Psychological Methods
2010, Vol. 15, No. 4, 309–334

1082-989X/10/$12.00 DOI: 10.1037/a0020761

A General Approach to Causal Mediation Analysis

Kosuke Imai
Princeton University

Luke Keele
Ohio State University

Dustin Tingley
Harvard University

Traditionally in the social sciences, causal mediation analysis has been formulated, understood, and implemented within the framework of linear structural equation models. We argue and demonstrate that this is problematic for 3 reasons: the lack of a general definition of causal mediation effects independent of a particular statistical model, the inability to specify the key identification assumption, and the difficulty of extending the framework to nonlinear models. In this article, we propose an alternative approach that overcomes these limitations. Our approach is general because it offers the definition, identification, estimation, and sensitivity analysis of causal mediation effects without reference to any specific statistical model. Further, our approach explicitly links these 4 elements closely together within a single framework. As a result, the proposed framework can accommodate linear and nonlinear relationships, parametric and nonparametric models, continuous and discrete mediators, and various types of outcome variables. The general definition and identification result also allow us to develop sensitivity analysis in the context of commonly used models, which enables applied researchers to formally assess the robustness of their empirical conclusions to violations of the key assumption. We illustrate our approach by applying it to the Job Search Intervention Study. We also offer easy-to-use software that implements all our proposed methods.

Keywords: causal inference, causal mechanisms, direct and indirect effects, linear structural equation models, sensitivity analysis

Causal inference is a central goal of social science research. In this context, randomized experiments are typically seen as a gold standard for the estimation of causal effects, and a number of statistical methods have been developed to make adjustments for methodological problems in both experimental and observational settings. However, one common criticism of experimentation and statistics is that they can provide only a black-box view of causality. The argument is that although the estimation of causal effects allows researchers to examine whether a treatment causally affects an outcome, it cannot tell us how and why such an effect arises. This is an important limitation because the identification of causal mechanisms is required to test competing theoretical explanations of the same causal effects. Causal mediation analysis plays an essential role in potentially overcoming this limitation by helping to identify intermediate variables (or mediators) that lie in the causal pathway between the treatment and the outcome.

Traditionally, causal mediation analysis has been formulated, understood, and implemented within the framework of linear structural equation modeling (LSEM; e.g., Baron & Kenny, 1986; Hyman, 1955; James, Mulaik, & Brett, 1982; Judd & Kenny, 1981; MacKinnon, 2008; MacKinnon & Dwyer, 1993). We argue and demonstrate that this is problematic for two reasons. First, by construction, the LSEM framework cannot offer a general definition of causal mediation effects that are applicable beyond specific statistical models. This is because the key identification assumption is stated in the context of a particular model, making it difficult to separate the limitations of research design from those of the specific statistical model.[1] Second, the methods developed in the LSEM framework are not generalizable to nonlinear models,

This article was published Online First October 18, 2010.

Kosuke Imai, Department of Politics, Princeton University; Luke Keele, Department of Political Science, Ohio State University; Dustin Tingley, Department of Government, Harvard University.

Funding for this study was supported in part by National Science Foundation Grant SES-0918968. The two companion articles, one forming the theoretical basis of this article and the other providing the details of software implementation of the proposed methods, are available as Imai, Keele, and Yamamoto (2010) and Imai et al. (2010a), respectively. The easy-to-use software, mediation, is freely available at the Comprehensive R Archive Network (http://cran.r-project.org/web/packages/mediation). The replication materials for all the empirical results presented in this article are available at the authors' dataverse (Imai et al., 2010). We thank Matt Incantalupo, Dave Kenny, Dave MacKinnon, Keith Markus, Scott Maxwell, Walter Mebane, and Kris Preacher for useful comments that significantly improved the presentation of this article.

Correspondence concerning this article should be addressed to Kosuke Imai, Department of Politics, Princeton University, Princeton, NJ 08544. E-mail: kimai@princeton.edu

[1] By "identification," we mean whether the causal mediation effects can be consistently estimated. Thus, identification is a minimum requirement for valid statistical inference and precedes the issue of statistical estimation, which is about how to make inferences from a finite sample. See below for the formal discussion in the context of causal mediation analysis and Manski (2007) for a general discussion.

including logit and probit models, for discrete mediators and outcomes as well as non- or semiparametric models.

In this article, we propose a general approach that overcomes these limitations. We use a single framework for the definition, identification, estimation, and sensitivity analysis of causal mediation effects without reference to any specific statistical model. First, following the recently published work (e.g., Jo, 2008; Sobel, 2008), we place causal mediation analysis within the counterfactual framework of causal inference and offer the formal definition of causal mediation effects. This definition formalizes, independent of any specific statistical models, the intuitive notion about mediation held by applied researchers that the treatment indirectly influences the outcome through the mediator.

Second, we slightly extend the result of Imai, Keele, and Yamamoto (2010), who proved that under the sequential ignorability assumption the average causal mediation effects are *nonparametrically* identified (i.e., can be consistently estimated without any functional form and distributional assumptions). Sequential ignorability consists of two assumptions: (a) Conditional on the observed pretreatment covariates, the treatment is independent of all potential values of the outcome and mediating variables, and (b) the observed mediator is independent of all potential outcomes given the observed treatment and pretreatment covariates. Such a nonparametric identification analysis is important because it establishes a minimum set of assumptions required for mediation effects to be interpreted as causal without respect to statistical models used by researchers.

Third, using our nonparametric identification result, we develop general estimation procedures for causal mediation effects that can accommodate linear and nonlinear relationships, parametric and nonparametric models, continuous and discrete mediators, and various types of outcome variables. In the literature, some have extended the LSEM framework to these settings (e.g., Li, Schneider, & Bennett, 2007; MacKinnon, 2008; MacKinnon, Lockwood, Brown, Wang, & Hoffman, 2007; Wang & Taylor, 2002). Our approach encompasses many of the existing methods as special cases, thereby accomplishing many of future statistical tasks identified in a recent review article by MacKinnon and Fairchild (2009).

The last and yet perhaps most important contribution of our proposed approach is a set of sensitivity analyses we develop for statistical models commonly used by applied researchers. Sensitivity analysis allows researchers to formally quantify the robustness of their empirical conclusions to the potential violation of sequential ignorability, which is the key and yet untestable assumption needed for identification. The fundamental difficulty in the causal mediation analysis is that there may exist unobserved confounders that causally affect both the mediator and the outcome even after conditioning on the observed treatment and pretreatment covariates. Therefore, assessing the sensitivity of one's empirical findings to the possible existence of such confounders is required in order to evaluate the validity of any mediation study. In the LSEM framework, Imai, Keele, and Yamamoto (2010) proposed a straightforward way to check how severe the violation of the key identifying assumption would need to be for the original conclusions to be reversed. We generalize this sensitivity analysis so that it can be applied to other settings.

Because our approach is developed without any reference to a particular statistical model, it is applicable across a wide range of situations. In this article, we illustrate its applicability using a variety of cross-section settings. Our general approach also allowed us to develop the easy-to-use software, mediation, which is freely available as an R package (R Development Core Team, 2009) at the Comprehensive R Archive Network.[2] All the analyses presented in this article are conducted with this software. The details about the software implementation and its usage are given in a companion article (Imai, Keele, Tingley, & Yamamoto, 2010a). Future research should address the application of our approach to the panel data settings (e.g., Cole & Maxwell, 2003; MacKinnon, 2008, Chapter 8), and multiple (e.g., MacKinnon, 2000; Preacher & Hayes, 2008) and multilevel (e.g., Krull & MacKinnon, 1999) mediators, all of which are beyond the scope of the current article.

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

To motivate the concepts and methods that we present, we rely on an example from the psychology literature on mediation and use the JOBS II for our illustration. JOBS II is a randomized field experiment that investigates the efficacy of a job training intervention on unemployed workers. The program is designed not only to increase reemployment among the unemployed but also to enhance the mental health of the job seekers. In the experiment, 1,801 unemployed workers received a prescreening questionnaire and were then randomly assigned to treatment and control groups. Those in the treatment group participated in job skills workshops in which participants 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.

Researchers who originally analyzed this experiment hypothesized that workshop attendance leads to better mental health and employment outcomes by enhancing participants' confidence in their ability to search for a job (Vinokur, Price, & Schul, 1995; Vinokur & Schul, 1997). In the JOBS II data, a continuous measure of job search self-efficacy represents this key mediating variable. The data also include baseline covariates measured before administering the treatment. The most important of these is the pretreatment level of depression, which is measured with the same methods as the continuous outcome variable. There are also several other covariates that are included in our analysis (as well as in the original analysis) to strengthen the validity of the key identifying assumption of causal mediation analysis. They include measures of education, income, race, marital status, age, sex, previous occupation, and the level of economic hardship.

Statistical Framework for Causal Mediation Analysis

In this section, we describe the counterfactual framework of causal inference, which is widely used in the statistical literature

[2] The web address is http://cran.r-project.org/web/packages/mediation.

(e.g., Holland, 1986) and is beginning to gain acceptance in psychology (e.g., Jo, 2008; Little & Yau, 1998; MacKinnon, 2008, Chapter 13; Schafer & Kang, 2008). Following prior work (e.g., Imai, Keele, & Yamamoto, 2010; Pearl, 2001; Robins & Greenland, 1992), we define causal mediation effects using the potential outcomes notation. We then review the key result of Imai, Keele, and Yamamoto (2010) and show a minimum set of the conditions under which the product of coefficients method (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002) and its variants yield valid estimates of causal mediation effects. Finally, we briefly explain how our approach differs from the existing approach based on the instrumental variable methods of Angrist, Imbens, and Rubin (1996). As we noted earlier, the strength of this framework is that it helps to clarify the assumptions needed for causal mediation effects without reference to specific statistical models.

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 that would be realized if the worker participates in the job training program and the other 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 (see later sections for an extension to nonbinary treatment). Then, we can use $Y_i(t)$ to denote the potential employment status that would result under the treatment status t. For example, $Y_i(1)$ measures worker i's employment status if the worker 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 $Y_i(1) - Y_i(0)$. 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 potential outcomes; formally, we write $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i$. When this is true, the average causal effect can be identified by the observed mean difference 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 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 research design. For example, analysts would want to ensure that participants in the experiment were not from the same household. The analyses that follow were conducted under this assumption, and the extension of our approach to the situation where the assumption is violated is left for future research.

Defining Causal Mediation Effects

In the statistics literature, the counterfactual framework and notation have been extended to define causal mediation effects. We relate this notation to 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 depression in which the mediating variable is workers' level of confidence in their ability to perform essential job search activities such as completing an employment application.

One possible hypothesis is that the participation in the job training program reduces the level of depression by increasing 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. 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, that is, $M_i = M_i(T_i)$. 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 define the potential outcomes. Previously, the potential outcomes were only a function of the treatment, but in a causal mediation analysis the potential outcomes depend on the mediator as well as the treatment variable. Therefore, 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 degree of depressive symptoms 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 observe only one of multiple potential outcomes, and the observed outcome Y_i equals $Y_i(T_i, M_i(T_i))$. Lastly, recall that no interference between units is assumed throughout; 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.

We now define causal mediation effects or indirect effects for each unit i as follows:

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

for $t = 0, 1$. Thus, the causal mediation effect represents the indirect effect of the treatment on the outcome through the mediating variable (Pearl, 2001; Robins, 2003; Robins & Greenland, 1992). The key to understanding Equation 1 is the following counterfactual question: What change would occur to the outcome if one changes the mediator from the value that would be realized under the control condition, $M_i(0)$, to the value that would be observed under the treatment condition, $M_i(1)$, while holding the treatment status at t? If the treatment has no effect on the mediator, that is, $M_i(1) = M_i(0)$, then the causal mediation effect is zero. 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 depression levels for worker i who participates in the training program. For this worker, $Y_i(1, M_i(1))$ equals an observed depression level if the worker actually participated in the program, whereas $Y_i(1, M_i(0))$ represents the depression level that would result if worker i participates but the mediator takes the value that would result under no participation. Similarly, $\delta_i(0)$ represents an impact worker i's depression level due to the change in the mediator induced by the participation in the program while suppressing the direct effect of program participation. Therefore, this definition formalizes, independent of any specific statistical models, the intuitive notion about mediation held by applied researchers that the treatment indirectly influences the outcome through the mediator.

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

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

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 depression level while holding the level of his or her job search self-efficacy constant at the level that would be realized under the program participation.[3] Then, the total effect of the treatment can be decomposed into the causal mediation and direct effects:

$$\tau_i \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \frac{1}{2}\sum_{t=0}^{1}\{\delta_i(t) + \zeta_i(t)\}.$$

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, that is, $\tau_i = \delta_i + \zeta_i$.

Finally, in causal mediation analysis, we are typically interested in the following average causal mediation effect:

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

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 and total effects as

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

and

$$\bar{\tau} \equiv \mathbb{E}(Y_i(1, M_i(1)) - Y_i(0, M_i(0))) = \frac{1}{2}\sum_{t=0}^{1}\{\bar{\delta}(t) + \bar{\zeta}(t)\},$$

respectively. As before, under the 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, that is, $\bar{\tau} = \bar{\delta} + \bar{\zeta}$, yielding the simple decomposition of the total effect into direct and indirect effects.

Note that the average total effect may be close to zero in some cases, but this does not necessarily imply that the average causal mediation effects are also small. It is possible that the average causal mediation and average direct effects have opposite signs and thus offset each other, yielding a small average total effect. In the context of program evaluation, this is an important circumstance because it implies that a policy can be improved by modifying it so that an effective mediator plays a larger role to increase its overall efficacy.

Sequential Ignorability Assumption

We now turn to the key assumption, which allows us to make valid inferences about the causal mediation effects defined above. The question is, what assumptions are needed to give the average mediation effect a causal interpretation? For randomized experiments, we only need to assume no interference between units to estimate the average treatment effect without bias. Causal mediation analysis, however, requires an additional assumption. In particular, we rely on the following assumption introduced by Imai, Keele, and Yamamoto (2010). Let X_i be a vector of the observed pretreatment confounders for unit i where $\mathcal{X}$ denotes the support of the distribution of X_i (i.e., the range of values X_i can take on). In the JOBS II data, X_i includes for each unemployed worker the pretreatment level of depressive symptoms as well as some demographic characteristics such as education, race, marital status, sex, previous occupation, and the level of economic hardship. Given these observed pretreatment confounders, the assumption can be formally written as

> *Assumption 1 (Sequential Ignorability; Imai, Keele, & Yamamoto, 2010):* We assume that the following two statements of conditional independence hold:
>
> $$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x, \tag{3}$$
>
> $$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid T_i = t, X_i = x, \tag{4}$$
>
> where $0 < \Pr(T_i = t \mid X_i = x)$ and $0 < p(M_i(t) = m \mid T_i = t, X_i = x)$ for $t = 0, 1$, and all $x \in \mathcal{X}$ and $m \in \mathcal{M}$.

Imai, Keele, and Yamamoto (2010) discussed how this assumption differs from those proposed in the prior literature. The main advantage of this assumption over other alternatives is its ease of interpretation. Assumption 1 is called sequential ignorability because two ignorability assumptions are made sequentially. First, given the observed pretreatment confounders, the treatment assignment is assumed to be ignorable, that is, 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 in which subjects may self-select into the treatment group. In such situations, a common strategy of empirical researchers is to collect as many pretreatment confounders as possible so that the ignorability of treatment assignment is more

[3] Pearl (2001) called $\zeta_i(t)$ a *natural direct effect* to distinguish it from a *controlled direct effect* of the treatment. Imai et al. (2009) argued that the former corresponds to causal mechanisms, whereas the latter represents the causal effect of direct manipulation. Imai et al. also discussed the implications of this distinction for experimental designs.

credible once the observed differences in these confounders between the treatment and control groups are appropriately adjusted.

