

# Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects\*

Kosuke Imai<sup>†</sup>   Luke Keele<sup>‡</sup>   Teppei Yamamoto<sup>§</sup>

First Draft: November 4, 2008

This Draft: July 13, 2009

## Abstract

Causal mediation analysis is routinely conducted by applied researchers in a variety of disciplines including epidemiology, political science, psychology, and sociology. The goal of such an analysis is to investigate alternative causal mechanisms by examining the roles of intermediate variables that lie in the causal path between the treatment and outcome variables. In this paper, we first prove that under a particular version of sequential ignorability assumption, the average causal mediation effect (ACME) is nonparametrically identified. We compare our identifying assumption with those proposed in the literature. Some practical implications of our identification result are also discussed. In particular, the popular estimator based on the linear structural equation model (LSEM) can be interpreted as an ACME estimator if the linearity and no-interaction assumptions are satisfied in addition to the proposed assumption. We show that this assumption can easily be relaxed within the framework of LSEM. Second, we consider a simple nonparametric estimator of the ACME in order to relax distributional and functional form assumptions. We also discuss a more general nonparametric approach. Third, we propose a new sensitivity analysis that can be easily implemented by applied researchers within the standard LSEM framework. Like the existing identifying assumptions, the proposed assumption may be too strong in many applied settings. Thus, sensitivity analysis is essential in order to examine the robustness of empirical findings to the possible existence of an unmeasured confounder. Finally, we apply the proposed methods to a randomized experiment from political psychology.

**Key Words:** causal inference, causal mediation analysis, direct and indirect effects, linear structural equation models, sequential ignorability, unmeasured confounders

---

\*A previous version of this paper was circulated under the title of “Identification and Inference in Causal Mediation Analysis.” We thank Brian Eggleston, Adam Glynn, Guido Imbens, Gary King, Dave McKinnon, Judea Pearl, Marc Ratkovic, Jas Sekhon, Dustin Tingley, Tyler VanderWeele, and seminar participants at Columbia University, Harvard University, New York University, Notre Dame, University of North Carolina, University of Colorado – Boulder, University of Pennsylvania, and University of Wisconsin – Madison for useful suggestions. Financial support from the National Science Foundation (SES-0752050) and the Princeton University Committee on Research in the Humanities and Social Sciences is acknowledged.

<sup>†</sup>Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609-258-6610, Email: kimai@princeton.edu, URL: <http://imai.princeton.edu>

<sup>‡</sup>Assistant Professor, Department of Political Science, 2140 Derby Hall, Ohio State University, Columbus, OH 43210 Phone: 614-247-4256, Email: keele.4@polisci.osu.edu

<sup>§</sup>Ph.D. student, Department of Politics, Princeton University.

# 1 Introduction

Causal mediation analysis is routinely conducted by applied researchers in a variety of scientific disciplines including epidemiology, political science, psychology, and sociology (see MacKinnon, 2008). The goal of such analysis is to investigate alternative causal mechanisms by examining the role of intermediate variables that lie in the causal paths between the treatment and outcome variables. The statistical literature began to formally study causal mediation analysis about fifteen years ago (Robins and Greenland, 1992), and a number of articles have appeared in more recent years (e.g., Pearl, 2001; Robins, 2003; Rubin, 2004; Petersen *et al.*, 2006; Geneletti, 2007; Joffe *et al.*, 2007; Ten Have *et al.*, 2007; Albert, 2008; Jo, 2008; Joffe *et al.*, 2008; Glynn, 2008; Sobel, 2008; VanderWeele, 2008, 2009).

In this paper, we contribute to this fast-growing literature in several ways. In Section 2, we prove that under a particular version of sequential ignorability assumption, the average causal mediation effect (ACME) is nonparametrically identified. We compare our identifying assumption with those proposed in the literature, and discuss practical implications of our result. In particular, Baron and Kenny (1986)'s popular estimator, which is based on a linear structural equation model (LSEM), can be interpreted as an ACME estimator under the proposed assumption if an additional assumption is satisfied. We show that this additional assumption can be easily relaxed within the standard LSEM framework.

In Section 3, after briefly discussing parametric estimation procedures, we relax distributional and functional-form assumptions by considering a simple nonparametric estimator. We conduct a Monte Carlo experiment to investigate the finite-sample performance of the proposed nonparametric estimator and its asymptotic confidence interval. We also discuss how this estimator can be generalized to small sample situations.

Like the existing identifying assumptions, the proposed assumption may be too strong in typical situations in which causal mediation analysis is employed. In particular, in experiments where the treatment is randomized but the mediator is not, the ignorability of the treatment assignment holds but the ignorability of the mediator may not. In Section 4, we propose a new sensitivity analysis that can be easily implemented by applied researchers within the standard LSEM framework. This method directly evaluates the robustness of empirical findings to the possible existence of unmeasured pre-treatment variables that confound the relationship between the mediator and the outcome. Finally, in Section 5, we use the proposed methods to analyze a randomized experiment from political psychology. Section 6 gives concluding remarks.

## 2 Identification

### 2.1 The Framework

Consider a simple random sample of size  $n$  from a population where for each unit  $i$  we observe  $(T_i, M_i, X_i, Y_i)$ . We use  $T_i$  to denote the binary treatment variable where  $T_i = 1$  ( $T_i = 0$ ) implies unit  $i$  receives (does not receive) the treatment. The mediating variable of interest, i.e., the mediator, is represented by  $M_i$ , whereas  $Y_i$  represents the outcome variable. Finally,  $X_i$  denotes the vector of observed pre-treatment covariates, and we use  $\mathcal{M}$ ,  $\mathcal{X}$ , and  $\mathcal{Y}$  to denote the support of the distributions of  $M_i$ ,  $X_i$ , and  $Y_i$ , respectively.

To define the causal mediation effects, we use the potential outcomes framework. Let  $M_i(t)$

denote the potential value of the mediator for unit  $i$  under the treatment status  $T_i = t$ . Similarly, we use  $Y_i(t, m)$  to represent the potential outcome for unit  $i$  when  $T_i = t$  and  $M_i = m$ . Then, the observed variables can be written as  $M_i = M_i(T_i)$  and  $Y_i = Y_i(T_i, M_i(T_i))$ . Similarly, if the mediator takes  $J$  different values, there exist  $2J$  potential values of the outcome variable, only one of which can be observed.

Using the potential outcomes notation, we can define the causal mediation effect for unit  $i$  under treatment status  $t$  as (see Robins and Greenland, 1992; Pearl, 2001),

$$\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0)), \quad (1)$$

for  $t = 0, 1$ . Pearl (2001) called  $\delta_i(t)$  the *natural indirect effect*, while Robins (2003) used the term the *pure indirect effect* for  $\delta_i(0)$  and the *total indirect effect* for  $\delta_i(1)$ . In words,  $\delta_i(t)$  represents the difference between the potential outcome that would result under treatment status  $t$ , and the potential outcome that would occur if the treatment status is the same and yet the mediator takes a value that would result under the different treatment status. Note that the former is observable (if the treatment variable is actually equal to  $t$ ) whereas the latter is by definition unobservable (under the treatment status  $t$  we never observe  $M_i(1-t)$ ). This notation implicitly assumes that the potential outcome depends only on the values of the treatment and mediating variables regardless of how they are realized, e.g., for  $t = 0, 1$  and all  $m \in \mathcal{M}$ ,  $Y_i(t, M_i(t)) = Y_i(t, M_i(1-t)) = Y_i(t, m)$  if  $M_i(1) = M_i(0) = m$ .

Thus, equation (1) formalizes the idea that the mediation effects represent the indirect effects of the treatment through the mediator. In this paper, we focus on the identification and inference of the average causal mediation effect (ACME), which is defined as,

$$\bar{\delta}(t) \equiv \mathbb{E}(\delta_i(t)) = \mathbb{E}\{Y_i(t, M_i(1)) - Y_i(t, M_i(0))\}, \quad (2)$$

for  $t = 0, 1$ . In the potential outcomes framework, the causal effect of the treatment on the outcome for unit  $i$  is defined as  $\tau_i \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0))$ , which is typically called the *total causal effect*. Therefore, the causal mediation effect and the total causal effect have the following relationship,

$$\tau_i = \delta_i(t) + \zeta_i(1-t), \quad (3)$$

where  $\zeta_i(t) = Y_i(1, M_i(t)) - Y_i(0, M_i(t))$  for  $t = 0, 1$ . This quantity  $\zeta_i(t)$  is called the *natural direct effect* by Pearl (2001) and the *pure/total direct effect* by Robins (2003). This represents the causal effect of the treatment on the outcome when the mediator is set to the potential value that would occur under treatment status  $t$ . In other words,  $\zeta_i(t)$  is the direct effect of the treatment when the mediator is held constant. Equation (3) shows an important relationship where the total causal effect is equal to the sum of the mediation effect under one treatment condition and the natural direct effect under the other treatment condition. Clearly, this equality also holds for the average total causal effect so that,  $\bar{\tau} \equiv \mathbb{E}\{Y_i(1, M_i(1)) - Y_i(0, M_i(0))\} = \bar{\delta}(t) + \bar{\zeta}(1-t)$  for  $t = 0, 1$  where  $\bar{\zeta}(t) = \mathbb{E}(\zeta_i(t))$ .

