# A General Approach to Causal Mediation Analysis[*]

Kosuke Imai[†] Luke Keele[‡] Dustin Tingley[§]

First Draft: June 13, 2009
This Draft: August 5, 2009

Approximately 10,638 words
(excluding references and appendecies)

## Abstract

In a highly influential paper, Baron and Kenny (1986) proposed a statistical procedure to conduct a causal mediation analysis and identify possible causal mechanisms. This procedure has been widely used across many branches of the social and medical sciences and especially in psychology and epidemiology. However, one major limitation of this approach is that it is based on a set of linear regressions and cannot be easily extended to more complex situations that are frequently encountered in applied research. In this paper, we propose an approach that generalizes the Baron-Kenny procedure. Our method can accommodate linear and nonlinear relationships, parametric and nonparametric models, continuous and discrete mediators, and various types of outcome variables. We also provide a formal statistical justification for the proposed generalization of the Baron-Kenny procedure by placing causal mediation analysis within the widely-accepted counterfactual framework of causal inference. Finally, we develop a set of sensitivity analyses that allow applied researchers to quantify the robustness of their empirical conclusions. Such sensitivity analysis is important because as we show the Baron-Kenny procedure and our generalization of it rest on a strong and untestable assumption even in randomized experiments. We illustrate the proposed methods by applying them to a randomized field experiment, the Job Search Intervention Study (JOBS II). We also offer easy-to-use software that implements all of our proposed methods.

**Key Words:** causal inference, causal mechanisms, sensitivity analysis, sequential ignorability, structural equation modeling, unobserved confounder

[*] The companion paper, which forms the theoretical basis of this paper, is available as Imai, Keele, and Yamamoto (2008). All of our proposed methods can be implemented via easy-to-use software, **mediation**, which is freely available at the Comprehensive R Archive Network (http://cran.r-project.org/web/packages/mediation). The details about this software are described in Imai, Keele, Tingley, and Yamamoto (2009). We thank Matt Incantalupo, Dave Kenny, Dave MacKinnon, Scott Maxwell, Walter Mebane, and Kris Preacher for useful comments. Financial support from the National Science Foundation (SES-0752050) and the Princeton University Committee on Research in the Humanities and Social Sciences is acknowledged.

[†] Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6610, Email: kimai@princeton.edu, URL: http://imai.princeton.edu

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

[§] Ph.D. candidate, Department of Politics, Princeton University, Princeton NJ 08544, Email: dtingley@princeton.edu

# 1 Introduction

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 randomized experiments and these statistical methods is that they can only provide 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 since the identification of causal mechanisms is required to test competing theoretical explanations of the same causal phenomena. While policy makers may only be interested in the causal effects of policy interventions, social scientists are generally interested in understanding the mechanisms that underlie such causal effects and testing alternative theories that explain them.

It is in this setting causal mediation analysis plays an essential role by establishing the influence of intermediate variables (or mediators) that lie in the causal pathway between the treatment and outcome variables. The existence of such mediators provides key evidence in the identification of causal mechanisms because it implies that the treatment affects the outcome by changing the level of mediators and consequently the outcome. In a highly influential paper, Baron and Kenny (1986) proposed a statistical procedure to conduct a causal mediation analysis. This procedure has been widely used across many branches of the social and natural sciences and especially in psychology and epidemiology (As of June 2009, the original paper has been cited more than 14,000 times according to Google Scholar; for an earlier formulation see Judd and Kenny (1981) and for the explicit statistical development see (MacKinnon and Dwyer, 1993; MacKinnon *et al.*, 1995)). However, one major limitation of this approach is that it is based on a set of linear regressions and cannot be easily extended to more complex situations that are frequently encountered in applied research (for other limitations see Green *et al.*, 2010).

In this paper, we propose an approach that generalizes the Baron-Kenny (hereafter, BK) procedure. Our method is relatively simple and yet can accommodate linear and nonlinear relationships, parametric and nonparametric models, continuous and discrete mediators, and various types of outcome variables. In the methodological literature, a number of scholars have considered the extension of the BK procedure to some of these settings (e.g., Wang and Taylor, 2002; MacKinnon *et al.*, 2007; Li *et al.*, 2007; MacKinnon, 2008). Our approach encompasses many of these existing methods as special cases, while we also show that other prior methods yield invalid estimates of causal mediation effects. Thus, our proposed generalization accomplishes many of future statistical tasks identified in a recent review paper by MacKinnon and Fairchild (2009).

Another contribution of this paper is to provide a formal statistical justification for the proposed generalization of the BK procedure. Following the recently published papers (e.g., Jo, 2008; Sobel, 2008), we place causal mediation analysis within the widely-accepted counterfactual framework of causal inference. Imai *et al.* (2008) prove that under the sequential ignorability assumption the average causal mediation effects are nonparametrically identified. Sequential ignorability consists of two assumptions: (1) conditional on the observed pre-treatment covariates the treatment is independent of all potential values of the outcome and mediating variables, and (2) the observed mediator is independent of all potential outcomes given the observed treatment and pre-treatment covariates. Imai *et al.* (2008) also show that under this assumption the BK procedure yields a valid estimate of the average causal mediation effect if the linearity assumption is satisfied. In this paper, we show how to estimate causal mediation effects without the linearity assumption.

Finally and perhaps most importantly, we develop a set of sensitivity analyses that allow applied researchers to quantify the robustness of their empirical conclusions. Such sensitivity analysis is essential for causal mediation analysis because, as we show, the Baron-Kenny procedure and our generalization of it rest on a strong and untestable assumption even in randomized experiments. In particular, there may exist unobserved confounders that causally affect both the mediator and the outcome even after conditioning on the observed treatment and pre-treatment 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 context of the BK procedure, Imai *et al.* (2008) propose a straightforward way to check how severe the violation of the key identifying assumption would need to be in order for the original conclusions to be reversed. We generalize this sensitivity analysis so that it can be applied to a variety of other settings. We conclude with a brief outline of the software that implements our proposed methods.

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

To motivate the concepts and methods that we present, we rely on a running example from the psychology literature on mediation. We use the Job Search Intervention Study (JOBS II) as our example, though other data sets used for mediation analysis would also fulfill our purposes. JOBS II is a randomized field experiment that investigates the efficacy of a job training intervention on unemployed workers. The program is designed to not only increase reemployment among the unemployed but also enhance the mental health of the job seekers. In the JOBS II field experiment, 1,801 unemployed workers received a pre-screening questionnaire and were then randomly assigned to treatment and control groups. Those in the treatment group participated in job-skills workshops. In the workshops, respondents learned job-search skills and coping strategies for dealing with setbacks in the job-search process. Those in the control condition received a booklet describing job-search tips. In follow-up interviews, the 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 *et al.*, 1995; Vinokur and Schul, 1997). In the JOBS II data, a continuous measure of job-search self-efficacy represents this key mediating variable. In addition to the outcome and mediators, the JOBS II data also include baseline covariates that were measured prior to the administration of the treatment. The most important of these is the pre-treatment level of depression which is measured using 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.

# 3 Statistical Framework for Causal Mediation Analysis

In this section, we describe the counterfactual framework of causal inference which is widely accepted in the statistical literature. Following the prior work (e.g., Robins and Greenland, 1992; Pearl, 2001; Glynn, 2008; Imai *et al.*, 2008), we define causal mediation effects using the potential outcomes notation. We then review the key result of Imai *et al.* (2008) and show the condition under which the BK procedure and its variants yield valid estimates of causal mediation effects. This establishes a clear connection between the modern statistical framework of causal inference and the traditional structural equation modeling approach used in the social sciences. Finally, we explain how our approach differs from the existing approach based on the instrumental variable methods of Angrist *et al.* (1996).

