Causal Mediation Analysis with an Application to a Job Search Intervention[*]

Luke Keele[†] Dustin Tingley[‡] Teppei Yamamoto[§] Kosuke Imai[¶]

Abstract

Causal mechanisms are often of interest in the social sciences. That is, researchers seek to study not only whether one variable affects another but also how such a causal relationship arises. Yet, commonly used statistical methods for identifying causal mechanisms rely upon untestable assumptions and are often inappropriate even under those assumptions. Randomizing treatment and intermediate variables is also insufficient. We make three contributions to improve research on causal mechanisms. First, we present a minimum set of assumptions required under standard designs of experimental and observational studies and develop a general algorithm for estimating causal mediation effects. Second, we provide a method to assess sensitivity of conclusions to potential violations of a key assumption. Third, we offer alternative research designs for identifying causal mechanisms under weaker assumptions. The proposed approach is illustrated using an intervention designed to increase employment.

Key Words: causal inference, direct and indirect effects, mediation, potential outcomes, sensitivity analysis

[*] The companion papers that present technical aspects of the methods introduced here are available as Imai *et al.* (2010a,c, 2013). All of our proposed methods can be implemented via an R package `mediation` (Imai *et al.*, 2010b), which is freely available for download at the Comprehensive R Archive Network (http://cran.r-project.org/web/packages/mediation). A Stata package `mediation` by Raymond Hicks and Tingley is available at IDEAS (http://ideas.repec.org/c/boc/bocode/s457294.html). Imai acknowledges financial support from the National Science Foundation (SES-0849715 and SES-0918968).

[†] Associate Professor, Department of Political Science, 211 Pond Lab, Penn State University, University Park, PA 16801 Phone: 814-863-1592, Email: ljk20@psu.edu

[‡] Assistant Professor, Department of Government, Harvard University, Cambridge MA 02138, Email: dtingley@gov.harvard.edu

[§] 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

[¶] Associate Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: http://imai.princeton.edu

1 Introduction

Federal agencies evaluating social interventions as well as state and local governments and foundation sponsors have typically focused on the identification of average treatment effects. These studies usually focus on establishing whether one variable affects another, but do not attempt to explain how a causal relationship arises. This approach to studying interventions has, at times, been criticized across disciplines for being atheoretical and even unscientific (e.g., Heckman and Smith, 1995; Brady and Collier, 2004; Deaton, 2010a,b).

One alternative to simply estimating average treatment effects is the investigation of causal mechanisms. A causal mechanism is a process where a causal variable of interest, i.e., treatment variable, influences an outcome both directly and through an intermediate variable often referred to as a mediator. As such, the examination of a causal mechanism requires the specification of an intermediate variable or a mediator that lies on the causal pathway between the treatment and outcome variables. Although statistical analysis of causal mechanisms is not widespread in the field of policy analysis and program evaluation, it is quite common in some other fields of social science. In these fields, the standard approach to causal mediation analysis has been to use structural equation models (e.g., Shadish *et al.*, 2001; MacKinnon, 2008).

In this paper, we outline what assumptions are necessary to identify a causal mechanism. We show that commonly used statistical methods rely upon untestable assumptions and are often inappropriate even under those assumptions. Below we outline three important aspects of investigating causal mechanisms in the evaluation of social interventions. First, we present a minimal set of assumptions required for identification. Using the potential outcomes framework, we demonstrate why conventional exogeneity assumptions alone are insufficient for identification of causal mechanisms.[1] In particular, we show that randomization, which is often seen as the gold standard for estimating causal effects from social interventions, cannot by itself identify a causal mechanism. Second, our formal framework allows us to develop a general algorithm for estimating causal mediation effects, which is applicable to any statistical model under these assumptions. We also discuss linkages to instrumental variables estimation. Third, we outline a method to assess the sensitivity of conclusions to potential violations of key assumptions. Identification of causal mechanisms

[1] This fact is well known in the methodological literature on causal inference (e.g., Robins and Greenland, 1992; Pearl, 2001; Robins, 2003; Petersen *et al.*, 2006; Imai *et al.*, 2010c, 2013), but has not received much attention among social scientists until recently (e.g., Bullock *et al.*, 2010; Glynn, 2010).

requires strong untestable assumptions. The assumptions needed for causal mechanisms are similar to the assumptions needed to estimate treatment effects in observational studies. Given the strength of the necessary assumptions, we argue that sensitivity analysis must play an essential role by formally quantifying the degree to which empirical findings rely upon the key assumption (e.g., Rosenbaum, 2002b; Imai and Yamamoto, 2010). We demonstrate these concepts and methods through an application to a job training intervention.