The causal mediation effects and natural direct effects differ from the *controlled direct effect* of the mediator, i.e.,  $Y_i(t, m) - Y_i(t, m')$  for  $t = 0, 1$  and  $m \neq m'$ , and that of the treatment, i.e.,  $Y_i(1, m) - Y_i(0, m)$  for all  $m \in \mathcal{M}$  (Pearl, 2001; Robins, 2003). Unlike the mediation effects, the

controlled direct effects of the mediator are defined in terms of specific values of the mediator,  $m$  and  $m'$ , rather than its potential values,  $M_i(1)$  and  $M_i(0)$ . While causal mediation analysis is used to identify possible causal paths from  $T_i$  to  $Y_i$ , the controlled direct effects may be of interest, for example, if one wishes to understand how the causal effect of  $M_i$  on  $Y_i$  changes as a function of  $T_i$ . In other words, the former examines whether  $M_i$  *mediates* the causal relationship between  $T_i$  and  $Y_i$  whereas the latter investigates whether  $T_i$  *moderates* the causal effect of  $M_i$  on  $Y_i$  (Baron and Kenny, 1986).

## 2.2 The Main Identification Result

We now present our main identification result using the potential outcomes framework described above. We show that under a particular version of sequential ignorability assumption, the ACME is nonparametrically identified. We first define our identifying assumption,

ASSUMPTION 1 (SEQUENTIAL IGNORABILITY)

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x, \quad (4)$$

$$Y_i(t', m) \perp\!\!\!\perp M_i \mid T_i = t, X_i = x, \quad (5)$$

for  $t, t' = 0, 1$ , and all  $x \in \mathcal{X}$  where it is also assumed that  $0 < \Pr(T_i = t \mid X_i = x)$  and  $0 < p(M_i = m \mid T_i = t, X_i = x)$  for  $t = 0, 1$ , and all  $x \in \mathcal{X}$  and  $m \in \mathcal{M}$ .

Thus, the treatment is first assumed to be ignorable given the pre-treatment covariates, and then the mediator variable is assumed to be ignorable *given* the observed value of the treatment as well as the pre-treatment covariates. We emphasize that unlike the standard sequential ignorability assumption in the literature (e.g., Robins, 1999), the conditional independence given in equation (5) of Assumption 1 must hold without conditioning on the observed values of post-treatment confounders. This issue is discussed further below.

The following theorem presents our main identification result, showing that under this assumption the ACME is nonparametrically identified.

THEOREM 1 (NONPARAMETRIC IDENTIFICATION) *Under Assumption 1, the ACME and the average natural direct effects are nonparametrically identified as follows for  $t = 0, 1$ ,*

$$\bar{\delta}(t) = \int \int \mathbb{E}(Y_i \mid M_i = m, T_i = t, X_i = x) \{dF_{M_i|T_i=1, X_i=x}(m) - dF_{M_i|T_i=0, X_i=x}(m)\} dF_{X_i}(x),$$

$$\bar{\zeta}(t) = \int \int \{\mathbb{E}(Y_i \mid M_i = m, T_i = 1, X_i = x) - \mathbb{E}(Y_i \mid M_i = m, T_i = 0, X_i = x)\} dF_{M_i|T_i=t, X_i=x}(m) dF_{X_i}(x).$$

where  $F_Z(\cdot)$  and  $F_{Z|W}(\cdot)$  represent the distribution function of a random variable  $Z$  and the conditional distribution function of  $Z$  given  $W$ .

A proof is given in Appendix A.1. Theorem 1 is quite general and can be easily extended to any types of treatment regimes, e.g., a continuous treatment variable. In fact, the proof requires no change except letting  $t$  and  $t'$  take values other than 0 and 1. Assumption 1 can also be somewhat relaxed by replacing equation (5) with its corresponding mean independence assumption. However, as mentioned above, this identification result does not hold under the standard sequential ignorability assumption. As shown by Avin *et al.* (2005) and also pointed out by Robins (2003),

the nonparametric identification of natural direct and indirect effects is not possible without an additional assumption if equation (5) holds only after conditioning on the post-treatment confounders  $Z_i$  as well as the pre-treatment covariates  $X_i$ , i.e.,  $Y_i(t', m) \perp\!\!\!\perp M_i \mid T_i = t, Z_i = z, X_i = x$ , for  $t, t' = 0, 1$ , and all  $x \in \mathcal{X}$  and  $z \in \mathcal{Z}$  where  $\mathcal{Z}$  is the support of  $Z_i$ . This is an important limitation since assuming the absence of post-treatment confounders may not be credible in many applied settings. In some cases, however, it is possible to address the main source of confounding by conditioning on pre-treatment variables alone (see Section 5 for an example).

### 2.3 Comparison with the Existing Results in the Literature

Next, we compare Theorem 1 with the related identification results in the literature. First, Pearl (2001, Theorem 1) makes the following set of assumptions in order to identify  $\bar{\delta}(t^*)$ ,

$$\mathbb{E}(Y(t, m) \mid X_i = x) \text{ and } \mathbb{E}(M_i(t^*) \mid X_i = x) \text{ are identifiable,} \quad (6)$$

$$Y_i(t, m) \perp\!\!\!\perp M_i(t^*) \mid X_i = x, \quad (7)$$

for all  $t = 0, 1$ ,  $m \in \mathcal{M}$ , and  $x \in \mathcal{X}$ . Under these assumptions, Pearl arrives at the same expressions for the ACME as the ones given in Theorem 1.

The direct comparison of Assumption 1 and Pearl's assumptions is difficult since there are many ways to achieve the identification condition given in equation (6). In particular, equation (4) implies equation (6), while the converse is not necessarily true. For example, equation (4) can be relaxed as follows while still implying equation (6),

$$Y_i(t, m) \perp\!\!\!\perp T_i \mid X_i = x, \text{ and } M_i(t) \perp\!\!\!\perp T_i \mid X_i = x, \quad (8)$$

for  $t = 0, 1$ , and all  $m \in \mathcal{M}$ . Equation (8) is slightly weaker than equation (4) because the former does not require the joint independence between  $\{Y_i(t', m), M_i(t)\}$  and  $T_i$  given  $X_i$ . Equation (4) also assumes that this joint independence must hold even when  $t \neq t'$ .

While equation (6) is implied by Assumption 1, equation (7) does not hold in general under Assumption 1 alone. Equation (7) requires conditional independence between the potential values of the outcome variable and the potential values of the mediating variable, whereas equation (5) of Assumption 1 is based on the conditional independence between the potential outcomes and the realized value of the mediator. In addition, equation (7) does not condition on the realized treatment status whereas equation (5) does.

To further facilitate the comparison, we consider a situation of practical importance where the treatment is randomized and researchers are interested in the identification of both  $\bar{\delta}(1)$  and  $\bar{\delta}(0)$ . In this case, Assumption 1 is weaker than Pearl's conditions. To see this, we show that the randomization of the treatment implies equation (4), which together with equation (7) implies equation (5). That is, for any  $t, t'$ , we have,  $p(Y_i(t', m) \mid M_i, T_i = t) = p(Y_i(t', m), M_i(t) \mid T_i = t) / p(M_i(t) \mid T_i = t) = p(Y_i(t', m), M_i(t)) / p(M_i(t)) = p(Y_i(t', m)) = p(Y_i(t', m) \mid T_i = t)$ , where the second and third equalities follow from equations (4) and (7), respectively. However, equations (4) and (5) do not imply  $Y_i(t, m) \perp\!\!\!\perp M_i(t') \mid X_i$  for  $t \neq t'$ . Section 3.3 provides such an example where equation (5) holds but equation (7) does not.

Second, Robins (2003) considers the identification under what he calls a FRCISTG model, which satisfies equation (4) as well as,

$$Y_i(t, m) \perp\!\!\!\perp M_i(t) \mid T_i = t, Z_i = z, X_i = x, \quad (9)$$

for  $t = 0, 1$  where  $Z_i$  is a vector of the observed values of post-treatment variables that confound the relationship between the mediator and outcome. The key difference between Assumption 1 and a FRCISTG model is that the latter allows conditioning on  $Z_i$  while the former does not. Robins (2003) argued that this is an important advantage over Pearl’s conditions, in that it makes the ignorability of the mediator more credible.

Under this model, Robins (2003, Theorem 2.1) shows that the following additional assumption is sufficient to identify the ACME,

$$Y_i(1, m) - Y_i(0, m) = B_i, \tag{10}$$

where  $B_i$  is a random variable independent of  $m$ . This assumption, called the no-interaction assumption, states that the controlled direct effect of the treatment does not depend on the value of the mediator. This result contrasts with Theorem 1, which shows that under the sequential ignorability assumption that does not condition on the post-treatment covariates, the no-interaction assumption is not required for the nonparametric identification.

Third, Petersen *et al.* (2006) presents yet another set of identifying assumptions, consisting of equations (8) and (5) as well as the following additional assumption,  $\mathbb{E}\{Y_i(1, m) - Y_i(0, m) \mid M_i(t^*) = m, X_i = x\} = \mathbb{E}\{Y_i(1, m) - Y_i(0, m) \mid X_i = x\}$  for all  $m \in \mathcal{M}$ . Theorem 1 shows that if equation (8) is replaced with equation (4), which is possible when the treatment is randomized, then this additional assumption is unnecessary for the nonparametric identification.