## 3.1 The Counterfactual Framework

In the counterfactual framework of causal inference, the causal effect of the job training program for each worker can be defined as the difference between two potential outcomes; one potential outcome that would be realized if the worker participates in the job training program, and the other potential outcome that would be realized if the worker does not participate. Suppose that we use $T_i$ to represent the binary treatment variable, which is equal to $1$ if worker $i$ participated in the program and to $0$ otherwise (see Section 5.5 for an extension to non-binary 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 the worker $i$'s employment status if she participates in the job training program. Although there are two such potential values for each worker, only one of them is observed; for example, if worker $i$ actually did not participate in the program, then only $Y_i(0)$ is observed. Thus, if we use $Y_i$ to denote the observed value of employment status, then we have $Y_i = Y_i(T_i)$ for all $i$.

Given this setup, the causal effect of the job training program on worker $i$'s employment status can be defined as $Y_i(1) - Y_i(0)$. Of course, since 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 defined as $\mathbb{E}(Y_i(1) - Y_i(0))$ where the expectation is taken with respect to the random sampling of units from a target population. If the treatment is randomized as done in JOBS II, then $T_i$ is statistically independent of $(Y_i(1), Y_i(0))$ because the probability of receiving the treatment is identical for every observation; formally, we write $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i$. Under this setting, it is immediate that the average causal effect can be identified as the observed difference in means between the treatment and control groups, $\mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(Y_i(1) \mid T_i = 1) - \mathbb{E}(Y_i(0) \mid T_i = 0) = \mathbb{E}(Y_i \mid T_i = 1) - \mathbb{E}(Y_i \mid T_i = 0)$, which is a 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 interaction between units. In the current context, this means for example that worker $i$'s employment status is not influenced by whether or not another worker $j$ participates in the training program. This assumption is apparent from the fact that the potential values of $Y_i$ are written as a function of $T_i$ which does not depend on $T_j$ for $i \neq j$. This assumption is best addressed through design. For example, analysts would want to ensure that participants in the experiment were not from the same household. The analyses that follow are conducted under this assumption.

## 3.2 Defining Causal Mediation Effects

The existing statistics literature uses the counterfactual framework and notation above 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 where the mediating variable is worker's level of confidence in their ability to perform essential job-search activities such as completing an employment application.

The 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. Since the level of job-search self-efficacy can be affected by the program participation, there exist two potential values, $M_i(1)$ and $M_i(0)$, only one of which will be observed, i.e., $M_i = M_i(T_i)$. For example, if worker $i$ actually participates in the program $T_i = 1$, then we observe $M_i(1)$ but not $M_i(0)$.

Next, we define the potential outcomes. Previously, the potential outcomes were only a function of the treatment, but in a mediation analysis the potential outcomes depend on the mediator as well as the treatment variable. 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 only observe one of multiple potential outcomes, and the observed outcome $Y_i$ equals $Y_i(T_i, M_i(T_i))$. Lastly, no interference between units is assumed; the potential mediator values for each unit do not depend on the treatment status of the other units, and the potential outcomes of each unit also do not depend on the treatment status and the mediator value of the other units.

We can now define causal mediation 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 $t = 0, 1$. As it is clear from this definition, causal mediation effects represent the indirect effects of the treatment on the outcome through the mediating variable (Pearl, 2001; Robins, 2003).

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 treatment condition, i.e., $M_i(1)$, to the value that would be observed under the control condition, i.e., $M_i(0)$, while holding the treatment status at $t$? Clearly, if the treatment has no effect on the mediator, i.e., $M_i(1) = M_i(0)$, then the causal mediation effects are zero. While $Y_i(t, M_i(t))$ 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, we can define the direct effects 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 effects of the program on worker $i$'s depression level when holding the level of his job-search self-efficacy constant at the level that would be realized under the program participation. Pearl (2001) calls $\zeta_i(t)$ as *natural direct effects* to distinguish them from *controlled direct effects* of the treatment. The controlled direct effect of the treatment is defined as $Y_i(1, m) - Y_i(0, m)$ for each $m \in \mathcal{M}$ where $\mathcal{M}$ is the support of the distribution of the mediator. It is important to note that causal mediation effects $\delta_i(t)$ fundamentally differ from controlled direct effects of the mediator, i.e., $Y_i(t, m) - Y_i(t, m')$ for $m \neq m'$ and $m, m' \in \mathcal{M}$, which represent the difference between the two potential outcomes; one under the counterfactual scenario where the mediator is set to a particular value $m$ and the other under the scenario where the mediator is set to a different value $m'$ (the treatment is held constant at $t$ in both scenarios).

The controlled direct effects of mediation are appropriate quantities of interest if researchers are interested in the causal effect of the mediator while holding the treatment variable is constant. The causal mediation effects, on the other hand, represent the indirect causal effect of the treatment through the mediator of interest. To identify the controlled direct effects, one can randomize both the treatment and mediator variables. The identification of causal mediation effects, however, is not straightforward because causal mediation effects cannot be even defined if researchers manipulate the mediator. In psychological experiments, researchers are typically interested in causal mediation effects because one of their primary goal is to identify psychological mechanisms that explain the causal effects of the treatment intervention on behavioral outcomes. Also, in such situations, the mediator is a psychological factor that is difficult to manipulate (e.g., Spencer *et al.*, 2005). Lastly, we note that direct effects represent all possible causal mechanisms except the one accounted for by the mediator $M_i$.

As expected, the total effect of the treatment, $\tau_i$, can be decomposed into the causal mediation and direct effects in the following manner, $\tau_i \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \frac{1}{2}\sum_{t=0}^{1}\{\delta_i(t) + \zeta_i(t)\}$. In addition, if we make the no-interaction assumption 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 $\zeta_i = \zeta_i(1) = \zeta_i(0)$), then we see that the causal mediation and direct effects sum to the total effect, i.e., $\tau_i = \delta_i + \zeta_i$.

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

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

for $t = 0, 1$. Similarly, we can define the average direct and total effects.

$$\begin{aligned}
\bar{\zeta}(t) &\equiv \mathbb{E}(Y_i(1, M_i(t)) - Y_i(0, M_i(t))), \\
\bar{\tau} &\equiv \mathbb{E}(Y_i(1, M_i(1)) - Y_i(0, M_i(0))) = \frac{1}{2}\{\bar{\delta}(0) + \bar{\delta}(1) + \bar{\zeta}(0) + \bar{\zeta}(1)\}.
\end{aligned}$$

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, i.e., $\bar{\tau} = \bar{\delta} + \bar{\zeta}$.

## 3.3 Sequential Ignorability Assumption

To estimate various causal effects of interest defined above, we use the following identification assumption introduced by Imai *et al.* (2008). Let $X_i$ be a vector of the observed pre-treatment confounders for unit $i$ where $\mathcal{X}$ denotes the support of the distribution of $X_i$. In the JOBS II data, it includes for each unemployed worker the pre-treatment 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 pre-treatment confounders, the assumption can be formally written as,

ASSUMPTION 1 (SEQUENTIAL IGNORABILITY (IMAI *et al.*, 2008)) *We assume that the following two statements of conditional independence hold,*

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

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

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

Imai *et al.* (2008) discuss how this assumption differs from those proposed in the prior literature. Assumption 1 is called sequential ignorability because two ignorability assumptions are made sequentially. First, given the observed pre-treatment confounders, the treatment assignment is assumed to be ignorable, i.e., statistically independent of potential outcomes and potential mediators. In the JOBS II study, this first ignorability assumption is satisfied because workers were randomly assigned to the treatment and control groups. In contrast, the assumption is not guaranteed to hold in observational studies where subjects may self-select into the treatment group. In such situations, a common strategy of empirical researchers is to collect as many pre-treatment confounders as possible so that the ignorability of treatment assignment is more credible once the observed differences in these confounders between the treatment and control groups are appropriately adjusted.