2 A Running Example: Job Search Intervention Study (JOBS II)

To motivate the concepts and methods that we present, we use data from the Job Search Intervention Study (JOBS II) (Vinokur *et al.*, 1995; Vinokur and Schul, 1997). JOBS II was a randomized job training intervention for unemployed workers. The program was designed to not only increase reemployment among the unemployed but also enhance the mental health of the job seekers. For the JOBS II intervention, 1,801 unemployed workers received a pre-screening questionnaire and were then randomly assigned to treatment and control groups. Those in the treatment group participated in job-skills workshops. In the workshops, respondents learned job-search skills and coping strategies for dealing with setbacks in the job-search process. Those in the control condition received a booklet describing job-search tips. In follow-up interviews, two key outcome variables were measured; a continuous measure of depressive symptoms based on the Hopkins Symptom Checklist, and a binary variable, representing whether the respondent had become employed. Besides being interested in average treatment effects, the study analysts also hypothesized that workshop attendance would lead to better mental health and employment outcomes by enhancing participants' confidence in their ability to search for a job (Vinokur *et al.*, 1995; Vinokur and Schul, 1997). In the JOBS II data, a continuous measure of job-search self-efficacy represents this key mediating variable. In addition to the outcome and mediator, data were collected on baseline covariates prior to the administration of the treatment. These baseline covariates include measures of education, income, race, marital status, age, sex, previous occupation, and the level of economic hardship. The most important of these is the pre-treatment level of depression which is measured using the same methods as the continuous outcome variable.

3 Statistical Framework for Causal Mediation Analysis

Following prior work (e.g., Robins and Greenland, 1992; Pearl, 2001; Glynn, 2008; Imai *et al.*, 2010c), we define causal mediation effects using the potential outcomes notation (e.g., Holland, 1986). We then review the key result of Imai *et al.* (2010c) and show a minimum set of the conditions under which the product of coefficient method (MacKinnon *et al.*, 2002) and its variants yield valid estimates of causal mediation effects. As we demonstrate, this establishes a clear connection between the modern statistical framework of causal inference and the traditional (single mediator) LSEM approach used in the social sciences. Finally, we briefly explain how this approach differs from the approach based on the instrumental variable methods of Angrist *et al.* (1996). As we noted earlier, the strength of the potential outcomes framework is that it helps to clarify the assumptions needed for causal mediation effects without reference to specific statistical models.

3.1 The Counterfactual Framework

In the counterfactual framework of causal inference, the causal effect of the job training program for each worker can be defined as the difference between two potential outcomes; one potential outcome that would be realized if the worker participates in the job training program, and the other potential outcome that would be realized if the worker does not participate. Suppose that we use T_i to represent the binary treatment variable, which is equal to 1 if worker i participated in the program and to 0 otherwise. We use $Y_i(t)$ to denote the potential employment status that would result under the treatment status t. For example, $Y_i(1)$ measures the worker i's employment status if she participates in the job training program. Although there are two such potential values for each worker, only one of them is observed; for example, if worker i actually did not participate in the program, then only $Y_i(0)$ is observed. Thus, if we use Y_i to denote the observed value of employment status, then we have $Y_i = Y_i(T_i)$ for all i.

Given this setup, the causal effect of the job training program on worker i's employment status can be defined as $Y_i(1) - Y_i(0)$. Of course, because only either $Y_i(1)$ or $Y_i(0)$ is observable, even randomized experiments cannot identify this unit-level causal effect. Thus, researchers often focus on the identification and estimation of the average causal effect, which is defined as $\mathbb{E}(Y_i(1) - Y_i(0))$ where the expectation is taken with respect to the random sampling of units from a target population. If the treatment is randomized as done in JOBS II, then T_i is statistically independent of $(Y_i(1), Y_i(0))$ because the probability of receiving the treatment is identical for every observation; formally, we write $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i$. When this is true, the average causal effect can be identified as the observed difference in means between the treatment and control groups, $\mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(Y_i(1) \mid T_i = 1) - \mathbb{E}(Y_i(0) \mid T_i = 0) = \mathbb{E}(Y_i \mid T_i = 1) - \mathbb{E}(Y_i \mid T_i = 0)$, which is the familiar result that the difference-in-means estimator is unbiased for the average causal effect in randomized experiments.

Finally, we note that the above notation implicitly assumes no interference between units. In the current context, this means for example that worker i's employment status is not influenced by whether or not another worker j participates in the training program. This assumption is apparent from the fact that the potential values of Y_i are written as a function of T_i which does not depend on T_j for $i \neq j$. The assumption is best addressed through design. For example, analysts would want to ensure that participants in the experiment were not from the same household. The analyses that follow are conducted under this assumption, and the extension of our approach to the situation where this assumption is violated is left for future research.

3.2 Defining Causal Mediation Effects

In the statistics literature, the counterfactual notation from above has been extended to define causal mediation effects. We, next, outline this notation for the quantities of interest in the JOBS II study. For example, suppose we are interested in the mediating effect of the job training program on employment status where the mediating variable is worker's level of confidence in their ability to perform essential job-search activities. One possible hypothesis is that the participation in the job training program increases the level of workers' self-confidence to search for a job. We use M_i to denote the observed level of job-search self-efficacy, which was measured after the implementation of the training program but before measuring the outcome variable.

Next, we define the potential outcomes. Previously, the potential outcomes were only a function of the treatment, but in a mediation analysis the potential outcomes depend on the mediator as well as the treatment variable. Because the level of job-search self-efficacy can be affected by the program participation, there exist two potential values, $M_i(1)$ and $M_i(0)$, only one of which will be observed, i.e., $M_i = M_i(T_i)$. For example, if worker i actually participates in the program ($T_i = 1$), then we observe $M_i(1)$ but not $M_i(0)$. Next, we use $Y_i(t, m)$ to denote the potential outcome that would result if the treatment and mediating variables equal t and m, respectively. For example, in the JOBS II study, $Y_i(1, 1.5)$ represents the employment status that would be observed if worker i participates in the training program and then has a job-search self-efficacy score of 1.5. As before, we only observe one of multiple potential outcomes, and the observed outcome Y_i equals $Y_i(T_i, M_i(T_i))$. Lastly, no interference between units is still assumed; the potential mediator values for each unit do not depend on the treatment status of the other units, and the potential outcomes of each unit also do not depend on the treatment status and the mediator value of the other units.