Finally, in the appendix of a recent working paper, Hafeman and VanderWeele (2008) show that if the mediator is binary, the ACME can be identified with a weaker set of assumptions than Assumption 1. However, it is unclear whether this result can be generalized to cases where the mediator is non-binary. In contrast, the identification result given in Theorem 1 holds for any type of mediator, whether discrete or continuous. Both identification results hold for general treatment regimes, unlike some of the previous results.

## 2.4 Implications for Linear Structural Equation Model

Next, we discuss the implications of Theorem 1 for LSEM, which is a popular tool among applied researchers who conduct causal mediation analysis. In an influential article, Baron and Kenny (1986) proposed a framework for mediation analysis which was later more rigorously developed by other researchers (e.g. MacKinnon and Dwyer, 1993; MacKinnon *et al.*, 1995). This framework is based on the following system of linear equations,

$$Y_i = \alpha_1 + \beta_1 T_i + \epsilon_{i1}, \tag{11}$$

$$M_i = \alpha_2 + \beta_2 T_i + \epsilon_{i2}, \tag{12}$$

$$Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \epsilon_{i3}. \tag{13}$$

Although we adhere to their original model, one may further condition on any observed pre-treatment covariates by including them as additional regressors in each equation. This will change none of the results given below so long as the model includes no post-treatment confounders.

Under this model, Baron and Kenny (1986) suggested that the existence of mediation effects can be tested by separately fitting the three linear regressions and testing the null hypotheses (1)  $\beta_1 = 0$ , (2)  $\beta_2 = 0$ , and (3)  $\gamma = 0$ . If all of these null hypotheses are rejected, they argued, then  $\beta_2 \gamma$  could be interpreted as the mediation effect.

We note that equation (11) is redundant given equations (12) and (13). To see this, substitute equation (12) into equation (13) to obtain,

$$Y_i = (\alpha_3 + \alpha_2\gamma) + (\beta_3 + \beta_2\gamma)T_i + (\gamma\epsilon_{i2} + \epsilon_{i3}). \quad (14)$$

Thus, as pointed out by other researchers (e.g. MacKinnon, 2008, Section 3.13), testing  $\beta_1 = 0$  is unnecessary since the ACME can be non-zero even when the average total causal effect is zero. This happens when the mediation effect offsets the direct effect of the treatment.

The next theorem proves that within the LSEM framework, Baron and Kenny’s interpretation is valid if the Assumption 1 holds.

**THEOREM 2 (IDENTIFICATION UNDER THE LSEM)** *Consider the LSEM defined in equations (11), (12), and (13). Under Assumption 1, the ACME is identified and given by,  $\bar{\delta}(0) = \bar{\delta}(1) = \beta_2\gamma$ , where the equality between  $\bar{\delta}(0)$  and  $\bar{\delta}(1)$  is also assumed.*

A proof is in Appendix A.2. The theorem implies that under the same set of assumptions, the average natural direct effects are identified as  $\bar{\zeta}(0) = \bar{\zeta}(1) = \beta_3$  where the average total causal effect is  $\bar{\tau} = \beta_3 + \beta_2\gamma$ . Thus, Assumption 1 enables the identification of the ACME under the LSEM. Eggleston *et al.* (2006) obtain a similar result under the assumptions of Pearl (2001) and Robins (2003), which were reviewed in Section 2.3.

It is important to note that under Assumption 1, the standard LSEM defined in equations (12) and (13) makes the following no-interaction assumption about the ACME,

**ASSUMPTION 2 (NO-INTERACTION BETWEEN THE TREATMENT AND THE ACME)**

$$\bar{\delta}(1) = \bar{\delta}(0).$$

This assumption is equivalent to the no-interaction assumption for the average natural direct effects,  $\bar{\zeta}(1) = \bar{\zeta}(0)$ . Although Assumption 2 is related to and implied by Robins’ no-interaction assumption given in equation (10), the key difference is that Assumption 2 is written in terms of the ACME rather than *controlled* direct effects.

As Theorem 1 suggests, Assumption 2 is not required for the identification of the ACME under the LSEM. We extend the outcome model given in equation (13) to,

$$Y_i = \alpha_3 + \beta_3T_i + \gamma M_i + \kappa T_i M_i + \epsilon_{i3}, \quad (15)$$

where the interaction term between the treatment and mediating variables is added to the outcome regression while maintaining the linearity in parameters. This formulation was first suggested by Judd and Kenny (1981) and more recently advocated by Kraemer *et al.* (2002, 2008) as an alternative to Barron and Kenny’s approach. Under Assumption 1 and the model defined by equations (12) and (15), we can identify the ACME as  $\bar{\delta}(t) = \beta_2(\gamma + t\kappa)$  for  $t = 0, 1$ . The average natural direct effects are identified as  $\bar{\zeta}(t) = \beta_3 + \kappa(\alpha_2 + \beta_2t)$ , and the average total causal effect is equal to  $\bar{\tau} = \beta_2\gamma + \beta_3 + \kappa(\alpha_2 + \beta_2)$ . This contrasts with the proposal by Kraemer *et al.* (2008) that the existence of mediation effects can be established by testing either  $\gamma = 0$  or  $\kappa = 0$ .

The connection between the parametric and nonparametric identification becomes clearer when both  $T_i$  and  $M_i$  are binary. To see this, note that  $\bar{\delta}(t)$  can be equivalently expressed as (dropping the integration over  $P(X_i)$  for notational simplicity),  $\bar{\delta}(t) = \sum_{m=0}^{J-1} \mathbb{E}(Y_i \mid M_i = m, T_i =$

$t, X_i) \{\Pr(M_i = m \mid T_i = 1, X_i) - \Pr(M_i = m \mid T_i = 0, X_i)\}$ , when  $M_i$  is discrete. Furthermore, when  $J = 2$ , this reduces to,  $\bar{\delta}(t) = \{\Pr(M_i = 1 \mid T_i = 1, X_i) - \Pr(M_i = 1 \mid T_i = 0, X_i)\} \{\mathbb{E}(Y_i \mid M_i = 1, T_i = t, X_i) - \mathbb{E}(Y_i \mid M_i = 0, T_i = t, X_i)\}$ . Thus, the ACME equals the product of two terms representing the average effect of  $T_i$  on  $M_i$  and that of  $M_i$  on  $Y_i$  (holding  $T_i$  at  $t$ ), respectively.

Finally, in the existing methodological literature, Sobel (2008) explores the identification problem of mediation effects under the framework of LSEM without assuming the ignorability of the mediator (see also Albert, 2008; Jo, 2008). However, Sobel (2008) maintains, among others, the assumption that the causal effect of the treatment is entirely through the mediator and applies the instrumental variables technique of Angrist *et al.* (1996). That is, the natural direct effect is assumed to be zero for all units *a priori*, i.e.,  $\zeta_i(t) = 0$  for all  $t = 0, 1$  and  $i$ . This assumption may be undesirable from the perspective of applied researchers, because the existence of the natural direct effect itself is often of interest in causal mediation analysis. See Joffe *et al.* (2008) for an interesting application.

### 3 Estimation and Inference

#### 3.1 Parametric Estimation and Inference

Under the LSEM given by equations (12) and (13) and Assumption 1, the estimation of the ACME is straightforward since the error terms are independent of each other. Thus, one can follow the proposal of Baron and Kenny (1986) and estimate equations (12) and (13) by fitting two separate linear regressions. The standard error for the estimated ACME, i.e.,  $\hat{\delta}(t) = \hat{\beta}_2 \hat{\gamma}$ , can be calculated either using the Delta method (Sobel, 1982), i.e.,  $\text{Var}(\hat{\delta}(t)) \approx \beta_2^2 \text{Var}(\hat{\gamma}) + \gamma^2 \text{Var}(\hat{\beta}_2)$ , or the exact variance formula (Goodman, 1960), i.e.,  $\text{Var}(\hat{\delta}(t)) = \beta_2^2 \text{Var}(\hat{\gamma}) + \gamma^2 \text{Var}(\hat{\beta}_2) + \text{Var}(\hat{\gamma}) \text{Var}(\hat{\beta}_2)$ . For the natural direct and total effects, standard errors can be obtained via the regressions of  $Y_i$  on  $T_i$  and  $M_i$  (equation 13) and  $Y_i$  on  $T_i$  (equation 11), respectively.

When the model contains the interaction term as in equation (15) (so that Assumption 2 is relaxed), the asymptotic variance can be computed in a similar manner. For example, using the delta method, we have  $\text{Var}(\hat{\delta}(t)) \approx (\gamma + t\kappa)^2 \text{Var}(\hat{\beta}_2) + \beta_2^2 \{\text{Var}(\hat{\gamma}) + t \text{Var}(\hat{\kappa}) + 2t \text{Cov}(\hat{\gamma}, \hat{\kappa})\}$  for  $t = 0, 1$ . Similarly,  $\text{Var}(\hat{\zeta}(t)) \approx \text{Var}(\hat{\beta}_3) + (\alpha_2 + t\beta_2)^2 \text{Var}(\hat{\kappa}) + 2(\alpha_2 + t\beta_2) \text{Cov}(\hat{\beta}_3, \hat{\kappa}) + \kappa^2 \{\text{Var}(\hat{\alpha}_2) + t \text{Var}(\hat{\beta}_2) + 2t \text{Cov}(\hat{\alpha}_2, \hat{\beta}_2)\}$ . (See the online appendix for a detailed derivation.) For the average total causal effect, the variance can be obtained from the regression of  $Y_i$  on  $T_i$ .