The second part of Assumption 1 is that the mediator is ignorable given the observed treatment and pre-treatment confounders. That is, the second part of the sequential ignorability assumption is made conditional on the observed value of the ignorable treatment and the observed pre-treatment confounders. Unlike the ignorability of treatment assignment, however, the ignorability of the mediator may not hold even in randomized experiments. In the JOBS II study, for example, the randomization of the treatment assignment does not justify this second ignorability assumption since the post-treatment 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 pre-treatment characteristics the mediator can be regarded as if it is 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 pre-treatment variables. Indeed, without an additional assumption, we cannot condition on the post-treatment confounders even if such variables are observed by researchers (e.g., Avin *et al.*, 2005). This means that, similar to the ignorability of treatment assignment in observational studies, it is difficult to know for certain whether or not the ignorability of the mediator holds even after researchers collect as many pre-treatment confounders as possible.

Such an assumption is often referred to as nonrefutable since one cannot disprove it with some possible configuration of the data (Manski, 2007). Thus, in Section 6, 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 since they allow the researcher to probe whether a substantive conclusion is robust to violations of the assumption.

## 3.4 Causal Interpretation of the Baron-Kenny and Other Related Estimates

In a highly influential article, Baron and Kenny (1986) proposed a procedure to estimate mediation effects based on the linear structural equation modeling (LSEM). Here, we use a more general version of their model by including the observed pre-treatment confounders as additional linear predictors,

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

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

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

The BK procedure consists of testing if $\beta_1, \beta_2, \gamma$ are all statistically significant in the expected directions when each of the three equations is separately fitted via the least squares method, and, then if they are, using $\hat{\beta}_2\hat{\gamma}$ as an estimated mediation effect.

Is this a valid estimate for the causal mediation effect under the potential outcomes framework? Using the counterfactual framework outlined above, Imai *et al.* (2008) prove that under sequential ignorability and the additional no-interaction assumption, i.e., $\bar{\delta}(1) = \bar{\delta}(0)$, the BK estimate can be interpreted as a valid estimate (i.e., asymptotically consistent) of the causal mediation effect so long as the linearity assumption holds (see also Jo, 2008). To understand the connection between the quantities estimated in the BK approach and the causal mediation effect defined earlier, we first write each potential outcome within the LSEM framework,

$$\begin{aligned} Y_i(T_i, M_i(T_i)) &= \alpha_1 + \beta_1 T_i + \xi_1^\top X_i + \epsilon_{i1}(T_i, M_i(T_i)), \\ M_i(T_i) &= \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \epsilon_{i2}(T_i), \\ Y_i(T_i, M_i(T_i)) &= \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \epsilon_{i3}(T_i, M_i(T_i)). \end{aligned}$$

An argument that is nearly identical to the proof of Theorem 2 in Imai *et al.* (2008) 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 $t = 0, 1$. Finally, the average total effect is given by $\beta_1$ which equals $\beta_2\gamma + \beta_3$.

Since $\hat{\beta}_1 = \hat{\beta}_2\hat{\gamma} + \hat{\beta}_3$ always holds under this model, equation (5) is redundant given equations (6) and (7). Thus, one of steps in the BK procedure is not necessary; $\hat{\beta}_1$ does not have to be statistically significant in order to conclude that average mediation effects exist. If fact, the average causal mediation and average direct effects can offset each other such that the average total effects are close to being non-existent.

Thus so long as an analyst is willing to adopt the linearity and no-interaction assumptions along with sequential ignorability, the BK procedure provides a valid estimate of the causal mediation effect.

**Relaxing the No-interaction Assumption.** In a series of recent papers, Kraemer and her colleagues (Kraemer *et al.*, 2002, 2008) propose an alternative to the standard BK procedure, which they term the MacArthur approach that relaxes the no-interaction assumption (see also Judd and Kenny, 1981). They argue that assuming there is no interaction between the treatment and mediator is often unrealistic and replace equation (7) with the following alternative specification,

$$Y_i \;=\; \alpha_3 + \beta_3 T_i + \gamma M_i + \kappa T_i M_i + \xi_3^\top X_i + \epsilon_{i3}. \tag{8}$$

Kraemer *et al.* (2008) propose that in addition to $\hat{\beta}_2$, either $\hat{\gamma}$ or $\hat{\kappa}$ must be statistically indistinguishable 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 rather than on the coefficient of this interaction term. Following Imai *et al.* (2008), it is easy to show that under this alternative model specification and Assumption 1, the average causal mediation effects are given by,

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

for $t = 0, 1$. In addition, the average direct effects and the average total effects are given by,

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

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

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 $\overline{X}$. Finally, to estimate the average total effect, we may fit the following model,

$$Y_i \;\; = \;\; \alpha_1 + \beta_1 T_i + \xi_1^\top X_i + \eta^\top T_i X_i + \epsilon_{i1},$$

where the average total effect is given by $\bar{\tau} = \beta_1 + \eta^\top \mathbb{E}(X_i)$. Appendix A derives the asymptotic variance for each of these quantities of interest.

As we have shown above, under the sequential ignorability assumption, the causal interpretation can be given to the BK procedure so long as the linearity and no-interaction assumptions hold. The MacArthur approach can also be placed within the counterfactual framework of causal inference.

## 3.5 Relationship with Instrumental Variables

Recently, some scholars have considered the use of instrumental variables for causal mediation analysis as an alternative to the BK method (e.g., Albert, 2008; Jo, 2008; Sobel, 2008). The instrumental variable approach to mediation is at times placed within a statistical framework called principal stratification. 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, while 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 forms of causal mediation analysis that is used in psychology and other social science research. A more general and promising approach is the causal mediation analysis based on principal stratification (see e.g., Gallop *et al.*, 2009)

## 3.6 An Application to JOBS II

We illustrate the BK 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 the sequential ignorability assumption more credible, we include the full set of covariates described in Section 2. Table 1 present the estimated causal quantities of interest based on the BK procedure and the instrumental variables method. First, we use the standard BK procedure assuming no interaction between the mediator and the treatment. Under this model, we find a small (but statistically significant at the 95% level) negative mediation effect (the first column). Since 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 column of the table present the results. The basic findings resemble the ones based on the standard BK procedure, and there is little evidence for the presence of the interaction effects for both mediation and direct effects. 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 significant.

# 4 A Generalization of the Baron-Kenny Procedure

In this section, we propose a generalization of the BK procedure that is applicable beyond the original linear structural equation models. 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. We first extend the nonparametric identification result of Imai *et al.* (2008) and then formalize their suggestion by proposing the two general algorithms to estimate the average causal mediation effects under the sequential ignorability assumption (see Wang and Taylor, 2002; VanderWeele, 2009; Huang *et al.*, 2004; Glynn, 2008, for related methods).

## 4.1 Nonparametric Identification under Sequential Ignorability

Imai *et al.* (2008) prove that under Assumption 1 the average causal mediation effects are nonparametrically identified. That is, without any additional distributional or functional-form assumptions (like the commonly assumed linearity assumption from OLS regression as done in the BK procedure), the average causal mediation effects can be inferred from the observed data. This identification result is important since it suggests the possibility of constructing a general method of estimating the average treatment effect for outcome and mediating variables of any types and using any parametric or nonparametric models. Moreover, it implies that we may estimate causal mediation effects while imposing weaker assumptions about the correct functional form or distribution of the observed data. Here, we slightly generalize this result to show that the distribution of any counterfactual outcome is identified under Assumption 1.

THEOREM 1 (NONPARAMETRIC IDENTIFICATION) *Under Assumption 1, we can identify,*

$$f(Y_i(t, M_i(t')) \mid X_i = x) \quad = \quad \int_{\mathcal{M}} f(Y_i \mid M_i = m, T_i = t, X_i = x)\, dF_{M_i}(m \mid T_i = t', X_i = x),$$

*for any $x \in \mathcal{X}$ and $t, t' = 0, 1$.*

Proof is omitted since it is a a straightforward generalization of Theorem 1 of Imai *et al.* (2008).