The second part of Assumption 1 states that the mediator is ignorable given the observed treatment and pretreatment 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 pretreatment confounders. Unlike the ignorability of treatment assignment, however, the ignorability of the mediator may not hold even in randomized experiments. In the JOBS II study, for example, the randomization of the treatment assignment does not justify this second ignorability assumption because the posttreatment level of workers' job search self-efficacy is not randomly assigned by researchers. In other words, the ignorability of the mediator implies that among those workers who share the same treatment status and the same pretreatment characteristics, the mediator can be regarded as if it were randomized.

We emphasize that the second stage of sequential ignorability is a strong assumption and must be made with care. 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. Moreover, the conditioning set of covariates must be pretreatment variables. Indeed, without an additional assumption, we cannot condition on the posttreatment confounders even if such variables are observed by researchers (e.g., Avin, Shpitser, & Pearl, 2005). This means that similar to the ignorability of treatment assignment in observational studies, it is difficult to know for certain whether the ignorability of the mediator holds even after researchers collect as many pretreatment confounders as possible.

Such an assumption is often referred to as nonrefutable because it cannot be directly tested from the observed data (Manski, 2007). Thus, 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.

Nonparametric Identification Under Sequential Ignorability

We now turn to the issue 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. This result is important for three reasons. First, it suggests the possibility of constructing a general method of estimating the average treatment effect for outcome and mediating variables of any type and using any parametric or nonparametric models. Second, it implies that we may estimate causal mediation effects while imposing weaker assumptions about the correct functional form or distribution of the observed data. Third, nonparametric identification analysis reveals the key role of the sequential ignorability assumption irrespective of the statistical models used by researchers.

We first slightly generalize the nonparametric identification result of Imai, Keele, and Yamamoto (2010). The following result states that under Assumption 1 the distribution of any counterfactual outcome is identified.

> *Theorem 1 (Nonparametric Identification):* Under Assumption 1, we can identify
>
> $$f(Y_i(t, M_i(t')) \mid X_i = x) = \int_{\mathcal{M}} f(Y_i \mid M_i = m, T_i = t, X_i = x) dF_{M_i}(m \mid T_i = t', X_i = x),$$
>
> for any $x \in \mathcal{X}$ and $t, t' = 0, 1$.

The proof is a generalization of Theorem 1 of Imai, Keele, and Yamamoto (2010) and thus is omitted. Theorem 1 shows that under sequential ignorability, the distribution of the required potential outcome (i.e., the quantity in the left-hand side of the equation) can be expressed as a function of the distributions of the observed data, that is, the conditional distribution of M_i given (T_i, X_i) and that of Y_i given (M_i, T_i, X_i). Thus, the assumption lets us make inferences about the counterfactual quantities we do not observe (i.e., the potential outcomes and mediators of workers in the opposite treatment status) using the quantities we do observe (i.e., observed outcomes and mediators for workers in a particular treatment status). As we show next, in the LSEM framework, for example, these conditional distributions are given by a set of the linear regression models. Because Theorem 1 is not based on any specific model, however, it enables us to develop a general estimation procedure for causal mediation effects under various nonlinear conditions.

Causal Interpretation of the Product of Coefficients and Related Methods

Before turning to our general method, 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:

$$Y_i = \alpha_1 + \beta_1 T_i + \xi_1^{\top} X_i + \varepsilon_{i1}, \tag{5}$$

$$M_i = \alpha_2 + \beta_2 T_i + \xi_2^{\top} X_i + \varepsilon_{i2}, \tag{6}$$

$$Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^{\top} X_i + \varepsilon_{i3}. \tag{7}$$

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). Similarly, the difference of coefficient methods yields the numerically identical estimate by computing $\hat{\beta}_1 - \hat{\beta}_3$ in this linear case (MacKinnon et al., 2007, 2002). Because $\hat{\beta}_1 = \hat{\beta}_2\hat{\gamma} + \hat{\beta}_3$ and $\beta_1 = \beta_2\gamma + \beta_3$ always holds, Equation 5 is redundant given Equations 6 and 7.

Does the product of coefficients method yield a valid estimate for the causal mediation effect under the potential outcomes framework? Imai, Keele, and Yamamoto (2010) prove that under sequential ignorability and the additional no-interaction assumption, that is, $\bar{\delta}(1) = \bar{\delta}(0)$, 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 as long as the linearity assumption holds (see also Jo, 2008). To understand the connection between the product of coefficients method and the causal mediation effect defined earlier, we first write each potential outcome within the LSEM framework:

$$Y_i(T_i, M_i(T_i)) = \alpha_1 + \beta_1 T_i + \xi_1^\top X_i + \varepsilon_{i1}(T_i, M_i(T_i)),$$

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

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

An argument that is nearly identical to the proof of Theorem 2 in Imai, Keele, and Yamamoto shows that under Assumption 1 the average causal mediation and direct effects are identified as $\bar{\delta}(t) = \beta_2\gamma$ and $\bar{\zeta}(t) = \beta_3$, respectively, for $t = 0, 1$.[4]

Thus, as long as an analyst is willing to adopt the linearity and no-interaction assumptions along with sequential ignorability, the product of coefficients method provides a valid estimate of the causal mediation effect. However, Theorem 1 implies that only the sequential ignorability assumption is needed and neither the linearity nor the no-interaction assumption is required for the identification of causal mediation effects. To illustrate this point, we next explore relaxing the no-interaction assumption.

Relaxing the No-Interaction Assumption

Judd and Kenny (1981), and more recently Kraemer and colleagues (Kraemer, Kiernan, Essex, & Kupfer, 2008; Kraemer, Wilson, Fairburn, and Agras, 2002), proposed an alternative to the standard product of coefficients method by relaxing the no-interaction assumption. They argued that assuming no interaction between the treatment and the mediator is often unrealistic and replaced Equation 7 with the following alternative specification:

$$Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \kappa T_i M_i + \xi_3^\top X_i + \varepsilon_{i3}. \qquad (8)$$

Such an interaction might arise in the JOBS II study if, for example, the average mediation effect via the improvement of the workers' mental health depends on whether they receive the job training program. Kraemer et al. (2008) argued that in addition to $\hat{\beta}_2$, either $\hat{\gamma}$ or $\hat{\kappa}$ must be statistically ~~in~~distinguishable from zero in order to conclude that average mediation effects exist.

Although the inclusion of the interaction term, $T_i M_i$, is a reasonable suggestion, the proposed procedure can be improved so that the hypothesis test is conducted directly on the average causal mediation effects. Following Imai, Keele, and Yamamoto (2010), one can show that under Assumption 1, the average causal mediation effects are given by

$$\bar{\delta}(t) = \beta_2(\gamma + \kappa t), \qquad (9)$$

for $t = 0, 1$. Estimation of this quantity in the JOBS II example would give different mediation effects, taking into account an interaction between program participation and job search self-efficacy. In addition, the average direct and total effects are given by

$$\bar{\zeta}(t) = \beta_3 + \kappa\{\alpha_2 + \beta_2 t + \xi_2^\top \mathbb{E}(X_i)\}, \qquad (10)$$

$$\bar{\tau} = \beta_2\gamma + \beta_3 + \kappa\{\alpha_2 + \beta_2 + \xi_2^\top \mathbb{E}(X_i)\}, \qquad (11)$$

for $t = 0, 1$. The consistent estimates of $\bar{\zeta}(t)$ and $\bar{\tau}$ can be obtained by replacing the coefficients of Equations 10 and 11 with their least squares estimates and $\mathbb{E}(X_i)$ by the sample average of X_i, which we denote by $\bar{X}$.[5] In Appendix A, we derive the asymptotic variance for each of these quantities of interest.

Although analysts can easily relax the no-interaction assumption, extending the LSEM mediation framework to models for discrete outcomes, for example, is much more difficult. In contrast, as we demonstrate later, our general approach can handle nonlinear models such as logit and probit in a straightforward manner.

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. In particular, although the existence of unobserved confounders is allowed, the direct effect is assumed to be zero (see Appendix B for details). This means that the instrumental variables approach eliminates, a priori, alternative causal mechanisms. For this reason, we believe that it is less than ideal for the causal mediation analysis used in social science research. A more general and promising approach is the causal mediation analysis based on principal stratification (e.g., Gallop et al., 2009).

An Application to JOBS II

We illustrate the product of coefficients method and related approaches using the JOBS II data. Here the outcome variable is a measure of depression and the mediator is the level of worker's job search self-efficacy. Both measures range from one to five. To make sequential ignorability more credible, we include the full set of covariates described in the outline of JOBS II. Table 1 presents the estimated quantities of interest based on the product of coefficients and instrumental variables methods. First, we use the standard product of coefficients method, assuming no interaction between the mediator and the treatment. We find a small (but statistically significant at the 95% level) negative mediation effect (the first column). Because the average treatment effect on the mediator is negative, the results imply that the program participation on average decreases slightly the depressive symptoms by increasing the level of job search self-efficacy. The average direct and total effects are estimated to be negative as well, and their effect sizes are larger. However, these estimates are statistically indistinguishable from zero.

Next, we relax the no-interaction assumption by allowing the average causal mediation effect to depend on the treatment status. The second and third columns of the table present the results. The basic findings resemble the ones based on the standard product of coefficients method, and there is little evidence for the presence of

[4] The average total effect is given by β_1, which equals $\beta_2\gamma + \beta_3$.

[5] To estimate the average total effect, we may fit the following model, $Y_i = \alpha_1 + \beta_1 T_i + \xi_1^\top X_i + \eta^\top T_i X_i + \varepsilon_{i1}$, where the average total effect is given by $\bar{\tau} = \beta_1 + \eta^\top \mathbb{E}(X_i)$.

Table 1
Estimated Causal Effects of Interest Based on the Product of Coefficients and Instrumental Variables Methods Using the Job Search Intervention Study Data

	Product of coefficients			
		With interaction		
Average effect	No interaction	Under treatment ($t = 1$)	Under control ($t = 0$)	Instrumental variable
Mediation	−.016	−.014	−.021	−.047
$\bar{\delta}(t)$	[−.03, −.002]	[−.026, −.001]	[−.040, −.002]	[−.120, .024]
Direct	−.032	−.027	−.034	.000[a]
$\bar{\zeta}(t)$	[−.107, .046]	[−.115, .048]	[−.114, .046]	
Total	−.047	−.047		−.047
$\bar{\tau}$	[−.120, .023]	[−.120, .024]		[−.120, .023]

Note. The outcome variable is a measure of depression for each worker, and the mediator represents the level of workers' job search self-efficacy. Each cell shows a point estimate and its corresponding 95% confidence intervals. The average treatment effect on the mediator is estimated to be .100 with the 95% confidence interval [.018, .182] (based on the delta method). The results in Columns 2–4 are based on the product of coefficients method without and with the interaction between the treatment and the mediator.
[a] Assumed to be zero.

the interaction effects. Although these calculations were done with the formulas described in the previous sections, the simulation approach to estimation described subsequently and implemented in our software returns nearly identical results. This correspondence is important because it shows how our approach encompasses the product of coefficients method as a special case. In later sections, we extend our approach to nonlinear models. Finally, we apply the instrumental variables method, though the assumption of no direct effect is unlikely to hold in this application. Under this method, the average mediation effect equals the average total effect. Thus, the results are somewhat different from those based on the other two methods; the average causal mediation effect is estimated to be negative but is not statistically significantly different from zero.

Sensitivity Analysis

We now turn to the final and important component of our approach. One advantage of the potential outcomes framework is that it helps to clarify the role of key identification assumptions. As we demonstrated above, randomization of the treatment alone does not identify causal mediation effects. This means that even in randomized experiments, an additional assumption (e.g., sequential ignorability in our approach) is required for identification. As such, this assumption is of particular interest, because if it is not satisfied, the estimated quantity cannot be given a causal interpretation. In particular, the second part of Assumption 1 is nonrefutable in the sense that it cannot be directly tested with the observed data.

As in many applications, it is reasonable to think that the sequential ignorability may have been violated in the JOBS II study. For example, Jo (2008) pointed out that the second part of Assumption 1 might be violated by stating 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" (p. 317).

Sensitivity analysis is an effective method for probing the plausibility of a nonrefutable assumption and thus is an important element of our approach. The goal of sensitivity analysis is to quantify the degree to which the key identification assumption must be violated for a researcher's original conclusion to be reversed. If an inference is sensitive, a slight violation of the assumption may lead to substantively different conclusions. The degree of sensitivity can be calibrated either in comparison to other studies (Rosenbaum, 2002, p. 325) or in conjunction with expert opinion (e.g., White, Carpenter, Evans, & Schroter, 2007). Given the importance of sequential ignorability, we argue that a mediation analysis is not complete without a sensitivity analysis. Our software, mediation, also allows users to conduct sensitivity analyses with only a single additional line of syntax, as illustrated in Imai et al. (2010a).

The Linear Structural Equation Models

In the context of the standard LSEM framework, Imai, Keele, and Yamamoto (2010) proposed a sensitivity analysis for causal mediation analysis based on the correlation between the error for the mediation model, ϵ_{i2}, and the error for the outcome model, ϵ_{i3}. They denoted this correlation across the two error terms as ρ, which serves as the sensitivity parameter. Such a correlation can arise if there exist omitted variables that affect both mediator and outcome variables because these omitted variables will be part of the two error terms. Thus, under sequential ignorability, ρ equals zero, and nonzero values of ρ imply departures from the ignorability assumption. Imai et al. showed that it is possible to express the average causal mediation effect as a function of ρ and model parameters that can be consistently estimated even though ρ is nonzero.

With this fact, the proposed sensitivity analysis asks the question of how large does ρ have to be for the causal mediation effect to go away. If small departures from zero in ρ produce an average causal mediation effect that is substantively different from the estimate obtained under sequential ignorability, this suggests that the study is sensitive to the potential violation of the sequential ignorability assumption. Moreover, we can also observe whether the confidence interval for the mediation effect contains zero, which provides a more nuanced analysis.

Here we extend the result in Imai, Keele, and Yamamoto (2010) to the LSEM with the no-interaction assumption relaxed. As such, we derive the mediation effect as a function of ρ and other quantities that can be consistently estimated.

> *Theorem 2 (Identification With a Given Error Correlation):* Consider the LSEM defined in Equations 6 and 8. Suppose that Equation 3 of Assumption 1 holds but Equation 4 may not. Assume that the correlation between ϵ_{2i} and ϵ_{3i}, that is, ρ, is given (and is assumed to be constant across the treatment and control groups) where $-1 < \rho < 1$. Then, the average causal mediation effects are identified and given by
>
> $$\bar{\delta}(t) = \frac{\beta_2 \sigma_{1t}}{\sigma_{2t}} \left\{ \tilde{\rho}_t - \rho \sqrt{(1 - \tilde{\rho}_t^2)/(1 - \rho^2)} \right\},$$
>
> where $\sigma_{jt}^2 \equiv \mathrm{Var}(\epsilon_{ij} \mid T_i = t)$ and $\tilde{\rho}_t \equiv \mathrm{Corr}(\epsilon_{i1}, \epsilon_{i2} \mid T_i = t)$ for $j = 1, 2$ and $t = 0, 1$.

A proof is given in Appendix C. Theorem 2 establishes the linkage between the average causal mediation effect and the degree of correlation of the two error terms given the identifiable model parameters.[6] The iterative procedure described in Imai, Keele, and Yamamoto (2010) can be used to obtain the confidence intervals under various values of ρ.[7]

How small is small enough for researchers to conclude the resulting conclusion is valid? Unfortunately, there is no absolute threshold. In principle, the magnitude of ρ from one study can be interpreted only relative to the corresponding ρ of another study. However, Imai, Keele, and Yamamoto (2010) showed that the magnitude of ρ from one study can be interpreted independently via the coefficients of determination, or R^2. This alternative parameterization facilitates the interpretation because one can understand the influence of potential omitted variables in terms of its explanatory power. We discuss and extend this alternative parameterization later in this article in the context of nonlinear models.