Using this notation, we define causal mediation effects for each unit i as follows,

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

for $t = 0, 1$. In this definition, the causal mediation effect represents the indirect effects of the treatment on the outcome through the mediating variable (Pearl, 2001; Robins, 2003). The indirect effect asks the following counterfactual question: What change would occur to employment status if one changes the mediator from the value that would be realized under the treatment condition, i.e., $M_i(1)$, to the value that would be observed under the control condition, i.e., $M_i(0)$, while holding the treatment status at t? Although $Y_i(t, M_i(t))$ is observable for units with $T_i = t$, $Y_i(t, M_i(1-t))$ can never be observed for any unit. In the JOBS II study, for example, $\delta_i(1)$ represents the difference between the two potential employment statuses for worker i who participates in the training program. For this worker, $Y_i(1, M_i(1))$ equals the employment status that is actually observed, whereas $Y_i(1, M_i(0))$ represents the employment status that would result if worker i participates but the mediator takes the value that would result under no participation.

Similarly, we can define the direct effects of the treatment for each unit as follows,

$$\zeta_i(t) \quad \equiv \quad Y_i(1, M_i(t)) - Y_i(0, M_i(t)), \tag{2}$$

for $t = 0, 1$. In the JOBS II study, for example, $\zeta_i(1)$ represents the direct effect of the job-training program on worker i's employment status while holding the level of his or her job-search self-efficacy constant at the level that would be realized under the program participation.[2] Then, the total effect of the treatment, τ_i, can be decomposed into the causal mediation and direct effects in the following manner, $\tau_i \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \frac{1}{2}\sum_{t=0}^{1}\{\delta_i(t) + \zeta_i(t)\}$, where we simply average over the treatment assignment. In addition, if we assume that causal mediation and direct effects do not vary as functions of treatment status (i.e., $\delta_i = \delta_i(1) = \delta_i(0)$ and $\zeta_i = \zeta_i(1) = \zeta_i(0)$ called the no-interaction assumption), then the mediation and direct effects sum to the total effect, i.e., $\tau_i = \delta_i + \zeta_i$.

Finally, in causal mediation analysis, we are typically interested in the following *average causal mediation effects* (ACME),

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

[2] Pearl (2001) calls $\zeta_i(t)$ as *natural direct effects* to distinguish them from *controlled direct effects* of the treatment. Imai, Tingley, and Yamamoto (2013) argue that the former corresponds to causal mechanisms whereas the latter represents the causal effects of direct manipulation. They also discuss the implications of this distinction for experimental designs.

for $t = 0, 1$. For the JOBS II study, this would represent the average causal mediation effect among all workers of the population, of which the analysis sample can be considered as representative. Similarly, averaging over the relevant population of workers, we can define the average direct (ADE) and total effects as $\bar{\zeta}(t) \equiv \mathbb{E}(Y_i(1, M_i(t)) - Y_i(0, M_i(t)))$, and $\bar{\tau} \equiv \mathbb{E}(Y_i(1, M_i(1)) - Y_i(0, M_i(0))) = \frac{1}{2}\{\bar{\delta}(0) + \bar{\delta}(1) + \bar{\zeta}(0) + \bar{\zeta}(1)\}$, respectively. Further, if we make the following no-interaction assumption (i.e., $\bar{\delta} = \bar{\delta}(1) = \bar{\delta}(0)$ and $\bar{\zeta} = \bar{\zeta}(1) = \bar{\zeta}(0)$), the average causal mediation and average direct effects sum to the average total effect, i.e., $\bar{\tau} = \bar{\delta} + \bar{\zeta}$, yielding the decomposition of the total effect into direct and indirect effects.

3.3 Nonparametric Identification under Sequential Ignorability

We now turn to the question of identification and specifically that of nonparametric identification. By nonparametric identification, we mean that without any additional distributional or functional-form assumptions, the average causal mediation effects can be consistently estimated. For randomized experiments, we need to assume treatment is independent of the potential outcomes and that there is no interaction between units to identify the average treatment effect. Causal mediation analysis, however, requires an additional assumption.

In particular, we rely on the following assumption introduced by Imai *et al.* (2010c). Let X_i be a vector of the observed pre-treatment confounders for unit i where $\mathcal{X}$ denotes the support of the distribution of X_i. In the JOBS II data, X_i includes for each unemployed worker the pre-treatment level of employment status as well as some demographic characteristics such as education, race, marital status, sex, previous occupation, and the level of economic hardship. Given these observed pre-treatment confounders, the identification assumption can be formally written as,

ASSUMPTION 1 (SEQUENTIAL IGNORABILITY (IMAI *et al.*, 2010C)) *We assume that the following two statements of conditional independence hold,*

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x, \tag{3}$$

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

where $0 < \Pr(T_i = t \mid X_i = x)$ *and* $0 < p(M_i = m \mid T_i = t, X_i = x)$ *for* $t = 0, 1$, *and all* $x \in \mathcal{X}$ *and* $m \in \mathcal{M}$.