Theorem 1 implies that under the sequential ignorability assumption and conditional on a certain value of the pre-treatment confounders $X_i = x$, one can estimate the marginal distribution of counterfactual outcome $Y_i(t, M_i(t'))$ once the two conditional distributions of observed variables, i.e., $M_i$ given $(T_i, X_i)$ and $Y_i$ given $(M_i, T_i, X_i)$, are identified. This makes sense since in the original BK procedure, these conditional distributions are given by the linear regression models of equations (6) and (7).

Since Theorem 1 is based on no specific model, it allows us to develop a general estimation procedure for various causal mediation effects of interest. This result allows us to develop general alogrithms for estimating causal mediation effects, to which we now turn.

## 4.2 The Proposed Algorithms

Theorem 1 suggests that in order to obtain one Monte Carlo draw of a counterfactual outcome $Y_i(t, M_i(t'))$ for any $t, t'$ and given $X_i = x$, we can first sample $M_i(t')$ from $f(M_i \mid T_i = t', X_i = x)$ and then, given this draw, sample $Y_i(t, M_i(t'))$ from $f(Y_i \mid T_i = t, M_i(t'), X_i = x)$. Once we obtain these Monte Carlo draws, we can in principle compute any quantities of interest in causal mediation analysis which are functions of these counterfactual outcomes (so long as they do not involve the joint distribution of $Y_i(t, M_i(1))$ and $Y_i(t, M_i(0))$ since only marginal distributions are identified).

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 where parametric models are specified for the mediator and the outcome variable. To make the exposition concrete, we describe the algorithm to estimate the average causal mediation effects. The proposed algorithm is based on the quasi-Bayesian approximation of King *et al.* (2000) where the posterior distribution of quantities of interest is approximated by their sampling distribution.

ALGORITHM 1 (PARAMETRIC INFERENCE) *Suppose that the quantity of interest is the average causal mediation effect, i.e.,* $\bar{\delta}(t)$.

**Step 1:** *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* $\theta_M$ *and* $\theta_Y$ *represent model parameters. This step simply entails fitting models for the observed outcome and mediator variables.*

**Step 2:** *Sample* $J$ *copies of* $\theta_M$ *and* $\theta_Y$ *from their sampling distributions and denote them* $\theta_M^{(j)}$ *and* $\theta_Y^{(j)}$*, respectively. The multivariate normal distribution is often used as an asymptotic approximation. That is, we take random draws from a multivariate normal distribution based on the parameter vectors and covariance matrices from the parametric models fit in Step 1.*

**Step 3:** *For each $j = 1, 2, \ldots, J$, repeat the following three steps.*

1. *For each $t = 0, 1$ and each $i = 1, 2, \ldots, n$, 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 $k = 1, 2, \ldots, K$. Here, we generate two predictions for each units' mediator status for each of parameter draws from $\theta_M^{(j)}$. The first prediction is generated with $T_i$ fixed at the treatment status, $t = 1$, and the second prediction is generated with $T_i$ fixed at the control status, $t = 0$.*

2. *For each $t = 0, 1$ and each $i = 1, 2, \ldots, n$, sample one copy of $Y_i(t, M_i^{(jk)}(t'))$ from $f_{\theta_M^{(j)}}(Y_i \mid t, M_i^{(jk)}(t'), X_i)$ and denote it as $Y_i^{(jk)}(t, M_i^{jk}(t'))$ for $k = 1, 2, \ldots, K$. In this step, we generate two outcome predictions holding the treatment status fixed. With $t$ fixed at 1, we generate outcome predictions with the mediator predictions under treatment from Step 3.1 Then with $t$ still fixed at 1, we generate outcome predictions with mediator predictions under control from Step 3.1. We repeat this for the outcome under $t = 0$.*

3. *Compute the average causal mediation effect as,*

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

   *Here, 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 with $\bar{\delta}^{(j)}(t)$, which is a distribution of predicted average mediation effects one for each of the $J$ draws from Step 2.*

**Step 4:** *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 while 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 of interest other than the average causal mediation effects. For example, the $\alpha$-quantile average causal mediation effects defined in Section 5.1 can be estimated by computing the sample quantile of $Y_i^{(jk)}(t, M_i^{(jk)}(t'))$ across treatment and control rather than its sample average.

Furthermore, depending on the parametric models chosen by researchers, further simplifications of Algorithm 1 may be possible. In particular, 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,

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

For example, if the outcome variable is binary and modeled using a logistic regression, this simplification is possible. Another example would be a situation where the median causal mediation effects are quantities of interest and a quantile regression model is used for the outcome variable.

Finally, suppose that one wishes to use non/semi-parametric model or a quantile regression for either the outcome or mediator model (or both). For these models, we propose to use a nonparametric bootstrap procedure to obtain a distribution of causal mediation. Although this algorithm is applicable to parametric inference as well, Algorithm 1 is typically much more computationally efficient. We describe the nonparametric bootstrap algorithm below.

ALGORITHM 2 (NONPARAMETRIC INFERENCE) *Suppose that the quantity of interest is the average causal mediation effect, i.e.,* $\bar{\delta}(t)$.

**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 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 which are now allowed to be nonparametric or semiparametric models.*
2. *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. *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* $k = 1, 2, \ldots, K$*. Outcome predictions are generated for each treatment status and two mediator predictions.*

4. *Compute the average causal mediation effect as,*

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

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

**Step 2:** *Compute the point estimate of $\bar{\delta}(t)$ and its uncertainty estimates using the $J$ estimates from the bootstrap sampling distribution. As before, 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 and 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 detail here. The above algorithms are much more complex than the usual product of coefficients method. However, it allows researchers to go beyond the BK procedure by relaxing the linearity assumption and modeling non-continuous outcome and mediator variables. If the algorithms are applied to the LSEM, then one would obtain the results that are approximately equal to those based on the BK procedure (e.g., the so-called Sobel test). Another advantage of these algorithms is that they allow us to develop an easy-to-use software which computes the estimates of causal mediation effects and uncertainty estimates under various modeling assumptions (Imai *et al.*, 2009).

# 5 Theoretical and Empirical Illustrations

In this section, we theoretically and empirically illustrate how our proposed algorithms can handle a variety of situations that often arise in causal mediation analysis. We also discuss how our methods relate to the existing approaches in the literature.

## 5.1 Quantile Mediation Effects

The BK procedure, which is based on the LSEM, provides 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 mean. In the JOBS II example, policy makers might be concerned about individuals with high levels of depression rather than those with the average level of depression. It may also be possible that a few individuals respond to the intervention in dramatic fashion making the average a poor description of how most individuals respond to the treatment. In such instances, *quantile mediation effects*, which represents the difference between a certain quantile (e.g., median) of two relevant potential outcomes, may be of interest. Formally, $\alpha$-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, $\tilde{\zeta}_\alpha(t) \equiv q_{1t}(\alpha) - q_{0t}(\alpha)$.

The quantile regression allows for a convenient way to model the quantiles of the outcome distribution while adjusting for a variety of covariates (Koenker, 2008). Specifically, we replace equation (7) with the quantile regression model. The usual product of coefficients method is not defined for the quantiles, but Algorithm 2 generalizes to the quantile regression and provides estimates of uncertainty.

**Empirical Illustration.** 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 1 presents the estimated quantile causal mediation and direct effects and their 95% confidence intervals using Algorithm 2. Both the direct and indirect effects in the figure are conditional on the pretreatment covariates included in the quantile regression models. The left panel demonstrates the effect of the intervention that occurs through the mediator job-search self-efficacy. The right panel shows how the intervention affects quantiles of depression directly. We see that the magnitude of the estimated mediation effects increases slightly as we move 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 widely across the quantiles although the confidence intervals are wide and always include zero.