Empirical Illustration

We return to the example with continuous outcome and mediator measures and ask whether the finding is sensitive to a potential violation of sequential ignorability. Here we relax the no-interaction assumption for the sensitivity analysis. We find $\bar{\delta}(1) = 0$ when ρ is equal to $-.165$ and $\bar{\delta}(0) = 0$ when ρ is $-.245$. Figure 1 graphically illustrates this point by plotting the estimated average mediation effects and their 95% confidence intervals as a function of ρ. We find that for $\bar{\delta}(0)$ the confidence intervals include zero for a ρ value of $-.09$ and for $\bar{\delta}(1)$ at $-.06$, which further underscores the sensitivity of the estimate. Imai, Keele, and Yamamoto (2010) found in another study that the mediation effects are zero for a ρ value of .48. Thus, the estimated mediation effects here are considerably more sensitive than in that study. This means that it would take a smaller unobserved confounder to overturn the conclusion obtained under sequential ignorability for the JOBS II study.

Generalization to Nonlinear Models

In this section, we show that the above methodology can be generalized to nonlinear models. The difficulty with the LSEM approach is that it does not readily extend to nonlinear models. For example, suppose that in the JOBS II study the outcome variable of interest is a binary measure about whether a subject is employed or not. The LSEM approach using the product of coefficients no longer generalizes in a straightforward way if researchers wish to use the logistic regression to model the outcome variable.

As we demonstrate below, our generalization can accommodate linear and nonlinear relationships, parametric and nonparametric models, continuous and discrete mediators, and various types of outcome variables (for related methods, see Glynn, 2008; Huang, Sivaganesan, Succop, & Goodman, 2004; VanderWeele, 2009; Wang & Taylor, 2002). This is possible because our methodology is not tied to specific statistical models. Using the nonparametric identification result of Theorem 1, we first develop two algorithms based on Monte Carlo simulation to estimate causal mediation effects that are applicable to any statistical models. We then demonstrate how these algorithms and sensitivity analysis can be applied to some frequently used nonlinear models.

The Estimation Algorithms

Recall that for each subject we observe $Y_i(T_i, M_i(T_i))$ but we need to infer the following counterfactual quantity: $Y_i(T_i, M_i(1 - T_i))$. Theorem 1 suggests that researchers can obtain one Monte Carlo draw of the potential outcome $Y_i(t, M_i(t'))$ for any t, t' using model predictions given the subject's pretreatment covariates $X_i = x$. To do this, we first sample $M_i(t')$ from the selected mediator model, $f(M_i \mid T_i = t', X_i = x)$, and then given this draw of the mediator, sample $Y_i(t, M_i(t'))$ from the outcome model, $f(Y_i \mid T_i = t, M_i(t'), X_i = x)$. Our nonparametric identification result implies that this procedure need not change regardless of statistical models used for the mediator and the outcome. Once we obtain these Monte Carlo draws of potential outcomes, we can compute the relevant quantities of interest that are functions of these potential outcomes.[8]

This observation leads to the following two general algorithms that can accommodate many situations researchers encounter in practice. First, we describe an algorithm for parametric inference in which parametric models (e.g., probit or logit) are specified for the mediator and the outcome variable. To make the exposition

[6] Thus, for $t = 0, 1$, $\bar{\delta}(t)$ becomes zero when ρ is equal to the correlation between ϵ_{i1} and ϵ_{i2} among those with $T_i = t$, which is denoted by $\tilde{\rho}_t$ and can be estimated by the sample correlation of the corresponding residuals.

[7] This use of residuals differs from the structural equation modeling tradition in that residuals are not being used to access the fit of statistical models as discussed in Mulaik (2007). Instead, they are used to assess the violation of the key identification assumption. Although others have pointed to assumptions about residuals (Bullock, Green, & Ha, 2010; MacKinnon et al., 2002; Mulaik, 2009) and the consequences of omitted variables (James et al., 1982, pp. 71–80), to our knowledge these scholars propose no formal sensitivity analysis about the identifying assumption. See VanderWeele (2010) for an alternative approach.

[8] This is true as long as they do not involve the joint distribution of $Y_i(t, M_i(1))$ and $Y_i(t, M_i(0))$ because only marginal distributions are identified under sequential ignorability.

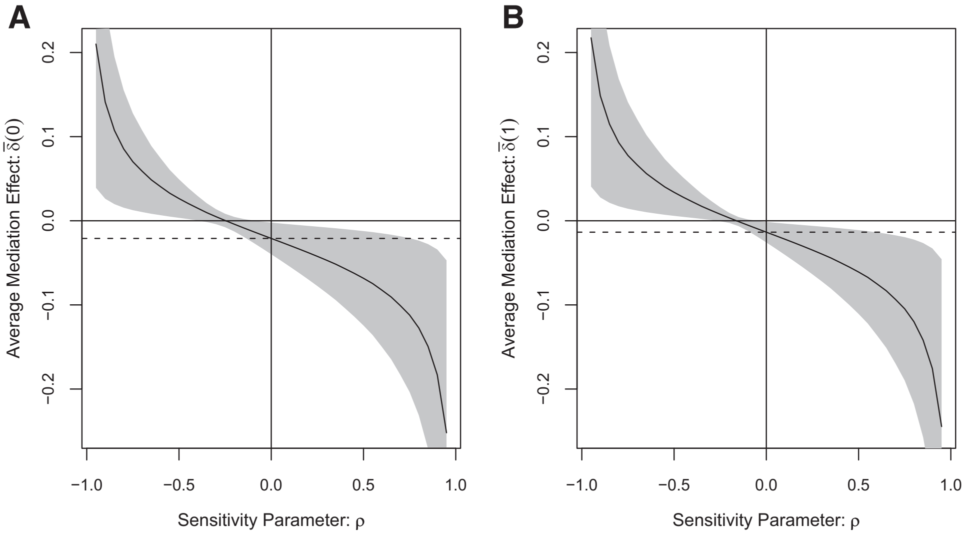

Figure 1. Sensitivity analysis with continuous outcome and mediator. Figure 1A is for the estimated average mediation effect under the control, and Figure 1B is for treatment. The dashed line represents the estimated mediation effect for $\rho = 0$. The gray areas represent the 95% confidence interval for the mediation effects at each value of ρ. The solid line represents the estimated average mediation effect at different values of ρ.

concrete, we describe the algorithm to estimate the average causal mediation effects. The proposed algorithm is based on the quasi-Bayesian Monte Carlo approximation of King, Tomz, and Wittenberg (2000), in which the posterior distribution of quantities of interest is approximated by their sampling distribution. A similar idea has been used for specific models (see Bauer, Preacher, & Gil, 2006; MacKinnon, Lockwood, & Williams, 2004). Our algorithm, in contrast, applies to any parametric statistical model of researchers' choice (see below for an algorithm that can be applied to nonparametric models). Here we provide an outline of the algorithms (their details are given in Appendix D).

> *Algorithm 1 (Parametric Inference):* This algorithm can be used for any parametric model.
>
> Step 1. Fit models for the observed outcome and mediator variables.
>
> Step 2. Simulate model parameters from their sampling distribution.
>
> Step 3. Repeat the following three steps: (a) simulate the potential values of the mediator, (b) simulate the potential outcomes given the simulated values of the mediator, (c) compute the causal mediation effects.
>
> Step 4. Compute summary statistics such as point estimates and confidence intervals.

The generality of this algorithm is attractive because the same algorithm can be applied to any parametric statistical model. This allows us to develop easy-to-use statistical software that computes point and uncertainty estimates under various statistical models. Imai et al. (2010a) illustrated the use of this software with the exact syntax used to produce the empirical results of this article. In these examples, we find that 1,000 simulations for each step (the default number of simulations in the software) are sufficient. However, for other applications with a smaller data set and/or more complex models, a greater number of simulations may be required.

Finally, our nonparametric identification result also permits the use of more complex models such as non- or semiparametric models and quantile regression models. For these models, we propose using a nonparametric bootstrap procedure to obtain a distribution of causal mediation effects. Although this algorithm is applicable to parametric inference as well, Algorithm 1 is typically much more computationally efficient. Each step of Algorithm 2 corresponds to that of Algorithm 1. Again, we outline the proposed algorithm here and provide the details in Appendix D.

> *Algorithm 2 (Nonparametric Inference):* The following algorithm is applicable to both parametric and nonparametric models.
>
> Step 1. For each of the bootstrapped samples, repeat the following steps: (a) fit models for the observed and mediator variables, (b) simulate the potential values of the mediator, (c) simulate the potential outcomes given the simulated values of the mediator, (d) compute the causal mediation effects.
>
> Step 2. Compute summary statistics such as point estimates and confidence intervals.

We evaluate our algorithms using two sets of simulations with varying sample sizes. In one set of simulations, we use a population model in which the mediator has a nonlinear effect on the outcome. We then compare the performance of a semiparametric model with that of a linear model with and without a quadratic term of the mediator. In the other set of simulations, we consider the case in which the mediator is continuous but the outcome is binary. We examine the performance of the probit model against the product of coefficients method. In both simulations, as expected, we find that our methods recover the population parameters with little bias. The results and details of the simulations are reported in Appendix E.

Next, we illustrate how our methods can handle a variety of situations that often arise in causal mediation analysis by applying them to the JOBS II study. This helps illustrate the generality of our methods. We also show theoretically how our methods relate to the existing approaches in the literature.

Quantile Causal Mediation Effects

We have shown that the methods based on the LSEM framework provide the average causal mediation effects under sequential ignorability. However, in some cases, researchers may be interested in distributional features of the outcome variable other than the average. In the JOBS II example, policymakers might be concerned about individuals with high levels of depression rather than those with the average level of depression. It might also be true that a few individuals respond to the intervention in dramatic fashion, making the average a poor description of how these individuals respond to the treatment. In such instances, *quantile causal mediation effects,* which represent the difference between a certain quantile (e.g., median) of two relevant potential outcomes, may be of interest.[9]

In this context, quantile regression allows for a convenient way to model the quantiles of the outcome distribution while adjusting for a variety of covariates (Koenker, 2005). Specifically, we replace Equation 7 with the quantile regression model. The usual product of coefficients method is not applicable here, but Algorithm 2 can be used.

Using the JOBS II data, we examine whether the job training program directly affected subjects' levels of depression and whether job search self-efficacy mediated the relationship between the outcome and the treatment. Figure 2 presents the estimated quantile causal mediation and direct effects and their 95% confidence intervals under Algorithm 2. Both direct and indirect effects in the figure are averaged over the observed pretreatment covariates included in the quantile regression models. Figure 2A demonstrates the effect of the intervention that occurs through the mediator job search self-efficacy. Figure 2B shows how the intervention affects quantiles of depression directly. One can see that the magnitude of the estimated mediation effects increases slightly as one moves from lower to higher quantiles, but the change is small, implying that the effects are relatively constant across the distribution. In contrast, the estimated direct effects vary substantially across the quantiles, although the confidence intervals are wide and always include zero.

Nonparametric and Semiparametric Regressions

In the LSEM framework, the estimation of the average causal mediation effects is based on a set of linear regressions. How might one relax the linearity assumption? In the JOBS II study, the change in depression mediated by increased job search self-efficacy may be very small among those with high levels of job search self-efficacy. For these subjects, the program participation is unlikely to further increase the mediating effects because of a diminishing effect of the treatment on the mediator. Conversely, mediation effects might be smallest among those with low job search skills, as they are unable to overcome societal and institutional thresholds that reinforce levels of depression.

Instead of assuming linear relationships between variables, non- or semiparametric regressions may be used to avoid linear functional form assumptions (e.g., Keele, 2008). These models attempt to recover the true relationship from the data while imposing much weaker functional form assumptions. Although one could use a quadratic term in the LSEM framework, it is well known that such transformation often provides a poor approximation in practice (see the simulation in Appendix F). Moreover, the product of coefficients method no longer applies to these and more complex situations. With Algorithm 2, however, the analyst can use nonparametric or semiparametric regression models, and the causal mediation effects can be easily estimated.

As an illustration, we allow the mediator to have a nonlinear effect on the outcome by applying a generalized additive model (GAM) to estimate the average causal mediation effects (Hastie & Tibshirani, 1990). We fit the following regression equation instead of Equation 7:

$$Y_i = \alpha_3 + \beta_3 T_i + s(M_i) + \xi^\top X_i + \varepsilon_{i3}, \tag{12}$$

where $s(\cdot)$ is a smooth and possibly nonlinear function that we estimate nonparametrically from the data (e.g., a spline). In the LSEM framework $s(\cdot)$ is assumed to be a linear function. We also relax the no-interaction assumption by fitting the following model, $Y_i = \alpha_3 + \beta_3 T_i + s_0(M_i)(1 - T_i) + s_1(M_i)T_i + \xi^\top X_i + \varepsilon_{i3}$, instead of Equation 8. For the semiparametric regression model, we used the package mgcv in R (Wood, 2006).

Figure 3 plots the estimated nonlinear relationship between the expected level of depression (the average outcome) and the level of job search self-efficacy (the mediator) with and without the interaction between the treatment and the mediator. Figure 3A (the no-interaction model) shows that there is a mild threshold effect between job search self-efficacy and depression. That is, self-efficacy must exceed the midpoint of the scale before there is any attendant decrease in depression. Figures 3B and 3C plot the estimated nonlinear relationships for the control and treatment groups separately under the model with the interaction. For both groups, the estimation is somewhat imprecise but is consistent with Figures 3B and 3C in that there is a negative relationship at higher levels of the mediator. For the treatment group, the pattern closely mirrors that observed in the no-interaction model, though the width of the confidence intervals must temper any decisive conclusions.

Table 2 presents the estimated average causal mediation effects based on the GAM and Algorithm 2 with 10,000 bootstrap resamples. First, we assume no interaction between the mediator and the treatment. Under this model (in the left column), we find a

[9] Formally, α-quantile causal mediation effects are defined as $\tilde{\delta}_\alpha(t) \equiv q_{t1}(\alpha) - q_{t0}(\alpha)$, for $t = 0, 1$ and $0 < \alpha < 1$, where $q_{tt'}(\alpha) \equiv \inf\{y; F(Y_i(t, M_i(t')) \leq y) \geq \alpha\}$ is the quantile function for the distribution of $Y_i(t, M_i(t'))$. Similarly, we can define quantile direct and total effects as $\tilde{\zeta}_\alpha(t) \equiv q_{1t}(\alpha) - q_{0t}(\alpha)$.