Imai *et al.* (2010c) prove that under Assumption 1 the average causal mediation effects are nonparametrically identified and discuss how this assumption differs from those proposed in the prior literature. Assumption 1 is called sequential ignorability because two ignorability assumptions are made sequentially.

First, given the observed pre-treatment confounders, the treatment assignment is assumed to be ignorable, i.e., statistically independent of potential outcomes and potential mediators. In the JOBS II study, this first ignorability assumption is satisfied because workers were randomly assigned to the treatment and control groups. In contrast, this part of the assumption is not guaranteed to hold in observational studies where subjects may self-select into the treatment group.

The second part of Assumption 1 states that the observed mediator is ignorable given the observed treatment and pre-treatment confounders. That is, the second part of the sequential ignorability assumption is made conditional on the observed value of the (ignorable) treatment and the observed pre-treatment confounders. Unlike the ignorability of treatment assignment, however, the ignorability of the mediator may not hold in randomized experiments. Randomization of treatment does not imply ignorability of the mediator. Ignorability of the mediator implies that among those workers who share the same treatment status and the same pre-treatment characteristics the observed values of the mediator can be regarded as if randomized.

We emphasize that the second stage of sequential ignorability is a strong assumption. Such an assumption is often referred to as nonrefutable because it cannot be directly tested from the observed data (Manski, 2007). Moreover, it is always possible that there might be unobserved variables that confound the relationship between the outcome and the mediator variables even after conditioning on the observed treatment status and the observed covariates. Furthermore, the conditioning set of covariates must be pre-treatment variables. Indeed, without an additional assumption, we cannot condition on the post-treatment confounders even if such variables are observed by researchers (e.g., Avin *et al.*, 2005). This means that, similar to the ignorability of treatment assignment in observational studies, it is difficult to know for certain whether or not the ignorability of the mediator holds even after researchers collect as many pre-treatment confounders as possible.

Thus, in Section 3.8, we develop a set of sensitivity analyses that will allow researchers to quantify the degree to which their empirical findings are robust to a potential violation of the sequential ignorability assumption. Sensitivity analyses are an appropriate approach to nonrefutable assumptions because they allow the researcher to probe whether a substantive conclusion is robust to potential violations of the assumption.

3.4 Identification Within the Structural Equation Framework

Later we discuss a general estimation framework for causal mechanisms. Here, we show that the potential outcomes framework encompasses the standard mediation analysis based on the single mediator LSEM as a special case. For illustration, consider the following set of linear equations,

$$M_i = \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \epsilon_{i2}, \qquad (5)$$

$$Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \epsilon_{i3}, \qquad (6)$$

After fitting each linear equation via least squares, the product of coefficients method uses $\hat{\beta}_2\hat{\gamma}$ as an estimated mediation effect (MacKinnon *et al.*, 2002).

Does the product of coefficients method yield a valid estimate for the causal mediation effect under the potential outcomes framework? Imai *et al.* (2010c) prove that under sequential ignorability and the additional no-interaction assumption, i.e., $\bar{\delta}(1) = \bar{\delta}(0)$, the estimate based on the product of coefficients method can be interpreted as a valid estimate (i.e., asymptotically consistent) of the causal mediation effect so long as the linearity assumption holds (see also Jo, 2008).

3.5 Sequential Ignorability and Conventional Exogeneity Assumptions

Importantly, sequential ignorability (Assumption 1) differs critically from the conventional exogeneity assumptions that are commonly understood to identify indirect effects in structural equation models as outlined in Section 3.4. Importantly, one might incorrectly conjecture that Assumption 1 is satisfied by the randomization of both treatment and mediator. For example, Spencer *et al.* (2005) propose the "causal chain" approach where researchers implement two randomized experiments, one in which the treatment is randomized to identify its effect on the mediator, and another in which the mediator is randomized to identify its effect on the outcome. Unfortunately, even though the treatment and mediator are each guaranteed to be exogenous in these two experiments, simply combining the two is not sufficient for identification.

We use a numerical example to illustrate why this is true. Consider the hypothetical population in Table 1, which describes the population proportion of "types" of units by the values of potential mediators and outcomes. While the values in Table 1 can never be jointly observed, the two randomized experiments will give sufficient information to identify the average causal effect of the treatment on the mediator as well as that of the mediator on the outcome. In this example, both of these effects are positive and equal to 0.2, and thus based on these results one might conclude that the ACME is positive. However, the ACME is actually *negative*. Thus, contrary to the commonly held belief, the conventional exogeneity assumptions do not necessarily identify the ACME.