#### 3.2 Nonparametric Estimation and Inference

Next, we consider a simple nonparametric estimator. Suppose that the mediator is discrete and takes  $J$  distinct values, i.e.,  $\mathcal{M} = \{0, 1, \dots, J-1\}$ . The case of continuous mediators is considered further below. First, we consider the cases where we estimate the ACME separately within each strata defined by the pre-treatment covariates  $X_i$ . One may then aggregate the resulting stratum-specific estimates to obtain the estimated ACME. In such situations, a nonparametric estimator can be obtained by plugging in sample analogues for the population quantities in the expression given in Theorem 1,

$$\hat{\delta}(t) = \sum_{m=0}^{J-1} \left\{ \frac{\sum_{i=1}^n Y_i \mathbf{1}\{T_i = t, M_i = m\}}{\sum_{i=1}^n \mathbf{1}\{T_i = t, M_i = m\}} \left( \frac{1}{n_1} \sum_{i=1}^n \mathbf{1}\{T_i = 1, M_i = m\} - \frac{1}{n_0} \sum_{i=1}^n \mathbf{1}\{T_i = 0, M_i = m\} \right) \right\}, \quad (16)$$

where  $n_t = \sum_{i=1}^n \mathbf{1}\{T_i = t\}$  and  $t = 0, 1$ . By law of large numbers, this estimator asymptotically converges to the true ACME under Assumption 1. The next theorem derives the asymptotic

variance of the nonparametric estimator defined in equation (16) given the realized values of the treatment variable.

**THEOREM 3 (ASYMPTOTIC VARIANCE OF THE NONPARAMETRIC ESTIMATOR)** *Suppose that Assumption 1 holds. Then, the variance of the nonparametric estimator defined in equation (16) is asymptotically approximated by,*

$$\begin{aligned} \text{Var}(\hat{\delta}(t)) \approx & \frac{1}{n_t} \sum_{m=0}^{J-1} \nu_{1-t,m} \left\{ \left( \frac{\nu_{1-t,m}}{\nu_{tm}} - 2 \right) \text{Var}(Y_i \mid M_i = m, T_i = t) + \frac{n_t(1 - \nu_{1-t,m})\mu_{tm}^2}{n_{1-t}} \right\} \\ & - \frac{2}{n_{1-t}} \sum_{m'=m+1}^{J-1} \sum_{m=0}^{J-2} \nu_{1-t,m} \nu_{1-t,m'} \mu_{tm} \mu_{tm'} + \frac{1}{n_t} \text{Var}(Y_i \mid T_i = t), \end{aligned}$$

for  $t = 0, 1$  where  $\nu_{tm} \equiv \Pr(M_i = m \mid T_i = t)$  and  $\mu_{tm} \equiv \mathbb{E}(Y_i \mid M_i = m, T_i = t)$ .

A proof is based on a tedious but simple application of the Delta method and thus is given in the online appendix. This asymptotic variance can be consistently estimated by replacing unknown population quantities with their corresponding sample counterparts. The estimated overall variance can be obtained by aggregating the estimated within-strata variances according to the sample size in each strata.

The second and perhaps more general strategy is to use nonparametric regressions to model  $\mu_{tm}(x) \equiv \mathbb{E}(Y_i \mid T_i = t, M_i = m, X_i = x)$  and  $\nu_{tm}(x) \equiv \Pr(M_i = m \mid T_i = t, X_i = x)$ , and then employ the following estimator,

$$\hat{\delta}(t) = \frac{1}{n} \left\{ \sum_{i=1}^n \sum_{m=0}^{J-1} \hat{\mu}_{tm}(X_i) (\hat{\nu}_{1m}(X_i) - \hat{\nu}_{0m}(X_i)) \right\}, \quad (17)$$

for  $t = 0, 1$ . This estimator is also asymptotically consistent for the ACME under Assumption 1 if  $\hat{\mu}_{tm}(x)$  and  $\hat{\nu}_{tm}(x)$  are consistent for  $\mu_{tm}(x)$  and  $\nu_{tm}(x)$ , respectively. Unfortunately, in general, there is no simple expression for the asymptotic variance of this estimator. Thus, one may use a nonparametric bootstrap (or a parametric bootstrap based on the asymptotic distribution of  $\hat{\mu}_{tm}(x)$  and  $\hat{\nu}_{tm}(x)$ ) to compute uncertainty estimates.

Finally, when the mediator is not discrete, we may nonparametrically model  $\mu_{tm}(x) \equiv \mathbb{E}(Y_i \mid T_i = t, M_i = m, X_i = x)$  and  $\psi_t(x) = p(M_i \mid T_i = t, X_i = x)$ . Then, one can use the following estimator,  $\hat{\delta}(t) = \frac{1}{nK} \sum_{i=1}^n \sum_{k=1}^K \left\{ \hat{\mu}_{t\tilde{m}_{1i}^{(k)}}(X_i) - \hat{\mu}_{t\tilde{m}_{0i}^{(k)}}(X_i) \right\}$ , where  $\tilde{m}_{ti}^{(k)}$  is the  $k$ th Monte Carlo draw of the mediator  $M_i$  from its predicted distribution based on the fitted model  $\hat{\psi}_t(X_i)$ .

### 3.3 A Simulation Study

Next, we conduct a small-scale Monte Carlo experiment in order to investigate the finite-sample performance of the nonparametric estimator defined in equation (16) as well as the proposed estimator of its asymptotic variance given in Theorem 3. We use a population model where the potential outcomes and mediators are given by  $Y_i(t, m) = \exp(Y_i^*(t, m))$ ,  $M_i(t) = \mathbf{1}\{M_i^*(t) \geq 0.5\}$  and  $Y_i^*(t, m)$ ,  $M_i^*(t)$  are jointly normally distributed. The population parameters are set to the following values:  $\mathbb{E}(Y_i^*(1, 1)) = 2$ ;  $\mathbb{E}(Y_i^*(1, 0)) = 0$ ;  $\mathbb{E}(Y_i^*(0, 1)) = 1$ ;  $\mathbb{E}(Y_i^*(0, 0)) = 0.5$ ;  $\mathbb{E}(M_i^*(1)) = 1$ ;  $\mathbb{E}(M_i^*(0)) = 0$ ;  $\text{Var}(Y_i^*(t, m)) = \text{Var}(M_i^*(t)) = 1$  for  $t \in \{0, 1\}$  and  $m \in \{0, 1\}$ ;

Estimator	Sample Size	Bias	RMSE	90% CI Coverage	95% CI Coverage
$\hat{\delta}(0)$	100	0.014	0.69	0.83	0.87
	500	0.013	0.29	0.88	0.93
	1000	0.013	0.20	0.89	0.94
	2000	0.016	0.14	0.90	0.95
$\hat{\delta}(1)$	100	0.080	1.46	0.87	0.92
	500	0.080	0.65	0.90	0.95
	1000	0.079	0.46	0.90	0.95
	2000	0.094	0.34	0.90	0.95

Table 1: Finite-Sample Performance of the Nonparametric Estimator and its Variance Estimator. The table presents the results of a Monte Carlo experiment with varying sample sizes and one million iterations. The upper half of the table represents the results for  $\hat{\delta}(0)$  and the bottom half  $\hat{\delta}(1)$ . The columns represent (from left to right): sample sizes, estimated biases of  $\hat{\delta}(t)$ , estimated root mean squared errors (RMSE), and the estimated coverage probabilities of the 90% and 95% confidence intervals. The true values of  $\bar{\delta}(0)$  and  $\bar{\delta}(1)$  are approximately 0.67 and 3.95, respectively. The results indicate that the point estimates are approximately unbiased even with small sample sizes and that the confidence intervals based on the estimated variances of  $\hat{\delta}(0)$  and  $\hat{\delta}(1)$  converge to their nominal coverage values when the sample size reaches about 2,000 and 500, respectively.

$\text{Corr}(Y_i^*(t, m), Y_i^*(t', m')) = 0.5$  for  $t, t' \in \{0, 1\}$  and  $m, m' \in \{0, 1\}$ ;  $\text{Corr}(Y_i^*(t, m), M_i^*(t'))$  equals 0.3 if  $T_i = 1 - t'$  and 0 otherwise for  $t \in \{0, 1\}$  and  $m \in \{0, 1\}$ ; and  $\text{Corr}(M_i^*(1), M_i^*(0)) = 0.3$ .

Under this setup, Assumption 1 is satisfied. In particular, for  $t, t' = 0, 1$ ,  $Y_i(t, m)$  is independent of  $M_i(t')$  when  $T_i = t'$  but they are correlated when  $T_i = 1 - t'$  so that one of Pearl’s assumptions, equation (7), is violated. We set the sample size  $n$  to 100, 500, 1000, and 2000 where half of the sample receives the treatment and the other half is assigned to the control group, i.e.,  $n_1 = n_0 = n/2$ . Through Monte Carlo approximation, we find that in this experiment the true ACMEs are given by  $\bar{\delta}(0) \approx 0.67$  and  $\bar{\delta}(1) \approx 3.95$ .

