)\big\} \Pr(P_i=c \mid X_i=x) \\
+\mathbb{E}[Y_i(1,m) \mid M_i(1)=m, X_i=x, P_i=a] p(M_i(1)=m \mid X_i=x, P_i=a) \\
\qquad \times \{1-\phi(m,x)\} \Pr(P_i=a \mid X_i=x) \\
+\mathbb{E}[Y_i(0,m) \mid M_i(0)=m, X_i=x, P_i=n] p(M_i(0)=m \mid X_i=x, P_i=n) \\
\qquad \times \{1-\phi(m,x)\} \Pr(P_i=n \mid X_i=x)
\end{array} \right\} dmdF_{X_i}(x) \qquad (28)
\end{aligned}$$

where $\phi(m,x) \equiv p(M_i(T_i(0))=m \mid X_i=x)/p(M_i(T_i(1))=m \mid X_i=x) = p(M_i=m \mid Z_i=0, X_i=x)/p(M_i=m \mid Z_i=1, X_i=x)$ under Assumption 3. Then, note that from Section A.1, we have $\mathbb{E}[Y_i(1,m) \mid X_i=x, P_i=c] = (S_{m11x}G_{m|11x}Q_{1|1x} - S_{m10x}G_{m|10x}Q_{1|0x})/(G_{m|11x}Q_{1|1x} - G_{m|10x}Q_{1|0x})$, $\mathbb{E}[Y_i(1,m) \mid M_i(1)=m, P_i=a, X_i=x] = S_{m10x}$, $\mathbb{E}[Y_i(0,m) \mid M_i(0)=m, P_i=n, X_i=x] = S_{m00x}$, $p(M_i(t)=m \mid X_i=x, P_i=c) = (G_{m|ttx}Q_{t|tx} - G_{m|t(1-t)x}Q_{t|(1-t)x})/\Pr(P_i=c \mid X_i=x)$ for $t \in \{0,1\}$, $p(M_i(1)=m \mid X_i=x, P_i=a) = G_{m|10x}$, $p(M_i(0)=m \mid X_i=x, P_i=n) = G_{m|01x}$, $\Pr(P_i=a \mid X_i=x) = Q_{1|0x}$, and $\Pr(P_i=n \mid X_i=x) = Q_{0|1x}$. Substituting these results into equation (28) yields the expression in equation (11) for $z=1$. The case of $z=0$ and the expressions for $\mu(z)$ can be derived by following analogous steps. ∎

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. Journal of Econometrics, **113**.

Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica, **70**(1), 91–117.

Albert, J. M. and Nelson, S. (2011). Generalized causal mediation analysis. Biometrics, **67**(3), 1028–1038.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, **91**(434), 444–455.

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. In Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, pages 357–363, Edinburgh, Scotland. Morgan Kaufmann.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical Association, **90**, 443–450.

Carlevaro, F., Croissant, Y., and Hoareau, S. (2012). mhurdle: Estimation of models with limited dependent variables. R package version 0.1-3.

Flores, C. A. and Flores-Lagunes, A. (2009). Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness. Technical Report 4237, IZA Discussion paper.

Flores, C. A. and Flores-Lagunes, A. (2010). Nonparametric partial identification of causal net and mechanism average treatment effects. Unpublished manuscript.

Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. Biometrika, **86**(2), 365–379.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. Biometrics, **58**(1), 21–29.

Heckman, J., Pinto, R., and Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. American Economic Review, page forthcoming.

Huber, M. (2013). Identifying causal mechanisms (primarily) based on inverse probability weighting. Journal of Applied Econometrics, page forthcoming.

Imai, K. and Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. Political Analysis, **21**(2), 141–171.

Imai, K., Keele, L., and Tingley, D. (2010a). A general approach to causal mediation analysis. Psychological Methods, **15**(4), 309–334.

Imai, K., Keele, L., and Yamamoto, T. (2010b). Identification, inference, and sensitivity analysis for causal mediation effects. Statistical Science, **25**(1), 51–71.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. Econometrica, **62**(2), 467–475.

Nagelkerke, N., Fidler, V., Bernsen, R., and Borgdorff, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. Statistics in Medicine, **19**, 1849–1864.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). Statistical Science, **5**, 465–480.

Orr, L. L., Bloom, H. S., Bell, S. H., Doolittle, F., Lin, W., and Cave, G. (1996). Does Training for the Disadvantaged Work? Evidence from the National JTPA Study. The Urban Institute Press, Washington, D.C.

Pearl, J. (2001). Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, pages 411–420, San Francisco, CA. Morgan Kaufmann.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In Highly Structured Stochastic Systems (eds., P.J. Green, N.L. Hjort, and S. Richardson), pages 70–81. Oxford University Press, Oxford.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. Epidemiology, **3**(2), 143–155.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. Journal of Educational Psychology, **66**, 688–701.

Rubin, D. B. (1990). Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. Statistical Science, **5**, 472–480.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized methods of moments. Journal of Business & Economic Statistics, **20**(4), 518–529.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2011). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. Technical report, Harvard University School of Public Health, Cambridge, MA.

Tchetgen Tchetgen, E. J. and VanderWeele, T. J. (2014). On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. Epidemiology, **25**, forthcoming.