## 5.2 Use of Nonparametric and Semiparametric Regressions

The BK procedure estimates the average causal mediation effects 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 amongst 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. Alternatively, mediation effects might be smallest amongst 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, nonparametric and semiparametric regressions may be used to avoid linear functional form assumptions (e.g., Keele, 2008). These models estimate the functional form from the data while imposing much weaker assumptions. While one could use the interaction term in the BK procedure, any more complex functional form would make the method infeasible. 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 and Tibshirani, 1990). In particular, we fit the following regression equation instead of equation (7), $Y_i = \alpha_3 + \beta_3 T_i + s(M_i) + \xi^\top X_i + \epsilon_{i3}$, where $s(\cdot)$ is assumed to be a smooth function that we estimate nonparametrically from the data. We also relax the no-interaction assumption by fitting the following model, $Y_i = \alpha_3 + \beta_3 T_i + s_0(M_i)(1 - T_i) + s_1(M_i)T_i + \xi^\top X_i + \epsilon_{i3}$, instead of equation (8). We use an implementation in the `R` package `mgcv` to fit GAM (Wood, 2006; Keele, 2008).

**Empirical Illustration.** Figure 2 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. The left panel (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 mid-point of the scale before there is any attendant decrease in depression. The middle and right panels 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 the right panel 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 conclusions drawn from the nonparametric estimate.

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 small, but statistically significant, negative mediation effect. This effect is estimated to be slightly larger in magnitude compared to 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. Like 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.

## 5.3 Discrete Mediator

Next, consider the situation where the mediator is discrete, a common occurrence for many applications in psychology where the measure for the mediator is often an ordered scale or binary. In this case, Theorem 1 reduces to,

$$h(Y_i(t, M_i(t')) \mid X_i = x) \ = \ \sum_{m \in \mathcal{M}} g(Y_i \mid M_i = m, T_i = t, X_i = x) \Pr(M_i = m \mid T_i = t', X_i = x). \tag{12}$$

When the mediator is binary and the quantity of interest is the average causal mediation effect, Equation 12 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). This implies that we can simplify Algorithm 1. Specifically, 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,

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

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

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 since discrete and binary measures are extremely common, but the standard BK method does not easily generalize.

**Empirical Illustration.** 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 recode the worker's job-search self-efficacy into two different 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 recode 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. We see that the treatment decreased depression by increasing job-search self-efficacy but it did have little direct causal effect. However, we see little differences in the average mediation effect across treatment status.

Next, we use the four-category measure for the mediator and fit an ordered probit model for the mediation equation. Table 4 presents the results. First, we assume no interaction between the mediator and the treatment. Consistent with the results given in Table 1, we find a small negative average mediation effect. We also relax the no-interaction assumption by including an interaction term $T_i M_i$ in the outcome regression model. The second and third column of the table present the results. As before, there is little evidence of an interaction effect. In general, we have shown that altering the model for the mediator in Step 1 of Algorithm 1 presents no complications for the estimation of the quantities of interest and provides estimates of statistical uncertainty.

## 5.4 Binary Outcome

One situation which 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., Freedman and Graubard, 1992; Wang and Taylor, 2002; Ditlevsen *et al.*, 2005; MacKinnon *et al.*, 2002, 2007;

MacKinnon, 2008). One important criticism of existing methods is that they lack a causal interpretation (Kaufman *et al.*, 2005). Here, we derive analytical expressions for causal mediation effect when the outcome is binary. We show that our general estimation 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 relatively simple model without the pre-treatment confounders (In Appendix C, we show that all of our analytical results will hold for the model with the observed pre-treatment covariates with some notational complexity),

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

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

where $\epsilon_{2i}$ and $\epsilon_{3i}$ are different i.i.d. random variables with zero mean, and $\mathrm{Var}(\epsilon_{2i}) = \sigma_2^2$ and $\mathrm{Var}(\epsilon_{3i}) = \sigma_3^2$. Note that under Assumption 1, we have the independence between two error terms. If $\epsilon_{2i}$ 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. Given the above model and Assumption 1, it can be shown that,

$$\int \mathbb{E}(Y_i \mid T_i = t, M_i)\, dF(M_i \mid T_i = t') = 1 - \Pr\{\gamma\epsilon_{2i} + \epsilon_{3i} \leq -\alpha_3 - \beta_3 t - \gamma(\alpha_2 + \beta_2 t')\},$$

for $t, t' = 0, 1$. Using this equality, we can derive the analytical expression for the average causal mediation effects. First, suppose $\epsilon_{3i}$ is an i.i.d. logistic random variable and $\epsilon_{2i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_2^2)$. Then, we can write the distribution function for $\epsilon_{1i} = \gamma\epsilon_{2i} + \epsilon_{3i}$ as,

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

which can be computed using a standard numerical integration techniques. Since this distribution function is symmetric around the origin, the average causal mediation effects can be written as,

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

whereas the average total effect equals $\bar{\tau} = H(\alpha_3 + \beta_3 + \gamma(\alpha_2 + \beta_2)) - H(\alpha_3 + \gamma\alpha_2)$.

Next, suppose $\epsilon_{3i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and $\epsilon_{2i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_2^2)$. Then, a similar calculation yields,

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

$$\bar{\tau} = \Phi\left(\frac{\alpha_3 + \beta_3 + \gamma(\alpha_2 + \beta_2)}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right) - \Phi\left(\frac{\alpha_3 + \gamma\alpha_2}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right), \tag{17}$$

for $t = 0, 1$. As in the LSEM case, we obtain the average total effect by fitting the probit model,

$$\Pr(Y_i = 1 \mid T_i) = \Phi(\alpha_1^* + \beta_1^* T_i), \tag{18}$$

where $\alpha_1^* = (\alpha_3 + \gamma\alpha_2)/\sqrt{\sigma_2^2\gamma^2 + 1}$ and $\beta_1^* = (\gamma\beta_2 + \beta_3)/\sqrt{\sigma_2^2\gamma^2 + 1}$.

**Exact Relationship with the Existing Methods.** There exist two commonly used methods for computing average mediation effects with binary outcomes. First, Freedman and Graubard (1992) suggest the "difference of coefficients" method, which is based on the difference between $\beta_1^*$ from equation (18) and $\beta_3$ from equation (14). Second, MacKinnon *et al.* (2007) advocate a "product of coefficients" method where the $\gamma$ parameter from equation (14) is multiplied by $\beta_2$ from equation (13) following the usual BK steps. As MacKinnon *et al.* (2007) correctly points out, since probit and logistic regressions are nonlinear models, the two methods give different estimates. Indeed, the nonlinearity of those models implies that unlike in the case of the BK procedure, neither of the two methods consistently estimates the average causal mediation effects given in equations (15) and (16).

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,

$$\upsilon \quad \equiv \quad \frac{\{\bar{\delta}(0) + \bar{\delta}(1)\}/2}{\tau}. \tag{19}$$

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 is used, (e.g., Freedman and Graubard, 1992; Ditlevsen *et al.*, 2005; MacKinnon *et al.*, 2007),

$$\tilde{\upsilon} \quad \equiv \quad \frac{\gamma\beta_2}{\gamma\beta_2 + \beta_3}. \tag{20}$$

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), i.e., $Y_i^*$. However, it does not generally equal the true proportion mediated $\upsilon$. Nevertheless, it is interesting to note that as shown in Appendix D, when the direct effect is small, $\tilde{\upsilon}$ may approximate $\upsilon$ well whether or not the logistic or probit regression is used and whether or not the pre-treatment covariates are included in the model.

**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 above to analyze whether mediation effects are present when the outcome is whether subjects were working more than 20 hours a week 6 months after the job training program. In estimating the average mediation effects with and without the interaction term, we include the same set of pretreatment covariates to bolster the sequential ignorability assumption.