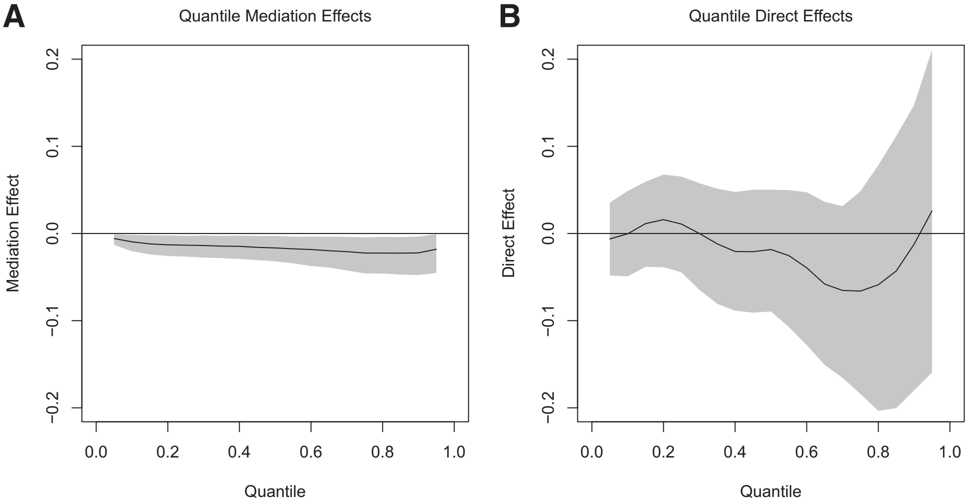

Figure 2. Estimated quantile causal mediation and direct effects based on the JOBS II data. The outcome variable is a measure of depression for each worker, and the mediator represents the level of workers' job search self-efficacy. The figure presents the estimated quantile mediation effect (A) and the estimated quantile direct effect (B), along with 95% confidence intervals. The quantile mediation effects are estimated to be negative and statistically significant, whereas the quantile direct effects have much wider confidence intervals.

small, but statistically significant, negative mediation effect. This effect is estimated to be slightly larger in magnitude compared with that in Table 1, though the difference is not statistically significant. The second and third columns of the table present the results without the no-interaction assumption. As in Table 1, the difference between the mediation effect for the treatment and control groups is small and is not statistically significant. Finally, as in Table 1, the estimated average direct and total effects for each specification are not statistically distinguishable from zero. In general, we find that modeling nonlinearity in the relationship does little to change our inference with the JOBS data, but the larger point is that Algorithm 2 allows us to relax basic model assumptions and still produce well-defined direct, mediation, and total effects under sequential ignorability.

Discrete Mediator

Next, consider the situation in which the mediator is discrete—a common occurrence for many applications in psychology in which the measure for the mediator is often an ordered scale or binary. In this case, Algorithm 1 can be simplified.[10] Modeling the mediator with either a probit–logit or ordered probit–logit model allows for straightforward parametric adjustment of pretreatment covariates. With these models, the proposed algorithms will provide the estimates of the average causal mediation effects and their estimation uncertainty. This is an important area of application because discrete and binary measures are extremely common, and yet the standard methods based on the LSEM framework are not directly applicable.

We demonstrate the flexibility of Algorithm 1 using the JOBS II data. The mediating variable in the original study, job search self-efficacy, is a continuous scale, as we noted previously. For demonstration purposes, we recoded the worker's job search self-efficacy into two discrete measures. In the first measure, we recoded the measure to be binary by splitting responses at the sample median. In a second measure, we recoded the scale to be a four-category ordered variable. Otherwise, we use the same set of variables as in Table 1.

Here we perform two analyses. In the first, we model the binary mediator with a probit model and estimate the average causal mediation effects with and without the no-interaction assumption. Table 3 presents the results that are largely consistent with the prior analysis when the mediator was measured with a continuous scale. One can see that the treatment decreased depression by increasing job search self-efficacy, but it had little direct causal effect. One can also see little differences in the average causal mediation effect across treatment status.[11]

Next, we use the four-category measure for the mediator and fit an ordered probit model for the mediator. Table 4 presents the

[10] First note that in this case Theorem 1 reduces to $h(Y_i\,(t, M_i(t')) \mid X_i = x) = \sum_{m \in \mathcal{M}} g(Y_i \mid M_i = m, T_i = t, X_i = x)\Pr(M_i = m \mid T_i = t', X_i = x)$. When the mediator is binary and the quantity of interest is the average causal mediation effect, the equation simplifies further to $\bar{\delta}(t) = \{\mathbb{E}(Y_i \mid M_i = 1, T_i = t) - \mathbb{E}(Y_i \mid M_i = 0, T_i = t)\}\,\{\Pr(M_i = 1 \mid T_i = 1) - \Pr(M_i = 1 \mid T_i = 0)\}$, which equals the expression derived by Li et al. (2007). If the support of the mediator distribution, $\mathcal{M}$, is bounded, then Step 3 of Algorithm 1 can be done with the following single calculation: For each $j = 1, 2, \ldots, J$ draw from the model sampling distributions, compute

$$\bar{\delta}^{(j)}(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{m \in \mathcal{M}} \mathbb{E}^{(j)}(Y_i \mid t, m, X_i) \times \{f^{(j)}(M_i = m \mid T_i = 1, X_i) - f^{(j)}(M_i = m \mid T_i = 0, X_i)\}.$$

Thus, we can complete Step 3 without sampling either $M_i(t)$ or $Y_i(t, M_i(t'))$.

[11] Nonlinear models with no interaction term can produce numerical estimates that differ for $\bar{\delta}(0)$ and $\bar{\delta}(1)$ due to the nonlinear link function. For simplicity we report $\bar{\delta}(1)$ for these models.

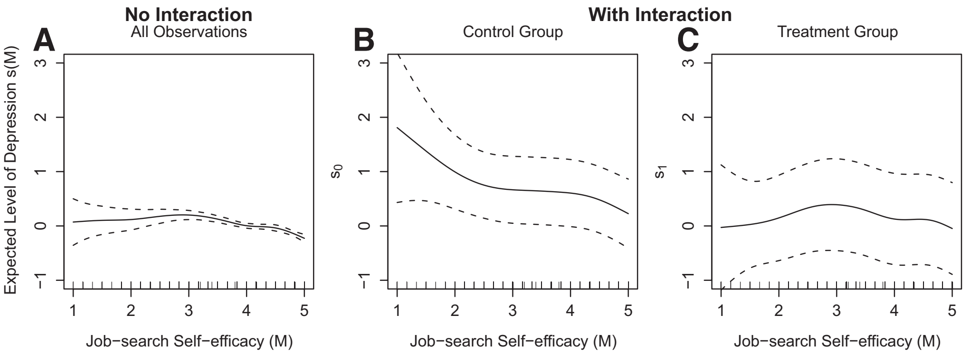

Figure 3. Generalized additive models and estimated nonlinear relationships between the mediator (job search self-efficacy) and the outcome (depression level). Figure 3A assumes no interaction between the treatment and the mediator, whereas Figures 3B and 3C allow for the interaction. The solid lines represent the estimated nonlinear relationships between the mediator (the horizontal axis) and the expected outcome (the vertical axis). Ninety-five percent confidence intervals (dashed lines) are based on nonparametric bootstrap. With the no-interaction assumption, changes in job search self-efficacy at lower levels have little effect on depression, though these effects are imprecisely estimated. Changes in the mediator have a negative effect at higher levels and are precisely estimated. When the interaction is allowed, the effect of the mediator is steadily decreasing in the control group, whereas for the treatment group the effect stays relatively constant.

results. First, we assume no interaction between the mediator and the treatment. In line with the results given in Table 1, there is a small negative average mediation effect. We also relax the no-interaction assumption by including an interaction term, T_iM_i, in the outcome regression model. The second and third columns of the table present the results. As before, there is little evidence of an interaction effect. We have shown that altering the model for the mediator presents no complications for Algorithm 1, which enables the estimation of the quantities of interest and provides estimates of statistical uncertainty.

Binary Outcome

One situation that has attracted the attention of many researchers is the case with the binary outcome and the continuous mediator. Many approaches have been proposed for such situations (e.g., Ditlevsen, Christensen, Lynch, Damsgaard, & Keiding, 2005; Freedman, Graubard, & Schatzkin, 1992; MacKinnon, 2008; MacKinnon et al., 2007, 2002; Wang & Taylor, 2002). One important criticism of existing methods is that they lack a causal interpretation (Kaufman, MacLehose, Kaufman, & Greenland, 2005). Here we derive analytical expressions for causal mediation effects when the outcome is binary. We show that our general approach can easily accommodate binary outcomes and examine the exact relationship between our proposed method and some of the existing approaches.

Analytical expressions for the average causal mediation effects. For the sake of notational and algebraic simplicity, we consider the following simple model without the pretreatment confounders. In Appendix F, we show that all our analytical results will hold for the model with the observed pretreatment covariates with some notational complexity. The model is given by

Table 2
Estimated Causal Quantities of Interest Based on the Generalized Additive Model

		With interaction	
Average effect	No interaction	Under treatment ($t = 1$)	Under control ($t = 0$)
Mediation	−.022	−.021	−.025
$\bar{\delta}(t)$	[−.041, −.004]	[−.042, −.004]	[−.048, −.004]
Direct	−.022	−.012	−.015
$\bar{\zeta}(t)$	[−.093, .048]	[−.081, .058]	[−.085, .055]
Total	−.044	−.037	
$\bar{\tau}$	[−.116, .028]	[−.111, .036]	

Note. The setup is identical to that of Table 1 except that the generalized additive model is used to model the outcome variable. Ninety-five percent confidence intervals are based on nonparametric bootstrap. The second column displays the results under the no-interaction assumption, whereas the third and fourth columns display the results without this assumption.

Table 3
Estimated Causal Quantities of Interest With the Binary Mediator

		With interaction	
Average effect	No interaction	Under treatment ($t = 1$)	Under control ($t = 0$)
Mediation	−.019	−.019	−.018
$\bar{\delta}(t)$	[−.033, −0.007]	[−.035, −.006]	[−.027, −.005]
Direct	−.026	−.031	−.029
$\bar{\zeta}(t)$	[−.098, .045]	[−.096, .040]	[−.099, −.039]
Total	−.045	−.048	
$\bar{\tau}$	[−.117, .027]	[−.118, .022]	

Note. The setup is identical to that of Table 1 except that a probit model is used to model the mediator. Ninety-five percent confidence intervals are based on Algorithm 1 with 1,000 Monte Carlo draws. The second column displays the results under the no-interaction assumption, whereas the third and fourth columns display the results without this assumption.

Table 4
Estimated Causal Quantities of Interest With the Discrete Mediator

Average effect	No interaction	With interaction Under treatment (t = 1)	 Under control (t = 0)
Mediation	−.013	−.017	−.011
$\bar{\delta}(t)$	[−.029, .003]	[−.040, .004]	[−.027, .004]
Direct	−.032	−.029	−.036
$\bar{\zeta}(t)$	[−.098, .037]	[−.104, .045]	[−.111, .036]
Total	−.044	−.047	
$\bar{\tau}$	[−.115, .028]	[−.122, .028]	

Note. The setup is identical to that of Table 1 except that an ordered probit model is used to model the mediator. Ninety-five percent confidence intervals are based on Algorithm 1 with 1,000 Monte Carlo draws. The second column displays the results under the no-interaction assumption, whereas the third and fourth columns display the results without this assumption.

$$M_i = \alpha_2 + \beta_2 T_i + \epsilon_{2i}, \tag{13}$$

$$Y_i = \mathbf{1}\{Y_i^* > 0\} \text{ where } Y_i^* = \alpha_3 + \beta_3 T_i + \gamma M_i + \epsilon_{3i}, \tag{14}$$

where ϵ_{2i} and ϵ_{3i} are independently and identically distributed (i.i.d.) random variables with zero mean and $\mathrm{Var}(\epsilon_{2i}) = \sigma_2^2$ and $\mathrm{Var}(\epsilon_{3i}) = \sigma_3^2$. The observed outcome variable Y_i is equal to 1 or 0 depending on whether the value of the latent variable Y_i^* is greater than zero. Note that under Assumption 1, we have the independence between two error terms. If ϵ_{2i} is an i.i.d. standard normal (logistic) random variate, then the model for the outcome variable is a probit (logistic) regression. Although a more complicated model (such as a model with interactions and a nonparametric model) is easily used within our general framework, this simple model establishes the clear relationship between our approach and the existing methods.

Here we focus on the estimation of the average causal mediation effects and briefly state the analytical results for the interested reader. The important point is that our general approach readily incorporates the calculation of causal mediation effects with binary outcome variables. Given the above model and Assumption 1, we can derive the analytical expression for the average causal mediation effects (see Appendix F for a proof). The functional form differs, depending on whether a probit or logit model is fitted to the outcome measure. This is due to the fact that different nonlinear link functions are used in each of these models.

First, consider a logit model. This implies that ϵ_{3i} is an i.i.d. logistic random variable and $\varepsilon_{2i} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_2^2)$. The average causal mediation effects can be written as

$$\bar{\delta}(t) = H(\alpha_3 + \beta_3 t + \gamma(\alpha_2 + \beta_2)) - H(\alpha_3 + \beta_3 t + \gamma\alpha_2), \tag{15}$$

where $H(\cdot)$ represents the following distribution function for $\epsilon_{1i} = \gamma\epsilon_{2i} + \epsilon_{3i}$:

$$H(\varepsilon_{1i}) = \int_{-\infty}^{\infty} \Phi\left(\frac{\varepsilon_{1i} - \varepsilon_{3i}}{\gamma\sigma}\right) \frac{\exp(\varepsilon_{3i})}{\{1 + \exp(\varepsilon_{3i})\}^2}\, d\varepsilon_{3i}.$$

This quantity can be computed with a standard numerical integration technique. The average total effect, on the other hand, equals $\bar{\tau} = H(\alpha_3 + \beta_3 + \gamma(\alpha_2 + \beta_2)) - H(\alpha_3 + \gamma\alpha_2)$.

Next, suppose a probit model is used for the outcome variable and hence $\varepsilon_{3i} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and $\varepsilon_{2i} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_2^2)$. Then, a similar calculation yields

$$\bar{\delta}(t) = \Phi\left(\frac{\alpha_3 + \beta_3 t + \gamma(\alpha_2 + \beta_2)}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right) - \Phi\left(\frac{\alpha_3 + \beta_3 t + \gamma\alpha_2}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right), \tag{16}$$

for t = 0, 1 and where $\Phi(\cdot)$ represents the standard normal cumulative density function. As in the LSEM case, we obtain the average total effect by fitting the probit model,

$$\Pr(Y_i = 1 \mid T_i) = \Phi(\alpha_1 + \beta_1 T_i), \tag{17}$$

where $\alpha_1 = (\alpha_3 + \gamma\alpha_2)/\sqrt{\sigma_2^2\gamma^2 + 1}$ and $\beta_1 = (\gamma\beta_2 + \beta_3)/\sqrt{\sigma_2^2\gamma^2 + 1}$. Then, the average total effect is given by $\bar{\tau} = \Phi(\alpha_1 + \beta_1) - \Phi(\alpha_1)$.

Analytical investigation of the relationship with the existing methods. Currently, there are two commonly used methods for computing average mediation effects with binary outcomes. First, Freedman et al. (1992) suggested the difference of coefficients method, which is based on the difference between β_1 from Equation 17 and β_3 from Equation 14. Second, MacKinnon et al. (2007) advocated the product of coefficients method in which the γ parameter from Equation 14 is multiplied by β_2 from Equation 13. There exist SPSS and SAS macros (SOBEL and INDIRECT) that implement the product of coefficients method for the binary outcome.

Unfortunately, as MacKinnon et al. (2007) correctly pointed out, because probit and logistic regressions are nonlinear models, the difference of coefficients and product of coefficients methods give different estimates. Indeed, the nonlinearity of those models implies that unlike in the case of the LSEM, neither of the two methods consistently estimates the average causal mediation effects given in Equations 15 and 16. In Appendix E, we show via a simulation study that the bias of product of coefficients method can be substantial in certain situations. This illustrates the important advantage of our proposed approach over the existing methods.

Proportion mediated. Another quantity of interest considered in the literature is the proportion mediated, or the magnitude of the average causal mediation effects relative to the average total effect. Using our notation, we can define this quantity as

$$\upsilon \equiv \frac{\{\bar{\delta}(0) + \bar{\delta}(1)\}/2}{\tau}, \tag{18}$$

which is the average causal mediation effect divided by the average total effect. Clearly, this quantity makes sense only when the sign of the sum of the two average causal mediation effects (i.e., the numerator) is the same as the sign of the average total effect (i.e., the denominator). In the literature, the following measure of the proportion mediated has been proposed (e.g., Ditlevsen et al., 2005; Freedman & Graubard, 1992; MacKinnon et al., 2007):

$$\tilde{\upsilon} \equiv \frac{\gamma\beta_2}{\gamma\beta_2 + \beta_3}. \tag{19}$$

It can be shown that $\tilde{\upsilon}$ is equal to $\gamma\beta_2/\{\beta_1\sqrt{\gamma^2\sigma_2^2 + 1}\}$ when the probit regression is used to model the outcome and is equal to $\gamma\beta_2/\{\tilde{\beta}_1\sqrt{\gamma^2\pi^2/3 + 1}\}$ when the logistic regression of the form, $\text{logit}^{-1}(\Pr(Y_i = 1 \mid T_i)) = \tilde{\alpha}_1 + \tilde{\beta}_1 T_i$, is used.