Table 1 shows the results of the experiments based on one million iterations. The performance of the point estimates turns out to be quite good in this particular setting. Even with sample size as small as 100, estimated biases are essentially zero for both  $\hat{\delta}(0)$  and  $\hat{\delta}(1)$ . In fact, for  $\hat{\delta}(0)$ , the ratio of the bias to root mean squared error stays well below 10% with a sample size of 1,000 and it only becomes about 11.4% when the sample size becomes as large as 2,000. For  $\hat{\delta}(1)$ , the ratio is about 5.5% when the sample size is 100 and increases to about 27.6% when the sample size is 2,000. The variance estimator also performs well. The estimated variances of  $\hat{\delta}(0)$  and  $\hat{\delta}(1)$  converge to their true values with the sample sizes of 2,000 and 500, respectively, as demonstrated by the coverage probabilities of the corresponding 90% and 95% confidence intervals.

## 4 Sensitivity Analysis

Although the ACME is nonparametrically identified under Assumption 1, this assumption, like other existing identifying assumptions, may be too strong in many applied settings. Consider randomized experiments where the treatment is randomized but the mediator is not. Causal mediation analysis is most frequently applied to such experiments. In this case, equation (4) Assumption 1 is satisfied but equation (5) may not hold for two reasons. First, there may exist unmeasured pre-treatment covariates that confound the relationship between the mediator and

the outcome. Second, there may exist observed or unobserved post-treatment confounders. These possibilities, along with other obstacles encountered in applied research, have led some scholars to warn against the abuse of mediation analyses (Green *et al.*, 2010).

In this section, we develop a method to assess the sensitivity of an estimated ACME to unmeasured pre-treatment confounding. The proposed sensitivity analysis, however, does not address the possible existence of post-treatment confounders. The method is based on the standard LSEM framework described in Section 2.4 and can be easily used by applied researchers to examine the robustness of their empirical findings. We derive the maximum departure from equation (5) that is allowed while maintaining their original conclusion about the direction of the ACME (see Imai and Yamamoto, 2008). For notational simplicity, we do not explicitly condition on the pre-treatment covariates  $X_i$ . However, the same analysis can be conducted by including them as additional covariates in each regression.

The proof of Theorem 2 implies that if equation (4) holds,  $\epsilon_{i2} \perp\!\!\!\perp T_i$  and  $\epsilon_{i3} \perp\!\!\!\perp T_i$  hold but  $\epsilon_{i2} \not\perp\!\!\!\perp \epsilon_{i3}$  does not unless equation (5) also holds. Thus, one way to assess the sensitivity of one's conclusions to the violation of equation (5) is to use the following sensitivity parameter,

$$\rho \equiv \text{Corr}(\epsilon_{i2}, \epsilon_{i3}), \quad (18)$$

where  $-1 < \rho < 1$ . In Appendix A.3, we show that Assumption 1 implies  $\rho = 0$ . (Of course, the contrapositive of this statement is also true;  $\rho \neq 0$  implies the violation of Assumption 1). A non-zero correlation parameter can be interpreted as the existence of omitted variables that are related to both the observed value of the mediator  $M_i$  and the potential outcomes  $Y_i$  even after conditioning on the treatment variable  $T_i$  (and the observed covariates  $X_i$ ). Note that these omitted variables must causally precede  $T_i$ . Then, we vary the value of  $\rho$  and compute the corresponding estimate of the ACME.

The next theorem shows that if the treatment is randomized, the ACME is identified given a particular value of  $\rho$ .

**THEOREM 4 (IDENTIFICATION WITH A GIVEN ERROR CORRELATION)** *Consider the LSEM defined in equations (11), (12), and (13). Suppose that equation (4) holds and the correlation between  $\epsilon_{i2}$  and  $\epsilon_{i3}$ , i.e.,  $\rho$ , is given. If we further assume  $-1 < \rho < 1$ , then the ACME is identified and given by,*

$$\bar{\delta}(0) = \bar{\delta}(1) = \frac{\beta_2 \sigma_1}{\sigma_2} \left\{ \tilde{\rho} - \rho \sqrt{(1 - \tilde{\rho}^2)/(1 - \rho^2)} \right\},$$

where  $\sigma_j^2 \equiv \text{Var}(\epsilon_{ij})$  for  $j = 1, 2$  and  $\tilde{\rho} \equiv \text{Corr}(\epsilon_{i1}, \epsilon_{i2})$ .

A proof is in Appendix A.4. We offer several remarks about Theorem 4. First, the unbiased estimates of  $(\alpha_1, \alpha_2, \beta_1, \beta_2)$  can be obtained by fitting the equation-by-equation least squares of equations (11) and (12). Given these estimates, the covariance matrix of  $(\epsilon_{i1}, \epsilon_{i2})$ , whose elements are  $(\sigma_1^2, \sigma_2^2, \tilde{\rho}\sigma_1\sigma_2)$ , can be consistently estimated by computing the sample covariance matrix of the residuals, i.e.,  $\hat{\epsilon}_{i1} = Y_i - \hat{\alpha}_1 - \hat{\beta}_1 T_i$  and  $\hat{\epsilon}_{i2} = M_i - \hat{\alpha}_2 - \hat{\beta}_2 T_i$ .

Second, the partial derivative of the ACME with respect to  $\rho$  is given by,  $\frac{\partial}{\partial \rho} \bar{\delta}(t) = -\frac{\beta_2 \sigma_1}{\sigma_2(1-\rho^2)} \sqrt{(1-\tilde{\rho}^2)/(1-\rho^2)}$  for  $t = 0, 1$ . This implies that the ACME is either monotonically increasing or decreasing in  $\rho$ , depending on the sign of  $\beta_2$ . The ACME is also symmetric about  $(\rho, \bar{\delta}(t)) = (0, \beta_2 \tilde{\rho} \sigma_1 / \sigma_2)$ .

Third, the ACME is zero if and only if  $\rho$  equals  $\tilde{\rho}$ . This implies that researchers can easily check the robustness of their conclusion obtained under the sequential ignorability assumption via correlation between  $\epsilon_{i1}$  and  $\epsilon_{i2}$ . For example, if  $\hat{\delta}(t) = \hat{\beta}_2 \hat{\gamma}$  is negative, the true ACME is also guaranteed to be negative if  $\rho < \tilde{\rho}$  holds.

Finally, the expression of the ACME given in Theorem 4 is cumbersome to use when computing the standard errors. A more straightforward and general approach is to apply the iterative feasible generalized least square algorithm of the seemingly unrelated regression (Zellner, 1962), and use the associated asymptotic variance formula. This strategy will also work when there is an interaction term between the treatment and mediating variables as in equation (15) and/or when there are observed pre-treatment covariates  $X_i$ .

The sensitivity parameter  $\rho$  can be given an alternative definition which allows it to be interpreted as the magnitude of an unobserved confounder. This alternative version of  $\rho$  is based on the following decomposition of the error terms in equations (12) and (13),

$$\epsilon_{ij} = \lambda_j U_i + \epsilon'_{ij},$$

for  $j = 2, 3$ , where  $U_i$  is an unobserved confounder and the sequential ignorability is assumed given  $U_i$  and  $T_i$ . Again, note that  $U_i$  has to be a pre-treatment variable so that the resulting estimates can be given a causal interpretation. In addition, we assume that  $\epsilon'_{ij} \perp U_i$  for  $j = 2, 3$ . We can then express the influence of the unobserved pre-treatment confounder using the following coefficients of determination,

$$R_M^{2*} \equiv 1 - \frac{\text{Var}(\epsilon'_{i2})}{\text{Var}(\epsilon_{i2})} \quad \text{and} \quad R_Y^{2*} \equiv 1 - \frac{\text{Var}(\epsilon'_{i3})}{\text{Var}(\epsilon_{i3})},$$

which represent the proportion of previously unexplained variance (either in the mediator or in the outcome) that is explained by the unobserved confounder (see Imbens, 2003). Another interpretation is based on the proportion of original variance that is explained by the unobserved confounder. In this case, we use the following sensitivity parameters,

$$\tilde{R}_M^2 \equiv \frac{\text{Var}(\epsilon_{i2}) - \text{Var}(\epsilon'_{i2})}{\text{Var}(M_i)} = (1 - R_M^2) R_M^{2*} \quad \text{and} \quad \tilde{R}_Y^2 \equiv \frac{\text{Var}(\epsilon_{i3}) - \text{Var}(\epsilon'_{i3})}{\text{Var}(Y_i)} = (1 - R_Y^2) R_Y^{2*},$$

where  $R_M^2$  and  $R_Y^2$  represent the coefficients of determination from the two regressions given in equations (12) and (13).