The results are given in Table 5. We begin by outlining the results under the no-interaction assumption. Unlike what we observed for the depression outcome, here it does not appear that mediation occurred. The average mediation effect is very 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. As a result, the estimated proportion mediated is a mere 6%. A 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. *Thus, 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 in this context.*

## 5.5 Non-binary Treatment

The counterfactual framework and our generalization of the BK procedure so far have assumed that the treatment variable is binary. Baron and Kenny (1986) did not make this assumption, and allowed the variable acting on the mediator and (in the case of partial mediation) outcome variable to be continuous. While some subsequent researchers expressed the Baron-Kenny procedure in terms of non-binary treatment variables (e.g., Li *et al.*, 2007), others retain the dichotomous treatment variable framework (e.g., Wang and Taylor, 2002; Ditlevsen *et al.*, 2005; MacKinnon *et al.*, 2007). Causal mediation analysis based on the potential outcomes framework also tends to consider the cases of binary treatment variables (e.g., Jo, 2008; Sobel, 2008; Albert, 2008).

Fortunately, our approach can be extended to the case of non-binary treatment at the cost of notational complexity. For example, the definition of mediation effects will be generalized to,

$$\delta_i(t; t_1, t_0) \quad \equiv \quad Y_i(t, M_i(t_1)) - Y_i(t, M_i(t_0)), \tag{21}$$

for any prespecified levels of the treatment, $t_1 \neq t_0$, which equals the definition given in equation (1) when $t_1 = 1$ and $t_0 = 0$. The corresponding average causal mediation effect is then defined as $\bar{\delta}(t; t_1, t_0) \equiv \mathbb{E}(\delta_i(t; t_1, t_0))$. Since the validity of Theorem 1 does not depend on the distribution of the treatment, the algorithms presented in Section 4 can be used to make inferences about this and other causal quantities of interest. Using our algorithms, one may, for example, plot the estimated value of $\int \bar{\delta}(t; t_1, 0) dF_{T_i}(t)$ (i.e., the average causal mediation effects averaged over the distribution of the treatment) against each value of $t_1$ in order to investigate how the average causal mediation effects change as the function of the treatment intensity.

# 6 Sensitivity Analysis

As shown above, randomization of the treatment alone does not identify mediation effects estimates as causal. This means that even in randomized experiments an additional assumption, e.g., the second part of Assumption 1, is required for identification. Moreover, this assumption is nonrefutable in the sense that it cannot be tested with the observed data (Manski, 2007). Sensitivity analysis is an effective method for probing the plausibility of a nonrefutable assumption. The goal of a sensitivity analysis is to quantify the exact degree to which the key identification assumption must be violated in order for a researcher's original conclusion to be reversed. If an inference is sensitive, a slight violation of the assumption may lead to substantively different conclusions. Thus, given the importance of the ignorability assumption in causal mediation analyses, we argue that a mediation study is not complete without a sensitivity analysis. For example, Jo (2008) points out that the second part of Assumption 1 might be violated in the JOBS II study and states that "individuals who improved their sense of mastery by one point in the intervention program may have different observed and unobserved characteristics from those of individuals who equally improved their sense of mastery in the control condition" (p.317).

Imai *et al.* (2008) propose a sensitivity analysis based on the correlation between $\epsilon_{i2}$, the errors for the mediation model, and $\epsilon_{i3}$, the errors for the outcome model. They denote this correlation across the two errors terms as $\rho$, which serves as the sensitivity parameter. Under sequential ignorability $\rho$ equals zero, and nonzero values of $\rho$ imply departures from the ignorability assumption. Using this fact, they devise the following strategy for sensitivity analyses in mediation studies. Assume that $\rho$ is nonzero, the question of interest is whether the average causal mediation effect is still nonzero. Imai *et al.* (2008) show that it is possible to write the average causal mediation effect as a function of $\rho$ and model parameters that can be consistently estimated even though $\rho$ is nonzero. This allows one to calculate the average causal mediation effect under various values of $\rho$ and observe when it becomes zero. Clearly if small departures from zero in $\rho$ dramatically alter the average causal mediation effect, this suggests that the study is sensitivity to the ignorability assumption. Moreover, we can also observe when the confidence intervals contain zero, which in general provides a more conservative test. Below we extend the basic analysis in Imai *et al.* (2008) in two ways. First, we derive the average causal mediation effect as a function of $\rho$ for the continuous mediator and continous outcome case but with the no-interaction assumption relaxed. Second, we develop sensitivity analyses for when either the outcome or mediator is a binary variable.

## 6.1 The Baron-Kenny Procedure

Here, we extend the result in Imai *et al.* (2008) to the LSEM with an interaction term between the treatment and the mediator.

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 $\epsilon_{2i}$ and $\epsilon_{3i}$, i.e., $\rho$, is given (and is assumed to be constant across the treat-*

*ment 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 E. Thus, for $t = 0, 1$, $\bar{\delta}(t)$ becomes zero when $\rho$ is equal to the correlation between $\epsilon_{i1}$ and $\epsilon_{i2}$ 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. The iterative procedure described in Imai *et al.* (2008) can be used to obtain the confidence intervals under various values of $\rho$.

**Empirical Illustration.** We return to the example Section 3.4 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 $\rho$ is equal to $-.165$, and $\bar{\delta}(0) = 0$ when $\rho$ is $-.245$. Figure 3 graphically illustrates this point by plotting the estimated average mediation effects and their 95% confidence intervals as a function of $\rho$. We find that for $\bar{\delta}(0)$ the confidence intervals include zero for a $\rho$ value of $-.09$ and for $\bar{\delta}(1)$ at $-.06$, which further underscores the sensitivity of the estimate. Imai *et al.* (2008) find in another study that the mediation effects are zero for a $\rho$ value of 0.48. Thus the mediation effects, here, are considerably more sensitive than in that study.

## 6.2 Binary Mediator

Next, we extend the sensitivity analysis to the situation where the mediator is defined by,

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

where $\epsilon_{i2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, and the outcome is continuous and defined as in (8). We assume $(\epsilon_{i2}, \epsilon_{i3})$ are bivariate normal with the mean zero and covariance $\rho\sigma_3$, and $\rho$ remains the correlation between the two error terms. When $\rho$ is nonzero, the expectation and variance of the outcome are $\mathbb{E}(Y_i \mid T_i = t, M_i =$

$m, X_i = x) = \alpha_3 + \beta_3 t + \gamma m + \kappa t m + \xi_3^\top x + \rho\sigma_3\eta_m(t,x)$ and $\mathrm{Var}(Y_i \mid T_i = t, M_i = m, X_i = x) = \sigma_3^2\left\{1 - \rho^2\eta_m(t)(\eta_m(t) + \alpha_2 + \beta_2 t + \xi_2^\top x)\right\}$, where $\eta_1(t,x) = \phi(\alpha_2 + \beta_2 t + \xi_2^\top x)/\Phi(\alpha_2 + \beta_2 t + \xi_2^\top x)$ and $\eta_0(t,x) = -\phi(\alpha_2 + \beta_2 t + \xi_2^\top x)/\Phi(-\alpha_2 - \beta_2 t - \xi_2^\top x)$ are the inverse Mills ratios. Since $\alpha_2, \beta_2, \xi_2$ 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 using feasible generalized least squares with a known value for $\rho$. Finally, once these model parameters are estimated, we can now estimate the average causal mediation as a function of $\rho$ and other consistently estimated parameters: $\bar{\delta}(t) = (\gamma + \kappa t)\mathbb{E}\{\Phi(\alpha_2 + \beta_2 + \xi_2^\top X_i) - \Phi(\alpha_2 + \xi_2^\top X_i)\}$. Like the sample selection model of Heckman (1979), $\rho$ is actually identified given the nonlinearity of the model, but we will not use this fact since it only hinges on the functional form assumption.

**Empirical Illustration.** We now demonstrate the sensitivity analysis for the binary mediator case analyzed in Section 5.3. 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, i.e., $\{\bar{\delta}(1) + \bar{\delta}(0)\}/2$, under a series of $\rho$ 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 values of $\rho$ greater than $-.09$.