					Treatment effect	Mediator effect	
Population	Potential mediators and outcomes				on mediator	on outcome	Causal mediation effect
proportion	$M_i(1)$	$M_i(0)$	$Y_i(t,1)$	$Y_i(t,0)$	$M_i(1) - M_i(0)$	$Y_i(t,1) - Y_i(t,0)$	$Y_i(t, M_i(1)) - Y_i(t, M_i(0))$
0.3	1	0	0	1	1	-1	-1
0.3	0	0	1	0	0	1	0
0.1	0	1	0	1	-1	-1	1
0.3	1	1	1	0	0	1	0
Average	0.6	0.4	0.6	0.4	0.2	0.2	-0.2

Table 1: The Fallacy of the Causal Chain Approach. The left five columns of the table show a hypothetical population proportion of "types" of units defined by the values of potential mediators and outcomes. Note that these values can never be jointly observed. The last row of the table shows the population average value of each column. In this example, the average causal effects of the treatment on the mediator (the sixth column) is positive and equal to 0.2. Moreover, the average causal effect of the mediator on the outcome (the seventh column) is also positive and equals 0.2. And yet, the average causal mediation effect (ACME; final column) is negative and equals -0.2.

In the above example, causal heterogeneity exists in such a way that the units with a positive effect of treatment on mediator (the first row of the table) exhibit a negative effect of mediator on outcome. This particular deviation from sequential ignorability makes the causal mediation effects negative on average even though all other average effects are positive. The key point, beyond this specific example, is the fundamental difference between the causal mediation effect and the *causal effect of the mediator* itself. The latter refers to the average difference in the potential outcomes that would be realized if the mediator were manipulated to certain fixed values, i.e., the average value of $Y_i(t,1) - Y_i(t,0)$, which can be consistently estimated when the conventional exogeneity assumption holds about the mediator. However, this quantity crucially differs from the causal mediation effect in that the mediator is artificially manipulated to take particular values (1 or 0) as opposed to being hypothetically set to the values that would naturally arise in response to treatment ($M_i(1)$ or $M_i(0)$). Because a causal mechanism represents how the effect of *treatment* on outcome is transmitted through the mediator, identifying the effect of the mediator itself is not sufficient.

This exercise demonstrates one important facet of causal mechanisms. The effects in a causal mechanism do not correspond to any direct experimental intervention. That is, the indirect effect does not correspond to a contrast between treatment regimes of any randomized experiment performed via interventions on T_i, M_i, or Y_i. There is no experiment that will directly identify this effect without additional assumptions.

3.6 Relationship with Instrumental Variables

Recently, some scholars have considered the use of instrumental variables for causal mediation analysis (e.g., Albert, 2008; Jo, 2008; Sobel, 2008). Using instrumental variables to estimate causal mediation effects requires an alternative set of identification assumptions which differ from Assumption 1 in important ways. While the IV estimator does correspond to a known experimental intervention, the direct effect is assumed to be zero. In the jargon of IV, the exclusion restriction is relaxed in a mediation analysis. This means that the instrumental variables approach eliminates, *a priori*, alternative causal mechanisms. While the exclusion restriction is often plausible in settings with noncompliance, for many causal mechanisms that is unrealistic. For example, in the context of the JOBS II study invoking the exclusion restriction implies assuming that a job-training intervention has no direct effect on obtaining a job. As such, while IV can be used for estimating mediation effects in theory, ruling out the direct effect a priori often cannot be justified.

3.7 Estimation of Causal Mediation Effects

As we outline above, analysts can use LSEM to estimate causal mediation effects. The linearity assumptions required with LSEMs, however, are often inappropriate. For example, in the JOBS II data, the employment status outcome measure is a dummy variable. Here, use of LSEM is no longer appropriate. Imai *et al.* (2010a) show that the nonparametric identification result leads to a general algorithm for computing the ACME and the ADE, which is applicable to any statistical model so long as sequential ignorability holds. Here, we briefly describe the algorithm, which consists of two steps.

First, analysts must fit regression models for the mediator and outcome. The mediator model includes on the right hand side of the model the treatment and any relevant pre-treatment covariates. The outcome is modeled as a function of the mediator, the treatment, and the pre-treatment covariates. The form of these models is immaterial. The models can be nonlinear such as logit or probit models or even non/semiparametric models such as generalized additive models. Based on the mediator model, we then generate two sets of predictions for the mediator, one under the treatment and the other under the control. In the JOBS II study, this would correspond to predicted levels of job-search self-efficacy after either attending the training sessions or receiving the booklet.

Next, we use the outcome model to make potential outcome predictions. Suppose that we are interested in estimating the ACME under the treatment, i.e., $\bar{\delta}(1)$. First, the outcome is predicted under the treatment using the value of the mediator predicted in the treatment condition. Second, the outcome is predicted under the treatment condition but now uses the mediator prediction from the control condition. The ACME is then computed as the average difference between the outcome predictions using the two different values of the mediator. For the JOBS II data, this would correspond to the average difference in employment status from fixing the treatment status but changing the level of job-search self-efficacy between the level predicted after attending the training session versus reading the pamphlet given the control group. Finally, inference proceeds via the bootstrap.

3.8 Sensitivity Analysis

As we discussed, identification of causal mechanisms requires an assumption which cannot be tested with the observed data. Given that the identification of causal mechanisms relies upon an untestable assumption, it is important to evaluate whether empirical results are sensitive this assumption. Sensitivity analysis provides one way to do this. The goal in a sensitivity analysis is to quantify the exact degree to which the key identification assumption must be violated in order for a researcher's original conclusion to be reversed. If inference is sensitive, a slight violation of the assumption may lead to substantively different conclusions.