As shown by Wang and Taylor (2002), $\tilde{\upsilon}$ is a valid measure of the proportion mediated on the latent variable scale (logit or probit), that is, Y_i^*. However, it does not generally equal υ, which represents the relative magnitude of the average causal mediation effect with respect to the average causal effect. Nevertheless, it is interesting to note that as shown in Appendix G, when the direct effect is small, $\tilde{\upsilon}$ may approximate υ well whether or not the logistic or probit regression is used and whether or not the pretreatment covariates are included in the model. In Appendix G, we also empirically examine the magnitude of bias for the existing measure $\tilde{\upsilon}$ using the JOBS II study.

Empirical illustration. In the JOBS II study, a key question of interest is whether the program participation leads to a better employment outcome by increasing job search self-efficacy. Here we use our approach to estimate the average causal mediation effects when the outcome is a binary variable, indicating whether subjects were working more than 20 hr a week 6 months after the job training program. We use the model with and without the interaction term and include the same set of pretreatment covariates to bolster the credibility of sequential ignorability.

The results are given in Table 5. Under the no-interaction assumption, unlike what was observed for the depression outcome, the estimated average mediation effect is small, and the 95% confidence interval contains zero. The estimated average direct effect is larger than the estimated average mediation effect but is not statistically significant, either. As a result, the estimated proportion mediated is a mere 6%. A statistical test of the no-interaction assumption reveals that the mediation effect does not vary across levels of the treatment, and the results mirror those estimated under the no-interaction assumption. As illustrated in this example, our algorithms easily extend to the binary outcome case and provide a principled method for estimating causal quantities of interest along with their statistical uncertainty.

Table 5
Estimated Causal Quantities of Interest With a Binary Outcome

		With interaction	
Average effect	No interaction	Under treatment (t = 1)	Under control (t = 0)
Mediation	.004	.003	.007
$\bar{\delta}(t)$	[−.001, .012]	[−.002, .020]	[−.001, .020]
Direct	.057	.054	.059
$\bar{\zeta}(t)$	[−.008, .124]	[−.009, .118]	[−.005, .121]
Total	.061	.062	
$\bar{\tau}$	[−.006, .128]	[−.003, .125]	
Proportion mediated	.058	.072	
$\bar{\upsilon}$	[−.104, .300]	[−.145, .402]	

Note. Outcome is whether a respondent was working more than 20 hr per week after the training sessions. Ninety-five percent confidence intervals are based on Algorithm 2 with 1,000 resamples. The second column displays the results under the no-interaction assumption, whereas the third and fourth columns display the results without this assumption.

Nonbinary Treatment

Our proposed approach so far has assumed that the treatment variable is binary. Similarly, other research tends to consider the cases of binary treatment variables, regardless of whether it is based on the potential outcomes framework (e.g., Albert, 2008; Jo, 2008) or not (e.g., Ditlevsen et al., 2005; MacKinnon et al., 2007; Wang & Taylor, 2002). Fortunately, our approach can be extended to the case of nonbinary treatment only at the cost of notational complexity. For example, the causal mediation effects can be defined for any two levels of the treatment:

$$\delta_i(t;\, t_1, t_0) \equiv Y_i(t, M_i(t_1)) - Y_i(t, M_i(t_0)), \tag{20}$$

where $t_1 \neq t_0$. This equals the definition given in Equation 1 when $t_1 = 1$ and $t_0 = 0$. The corresponding average causal mediation effect is defined as $\bar{\delta}(t;\, t_1, t_0) \equiv \mathbb{E}(\delta_i(t;\, t_1, t_0))$.

Because the values of t, t_0, and t_1 need to be selected in order to compute $\bar{\delta}(t;\, t_1, t_0)$, a more comprehensive approach would be to choose the baseline treatment level ($t_0 = 0$) first and then compute the average causal mediation effects for different treatment values with respect to this baseline. One can further plot the estimated value of the average causal mediation effects (relative to the baseline treatment level, say, $t_0 = 0$) averaged over the distribution of the observed treatment, that is, $\int \bar{\delta}(t;\, t_1, 0)\, dF_{T_i}(t)$, against each chosen value of t_1 in order to investigate how the average causal mediation effects change as the function of the treatment intensity. Because the validity of Theorem 1 does not depend on the distribution of the treatment, the algorithms presented previously can be used to make inferences about this and other quantities of interest.

Sensitivity Analysis for Nonlinear Models

Finally, we show that sensitivity analysis discussed earlier can also be generalized to some of the most commonly used nonlinear models. Unlike our estimation algorithms, sensitivity analysis must be developed in the context of a specific statistical model. However, the same idea is applicable across models, and here we show that the extension is possible for the cases of binary mediator and binary outcome variables. We also discuss an alternative parameterization, which can facilitate the interpretation of sensitivity analysis.

Binary Mediator

First, we extend the sensitivity analysis to the situation in which the mediator is binary. We consider the latent binary variable model defined by

$$M_i = \mathbf{1}\{M_i^* > 0\} \text{ where } M_i^* = \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \varepsilon_{i2},$$

where $\varepsilon_{i2} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and the outcome is continuous and defined as in Equation 8 and hence an interaction between the treatment and the mediator is allowed. The process, though more complex in its mechanics, is essentially identical to the case for the LSEM. We simply derive the mediation effects as a function of ρ and other quantities that can be consistently estimated when ρ is nonzero. We assume $(\epsilon_{i2}, \epsilon_{i3})$ are i.i.d. bivariate normal random variables with mean zero and covariance $\rho\sigma_3$, and ρ remains the correlation between the two error terms. As before, we can identify the expectation and variance of the outcome variable when ρ is nonzero. Furthermore, the additional model parameters necessary for calculating the average causal mediation effect can be estimated for each value of ρ.[12]

We now demonstrate the sensitivity analysis for the binary mediator example analyzed earlier. We ask how sensitive this estimate is to the possible existence of an unobserved confounder that might explain the association between the mediating variable and the outcome. For the sensitivity analysis, we estimate the average causal mediation effect, that is, $\{\bar{\delta}(1) + \bar{\delta}(0)\}/2$, under a series of ρ values and use Algorithm 1 to compute 95% confidence intervals. Figure 4 presents the result. We find that the estimated average mediation effect is zero when $\rho = -.24$ and that the 95% confidence interval contains zero for ρ greater than $-.09$.

Binary Outcome

Finally, we extend the sensitivity analysis to binary outcomes. Here we assume a mediator model as defined by Equation 6 and outcome model defined as

$$Y_i = \mathbf{1}\{Y_i^* > 0\} \text{ where } Y_i^* = \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \varepsilon_{i3}, \tag{21}$$

where $\varepsilon_{i3} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. For the sake of simplicity, we do not include the interaction term between the treatment and the mediator, but a similar sensitivity analysis can be developed for the model with the interaction term. As before, the sensitivity parameter ρ represents the correlation between the two error terms, and $(\epsilon_{i2}, \epsilon_{i3})$ follows the bivariate normal distribution with the mean zero and the covariance $\rho\sigma_2$.

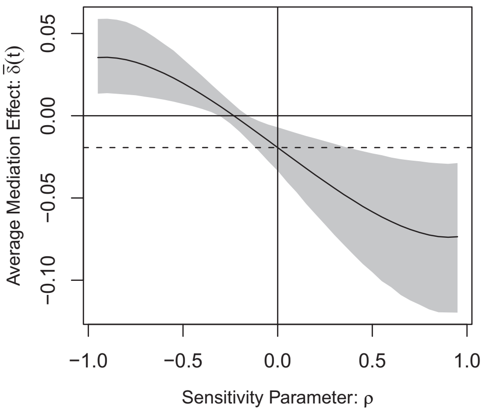

Figure 4. Sensitivity analysis with continuous outcome and binary mediator. The dashed line represents the estimated mediation effect. The gray areas represent the 95% confidence interval for the mediation effects at each value of ρ.

The result, given in Appendix H, parallels closely with the sensitivity analysis for LSEM. For example, taking the partial derivative with respect to ρ shows that $\bar{\delta}(t)$ is monotonically decreasing (increasing) with respect to ρ when $\beta_2 > 0$ ($\beta_2 < 0$). Moreover, when $\beta_2 \neq 0$, we have $\bar{\delta}(t) = 0$ if and only if $\gamma = 0$ or equivalently $\rho = \tilde{\rho}$, the same condition as for LSEM (see Theorem 4 of Imai, Keele, & Yamamoto, 2010). Because the average total effect, $\bar{\tau} = \mathbb{E}\{\Phi(\alpha_1 + \beta_1 + \xi_1^\top X_i) - \Phi(\alpha_1 + \xi_1^\top X_i)\}$, is identified, we can use this result to conduct sensitivity analysis for the proportion mediated. We omit an empirical illustration, although the analysis would proceed as in the last two examples.

Sensitivity Analysis Based on the Coefficients of Determination

Despite its simplicity, researchers may find it difficult to interpret the magnitude of the sensitivity parameter ρ. Here we present an alternative method for the interpretation of ρ. In the case of the LSEM, Imai, Keele, and Yamamoto (2010) showed how to interpret the same sensitivity analysis using the following decomposition of the error terms for Equations 6 and 7:

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

for $j = 2, 3$, where U_i is an unobserved pretreatment confounder (or linear combinations of confounders) that influences both the mediator and the outcome, and λ_j represents an unknown coefficient for each equation. Such omitted variables lead to the violation of the sequential ignorability assumption.

Under this assumption, ρ can be written as a function of the coefficients of determination (i.e., R^2). This allows for the sensitivity analysis to be based on the magnitude of an effect of the omitted variable. This can be done in two ways. First, ρ can be expressed as a function of the proportions of previously unexplained variances in the mediator and outcome regressions, that is, $R_M^{*2} \equiv 1 - \mathrm{Var}(\epsilon'_{i2})/\mathrm{Var}(\epsilon_{i2})$ and $R_Y^{*2} \equiv 1 - \mathrm{Var}(\epsilon'_{i3})/\mathrm{Var}(\epsilon_{i3})$, respectively. We can now express the sensitivity parameter as a function of these two quantities: $\rho = \mathrm{sgn}(\lambda_2\lambda_3)R_M^* R_Y^*$, where $\mathrm{sgn}(x)$ is the sign function (which equals 1 if x is positive, equals 0 if x is zero, and equals -1 if it is negative). Thus, sensitivity analysis can be conducted once researchers specify the direction of the effects

[12] When ρ is nonzero, the expectation and variance of the outcome are $\mathbb{E}(Y_i \mid T_i = t, M_i = m, X_i = x) = \alpha_3 + \beta_3 t + \gamma m + \kappa tm + \xi_3^\top x + \rho\sigma_3\eta_m(t, x)$ and $\mathrm{Var}(Y_i \mid T_i = t, M_i = m, X_i = x) = \sigma_3^2\,\{1 - \rho^2\eta_m(t, x)(\eta_m(t, x) + \alpha_2 + \beta_2 t + \xi_2^\top x)\}$, where $\eta_1(t, x) = \varphi(\alpha_2 + \beta_2 t + \xi_2^\top x)/\Phi(\alpha_2 + \beta_2 t + \xi_2^\top x)$ and $\eta_0(t, x) = -\varphi(\alpha_2 + \beta_2 t + \xi_2^\top x)/\Phi(-\alpha_2 - \beta_2 t - \xi_2^\top x)$ are the inverse Mills ratios. Because $(\alpha_2, \beta_2, \xi_2)$ can be estimated from a probit regression of M_i on $(1, T_i, X_i)$, the parameters $(\alpha_3, \beta_3, \gamma, \kappa, \xi_3, \sigma_3)$ can be consistently estimated with feasible generalized least squares with a known value for ρ. Finally, once these model parameters are estimated, the average causal mediation effect is calculated as a function of ρ and other ~~consistently estimated~~ parameters: $\bar{\delta}(t) = (\gamma + \kappa t)\mathbb{E}\{\Phi(\alpha_2 + \beta_2 + \xi_2^\top X_i) - \Phi(\alpha_2 + \xi_2^\top X_i)\}$. As in the sample selection model of Heckman (1979), ρ is actually identified given the nonlinearity of the model, but we will not use this fact because it only hinges on the functional form assumption. Using these results, one can conduct a sensitivity analysis for the case with a binary mediator, which is done automatically with our software.

of the unobserved confounder U_i in the mediator and outcome models, $\text{sgn}(\lambda_2\lambda_3)$,[13] as well as the relative magnitude of those effects, R_M^{*2} and R_Y^{*2}.

Similarly, the sensitivity analysis can be based on the proportions of original variances that are explained by the unobserved confounder in the mediator and outcome regressions, that is, $\tilde{R}_M^2 \equiv \{\text{Var}(\epsilon_{i2}) - \text{Var}(\epsilon'_{i2})\}/\text{Var}(M_i)$ and $\tilde{R}_Y^2 \equiv \{\text{Var}(\epsilon_{i3}) - \text{Var}(\epsilon'_{i3})\}/\text{Var}(Y_i)$, respectively. In this case, the expression for ρ is given by $\text{sgn}(\lambda_2\lambda_3)\tilde{R}_M\tilde{R}_Y/\sqrt{(1 - R_M^2)(1 - R_Y^2)}$, where R_M^2 and R_Y^2 are the usual coefficients of determination for the mediator and outcome regressions.

When the mediator or the outcome variable is binary, we use the pseudo-R^2 of McKelvey and Zavoina (1975). For example, in the binary mediator case, we redefine $\tilde{R}_M^2 = \{1 - \text{Var}(\epsilon'_{i2})\}/\{\text{Var}(\hat{M}_i^*) + 1\}$ and $R_M^2 = \text{Var}(\hat{M}_i^*)/\{\text{Var}(\hat{M}_i^*) + 1\}$ in the above formula, where $\hat{M}_i^*$ represents the predicted value of the latent mediator variable for the probit regression. Thus, in all cases considered here we can interpret ρ using two alternative coefficients of determination. This value can then be used to compare across studies or evaluated in reference to subject specific knowledge about the likely magnitude of effect from the confounding variable.

Next, we present an empirical example of a sensitivity analysis in terms of the coefficients of determination for the case of a continuous outcome (depression 6 months following the treatment) and a dichotomous mediator. The model used here is the same as the one that produces the results given in Table 3 except that for the purpose of illustration we use an alternative mediator that is a dichotomized index of several psychological measures such as the original job search self-efficacy variable and the internal locus of self-control. We call this variable *mastery.* The resulting estimate of the average mediation effect is $-.031$, with the 95% confidence interval of $[-.050, -.007]$.

How sensitive is this result to an unobserved confounder that influences both the mediator and the outcome? Consider the so-called ability bias in which participants with greater ability are likely to respond to the training, thereby increasing the level of their mastery, and yet they are also likely to have a relatively lower level of depression. Under this scenario, we assume the sign of the product of coefficients for the unobserved confounder is negative, that is, $\text{sgn}(\lambda_2\lambda_3) - 1$, because their effects are expected to operate in the opposite directions.