In either case, it is straightforward to show that the following relationship between  $\rho$  and these parameters holds, i.e.,  $\rho^2 = R_M^{2*} R_Y^{2*} = \tilde{R}_M^2 \tilde{R}_Y^2 / \{(1 - R_M^2)(1 - R_Y^2)\}$  or equivalently,

$$\rho = \text{sgn}(\lambda_2 \lambda_3) R_M^* R_Y^* = \frac{\text{sgn}(\lambda_2 \lambda_3) \tilde{R}_M \tilde{R}_Y}{\sqrt{(1 - R_M^2)(1 - R_Y^2)}},$$

where  $R_M^*$ ,  $R_Y^*$ ,  $\tilde{R}_M$  and  $\tilde{R}_Y$  are in  $[0, 1]$ . Thus, in this framework, researchers can specify the values of  $(R_M^{2*}, R_Y^{2*})$  or  $(\tilde{R}_M^2, \tilde{R}_Y^2)$  as well as the sign of  $\lambda_2 \lambda_3$  in order to determine values of  $\rho$  and estimate the ACME based on these values of  $\rho$ . Then, the analyst can examine variation in the estimated ACME with respect to change in these parameters.

## 5 An Empirical Application

In this section, we apply our proposed methods to an influential randomized experiment from political psychology.

### 5.1 Data

Nelson *et al.* (1997) examine how the framing of political issues in news media affects citizens' political opinions. While the authors are not the first to use causal mediation analysis in political science, their study is one of the most well-known examples in political psychology and also represents a typical application of causal mediation analyses in the social sciences. Media framing is the process by which news organizations define a political issue or emphasize its particular aspects. The authors hypothesize that differing frames for the same news story alter citizens' political tolerance by affecting more general political attitudes. They conducted a randomized experiment to test this mediation hypothesis.

Specifically, Nelson *et al.* (1997) used two different local news stories about a Ku Klux Klan rally held in central Ohio. In the experiment, student subjects were randomly assigned to watch two different segments of the local news. The two news clips were identical except for the final story on the Klan rally. In one newscast, the Klan rally was presented as a free speech issue. In a second newscast, the journalists presented the Klan rally as a disruption of public order that threatened to turn violent. The sample size is 136 with 67 subjects exposed to the free speech frame and 69 subjects assigned to the public order frame.

The outcome was measured using two different scales of political tolerance. Immediately after viewing the news broadcast, subjects were asked two seven-point scale questions measuring their tolerance for the Klan speeches and rallies. The hypothesis was that the causal effects of the media frame on tolerance are mediated by subjects' attitudes about the importance of the right to free speech and the maintenance of public order. The researchers used additional survey questions to measure these hypothesized mediating factors.

It is important to note that the researchers in this example are primarily interested in the mediating mechanism between media framing and political tolerance, not the causal effects of the hypothesized mediators *per se*. Indeed, in many social science experiments, researchers' interest lies in the identification of causal mediation effects rather than controlled direct effects. Causal mediation analysis is particularly appealing in such situations.

### 5.2 Analysis under Sequential Ignorability

In the original analysis, Nelson *et al.* (1997) used a LSEM similar to the one discussed in Section 2.4 and found that subjects who viewed the Klan story with the free speech frame were significantly more tolerant of the Klan than those who saw the story with the public order frame. The researchers also found evidence supporting their main hypothesis that subjects' general attitudes mediated the causal effect of the news story frame on tolerance for the Klan. In the analysis that follows, we only analyze the public order mediator, of which the researchers found a significant mediation effect.

As we showed in Section 2.4, the original results can be given a causal interpretation under sequential ignorability, *i.e.*, Assumption 1. Here, we first make this assumption and estimate causal effects based on our theoretical results. Table 2 presents the findings. The second and third columns of the table show the estimated ACME and average total effect based on the LSEM and the

	Parametric	Nonparametric
<i>Average Mediation Effects</i>		
Free speech frame $\hat{\delta}(0)$	-0.451 [-0.871, -0.031]	-0.374 [-0.823, 0.074]
Public order frame $\hat{\delta}(1)$	-0.566 [-1.081, -0.050]	-0.596 [-1.168, -0.024]
Average Total Effect $\hat{\tau}$	-0.540 [-1.207, 0.127]	-0.627 [-1.153, -0.099]
<i>With the no-interaction assumption</i>		
Average Mediation Effect	-0.510	
$\hat{\delta}(0) = \hat{\delta}(1)$	[-0.969, -0.051]	
Average Total Effect $\hat{\tau}$	-0.540 [-1.206, 0.126]	

Table 2: Parametric and Nonparametric Estimates of the ACME under Sequential Ignorability in the Media Framing Experiment. Each cell of the table represents an estimated average causal effect and its 95% confidence interval. The outcome is the subjects’ tolerance level for the free speech rights of the Ku Klux Klan, and the treatments are the public order frame ( $T_i = 1$ ) and the free speech frame ( $T_i = 0$ ). The second column of the table shows the results of the parametric LSEM approach, while the third column of the table presents those of the nonparametric estimator. The lower part of the table shows the results of parametric mediation analysis under the no-interaction assumption ( $\hat{\delta}(1) = \hat{\delta}(0)$ ), while the upper part presents the findings without this assumption thereby showing the estimated average mediation effects under the treatment and the control, i.e.  $\hat{\delta}(1)$  and  $\hat{\delta}(0)$ .

nonparametric estimator, respectively. The 95% asymptotic confidence intervals are constructed using the Delta method. For most of the estimates, the 95% confidence intervals do not contain zero, mirroring the finding from the original study that general attitudes about public order mediated the effect of the media frame.

As shown in Section 2.4, we can relax the no-interaction assumption (Assumption 2) that is implicit in the LSEM of Baron and Kenny (1986). The first and second rows of the table present estimates from the parametric and nonparametric analysis without this assumption. These results show that the estimated ACME under the public order condition ( $\hat{\delta}(1)$ ) is larger than the effect under the free speech condition ( $\hat{\delta}(0)$ ) for both the parametric and nonparametric estimators. In fact, the 95% confidence interval for the nonparametric estimate of  $\bar{\delta}(0)$  includes zero. However, we fail to reject the null hypothesis of  $\bar{\delta}(0) = \bar{\delta}(1)$  under the parametric analysis, with the  $p$ -value of 0.238.

Based on this finding, the no-interaction assumption could be regarded as appropriate. The last two rows in Table 2 contain the analysis based on the parametric estimator under this assumption. As expected, the estimated ACME is between the previous two estimates, and the 95% confidence interval does not contain zero. Finally, the estimated average total effect is identical to that without Assumption 2. This makes sense since the no-interaction assumption only restricts the way the treatment effect is transmitted to outcome and thus does not affect the estimate of the overall

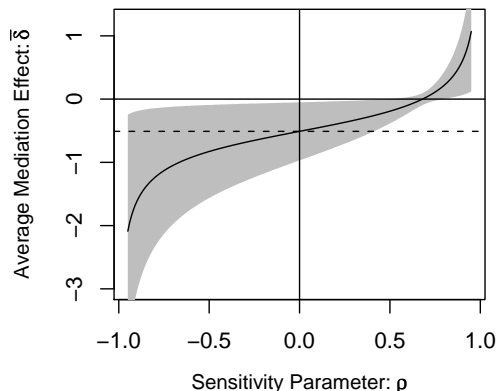


Figure 1: Sensitivity Analysis for the Media Framing Experiment. The figure presents the results of the sensitivity analysis described in Section 4. The solid line represents the estimated ACME for the attitude mediator for differing values of the sensitivity parameter  $\rho$ , which is defined in equation (18). The gray region represents the 95% confidence interval based on the Delta method. The horizontal dashed line is drawn at the point estimate of  $\bar{\delta}$  under Assumption 1.

treatment effect.

### 5.3 Sensitivity Analysis

The estimates in Section 5.2 are identified if the sequential ignorability assumption holds. However, since the original researchers were only able to randomize news stories but not subjects' attitudes, this assumption is rather unlikely to hold. For example, the assumption will be violated if subjects' underlying ideology affects both their public order attitude and their tolerance for the Klan rally within each treatment condition. This scenario is of particular concern since it is well established that politically conservative people tend to be more concerned about public order issues and also be more sympathetic to political groups like the Klan. Thus, we next ask how sensitive these estimates are to violations of this assumption using the methods proposed in Section 4. We consider political ideologies a pre-treatment confounder since in psychological literature they are usually thought to cause attitudes toward specific issues but not vice versa. We also maintain Assumption 2.

Figure 1 presents the results for the sensitivity analysis. We plot the estimated ACME of the attitude mediator against differing values of the sensitivity parameter  $\rho$ , which is equal to the correlation between the two error terms of equations (22) and (23) for each. The analysis indicates that the original conclusion about the direction of the ACME under Assumption 1 (represented by the dashed horizontal line) would be maintained unless  $\rho$  is greater than 0.68. This implies that the conclusion is plausible given even fairly large departures from the ignorability of the mediator. This result holds even after we take into account the sampling variability, as the confidence interval covers the value of zero only when  $0.50 < \rho < 0.78$ . Thus, the original finding about the negative ACME is relatively robust to the violation of equation (5) of Assumption 1 under the LSEM.