Imai *et al.* (2010a,c) propose a sensitivity analysis for causal mechanisms based on the correlation between ϵ_{i2}, the error for the mediation model, and ϵ_{i3}, the error for the outcome model, under a standard LSEM setting and several commonly used non-linear models. They use ρ to represent the correlation across the two error terms. If sequential ignorability holds, all relevant pre-treatment confounders have been conditioned on and thus ρ equals zero. Nonzero values of ρ imply departures from the sequential ignorability assumption and that some hidden confounder is biasing the ACME estimate.

For example, in the JOBS II study, the key concern is an unmeasured confounder that affects both the sense of job-search self-efficacy and either of the outcome measures. Any confounding of this type will be reflected in the data generating process as a correlation between ϵ_{i2} and ϵ_{i3}. Ignoring this and estimating the two models separately will lead to a biased estimate of the ACME. Thus, ρ can serve as a sensitivity parameter since more extreme values of ρ represent larger departures from the sequential ignorability assumption. In particular, while the true value of ρ is unknown, it is possible to calculate the values of ρ for which the ACME is zero or its confidence interval contains zero.

Researchers may find it difficult to interpret the sensitivity parameter ρ. To ease interpretation, Imai *et al.* (2010c) develop an alternative formulation of the sensitivity analysis based on how much the omitted variable would alter the R^2's of the mediator and outcome models. For example, if a confounder is important in determining job-search self-efficacy and the outcome measures, then the models excluding the confounder will have a much smaller value of R^2 compared to a model including the confounder. On the other hand, if the confounder is unimportant, R^2 will not be very different whether including or excluding the variable. Thus, this relative change in R^2 can be used as a sensitivity parameter. For example, the original results would be considered weak if the sensitivity analysis suggests that confounder would need to explain only small portion of the remaining variance in job-search self-efficacy and employment status for the ACME to lose statistical significance.

While sensitivity analysis can shed light on whether the estimates obtained under sequential ignorability are robust to possible hidden pre-treatment confounders, it is important to note the limitations of this method of sensitivity analysis. First, the proposed method is designed to probe for sensitivity to the presence of an unobserved *pre-treatment* confounder. In particular, it does not address the possible existence of confounders which are affected by the treatment and then confound the relationship between the mediator and the outcome. If such a confounder exists, we will need a different strategy for both identification and sensitivity analysis (e.g. Imai and Yamamoto, 2011). Second, unlike statistical hypothesis testing, sensitivity analysis does not provide an objective criterion which allows researchers to determine whether sequential ignorability is valid or not. This is not surprising given that sequential ignorability is an irrefutable assumption. Therefore, as suggested by Rosenbaum (2002a, p.325), a cross-study comparison is helpful for assessing the robustness of one's conclusion relative to those of other similar studies. Finally, the proposed framework rests on the more fundamental presumption that the causal ordering imposed by the analyst is correct (e.g., whether emotional reactions occur before policy preference is formed). This can only be verified by some appeal to scientific evidence not present in the data.

4 An Application to JOBS II

Next, we present estimates for the proposed causal mechanism in the JOBS II study. In the JOBS II study, a key question of interest is whether the program participation leads to a higher level of employment and reduced depression through increasing job-search self-efficacy. We focus on how basic interpretation differs from a focus on average treatment effects. Table 2 contains the estimated mechanism effects for two outcomes in the JOBS II study. One outcome measure is a continuous scale of depressive symptoms. The second outcome measure is an indicator variable for whether subjects were working more than 20 hours a week 6 months after the job training program. In order to make Assumption 1 plausible, models for both outcome variables as well as the mediator variable are specified with a number of pre-treatment covariates, including gender, age, marital status, race, education, income, perceived level of economic hardship, and occupational categories prior to participation in the training program. The outcome models include the interaction term between the treatment and the mediator variable so that the causal mediation and direct effects will be allowed to differ depending on the baseline treatment condition.

Table 2: *Estimated Causal Quantities of Interest for JOBS II Study.*

		Depression	Employment Status
Average Mediation Effects	$\bar{\delta}(1)$	$-.017$	$.002$
		$[-.041,\ .004]$	$[-.002,\ .009]$
	$\bar{\delta}(0)$	$-.026$	$.007$
		$[-.063,\ .006]$	$[-.002,\ .020]$
Average Direct Effects	$\bar{\zeta}(1)$	$-.036$	$.053$
		$[-.120,\ .048]$	$[-.010,\ .114]$
	$\bar{\zeta}(0)$	$-.045$	$.058$
		$[-.131,\ .039]$	$[-.005,\ .119]$
Average Total Effect	$\bar{\tau}$	$-.062$	$.060$
		$[-.153,\ .028]$	$[-.003,\ .121]$

Note: $N = 899$. Depression outcome is a continuous measure of depressive symptoms. Employment status outcome is whether a respondent was working more than 20 hours per week after the training sessions. In square brackets are 95% bootstrap percentile confidence intervals. Models for the outcome and mediator were specified with a number of covariates including measures of the outcomes measured prior to treatment.