Figure 5 presents our sensitivity analysis based on the coefficients of determination, $\tilde{R}_M^2$ and $\tilde{R}_Y^2$, which represent the proportions of original variances explained by the unobserved confounder for the mediator and the outcome, respectively. In the figure, the contour line of 0 corresponds to values of $\tilde{R}_M^2$ and $\tilde{R}_Y^2$ that yield zero average causal mediation effect. For example, when $\tilde{R}_M^2 = .6$ and $\tilde{R}_Y^2 = .3$, the estimated mediation effect would be approximately zero. This means that the unobserved confounder, ability, would have to explain 60% of the original variance in the (latent) mastery variable and 30% of the original variance in the depression variable for the estimate to be zero. At higher values of both $\tilde{R}_M^2$ and $\tilde{R}_Y^2$, the estimated average causal mediation effect would be positive, whereas at lower values the sign of the estimate remains negative. This implies that the values of $\tilde{R}_M^2$ and $\tilde{R}_Y^2$ must be relatively high for the original conclusion to be reversed.

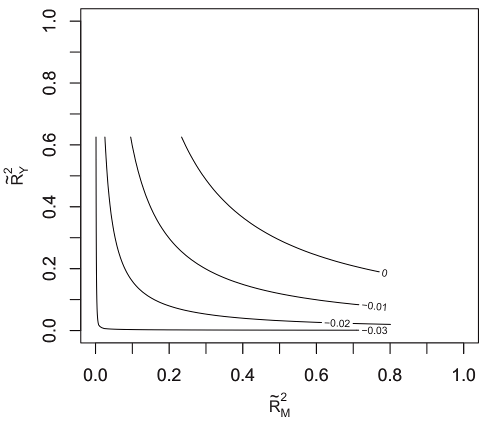

Figure 5. Sensitivity analysis with continuous outcome and binary mediator. The plot contains contour lines that represent estimated average mediation effect corresponding to unobserved pretreatment confounders of various magnitudes. These magnitudes are measured by the coefficients of determination, $\tilde{R}_M^2$ and $\tilde{R}_Y^2$, each of which represents the proportion of original variance explained by the unobserved confounder for the mediator and the outcome, respectively. Here we assume that the unobserved confounder, ability, influences the mediator and the outcome in the opposite directions.

Concluding Remarks

In this article, we proposed a general approach to causal mediation analysis. Our approach consisted of the following four elements: (a) general definitions of causal mediation effects given within the counterfactual framework, (b) nonparametric identification based on the sequential ignorability assumption, (c) estimation algorithms for parametric and nonparametric models, and (d) sensitivity analyses to assess the robustness of empirical results. The proposed approach is general because it is developed without any reference to a particular statistical model and therefore is applicable to a wide range of situations. In doing so, we developed easy-to-use software, mediation (Imai et al., 2010a), and give researchers access to estimation strategies that can handle a variety of data. Furthermore, our approach straightforwardly calculates estimates of uncertainty that allow for hypothesis testing and confidence interval construction. We also believe it is important to probe the extent to which an unverifiable assumption drives the results of the analysis. Thus, as a part of our general approach, we have developed a sensitivity analysis that allows researchers to quantify the exact degree of departure from the key identification assumption that is required for the original results to no longer hold.

We emphasize that in our approach each component is closely linked together. The potential outcomes definition of causal mediation effects allows us to conduct the nonparametric identification analysis, which reveals the key assumption required for researchers to interpret mediation effects as causal. The result of this identification analysis implies the general estimation algorithms that are applicable across a variety of statistical models. Further-

[13] For example, if the omitted variable affects the mediator and outcome variable in the same direction, then $\text{sgn}(\lambda_2\lambda_3) = 1$.

more, the potential violation of the key identification assumption motivates the development of the sensitivity analysis within the context of statistical models commonly used by applied researchers.

Future research should extend this general framework to more complex situations such as panel data, multilevel models, and multiple mediators. Like much of the existing literature, we have focused on the statistical methods that are applicable to either standard randomized experiments (in which the treatment is randomized) or observational studies in cross-section settings. However, we do not consider alternative experimental designs in which the mediator is either directly or indirectly manipulated by researchers. In our ongoing work (Imai, Tingley, & Yamamoto, 2009), we evaluated the advantages and disadvantages of the existing experimental designs and developed new experimental designs that may more effectively identify causal mediation effects. We also showed how these experimental designs can serve as templates for observational studies so that causal mediation analysis can be conducted with less stringent assumptions than sequential ignorability (Imai, Keele, Tingley, & Yamamoto, 2010b).

References

Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in Medicine, 27,* 1282–1304.

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

Avin, C., Shpitser, I., & Pearl, J. (2005). Identifiability of path-specific effects. In L. P. Kaelbling & A. Saffiotti (Eds.), *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (pp. 357–363). San Francisco, CA: Morgan Kaufmann.

Baron, R. M., & 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,* 1173–1182.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11,* 142–163.

Bohrnstedt, G. W., & Goldberger, A. S. (1969). On the exact covariance of products of random variables. *Journal of the American Statistical Association, 64*(328), 1439–1442.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology, 98,* 550–558.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112,* 558–577.

Ditlevsen, S., Christensen, U., Lynch, J., Damsgaard, M. T., & Keiding, N. (2005). The mediation proportion: A structural equation approach for estimating the proportion of exposure effect on outcome explained by an intermediate variable. *Epidemiology, 16*(1), 114–120.

Freedman, L. S., Graubard, B., & Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine, 11,* 167–178.

Gallop, R., Small, D. S., Lin, J. Y., Elliot, M. R., Joffe, M., & Ten Have, T. R. (2009). Mediation analysis with principal stratification. *Statistics in Medicine, 28,* 1108–1130.

Glynn, A. N. (2008). *Estimating and bounding mechanism specific causal effects.* Paper 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.

Hastie, T. J., & Tibshirani, R. (1990). *Generalized additive models.* London, England: Chapman & Hall.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–161.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Huang, B., Sivaganesan, S., Succop, P., & Goodman, E. (2004). Statistical assessment of mediational effects for logistic mediational models. *Statistics in Medicine, 23,* 2713–2728.

Hyman, H. (1955). *Survey design and analysis: Principles, cases, and procedures.* New York, NY: Free Press.

Imai, K., Keele, L., & Tingley, D. (2010). *Replication data for: A general approach to causal mediation analysis.* Retrieved from http://hdl.handle.net/1902.1/14801

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2010a). Advances in social science research using R. In H. D. Vinod (Ed.), *Causal mediation analysis using R* (pp. 129–154). New York, NY: Springer.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2010b). *Unpacking the black box: Learning about causal mechanisms from experimental and observational studies.* Retrieved from http://imai.princeton.edu/research/mediationP.html

Imai, K., Tingley, D., & Yamamoto, T. (2009). *Experimental designs for identifying causal mechanisms.* Retrieved from http://imai.princeton.edu/research/Design.html

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

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data.* Beverly Hills, CA: Sage.

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

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5,* 602–619.

Kaufman, J. S., MacLehose, R. F., Kaufman, S., & Greenland, S. (2005). The mediation proportion. *Epidemiology, 16*(5), 710.

Keele, L. J. (2008). *Semiparametric regression for the social sciences.* Chichester, England: Wiley.

King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science, 44,* 341–355.

Koenker, R. (2005). *Quantile regression.* New York, NY: Cambridge University Press.

Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology, 27,* S101–S108.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry, 59,* 877–883.

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review, 23,* 418–444.

Li, Y., Schneider, J. A., & Bennett, D. A. (2007). Estimation of the mediation effect with a binary mediator. *Statistics in Medicine, 26,* 3398–3414.

Little, R. J., & Yau, L. H. Y. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3,* 147–159.

MacKinnon, D. P. (2000). Contrasts in multiple mediator models. In J. S. Rose, L. Chassin, C. C. Presson, & S. J. Sherman (Eds.), *Multivariate* *applications in substance use research: New methods for new questions* (pp. 141–160). Mahwah, NJ: Erlbaum.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* New York, NY: Routledge.

MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review, 17,* 144–158.

MacKinnon, D., & Fairchild, A. (2009). Current directions in mediation analysis. *Current Directions in Psychological Sciences, 18*(1), 16–20.

MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W., & Hoffman, J. M. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials, 4,* 499–513.

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

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39,* 99–128.

Manski, C. F. (2007). *Identification for prediction and decision.* Cambridge, MA: Harvard University Press

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variable. *Journal of Mathematical Sociology, 4,* 103–120.

Mulaik, S. (2007). There is a place for approximate fit in structural equation modelling. *Personality and Individual Differences, 42,* 883–891.

Mulaik, S. (2009). *Linear causal modeling with structural equations.* Boca Raton, FL: CRC Press.

Pearl, J. (2001). Direct and indirect effects. In J. Breese & D. Koller (Eds.), *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufmann.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavioral Research Methods, 40,* 879–891.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42,* 185–227.

R Development Core Team. (2009). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In P. J. Green, N. L. Hjort, & S. Richardson (Eds.), *Highly structured stochastic systems* (pp. 70–81). New York, NY: Oxford University Press.

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

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies: Rejoinder. *Statistical Science, 17*(3), 321–327.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13,* 279–313.

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

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology, 20*(1), 18–26.

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology, 21*(4), 540–551.

Vinokur, A. D., Price, R. H., & 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. D., & 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,* 867–877.

Wang, Y., & Taylor, J. M. G. (2002). A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics, 58,* 803–812.

White, I. R., Carpenter, J., Evans, S., & Schroter, S. (2007). Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clinical Trials, 4*(2), 125–139.

Wood, S. N. (2006). *Generalized additive models: An introduction with R.* Boca Raton, FL: Chapman & Hall/CRC.

Appendix A

Asymptotic Variance Under the Model With the Interaction

Using the delta method and noting the independence between $\hat{\beta}_2$ and $(\hat{\gamma}, \hat{\kappa})$, we can write the asymptotic variance of estimated average causal mediation effect in Equation 9 as $\text{Var}(\hat{\beta}_2(\hat{\gamma} + \hat{\kappa}t) \mid T, X) \approx (\gamma + t\kappa)^2 \text{Var}(\hat{\beta}_2 \mid T, X) + \beta_2^2\{\text{Var}(\hat{\gamma} \mid T, X) + t^2\text{Var}(\hat{\kappa} \mid T, X) + 2t\text{Cov}(\hat{\gamma}, \hat{\kappa} \mid T, X)\}$ *for* $t = 0, 1$ (Preacher, Rucker, & Hayes, 2007). On the basis of this asymptotic variance, one can conduct the statistical test with the null hypothesis of the form $\bar{\delta}(t) = 0$ for $t = 0, 1$.

To derive the asymptotic variance for the estimated direct effect given in Equation 10, we first rewrite the variance as

$$\begin{aligned}
&\text{Var}(\hat{\beta}_3 + \hat{\kappa}\{\hat{\alpha}_2 + \hat{\beta}_2 t + \hat{\xi}_2^\top \overline{X}\} \mid T, X) \\
&\quad = \text{Var}(\hat{\beta}_3 \mid X, T) + \text{Var}(\hat{\kappa}\{\hat{\alpha}_2 + \hat{\beta}_2 t + \hat{\xi}_2^\top \overline{X}\} \mid T, X) + \\
&\qquad 2\text{Cov}(\hat{\beta}_3, \hat{\kappa}\{\hat{\alpha}_2 + \hat{\beta}_2 t + \hat{\xi}_2^\top \overline{X}\} \mid T, X).
\end{aligned}$$

Noting the independence between $(\hat{\beta}_3,\hat{\kappa})$ and $(\hat{\alpha}_2,\hat{\beta}_2,\hat{\xi}_2)$ and using the result of Goodman (1960), we can write the second term of the above equation as

$$\begin{aligned}
&\text{Var}(\hat{\kappa}\{\hat{\alpha}_2 + \hat{\beta}_2 t + \hat{\xi}_2^\top \overline{X}\} \mid T, X) = \\
&\quad (\alpha_2 + \beta_2 t + \xi_2^\top \overline{X})^2 \text{Var}(\hat{\kappa} \mid T, X) + \kappa^2 \text{Var}(\hat{\alpha}_2 + \hat{\beta}_2 t + \hat{\xi}_2^\top \overline{X} \mid T, X) \\
&\qquad + \text{Var}(\hat{\kappa} \mid T, X)\text{Var}(\hat{\alpha}_2 + \hat{\beta}_2 t + \hat{\xi}_2^\top \overline{X} \mid T, X),
\end{aligned}$$

where $\text{Var}(\hat{\alpha}_2 + \hat{\beta}_2 t + \hat{\xi}_2^\top \overline{X} \mid T, X) = \text{Var}(\hat{\alpha}_2 \mid T, X) + t^2\text{Var}(\hat{\beta}_2 \mid T, X) + \overline{X}^\top \text{Var}(\xi_2 \mid T, X)\overline{X} + 2t\text{Cov}(\hat{\alpha}_2, \hat{\beta}_2 \mid T, X) + 2t\overline{X}^\top \text{Cov}(\hat{\beta}_2, \hat{\xi}_2 \mid T, X) + 2\overline{X}^\top \text{Cov}(\hat{\alpha}_2, \hat{\xi}_2 \mid T, X)$. Finally, using the result of Bohrnstedt and Goldberger (1969) and noting the independence again, we write the final covariance term as

$$\begin{aligned}
&\text{Cov}(\hat{\beta}_3, \hat{\kappa}\{\hat{\alpha}_2 + \hat{\beta}_2 t + \hat{\xi}_2^\top \overline{X}\} \mid T, X) = \\
&\qquad (\alpha_2 + \beta_2 t + \xi_2^\top \overline{X})\text{Cov}(\hat{\beta}_3, \hat{\kappa} \mid T, X).
\end{aligned}$$

Finally, the variance for the estimated total effect is given by $\text{Var}(\hat{\beta}_1 + \hat{\eta}^\top \overline{X} \mid T, X) = \text{Var}(\hat{\beta}_1 \mid T, X) + \overline{X}^\top \text{Var}(\hat{\eta} \mid T, X)\overline{X} + 2\overline{X}^\top \text{Cov}(\hat{\beta}_1, \hat{\eta} \mid T, X)$.

Appendix B

The Assumption of Instrumental Variables Estimation

> *Assumption 2 (Instrumental Variables Assumption; Angrist et al., 1996):* The assumption consists of the following three parts:
>
> 1. Ignorability of treatment assignment: $\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x$ and $0 < \Pr(T_i = t \mid X_i = x)$ for $t, t' = 0, 1$ and all $x \in \mathcal{X}$;
> 2. Monotonic treatment effect on the mediator: $M_i(1) \geq M_i(0)$ (or $M_i(1) \leq M_i(0)$) for all $i = 1, 2, \ldots, n$;
> 3. No direct effect (exclusion restriction): $Y_i(1, m) = Y_i(0, m)$ for all $m \in \mathcal{M}$.

Like Assumption 1, Assumption 2 requires that the treatment assignment is ignorable given the observed pretreatment covariates. As noted before, the ignorability of treatment assignment is satisfied in experiments in which the treatment is randomized, a typical setting where causal mediation analysis is employed. More importantly, the instrumental variables method replaces the sequential ignorability assumption about the mediator with two alternative assumptions. Although allowing for the possibility that there may exist unobserved variables that confound the relationship between the outcome and mediating variables, the instrumental variables method assumes that the treatment monotonically affects the mediator and that the treatment has no direct effect on the outcome. Like the ignorability of the mediator, these two assumptions are not directly testable because we never observe $M_i(1)$ and $M_i(0)$ (or $Y_i(1, m)$ and $Y_i(0, m)$) jointly for any given unit.

Although these assumptions are not refutable, we can probe their plausibility. The monotonicity assumption may be plausible in some cases. In the context of the JOB II study, for example, the assumption implies that the program participation would help every worker in the study by improving his or her level of self-confidence in the search for a job. However, the assumption is violated if there are some workers whose self-confidence level is negatively affected by the job training program. Thus, the monotonicity assumption rules out any job seeker having a negative-

(*Appendices continue*)

reaction to the training programming and thus being less effective at finding a job.