## 6.3 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,

$$Y_i \;=\; \mathbf{1}\{Y_i^* > 0\} \quad \text{where} \quad Y_i^* \;=\; \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \epsilon_{i3}, \tag{22}$$

where $\epsilon_{i3} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. 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 $\rho$ 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 $\rho\sigma_2$.

The result, given in Appendix F, parallels closely with the sensitivity analysis for LSEM. For example, taking the partial derivative with respect to $\rho$ shows that $\bar{\delta}(t)$ is monotonically decreasing (increasing) with respect to $\rho$ 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 the one for LSEM (see Theorem 4 of Imai *et al.* (2008)). Since 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 also use this result to conduct sensitivity analysis for the proportion mediated. Due to the space limitaiton, we omit an empirical illustration although the analysis would proceed as before.

### 6.4 Sensitivity Analysis based on the Coefficients of Determination

Despite its simplicity, researchers may find it difficult to interpret the magnitude of the sensitivity parameter $\rho$. In the case of the Baron-Kenny procedure, Imai *et al.* (2008) show how to interpret the same sensitivity analysis using the following decomposition of the error terms,

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

for $j = 2, 3$ where $U_i$ is an unobserved pre-treatment confounder that influences both the mediator and the outcome, and $\lambda_j$ represents an unknown coefficient for each equation.

Under this setup, they demonstrate that $\rho$ can be written as a function of the coefficients of determination, i.e., $R^2$s. First, $\rho$ can be expressed using the proportions of previously unexplained variances in the mediator and outcome regressions, i.e., $R^{*2}_M \equiv 1 - \text{Var}(\epsilon'_{i2})/\text{Var}(\epsilon_{i2})$ and $R^{*2}_Y \equiv 1 - \text{Var}(\epsilon'_{i3})/\text{Var}(\epsilon_{i3})$, respectively. This expression is given by $\rho = \text{sgn}(\lambda_2\lambda_3) R^*_M R^*_Y$. Thus, sensitivity analysis can be conducted once researchers specify whether the unobserved confounder $U_i$ affects the mediator and outcome regressions in the same direction, i.e., $\text{sgn}(\lambda_2\lambda_3)$, as well as the relative magnitude of those effects, i.e.,

${R^*}^2_M$ and ${R^*}^2_Y$.

Similarly, sensitivity analysis can be based on the proportion of original variance that is explained by the unobserved confounder in the mediator and outcome regressions, i.e., $\widetilde{R}^2_M \equiv \{\mathrm{Var}(\epsilon_{i2}) - \mathrm{Var}(\epsilon'_{i2})\}/\mathrm{Var}(M_i)$ and $\widetilde{R}^2_Y \equiv \{\mathrm{Var}(\epsilon_{i3}) - \mathrm{Var}(\epsilon'_{i3})\}/\mathrm{Var}(Y_i)$, respectively. In this case, the expression is given by $\mathrm{sgn}(\lambda_2\lambda_3)\widetilde{R}_M\widetilde{R}_Y/\sqrt{(1-R^2_M)(1-R^2_Y)}$ where $R^2_M$ and $R^2_Y$ 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 $\widetilde{R}^2_M = \{1 - \mathrm{Var}(\epsilon'_{i2})\}/\{\mathrm{Var}(\widehat{M^*_i}) + 1\}$ and $R^2_M = \mathrm{Var}(\widehat{M^*_i})/\{\mathrm{Var}(\widehat{M^*_i}) + 1\}$ in the above formula where $\widehat{M^*_i}$ represents the predicted value of the latent mediator variable for the probit regression. Thus, in all cases considered in this section, we can interpret the value of $\rho$ using two alternative coefficients of determination.

**Empirical Illustration.** Next, we present a sensitivity analysis in terms of the coefficients of determination for the case of a continuous outcome (depression 6 months following the treatment) and the dichotomous mediator measure. 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 which 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 $[-.05, -.007]$.

How sensitive is this result to an unobserved confounder? Consider the so-called "ability bias" where 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, the sign of the product of coefficients for the unobserved confounder is negative, i.e., $\mathrm{sgn}(\lambda_2\lambda_3) = -1$.

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

# 7 Concluding Remarks

In this paper, we propose a general approach to causal mediation analysis. Our approach consists of the identification assumption, the estimation algorithms, and the sensitivity analysis. In doing so we give researchers access to a broad range of estimation strategies that handle a variety of different types of data. Furthermore, our approach straightforwardly calculates estimates of uncertainty which 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 causal mediation analysis. Thus, 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. In order to facilitate the use of the proposed methodology, we have developed easy-to-use software, **mediation**, which is publicly available at the Comprehensive R Archive Network (htpp://cran.r-project.org/web/packages/mediation). The details of this implementation and many examples are given in the companion paper (Imai *et al.*, 2009).

# References

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

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

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. *Proceedings of the International Joint Conference on Artificial Intelligence* .

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51**, 6, 1173–1182.

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

Ditlevsen, S., Christensen, U., Lynch, J., Damsgaard, M., and 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. and Graubard, B. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* **11**, 167–178.

Gallop, R., Small, D., Lin, J. Y., Elliot, M. R., Joffe, M., and Have, T. R. T. (2009). Mediation analysis with principal stratification. *Statistics in Medicine* **28**, 7, 1108–1130.

Glynn, A. N. (2008). Estimating and bounding mechanism specific causal effect. Unpublished manuscript, presented at the 25th Annual Summer Meeting of the Society for Political Methodology, Ann Arbor, Michigan.

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association* **55**, 292, 708–713.

Green, D. P., Ha, S. E., and Bullock, J. G. (2010). Enough already about black box experiments: Studying mediation is more difficult than most scholars suppose. *Annals of the American Academy of Political and Social Sciences* .

Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman Hall, London.

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

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

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2009). Causal mediation analysis in R. Tech. rep., Department of Politics, Princeton University. available at `http://imai.princeton.edu/research/mediationR.html`.

Imai, K., Keele, L., and Yamamoto, T. (2008). Identification, inference and sensitivity analysis for causal mediation effects. Tech. rep., Department of Politics, Princeton University. available at `http://imai.princeton.edu/research/mediation.html`.

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

Judd, C. and Kenny, D. (1981). Estimating mediation in treatment evaluations. *Evaluation Review* **5**, 5, 602–619.

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

Keele, L. (2008). *Semiparametric Regression for the Social Sciences*. Wiley and Sons, Chichester, UK.

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

Koenker, R. (2008). *Quantile Regression*. Cambridge University Press, Cambridge.

Kraemer, H., Kiernan, M., Essex, M., and Kupfer, D. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology* **27**, 2, S101–S108.

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

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

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

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

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

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

MacKinnon, D., Warsi, G., and Dwyer, J. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research* **30**, 41–62.

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

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

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

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

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

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

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

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

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

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

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

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

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.

# Mathematical Appendix

## A Asymptotic Variance under the Model with the Interaction

Using the Delta method and noting the independence between $\hat{\beta}_2$ and $(\hat{\gamma}, \hat{\kappa})$, the asymptotic variance of estimated average causal mediation effect in equation (9) is given by $\mathrm{Var}(\hat{\beta}_2(\hat{\gamma} + \hat{\kappa} t) \mid T, X) \approx (\gamma + t\kappa)^2 \mathrm{Var}(\hat{\beta}_2 \mid T, X) + \beta_2^2 \{\mathrm{Var}(\hat{\gamma} \mid T, X) + t\mathrm{Var}(\hat{\kappa} \mid T, X) + 2t\mathrm{Cov}(\hat{\gamma}, \hat{\kappa} \mid T, X)\}$ for $t = 0, 1$. Based on this asymptotic variance, one can easily 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,

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

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,

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

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

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

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

# 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 pre-treatment covariates. As noted before, the ignorability of treatment assignment is satisfied in experiments where 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. While 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 since 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.

While 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 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 since the main goal of causal mediation effect 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 since 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 which under the no direct effect assumption equals the average mediation effect. Specifically, under Assumption 2, $\beta_3$ 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 $\beta_1$ 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 $X_i$.