We begin with a discussion of the results for the depression outcome. The estimate of the average total effect is the same as the familiar average treatment effect. Here, we observe a slight decrease in depressive symptoms (about $-.062$ points on the scale of 1 to 5), but the estimate is not statistically significant at conventional levels ($p = .17$). In an analysis of the causal mechanism, we decompose the total effect into direct and indirect effects. The indirect or mediation effect is the portion of the treatment effect that is transmitted through job-search self-efficacy. In this case, that is a relatively small portion of the total effect ($-.017$ and $-.026$ points for the treatment and control baselines respectively, which correspond to about 27% and 42% of the total effect) and is not statistically significantly different from zero ($p = .18$ and $p = .13$, respectively).

In the second column, we present the results for the measure of employment status. Here, since the outcome is binary we might use a model such as logit or probit models to estimate the indirect effect. Logit and probit models, however, cannot be combined with the popular product of coefficients method for estimating mediation effects. The algorithm we developed can, however, accommodate a wide variety of models that might be used for either the outcome or mediator model. Again, we begin with the total effect or the average treatment effect. We observe that the treatment increased the probability of obtaining a job by about six percentage points. Again, the 95% confidence intervals include zero, albeit by a slim margin ($p = .05$). However, when we decompose this estimate into its indirect and direct components, we observe that the treatment effect was almost entirely transmitted directly from the treatment to the outcome with a minuscule amount (.002 and .007 for the treatment and control baselines, respectively) transmitted through the hypothesized mechanism.

4.1 Sensitivity Analysis

As we demonstrated in Section 3.5, randomization of the treatment alone does not identify causal mediation effects. This means that even in a randomized intervention like the JOBS II study, an additional assumption, e.g., sequential ignorability, is required for identification. It is not unreasonable to think the sequential ignorability may have been violated in the JOBS II study. For example, Jo (2008, p. 317) points out that the second part of the sequential ignorability might be violated in the JOBS II study and states that "individuals who improved their sense of mastery by one point in the intervention program may have different observed and unobserved characteristics from those of individuals who equally improved their sense of mastery in the control condition."

We next present results from sensitivity analyses for the results in Table 2. We focus on the average causal mediation effect for the treatment baseline (i.e., $\bar{\delta}(1)$) on the depression outcome, since the estimated magnitude of the mediation effect for the employment outcome is very small. Here, we attempt to understand how large a confounder would have to be for the sign of the point estimate to change since the estimated

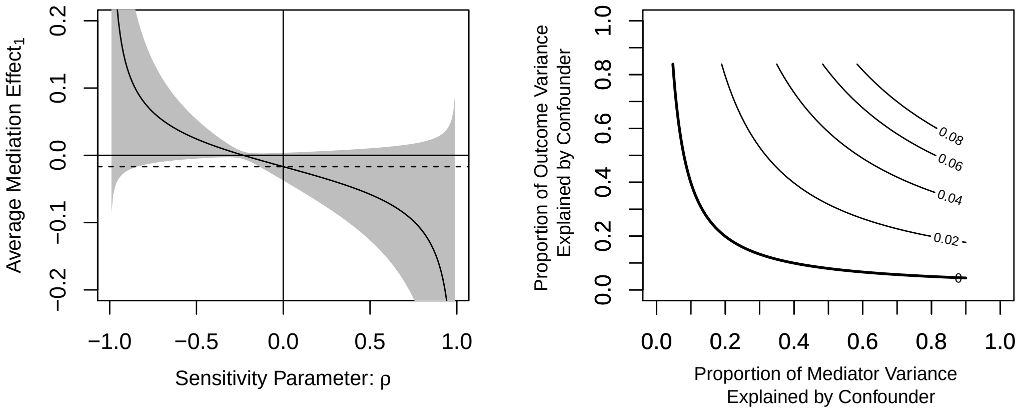

Figure 1: Sensitivity Analysis for the JOBS II study, with Depressive Symptoms as outcome. In the left panel, the true ACME is plotted against the sensitivity parameter ρ, which is the correlation between the error terms in the mediator and outcome regression models. The dashed line represents the estimated ACME when the sequential ignorability assumption is made. The shaded areas represent the 95% confidence interval for the mediation effects at each value of ρ. In the right panel, the contours represent the true ACME plotted as function of the proportion of the total mediator variance (horizontal axis) and the total outcome variance (vertical axis) that are each explained by the unobserved confounder included in the corresponding regression models. Here the unobserved confounder is assumed to affect the mediator and outcome in opposite directions.

effect is not statistically significant. We discuss the sensitivity analysis results both in terms of the parameter ρ and also using R^2 values.

We first ask how large ρ must be for the mediation effect to be zero. We find that for this outcome, the estimated ACME equals zero when ρ equals -0.23. Of course, if we take into account sampling uncertainty, we find that the 95% confidence intervals for the ACME include zero for all values of ρ. Our analysis indicates that for the true ACME to be zero, there must be an unobserved confounder that affects both job-search self-efficacy and depressive symptoms in opposite directions and makes the correlation between the two error terms greater than -0.23.

We acknowledge that analysts may have difficulty in interpreting the sensitivity analysis results in substantive terms. Repeated use of the sensitivity analysis across different studies will allow researchers to understand what are large and small values of ρ. We can also express the degree of sensitivity in terms of the importance of an unobserved confounder in explaining the observed variation in the mediator and outcome variables.