The assumption of no direct effect for every unit is more problematic, as the main goal of causal mediation analysis is to test alternative causal mechanisms. This assumption implies that there is no other causal pathway other than through the mediator of interest. In the JOBS II study, the assumption is difficult to justify because the increase in the level of job search self-efficacy is probably not the only reason why the job training program reduces the depressive symptoms. Thus, although Assumption 2 has an advantage of allowing for the existence of unobserved confounders, it a priori excludes the possibility of a direct effect from the treatment to the outcome.

If the instrumental variables assumption is maintained, there is no need to use the product of coefficients as the average mediation effect. Indeed, all one needs to do is to estimate the average total effect of the treatment that under the no-direct-effect assumption equals the average mediation effect. Specifically, under Assumption 2, β_3 in Equation 7 is assumed to equal zero, and the average causal mediation effect is identified as $\bar{\delta} = \beta_2\gamma = \beta_1$, where β_1 is given in Equation 5. Thus, the estimated average causal mediation effect and its variance are easily obtained by regressing Y_i on T_i and X_i.

Appendix C

Proof of Theorem 2

Equations 6 and 8 imply

$$\begin{aligned} Y_i = (\alpha_2\gamma + \alpha_3) + \{\beta_3 + (\gamma + \kappa)\beta_2 + \alpha_2\kappa\}T_i \\ + (\gamma\xi_2 + \xi_3)^\top X_i + \kappa\xi_2^\top T_i X_i + (\gamma + \kappa T_i)\varepsilon_{i2} + \varepsilon_{i3}. \end{aligned}$$

Let $\varepsilon_{i1} = (\gamma + \kappa T_i)\varepsilon_{i2} + \varepsilon_{i3}$. Then, $\mathbb{E}(\epsilon_{i1} \mid T_i) = (\gamma + \kappa T_i)\mathbb{E}(\epsilon_{i2} \mid T_i) + \mathbb{E}(\epsilon_{i3} \mid T_i) = 0$, where the second equality follows from Equation 3 of Assumption 1 (for details, see proof of Theorem 2 in Imai, Keele, & Yamamoto, 2010). Thus, the identifiable parameters are $(\alpha_1, \beta_1, \xi_1, \alpha_2, \beta_2, \xi_2, \kappa\xi_2, \sigma_{1t}^2, \sigma_{2t}^2, \tilde{\rho}_t)$ for $t = 0, 1$, where $\alpha_1 = \alpha_2\gamma + \alpha_3$, $\beta_1 = \beta_3 + (\gamma + \kappa)\beta_2 + \alpha_2\kappa$, and $\xi_1 = \gamma\xi_2 + \xi_3$. This means that if we identify (γ, κ), then $(\alpha_3, \beta_3, \xi_3)$ is also identifiable. To identify (γ, κ), we solve the following system of equations (note that κ can also be identified from $\kappa\xi_2$ as long as there is no interaction term between X_i and T_i in the outcome regression; for the sake of generality, however, we do not pursue this identification strategy here):

$$\sigma_{1t}^2 = (\gamma + t\kappa)^2\sigma_{2t}^2 + 2(\gamma + t\kappa)\rho\sigma_{2t}^2\sigma_{3t}^2 + \sigma_{3t}^2$$

$$\tilde{\rho}_t^2\sigma_{1t}\sigma_{2t} = (\gamma + t\kappa)\sigma_{2t}^2 + \rho\sigma_{2t}\sigma_{3t},$$

where $(\rho_t, \sigma_{1t}^2, \sigma_{2t}^2)$ is identifiable from the data and $(\gamma, \kappa, \sigma_{3t}^2)$ is the set of unknown parameters for $t = 0, 1$. The number of equations is four and is equal to the number of parameters, and thus one can express κ and γ as the functions of identifiable parameters. Thus, with Equation 9, the desired expression results.

Appendix D

The Estimation Algorithms

Parametric Inference

Suppose that the quantity of interest is the average causal mediation effect, $\bar{\delta}(t)$.

Step 1: Fit models for the observed outcome and mediator variables. Fit a parametric model, $f_{\theta_M}(M_i \mid T_i, X_i)$, for the mediator and another parametric model, $f_{\theta_Y}(Y_i \mid T_i, M_i, X_i)$, for the outcome where θ_M and θ_Y represent model parameters.

Step 2: Simulate model parameters from their sampling distribution. Sample J copies of θ_M and θ_Y from their sampling distributions and denote them $\theta_M^{(j)}$ and $\theta_Y^{(j)}$, respectively. We use the approximation based on the multivariate normal distribution (with mean and variance equal to the estimated parameters and their estimated asymptotic covariance matrix, respectively).

Step 3. For each $j = 1, \ldots, J$, repeat the following three steps:

1. Simulate the potential values of the mediator. For each $t = 0, 1$ and each $i = 1, \ldots, n$, sample K copies of $M_i(t)$ from $f_{\theta_M^{(j)}}(M_i \mid t, X_i)$ and denote them as $M_i^{(jk)}(t)$ for $k = 1, \ldots, K$. That is, we generate two potential values of the mediator each based on the mediator model, one under the treatment condition and the other under the control condition.

(*Appendices continue*)

2. Simulate the potential outcomes given the simulated values of the mediator. For each $t = 0, 1$ and each $i = 1, \ldots, n$, sample one copy of $Y_i(t, M_i^{(jk)}(t'))$ from $f_{\theta_M^{(j)}}(Y_i \mid t, M_i^{(jk)}(t'), X_i)$ and denote it as $Y_i^{(jk)}(t, M_i^{jk}(t'))$ for $k = 1, \ldots, K$. That is, for each treatment status, we generate two potential outcomes, again using model predictions, one using the mediator status generated under the treatment condition and the other using the mediator value under the control condition.

3. Compute the causal mediation effects. Compute the average causal mediation effect as

$$\bar{\delta}^{(j)}(t) = \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}\{Y_i^{(jk)}(t, M_i^{(jk)}(1)) - Y_i^{(jk)}(t, M_i^{(jk)}(0))\}.$$

That is, we simply take the difference across the two outcome predictions under treatment and the two outcome predictions under control and average across the predictions for each of the N units in the study. This provides us with $\bar{\delta}^{(j)}(t)$, which is one Monte Carlo draw of the average mediation effect.

Step 4: Compute summary statistics. Finally, we compute the point estimate of $\bar{\delta}(t)$ and its uncertainty estimates from the distribution of mediation effects: $\bar{\delta}^{(j)}(t)$. For example, the sample median and the sample standard deviation of the distribution can be used as the point estimate of $\bar{\delta}(t)$ and its standard error, whereas percentiles of this distribution can serve as confidence intervals for $\bar{\delta}(t)$.

In principle, one can modify Step 3.3 of the above algorithm to accommodate any quantities other than the average causal mediation effects. For example, the α-quantile average causal mediation effects defined in Footnote 9 can be estimated by computing the sample quantile of $Y_i^{(jk)}(t, M_i^{(jk)}(t'))$ across treatment and control. Although the proposed algorithm is applicable for any parametric model, algorithms specifically tailored for a particular model are typically more computationally efficient because, depending on the models chosen by researchers, further simplifications of Algorithm 1 may be possible.[14]

Nonparametric Inference

Suppose that the quantity of interest is the average causal mediation effect, $\bar{\delta}(t)$.

Step 1. Take a random sample with replacement of size n from the original data J times. For each of the J bootstrapped samples, repeat the following steps:

1. Fit models for the observed and mediator variables. Fit a possibly nonparametric model, $f(M_i \mid T_i, X_i)$, for the mediator and another possibly nonparametric model, $f(Y_i \mid T_i, M_i, X_i)$, for the outcome. Denote the estimates as $f^{(j)}(M_i \mid T_i, X_i)$ and $f^{(j)}(Y_i \mid T_i, M_i, X_i)$. Again, these are simply the mediator and outcome models that are now allowed to be nonparametric or semiparametric models.

2. Simulate the potential values of the mediator. For each $t = 0, 1$ and each $i = 1, 2, \ldots, n$, sample K copies of $M_i(t)$ from $f^{(j)}(M_i \mid t, X_i)$ and denote them as $M_i^{(jk)}(t)$ for $k = 1, 2, \ldots, K$. Once again, we generate a set of predictions for the mediator under each treatment status.

3. Simulate the potential outcomes given the simulated values of the mediator. For each $t = 0, 1$ and each $i = 1, 2, \ldots, n$, sample one copy of $Y_i(t, M_i^{(jk)}(t'))$ from $f^{(j)}(Y_i \mid t, M_i^{(jk)}(t'), X_i)$ and denote it as $Y_i^{(jk)}(t, M_i^{(jk)}(t'))$ for $k = 1, 2, \ldots, K$. Potential outcome predictions are generated for each treatment status and two mediator predictions.

4. Compute the causal mediation effects. Compute the average causal mediation effect as

$$\bar{\delta}^{(j)}(t) = \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}\{Y_i^{(jk)}(t, M_i^{(jk)}(1)) - Y_i^{(jk)}(t, M_i^{(jk)}(0))\},$$

which is the difference between the two outcome predictions under each treatment status.

Step 2: Compute summary statistics. Compute the point estimate of $\bar{\delta}(t)$ and its uncertainty estimates using the J estimates from the bootstrap sampling distribution. The sample median and the sample standard deviation of $\bar{\delta}^{(j)}(t)$ can be used as the point estimate of $\bar{\delta}(t)$ and its standard error. Percentiles may be used as confidence intervals.

As before, in some cases, the simplification of Algorithm 2 is possible, though we do not describe the details here.

[14] For example, there is no need to simulate the mediator and the outcome as well as model parameters in the LSEM framework, although point and uncertainty estimates will be approximately equal. Moreover, if the quantities of interest can be derived analytically from the selected parametric model for the outcome variable, $f_{\theta_M}(Y_i \mid T_i, M_i, X_i)$, then Step 3.2 can be skipped and Step 3.3 can be modified to compute the average causal mediation effect for each unit directly given a Monte Carlo draw $M_i^{(jk)}(t)$ for $t = 0, 1$ as follows:

$$\bar{\delta}^{(j)}(t) = \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}\{\mathbb{E}_{\theta_Y^{(j)}}(Y_i \mid t, M_i^{(jk)}(1), X_i) - \mathbb{E}_{\theta_Y^{(j)}}(Y_i \mid t, M_i^{(jk)}(0), X_i)\}.$$

One situation in which this simplification is possible is when the outcome variable is binary and modeled with logistic regression.

(*Appendices continue*)

Appendix E

Simulation Studies

Here we conduct a series of Monte Carlo simulations to demonstrate the superiority of our method over the existing methods based on the LSEM. Although there exist a variety of situations one can examine, we focus on two specific scenarios: (a) when nonlinearity is present in the outcome model and (b) when the outcome is binary.

Nonlinear Model

We begin by conducting a simulation study with nonlinearity in the outcome model. In such a situation, one common strategy is to add quadratic and/or cubic terms as additional linear predictors within the LSEM framework. For more complex forms of nonlinearity, however, this strategy can perform poorly especially in a large sample. Thus, we also examine the performance of the semiparametric model discussed in the section Nonparametric and Semiparametric Regressions. Specifically, we generate data from Equations 6 and 7 but make the relationship between the outcome and mediator variables nonlinear as shown in Figure E1. Thus, a quadratic model is a reasonable but incorrect approximation. We do not include any additional covariates in the assumed models beyond the binary treatment and the mediator. We simulate the error terms as independent normal random variables so that sequential ignorability holds.

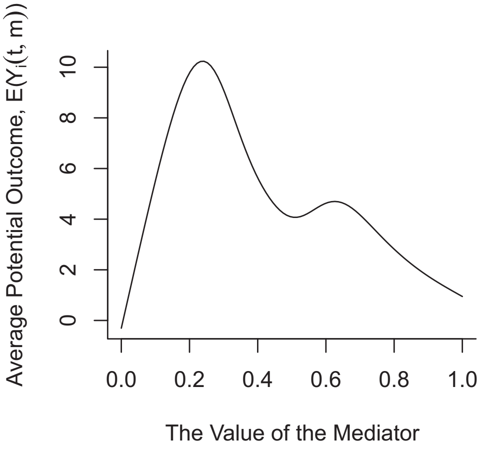

Figure E1. Nonlinear relationship used for a Monte Carlo study. The solid line represents the average potential outcome, $\mathbb{E}(Y_i(t, m))$, as a function of the mediator while the treatment variable is held constant.

Table E1
Simulation Results With the Nonlinear Outcome Model by Sample Size

	Large effect size		Small effect size	
Method	Bias	RMSE	Bias	RMSE
N = 100				
GAM	−0.53	3.67	−0.11	0.72
Quadratic	−4.29	4.63	−0.25	0.69
Linear	−6.26	6.45	−0.32	0.37
N = 500				
GAM	−0.47	2.08	−0.09	0.39
Quadratic	−7.19	7.40	−0.35	0.49
Linear	−6.66	6.71	−0.33	0.34

Note. In the generalized additive model (GAM), the estimation of the average causal mediation effect is based on Algorithm 2. The quadratic or linear structural equation model includes the square term of the mediator as an additional linear predictor in the outcome model. The estimation of the average causal mediation effect for this model is based on Algorithm 1. The linear model represents the standard product of coefficients method. The true average causal mediation effect is 4.4 for large effect size and 0.22 for small effect size. RMSE = root-mean-squared error.

Within these basic settings, we conduct two Monte Carlo simulations each with two different sample sizes (100 and 500) for a total of four sets of scenarios. Several population parameters were held constant across all four sets of results: The population values of α_2, β_2, α_3, and β_3 were all set to 0.25. In the first set of simulations, the amount of the total effect that was mediated is approximately 25%. In the second set of simulations, the amount of the total effect that was mediated is higher and about 75%. In sum, we compare how well each method performs when the mediation effects are large or small with two sample sizes (small and moderate).

For each of the four data-generating processes, we compare three estimation methods. The first method uses the GAM and Algorithm 2 as discussed in the main text. The second method we use is Algorithm 1 based on the LSEM, which includes the square term of the mediator as an additional linear predictor in the outcome model. Finally, we estimate average causal mediation effects using the standard product of coefficients method. The results are reported in Table E1 and are based on 2,500 simulations. We compute both bias and the root-mean-squared error (RMSE) for each method.

(*Appendices continue*)

Table E2
Performance of the Proposed Method With Binary Outcome Variable Relative to the Product of Coefficients Method

	Large effect size				Small effect size			
	Proposed method		Product of coefficients		Proposed method		Product of coefficients	
Sample size	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
50	0.014	0.172	0.213	0.682	0.002	0.097	0.092	0.349
100	0.008	0.127	0.198	0.451	0.000	0.068	0.092	0.228
500	0.001	0.065	0.184	0.257	−0.002	0.032	0.089	0.127
1,000	−0.001	0.046	0.180	0.218	−0.001	0.023	0.090	0.110

Note. The data-generating processes are based on the probit outcome (binary measure of employment) regression and the linear mediation (continuous measure of job search self-efficacy) regression from the Job Search Intervention Study example. No covariates are used for this simulation experiment. The true average causal mediation effect is 0.106 for large effect size and 0.052 for small effect size. RMSE = root-mean-squared error.

We first review the results when the average causal mediation effect is large and the sample size is small. Here conventional methods perform very poorly, with the bias being nearly an order of magnitude higher compared with the semiparametric model. This same basic pattern holds for the larger sample size as well. The parametric models fare better in terms of RMSE when the sample size is small. This is not surprising, as nonparametric methods tend to have higher levels of variance in small samples. Regardless, the RMSE for the GAM is significantly smaller. Once the sample size is moderate, the difference becomes more stark, as the RMSE for the parametric models is more than twice that of the GAM.