Next, we present the same sensitivity analysis using the alternative interpretation of  $\rho$  which is based on two coefficients of determination as defined in Section 4; (1) the proportion of unexplained variance that is explained by an unobserved pre-treatment confounder ( $R_M^{2*}$  and  $R_Y^{2*}$ ) and (2) the

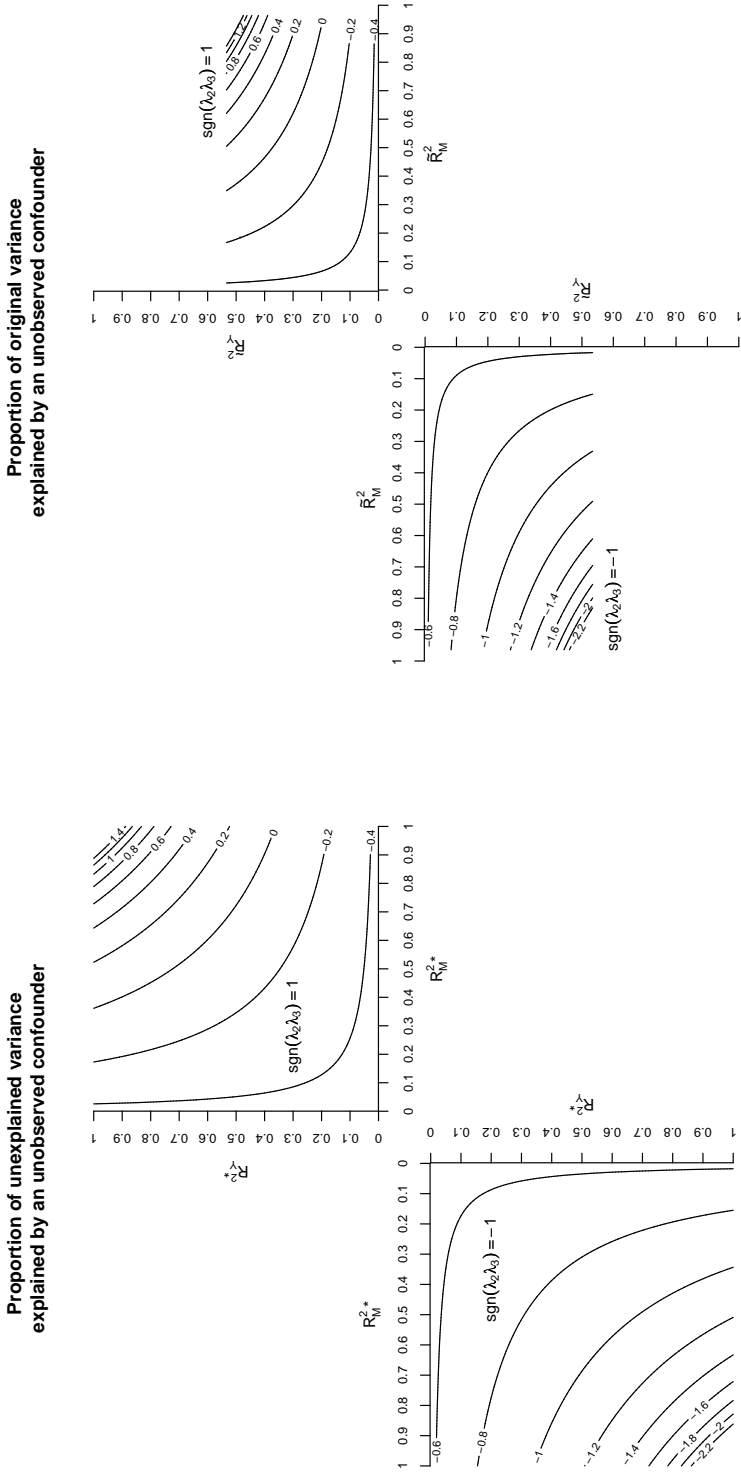


Figure 2: An Alternative Interpretation of the Sensitivity Analysis. The plot presents the results of the sensitivity analysis described in Section 4. Each plot contains various mediation effects under an unobserved pre-treatment confounder of various magnitudes. The left plot contains the contours for  $R_M^{2*}$  and  $R_Y^{2*}$  which represent the proportion of unexplained variance that is explained by the unobserved confounder for the mediator and outcome, respectively. The right plot contains the contours for  $\tilde{R}_M^2$  and  $\tilde{R}_Y^2$  which represent the proportion of the variance explained by the unobserved pre-treatment confounder. Each line represents the estimated ACME under proposed values of either  $(R_M^{2*}, R_Y^{2*})$  or  $(\tilde{R}_M^2, \tilde{R}_Y^2)$ . The term  $\text{sgn}(\lambda_2\lambda_3)$  represents the sign on the product of the coefficients of the unobserved confounder.

proportion of the original variance explained by the same unobserved confounder ( $\tilde{R}_M^2$  and  $\tilde{R}_Y^2$ ). Figure 2 shows two plots based on the types of coefficients of determination. The lower left quadrant of each plot in the figure represents the case where the product of the coefficients for the unobserved confounder is negative, while upper right quadrant represents the case where the product is positive.

For example, this product will be positive if the unobserved pre-treatment confounder represents subjects' political ideology, since conservatism is likely to be positively correlated with both public order importance and tolerance for the Klan. Under this scenario, the ACME is still guaranteed to be negative as long as the unobserved confounder explains less than 36% of the variance in the mediator or outcome that is left unexplained by the treatment alone, no matter how large the corresponding portion of the variance in the other variable may be. Similarly, the direction of the original estimate is maintained if the unobserved confounder explains less than 34.7% (19.2%) of the original variance in the mediator (outcome), regardless of the degree of confounding for the outcome (mediator). Finally, the original conclusion is perfectly robust to the violation of sequential ignorability if the product of the coefficients for the unobserved confounder is negative, since the estimated ACME is always negative in the lower left quadrant of each plot.

## 6 Concluding Remarks

In this paper, we study identification, inference, and sensitivity analysis for causal mediation effects. Causal mediation analysis is routinely conducted in various disciplines, and our paper contributes to this fast-growing methodological literature in several ways. First, we provide a new identification condition for the ACME, which is relatively easy to interpret in substantive terms and also weaker than existing results in some situations. Second, we prove that the estimates based on the standard LSEM can be given valid causal interpretations under our proposed framework. This provides a basis for formally analyzing the validity of empirical studies using the LSEM framework. Third, we propose a simple nonparametric estimator of ACME and discuss a more general nonparametric approach. This allows researchers to avoid the stronger functional form assumptions required in the standard LSEM. Finally, we offer a parametric sensitivity analysis that can be easily used by applied researchers in order to assess the sensitivity of estimates to the violation of this assumption. We view this as a significant contribution because the assumptions required for identifying causal mediation effects are often too strong to justify in applied settings.

Possible future generalizations include allowing multiple mediators in the identification analysis as well as extending the sensitivity analysis to nonlinear regression models. These extensions are discussed in detail by Imai *et al.* (2009). They also develop procedures for sensitivity analysis when the model includes an interaction term between the treatment and mediating variables (i.e., when the no-interaction assumption is relaxed). Finally, an important limitation of our framework is that it does not allow the presence of a post-treatment variable that confounds the relationship between mediator and outcome. As discussed in Section 2.3, some of the previous results avoid this problem by making additional identification assumptions (e.g. Robins, 2003). The exploration of alternative solutions is left for future research.

## A Proofs

### A.1 Proof of Theorem 1

First, note that equation (4) in Assumption 1 implies,

$$Y_i(t', m) \perp\!\!\!\perp T_i \mid M_i(t) = m', X_i = x. \quad (19)$$

for all  $t, t' = 0, 1$ ,  $m, m' \in \mathcal{M}$ , and  $x \in \mathcal{X}$ . Next, equation (5) can be rewritten as,

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid T_i = t, X_i = x. \quad (20)$$

Now, for any  $t, t'$ , we have,

$$\begin{aligned} & \mathbb{E}(Y_i(t, M_i(t')) \mid X_i = x) \\ &= \int \mathbb{E}(Y_i(t, m) \mid M_i(t') = m, X_i = x) dF_{M_i(t') \mid X_i = x}(m) \\ &= \int \mathbb{E}(Y_i(t, m) \mid M_i(t') = m, T_i = t', X_i = x) dF_{M_i(t') \mid X_i = x}(m) \\ &= \int \mathbb{E}(Y_i(t, m) \mid T_i = t', X_i = x) dF_{M_i(t') \mid X_i = x}(m) \\ &= \int \mathbb{E}(Y_i(t, m) \mid T_i = t, X_i = x) dF_{M_i(t') \mid T_i = t', X_i = x}(m) \\ &= \int \mathbb{E}(Y_i(t, m) \mid M_i(t) = m, T_i = t, X_i = x) dF_{M_i(t') \mid T_i = t', X_i = x}(m) \\ &= \int \mathbb{E}(Y_i \mid M_i = m, T_i = t, X_i = x) dF_{M_i(t') \mid T_i = t', X_i = x}(m) \\ &= \int \mathbb{E}(Y_i \mid M_i = m, T_i = t, X_i = x) dF_{M_i \mid T_i = t', X_i = x}(m), \end{aligned} \quad (21)$$

where the second equality follows from equation (19), equation (20) is used to establish the third and fifth equalities, equation (4) is used to establish the fourth and last equalities, and the sixth equality follows from the fact that  $M_i = M_i(T_i)$  and  $Y_i = Y_i(T_i, M_i(T_i))$ . Finally, equation (21) implies,  $\mathbb{E}(Y_i(t, M_i(t'))) = \int \int \mathbb{E}(Y_i \mid M_i = m, T_i = t, X_i = x) dF_{M_i \mid T_i = t', X_i = x}(m) dF_{X_i}(x)$ . Substituting this expression into the definition of  $\bar{\delta}(t)$  given by equations (1) and (2) yields the desired expression for the ACME. In addition, since  $\bar{\tau} = \bar{\zeta}(t) + \bar{\delta}(t')$  for any  $t, t' = 0, 1$  and  $t \neq t'$  under Assumption 1, the result for the average natural direct effects is also immediate.  $\square$