In the right panel of Figure 1, the true ACME is shown as contours with respect to the proportions of the variance in the mediator (horizontal axis) and in the outcome (vertical axis) that are each explained by the unobserved confounder in the true regression models. Here, we explore the case where the unobserved confounder affects the mediator and outcome in opposite directions, as we found in the preceding analysis that the ACME can only become negative and more distant from zero when the effects of the confounder were in the same direction. These two sensitivity parameters are each bounded above by one minus the R^2 of the observed models, which represents the proportion of the variance that is not yet explained by the observed predictors in each model. In this example, these upper bounds are 0.88 for the mediator model and 0.98 for the outcome model. Other things being equal, a low value of this upper bound indicates a more robust estimate of the ACME because there is less room for an unobserved confounder to bias the result.

We find that the true ACME changes sign if the product of these proportions are greater than 0.04 and the confounder affects both job-search self-efficacy and depressive symptoms in the same direction. For example, if an unobserved confounder explains more than 20 percent of the variance in self-efficacy and 20 percent of the variance of the depressive symptoms scale, then the true ACME is positive. Thus, the negative ACME reported in the original analysis is robust to confounding due to an unmeasured confounder when the latter explains less than about 20 percent ($\approx \sqrt{0.04}$) of the variance in the mediator and outcome.

5 Concluding Remarks about Mediation

Generally in program evaluation, analysts focus solely on average treatment effects. There is good reason for this given that with randomization we can estimate this parameter under relatively weak assumptions. Policymakers may, however, demand deeper explanations for why interventions matter. Analysts may be able to use causal mechanisms to provide such explanations.

Here, we have outlined the assumptions and methods needed for going beyond average treatment effects to the estimation of causal mechanisms. Researchers often estimate causal mechanisms without fully understanding the assumptions needed and awareness of the key assumption can help improve design, especially in terms of collecting a full set of possible pretreatment covariates that might confound the indirect effect. The sensitivity analysis we develop allows researchers to formally evaluate the robustness of their conclusions to the potential violations of those assumptions. Strong assumptions such as sequential ignorability deserve great care and require a combination of innovative statistical methods and research designs.

Recent work has explored how analysts can use experimental designs other than a single randomization to shed light on causal mechanisms. The problem with the single experiment design is that we cannot be sure that the observed mediator is ignorable conditional on the treatment and pre-treatment covariates. Imai *et al.* (2013) propose several different experimental designs and derive their identification power under a minimal set of assumptions. These alternative designs can often provide bounds on mediation effects under weaker assumptions than is true with a single experiment. As such, researchers have a number of tools available when they are interested in moving beyond the average treatment effect.

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. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, 357–363, Edinburgh, Scotland. Morgan Kaufmann.

Brady, H. E. and Collier, D. (2004). *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman & Littlefield Pub Inc.

Bullock, J., Green, D., and Ha, S. (2010). Yes, But What's the Mechanism? (Don't Expect an Easy Answer). *Journal of Personality and Social Psychology* **98**, 4, 550–558.

Deaton, A. (2010a). Instruments, randomization, and learning about development. *Journal of Economic Literature* **48**, 2, 424–455.

Deaton, A. (2010b). Understanding the mechanisms of economic development. *Journal of Economic Perspectives* **24**, 3, 3–16.

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.

Glynn, A. N. (2010). What can we learn with statistical truth serum?: Design and analysis of the list experiment. Tech. rep., Department of Government, Harvard University.

Heckman, J. J. and Smith, J. A. (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives* **9**, 2, 85–110.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81**, 945–960.

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., Tingley, D., and Yamamoto, T. (2010b). *Advances in Social Science Research Using R (ed. H. D. Vinod)*, chap. Causal Mediation Analysis Using R, 129–154. Lecture Notes in Statistics. Springer, New York.

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

Imai, K., Tingley, D., and Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of The Royal Statistical Society, Series A* **176**, 1, Forthcoming.

Imai, K. and Yamamoto, T. (2010). Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis. *American Journal of Political Science* **54**, 2, 543–560.

Imai, K. and Yamamoto, T. (2011). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. Working paper available at `http://imai.princeton.edu/research/medsens.html` Revise and Resubmit at *Political Analysis*.

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods* **13**, 4, 314–336.

MacKinnon, D. (2008). *Introduction to Statistical Mediation Analysis*. Routledge, New York, NY.

MacKinnon, D., Lockwood, C., Hoffman, J., West, S., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods* **7**, 1, 83–104.

Manski, C. F. (2007). *Identification For Prediction And Decision*. Harvard University Press, Cambridge, Mass.

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

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. 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.

Rosenbaum, P. R. (2002a). Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association* **97**, 457, 1–10.

Rosenbaum, P. R. (2002b). Covariance adjustment in randomized experiments and observational studies (with discussion). *Statistical Science* **17**, 286–327.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston.

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics* **33**, 2, 230–251.

Spencer, S., Zanna, M., and Fong, G. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology* **89**, 6, 845–851.

Vinokur, A., Price, R., and Schul, Y. (1995). Impact of the jobs intervention on unemployed workers varying in risk for depression. *American Journal of Community Psychology* **23**, 1, 39–74.

Vinokur, A. and Schul, Y. (1997). Mastery and inoculation against setbacks as active ingredients in the jobs intervention for the unemployed. *Journal of Consulting and Clinical Psychology* **65**, 5, 867–877.