When the mediation effect is small, the GAM results again clearly outperform conventional models in terms of bias. The bias for the mediation effect is typically 3 times smaller when the GAM is used. As expected, the RMSE is actually worse for small sample sizes but improves markedly for the large sample sizes. Together, this simulation exercise demonstrates why the generality of our proposed approach is valuable. Nonparametric and semiparametric regression models allow for robust and flexible modeling of nonlinearity, and our proposed method can accommodate these models in a straightforward manner.

Binary Outcome

In our second example, we study whether our estimation algorithm can accurately recover the average causal mediation effect in a case with a binary outcome and a continuous mediator. As in the previous simulation, we use two effect sizes. To make the data-generating process realistic, we set the parameters (α_2, α_3, β_3, γ) equal to the values generated from regression analysis of the JOBS II data with the dichotomous employment outcome variable (probit) and continuous job search self-efficacy mediator (ordinary least squares; i.e., 3.9, −0.9, 0.15, 0.19). For the sake of simplicity, we do not use covariates, though our approach easily accommodates them.

We set $\beta_2 = 3$ in our first simulation and $\beta_2 = 1.5$ in our second simulation so as to produce a larger (about 0.1) and smaller (about 0.05) average causal mediation effect, where β_2 is defined in Equation 6. We vary the sample size from a relatively small (50) to a relatively large sample (1,000). We compare our method (Algorithm 1) against the product of coefficients method in terms of bias and RMSE for the estimated average causal mediation effect. As before, each Monte Carlo experiment is based upon 2,000 simulations with our mediation software.

Table E2 presents the results from our simulations. The bias and RMSE of our method are small even with a sample size of only 50 and steadily decline as the sample size increases. Conversely, using the product of coefficients method to estimate the average causal mediation effect produces substantial bias and a large RMSE for all cases we examine. This demonstrates that blindly applying the product of coefficients method in practice can yield highly misleading inferences.

(*Appendices continue*)

Appendix F

The Average Causal Mediation Effects With Binary Outcome

We derive the general expression for the average causal mediation effects using the following model with the pretreatment confounders. The result given in the main text is a special case of this general result.

$$M_i = \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \epsilon_{2i},$$

$$Y_i = \mathbf{1}\{Y_i^* > 0\} \text{ where } Y_i^* = \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \epsilon_{3i},$$

where ε_{2i} and ε_{3i} are different i.i.d. random variables with zero mean and $\mathrm{Var}(\varepsilon_{2i}) = \sigma_2^2$ and $\mathrm{Var}(\varepsilon_{3i}) = \sigma_3^2$. Under Assumption 1, we have

$$\begin{aligned}
&\int \mathbb{E}(Y_i \mid T_i = t, M_i, X_i = x) dF(M_i \mid T_i = t', X_i = x) \\
&= \int \Pr(\alpha_3 + \beta_3 t + \gamma M_i + \xi_3^\top x + \epsilon_{3i} > 0) dF(M_i \mid T_i = t', X_i = x) \\
&= \int \Pr\{\alpha_3 + \beta_3 t + \xi_3^\top x + \gamma(\alpha_2 + \beta_2 t' + \xi_2^\top x + \epsilon_{2i}) + \epsilon_{3i} > 0\} \\
&\qquad \times dF(\epsilon_{2i}) \\
&= \int\int \mathbf{1}\{\alpha_3 + \beta_3 t + \xi_3^\top x + \gamma(\alpha_2 + \beta_2 t' + \xi_2^\top x + \epsilon_{2i}) + \epsilon_{3i} > 0\} \\
&\qquad \times dF(\epsilon_{3i}) dF(\epsilon_{2i}) \\
&= 1 - \Pr\{\gamma\epsilon_{2i} + \epsilon_{3i} \leq -\alpha_3 - \beta_3 t - \xi_3^\top x - \gamma(\alpha_2 + \beta_2 t' + \xi_2^\top x)\},
\end{aligned}$$

where the second equality follows from the change of variables formula as well as the fact that $T_i \perp\!\!\!\perp \varepsilon_{2i}$, and the fourth equality follows from the independence between ε_{2i} and ε_{3i}.

Suppose ε_{3i} is an i.i.d. logistic random variable and $\varepsilon_{2i} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_2^2)$. Using the results derived above, we obtain

$$\begin{aligned}
\bar{\delta}(t) = \mathbb{E}\{&H(\alpha_3 + \beta_3 t + \xi_3^\top X_i + \gamma(\alpha_2 + \beta_2 + \xi_2^\top X_i)) \\
&- H(\alpha_3 + \beta_3 t + \xi_3^\top X_i + \gamma(\alpha_2 + \xi_2^\top X_i))\},
\end{aligned}$$

$$\begin{aligned}
\bar{\tau} = \mathbb{E}\{&H(\alpha_3 + \beta_3 + \xi_3^\top X_i + \gamma(\alpha_2 + \beta_2 + \xi_2^\top X_i)) \\
&- H(\alpha_3 + \gamma(\delta_2 + \xi_2^\top X_i))\}.
\end{aligned}$$

Next, suppose $\varepsilon_{3i} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and $\varepsilon_{2i} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_2^2)$. Then, we have

$$\begin{aligned}
\bar{\delta}(t) = \mathbb{E}\Bigg\{&\Phi\left(\frac{\alpha_3 + \beta_3 t + \xi_3^\top X_i + \gamma(\alpha_2 + \beta_2 + \xi_2^\top X_i)}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right) \\
&- \Phi\left(\frac{\alpha_3 + \beta_3 t + \xi_3^\top X_i + \gamma(\alpha_2 + \xi_2^\top X_i)}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right)\Bigg\},
\end{aligned}$$

$$\bar{\tau} = \mathbb{E}\{\Phi(\alpha_1 + \beta_1 + \xi_1^\top X_i) - \Phi(\alpha_1 + \xi_1^\top X_i)\}.$$

Appendix G

The Proportion Mediated With Binary Outcome

Without Covariates

Suppose that the probit regression is used to model the binary outcome variable without pretreatment covariates, that is, the model defined in Equations 13 and 14, where $\varepsilon_{i3} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_2^2)$ and $\varepsilon_{i3} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. We derive the condition under which the common measure used in the literature, $\tilde{\upsilon}$ in Equation 19, approximates the true proportion mediated, υ. First, we use the following linear approximation to the average causal mediation effect:

$$\begin{aligned}
\bar{\delta}(t) &\approx \frac{\delta}{\delta t'}\Phi\left(\frac{\alpha_3 + \beta_3 t + \gamma(\alpha_2 + \beta_2 t')}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right)\Bigg|_{t'=0} \\
&= \phi\left(\frac{\alpha_3 + \beta_3 t + \gamma\alpha_2}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right)\frac{\gamma\beta_2}{\sqrt{\sigma_2^2\gamma^2 + 1}},
\end{aligned}$$

for $t = 0, 1$. Similarly, we can approximate the average total effect,

$$\bar{\tau} \approx \phi\left(\frac{\alpha_3 + \gamma\alpha_2}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right)\frac{\gamma\beta_2 + \beta_3}{\sqrt{\sigma_2^2\gamma^2 + 1}}.$$

With these results, the proportion mediated is approximated by

$$\upsilon \approx \frac{1}{2}\left\{1 + \exp\left(-\frac{\beta_3\{\beta_3 + 2(\alpha_3 + \gamma\alpha_2)\}}{2(\sigma_2^2\gamma^2 + 1)}\right)\right\}\frac{\gamma\beta_2}{\beta_3 + \gamma\beta_2}.$$

Thus, one condition under which $\tilde{\upsilon}$ approximately equals υ is that the average direct effect is small.

This result extends to the situation in which the logit regression is used, that is, the model defined in Equations 13 and 14, where $\varepsilon_{i3} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_2^2)$ and ϵ_{i3} is an i.i.d. standard logistic random variable.

(*Appendices continue*)

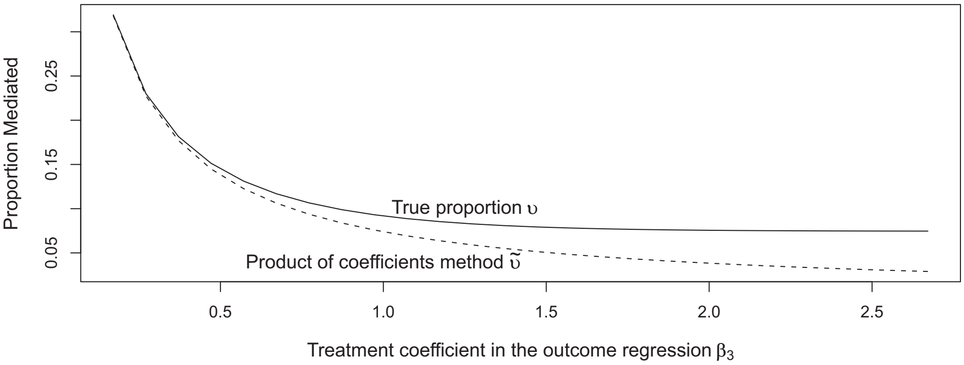

Figure G1. The bias of the product of coefficients method for estimating proportion mediated. The bias is calculated based on the JOBS II data. The outcome variable is whether someone finds employment, and the mediator represents the level of job search self-efficacy. The figure compares the true proportion mediated (solid line), υ defined in Equation 18, with the estimate based on the product of coefficients method (dashed line), $\tilde{\upsilon}$ defined in Equation 19. The bias increases as the coefficient for the treatment in the outcome regression, β_3, becomes greater.

The calculation similar to the above yields the following approximation to the true proportion mediated:

$$\upsilon \approx \frac{1}{2}\left\{1 + \int_{-\infty}^{\infty} \exp\left(-\frac{\beta_3\{\beta_3 + 2(\alpha_3 + \gamma\alpha_2 - \varepsilon_{3i})\}}{2\gamma^2\sigma_2^2}\right) \times \frac{\exp(\varepsilon_{3i})}{\{1 + \exp(\varepsilon_{3i})\}} d\varepsilon_{3i}\right\} \frac{\gamma\beta_2}{\beta_3 + \gamma\beta_2}.$$

Thus, when the average direct effect is small, the standard measure approximates the true proportion mediated well.

With Covariates

The above result can be generalized to the model with the observed pretreatment covariates. Because the analytical calculation is similar, we present the probit case, that is, $M_i = \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \varepsilon_{2i}$, with $\varepsilon_{2i} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_2^2)$ and $Pr(Y_i \mid T_i, M_i, X_i) = \Phi(\alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i)$. Using the same linear approximation as before, we obtain

$$\bar{\delta}(t) \approx \mathbb{E}\left\{\phi\left(\frac{\alpha_3 + \beta_3 t + \xi_3 X_i + \gamma(\alpha_2 + \xi_2 X_i)}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right)\right\} \frac{\gamma\beta_2}{\sqrt{\sigma_2^2\gamma^2 + 1}},$$

$$\bar{\tau} \approx \mathbb{E}\left\{\phi\left(\frac{\alpha_3 + \xi_3 X_i + \gamma(\alpha_2 + \xi_2 X_i)}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right)\right\} \frac{\gamma\beta_2 + \beta_3}{\sqrt{\sigma_2^2\gamma^2 + 1}},$$

for $t = 0, 1$. Thus, the proportion mediated is approximately equal to the following expression:

$$\upsilon \approx \frac{1}{2}\left[1 + \mathbb{E}\left\{\phi\left(\frac{\alpha_3 + \beta_3 + \xi_3 X_i + \gamma(\alpha_2 + \xi_2 X_i)}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right)\right\} \Bigg/ \mathbb{E}\left\{\phi\left(\frac{\alpha_3 + \xi_3 X_i + \gamma(\alpha_2 + \xi_2 X_i)}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right)\right\}\right] \frac{\gamma\beta_2}{\beta_3 + \gamma\beta_2}.$$

Thus, again, one condition under which ν approximately equals $\tilde{\upsilon}$ is that the direct effect is small.

An Example from the JOBS II Study

The difference between the proportion mediated measure used in the literature, $\tilde{\upsilon}$, and the true proportion mediated, υ, becomes larger as β_3 increases in magnitude. To illustrate this difference, we again use the JOBS II data and use the dichotomous outcome measure of whether someone was employed 6 months following the treatment. We first estimate the mediation and outcome models with our full set of covariates using linear and probit regression, respectively. For our illustration, we use these estimates, except that we fix the mediation effect to be 1 because the mediation effect in the JOBS data is quite small. Next, we calculate the proportion mediated at different values of β_3, beginning with the estimated value from the sample, .171, and plot the difference between $\tilde{\upsilon}$ and υ in Figure G1. Consistent with the analytical result derived above, the estimate based on the product of coefficients method diverges from the true proportion as β_3 becomes larger.

(*Appendices continue*)

Appendix H

The Details of the Sensitivity Analysis for Binary Outcome

For the binary outcome model given in Equation 21, identification of the average causal mediation effects under nonzero values of ρ requires several steps. For $t = 0, 1$, we can write the average causal mediation effects with nonzero ρ as

$$\bar{\delta}(t) = \mathbb{E}\left\{\Phi\left(\alpha_1 + \beta_1 t + \xi_1^\top X_i + \frac{\gamma\beta_2(1-t)}{\sqrt{\gamma^2\sigma_2^2 + 2\gamma\rho\sigma_2 + 1}}\right) - \Phi\left(\alpha_1 + \beta_1 t + \xi_1^\top X_i - \frac{\gamma\beta_2 t}{\sqrt{\gamma^2\sigma_2^2 + 2\gamma\rho\sigma_2 + 1}}\right)\right\}.$$

In the expression above, β_2 and σ_2^2 can be consistently estimated via the regression of M_i on $(1, T_i, X_i)$. The four parameters in this expression that we still need to identify are α_2, β_1, ξ_1, and γ. The first step toward identification requires estimating the probit regression: $Y_i = \mathbf{1}\{Y_i^* > 0\}$ with $Y_i^* = \alpha_1 + \beta_1 T_i + \xi_1^\top X_i + \varepsilon_{i1}$, where we assume $\varepsilon_{i1} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Second, we define $\tilde{\rho} = \mathrm{Corr}(\epsilon_{i1}, \epsilon_{i2})$, which identifies γ as

$$\gamma = \frac{-\rho + \tilde{\rho}\sqrt{(1-\rho^2)/(1-\tilde{\rho}^2)}}{\sigma_2}.$$

Next we estimate Equation 21, which gives a set of biased estimates when ρ is nonzero. We denote $\tilde{\gamma}$ as the biased coefficient for M_i in this probit model of the outcome, and we use it to obtain a consistent estimate of $\tilde{\rho} = \sigma_2\tilde{\gamma}/\sqrt{1 + \sigma_2^2\tilde{\gamma}^2}$. In fact, we can also obtain α_1 and β_1 from this probit equation: $\alpha_1 = \tilde{\alpha}_3\sqrt{1-\tilde{\rho}^2} + \alpha_2\tilde{\rho}/\sigma_2$, $\beta_1 = \tilde{\beta}_3\sqrt{1-\tilde{\rho}^2} + \beta_2\tilde{\rho}/\sigma_2$, and $\xi_1 = \tilde{\xi}_3\sqrt{1-\tilde{\rho}^2} + \xi_2\tilde{\rho}/\sigma_2$, where $(\tilde{\alpha}_3, \tilde{\beta}_3, \tilde{\xi}_3)$ are the intercept and the coefficients for T_i and X_i, respectively. We now have consistent estimates for all the quantities needed to estimate the average causal mediation effects when ρ is nonzero.

Received August 5, 2009
Revision received May 28, 2010
Accepted June 14, 2010 ■