### A.2 Proof of Theorem 2

We first show that under Assumption 1 the model parameters in the LSEM are identified. Rewrite equations (12) and (13) using the potential outcome notation as follows,

$$M_i(T_i) = \alpha_2 + \beta_2 T_i + \epsilon_{i2}(T_i), \quad (22)$$

$$Y_i(T_i, M_i(T_i)) = \alpha_3 + \beta_3 T_i + \gamma M_i(T_i) + \epsilon_{i3}(T_i, M_i(T_i)), \quad (23)$$

where the following normalization is used,  $\mathbb{E}(\epsilon_{i2}(t)) = \mathbb{E}(\epsilon_{i3}(t, m)) = 0$  for  $t = 0, 1$  and  $m \in \mathcal{M}$ . Then, equation (4) of Assumption 1 implies  $\epsilon_{i2}(t) \perp\!\!\!\perp T_i$ , yielding  $\mathbb{E}(\epsilon_{i2}(T_i) \mid T_i = t) = \mathbb{E}(\epsilon_{i2}(t)) = 0$  for any  $t = 0, 1$ . Similarly, equation (5) implies  $\epsilon_{i3}(t, m) \perp\!\!\!\perp M_i \mid T_i = t$  for all  $t$  and  $m$ , yielding

$\mathbb{E}(\epsilon_{i3}(T_i, M_i(T_i)) | T_i = t, M_i = m) = \mathbb{E}(\epsilon_{i3}(t, m) | T_i = t) = \mathbb{E}(\epsilon_{i3}(t, m)) = 0$  for any  $t$  and  $m$  where the second equality follows from equation (4). Thus, the parameters in equations (12) and (13) are identified under Assumption 1. Finally, under Assumption 1 and the LSEM, we can write  $\mathbb{E}(M_i | T_i) = \alpha_2 + \beta_2 T_i$ , and  $\mathbb{E}(Y_i | M_i, T_i) = \alpha_3 + \beta_3 T_i + \gamma M_i$ . Using these expressions and Theorem 1, the ACME can be shown to equal  $\beta_2 \gamma$ .  $\square$

### A.3 Proof that $\rho = 0$ under Assumption 1

First, as shown in Appendix A.2, Assumption 1 implies  $\mathbb{E}(\epsilon_{i2}(T_i) | T_i) = 0$  and  $\mathbb{E}(\epsilon_{i3}(T_i, M_i(T_i)) | T_i, M_i) = 0$  where the (potential) error terms are defined in equations (22) and (23). These mean independence relationships (together with the law of iterated expectations) imply,

$$\begin{aligned} 0 &= \mathbb{E}(\epsilon_{i3}(T_i, M_i(T_i))M_i) \\ &= \mathbb{E}\{\epsilon_{i3}(T_i, M_i(T_i))(\alpha_2 + \beta_2 T_i + \epsilon_{i2}(T_i))\} \\ &= \mathbb{E}\{\epsilon_{i3}(T_i, M_i(T_i))\epsilon_{i2}(T_i)\}. \end{aligned}$$

Thus, under Assumption 1, we have  $\rho = 0 \iff \mathbb{E}\{\epsilon_{i2}(T_i)\epsilon_{i3}(T_i, M_i(T_i))\} = 0$ .  $\square$

### A.4 Proof of Theorem 4

First, we write the LSEM in terms of equations (12) and (14). We omit possible pre-treatment confounders  $X_i$  from the model for notational simplicity, although the result below remains true even if such confounders are included. Since equation (4) implies  $\mathbb{E}(\epsilon_{ji} | T_i) = 0$  for  $j = 2, 3$ , we can consistently estimate  $(\alpha_1, \alpha_2, \beta_1, \beta_2)$ , where  $\alpha_1 = \alpha_3 + \alpha_2 \gamma$  and  $\beta_1 = \beta_3 + \beta_2 \gamma$ , as well as  $(\sigma_1^2, \sigma_2^2, \tilde{\rho})$ . Thus, given a particular value of  $\rho$ , we have  $\tilde{\rho} \sigma_1 \sigma_2 = \gamma \sigma_2^2 + \rho \sigma_2 \sigma_3$  and  $\sigma_1^2 = \gamma^2 \sigma_2^2 + \sigma_3^2 + 2\gamma \rho \sigma_2 \sigma_3$ . If  $\rho = 0$ , then  $\gamma = \tilde{\rho} \sigma_1 / \sigma_2$  provided that  $\sigma_3^2 = \sigma_1^2 (1 - \tilde{\rho}^2) \geq 0$ . Now, assume  $\rho \neq 0$ . Then, substituting  $\sigma_3 = (\tilde{\rho} \sigma_1 - \gamma \sigma_2) / \rho$  into the above expression of  $\sigma_1^2$  yields the following quadratic equation,  $\gamma^2 - 2\gamma \tilde{\rho} \sigma_1 / \sigma_2 + \sigma_1^2 (\tilde{\rho}^2 - \rho^2) / \{\sigma_2^2 (1 - \rho^2)\} = 0$ . Solving this equation and using  $\sigma_3 \geq 0$ , we obtain the following desired expression,  $\gamma = \frac{\sigma_1}{\sigma_2} \left\{ \tilde{\rho} - \rho \sqrt{(1 - \tilde{\rho}^2) / (1 - \rho^2)} \right\}$ . Thus, given a particular value of  $\rho$ ,  $\bar{\delta}(t)$  is identified.  $\square$

## References

- Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in Medicine* **27**, 1282–1304.
- Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* **91**, 434, 444–455.
- Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. *Proceedings of the International Joint Conference on Artificial Intelligence* .
- Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51**, 6, 1173–1182.

- Egleston, B., Scharfstein, D. O., Munoz, B., and West, S. (2006). Investigating mediation when counterfactuals are not metaphysical: Does sunlight UVB exposure mediate the effect of eye-glasses on cataracts? Working Paper 113, Department of Biostatistics, Johns Hopkins University.
- Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society, Series B (Methodological)* **69**, 2, 199–215.
- Glynn, A. N. (2008). Estimating and bounding mechanism specific causal effect. Unpublished manuscript, presented at the 25th Annual Summer Meeting of the Society for Political Methodology, Ann Arbor, Michigan.
- Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association* **55**, 292, 708–713.
- Green, D. P., Ha, S. E., and Bullock, J. G. (2010). Enough already about black box experiments: Studying mediation is more difficult than most scholars suppose. *Annals of the American Academy of Political and Social Sciences* .
- Hafeman, D. M. and VanderWeele, T. J. (2008). Alternative assumptions for the identification of direct and indirect effects. *Unpublished manuscript* .
- Imai, K., Keele, L., and Tingley, D. (2009). A general approach to causal mediation analysis. Tech. rep., Department of Politics, Princeton University. available at <http://imai.princeton.edu/research/BaronKenny.html>.
- Imai, K. and Yamamoto, T. (2008). Causal inference with measurement error: Nonparametric identification and sensitivity analyses of a field experiment on democratic deliberations. Tech. rep., Department of Politics, Princeton University.
- Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* **93**, 2, 126–132.
- Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods* **13**, 4, 314–336.
- Joffe, M. M., Small, D., Brunelli, S., Ten Have, T., and Feldman, H. I. (2008). Extended instrumental variables estimation for overall effects. *International Journal of Biostatistics* **4**, 1, Article 4.
- Joffe, M. M., Small, D., and Hsu, C.-Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science* **22**, 1, 74–97.
- Judd, C. M. and Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review* **5**, 5, 602–619.
- Kraemer, H. C., Kiernan, M., Essex, M., and Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology* **27**, 2, S101–S108.

- Kraemer, H. C., Wilson, T., Fairburn, C. G., and Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry* **59**, 877–883.
- MacKinnon, D. and Dwyer, J. (1993). Estimating mediated effects in prevention studies. *Evaluation Review* **17**, 144–158.
- MacKinnon, D., Warsi, G., and Dwyer, J. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research* **30**, 41–62.
- MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis*. Taylor & Francis, New York.
- Nelson, T. E., Clawson, R. A., and Oxley, Z. M. (1997). Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review* **91**, 3, 567–583.
- Pearl, J. (2001). Direct and indirect effects. In M. Kaufmann, ed., *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420, San Francisco, CA.
- Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology* **17**, 3, 276–284.
- Robins, J. (1999). *Statistical Models in Epidemiology, the Environment and Clinical Trials* (eds. M. E. Halloran and D. A. Berry), chap. Marginal Structural Models Versus Structural Nested Models as Tools for Causal Inference, 95–134. Springer, New York.
- 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), 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.
- Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes (with discussions). *Scandinavian Journal of Statistics* **31**, 2, 161–170.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology* **13**, 290–321.
- Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics* **33**, 2, 230–251.
- Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., and Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics* **63**, 3, 926–934.
- VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters* **78**, 2957–2962.
- VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* **20**, 1, 18–26.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* **57**, 348–368.