# C Binary Outcome with Covariates

Our analytical results in Section 5.4 can be extended to the model with the pre-treatment confounders,

$$\begin{aligned} M_i &= \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \epsilon_{2i}, \\ Y_i &= \mathbf{1}\{Y_i^* > 0\} \quad \text{where} \quad Y_i^* = \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \epsilon_{3i}, \end{aligned}$$

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

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

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

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

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

$$\begin{aligned} \bar{\delta}(t) &= \mathbb{E}\left\{\Phi\left(\frac{\alpha_3 + \beta_3 t + \xi_3^\top X_i + \gamma(\alpha_2 + \beta_2 + \xi_2^\top X_i)}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right) - \Phi\left(\frac{\alpha_3 + \beta_3 t + \xi_3^\top X_i + \gamma(\alpha_2 + \xi_2^\top X_i)}{\sqrt{\sigma_2^2\gamma^2 + 1}}\right)\right\}, \\ \bar{\tau} &= \mathbb{E}\left\{\Phi(\alpha_1 + \beta_1 + \xi_1^\top X_i) - \Phi(\alpha_1 + \xi_1^\top X_i)\right\} \end{aligned}$$

# D The Proportion Mediated with Binary Outcome

**Without Covariates.** Suppose that the probit regression is used to model the binary outcome variable without pre-treatment covariates, i.e., the model defined in equations (13) and (14) where $\epsilon_{i3} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_2^2)$ and $\epsilon_{i3} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. We derive the condition under which the common measure used in the literature, i.e., $\tilde{\upsilon}$ in equation (20) approximates the true proportion mediated $\upsilon$. First, we use the following linear approximation to the average causal mediation effect,

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

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

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

Using these results, the proportion mediated is approximated by,

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

which is equal to $\tilde{\upsilon}$ when $\beta_3 = 0$ or $\beta_3 = -2(\alpha_3+\gamma\alpha_2)$. Thus, when the average direct effect is small, $\tilde{\upsilon}$ approximately equals $\upsilon$.

This result extends to the situation where the logit regression is used, i.e., the model defined in equations (13) and (14) where $\epsilon_{i3} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_2^2)$ and $\epsilon_{i3}$ is an i.i.d. standard logistic random variable. The calculation similar to the above yields the following approximation to the true proportion mediated,

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

When $\beta_3 = 0$, we have $\upsilon \approx \tilde{\upsilon}$. Thus, when the average direct effect is close to zero, the standard measure approximates the true proportion mediated.

**With Covariates.** The above result can be generalized to the model with the observed pre-treatment covariates. Since the analytical calculation is similar, we present the probit case, i.e., $M_i = \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \epsilon_{2i}$ with $\epsilon_{2i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_2^2)$ 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,

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

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

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

which equals $\upsilon$ when $\beta_3 = 0$.

# E Proof of Theorem 2

Equations (6) and (8) imply,

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

Let $\epsilon_{i1} = (\gamma + \kappa T_i)\epsilon_{i2} + \epsilon_{i3}$. 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 (see Proof of Theorem 2 in Imai *et al.* (2008) for details). 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 $(\gamma, \kappa)$, then $(\alpha_3, \beta_3, \xi_3)$ is also identifiable. To identify $(\gamma, \kappa)$, we solve the following system of equations (Note that $\kappa$ can be also identified from $\kappa\xi_2$ so 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),

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

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 $\kappa$ and $\gamma$ as the functions of identifiable parameters. Thus, using equation (9), the desired expression results. □

# F The Details of the Sensitivity Analysis for Binary Outcome

For the binary outcome model given in equation (22), identification of the average causal mediation effects under nonzero values of $\rho$ requires several steps. For $t = 0, 1$, we can write the average causal mediation effects with nonzero $\rho$ 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, $\beta_2$ and $\sigma_2^2$ 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 $\alpha_1$, $\beta_1$, $\xi_1$, and $\gamma$. The first step toward identification requires estimating the probit regression: $Y_i = \mathbf{1}\{Y_i^* > 0\}$ with $Y_i^* = \alpha_1 + \beta_1 T_i + \xi_1^\top X_i + \epsilon_{i1}$ where we assume $\epsilon_{i1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$). Second, we define $\tilde{\rho} = \text{Corr}(\epsilon_{i1}, \epsilon_{i2})$, which identifies $\gamma$ as,

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

Next we estimate (22), which gives a set of biased estimates when $\rho$ 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 $\alpha_1$ and $\beta_1$ 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 $\rho$ is nonzero.

| | The Baron-Kenny Procedure | | | Instrumental |
|---|---|---|---|---|
| | No Interaction | With Interaction | | Variables |
| | | Under Treatment | Under Control | |
| | | $(t = 1)$ | $(t = 0)$ | |
| Average Mediation Effect | $-.016$ | $-.014$ | $-.021$ | $-.047$ |
| $\bar{\delta}(t)$ | $[-.03,\ -.002]$ | $[-.026,\ -.001]$ | $[-.040,\ -.002]$ | $[-.120,\ .024]$ |
| Average Direct Effect | $-.032$ | $-.027$ | $-.034$ | 0 |
| $\bar{\zeta}(t)$ | $[-.107,\ .046]$ | $[-.115,\ .048]$ | $[-.114,\ .046]$ | |
| Average Total Effect | $-.047$ | $-.047$ | | $-.047$ |
| $\bar{\tau}$ | $[-.120,\ .023]$ | $[-.120,\ .024]$ | | $[-.120,\ .023]$ |

Table 1: Estimated Causal Effects of Interest Based on the Baron-Kenny (BK) Procedure and the Instrumental Variables (IV) Method Using the JOBS II Data. The outcome variable is a measure of depression for each worker, and the mediator represents the level of their job-search self-efficacy. Each cell shows a point estimate and its corresponding 95% confidence interval. The average treatment effect on the mediator is estimated to be .100 with the 95% confidence interval $[.018, .182]$. The results in the first three columns are based on the BK procedure without and with the interaction between the treatment and the mediator. The final column presents the results based on the IV method which assumes zero direct effect and thus has the same estimate for the average mediation effect and the average total effect. Across all methods, the estimated average causal mediation effect is negative.

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

Table 2: Estimated Causal Quantities of Interest based on the Generalized Additive Model (GAM). The setup is identical to the one in Table 1 except that the GAM is used to model the outcome variable. $95\%$ confidence intervals are based on non-parametric bootstrap. The first column displays the results under the no-interaction assumption, whereas the other columns display the results without this assumption.

|  | No Interaction | With Interaction |  |
|---|---|---|---|
|  |  | Under Treatment<br>$(t = 1)$ | Under Control<br>$(t = 0)$ |
| Average Mediation Effect | $-.019$ | $-.019$ | $-.018$ |
| $\bar{\delta}(t)$ | $[-.033,\ -0.007]$ | $[-.035,\ -.006]$ | $[-.027,\ -.005]$ |
| Average Direct Effect | $-.026$ | $-.031$ | $-.029$ |
| $\bar{\zeta}(t)$ | $[-.098,\ .045]$ | $[-.096,\ .040]$ | $[-.099,\ -.039]$ |
| Average Total Effect | $-.045$ | $-.048$ |  |
| $\bar{\tau}$ | $[-.117,\ .027]$ | $[-.118,\ .022]$ |  |

Table 3: Estimated Causal Quantities of Interest with the Binary Mediator. The setup is identical to the one in Table 1 except that an probit model is used to model the mediator. $95\%$ confidence intervals are based on Algorithm 1 with 1000 Monte Carlo draws. The first column displays the results under the no-interaction assumption, whereas the other columns display the results without this assumption.

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

Table 4: Estimated Causal Quantities of Interest with the Discrete Mediator. The setup is identical to the one in Table 1 except that an ordered probit model is used to model the mediator. $95\%$ confidence intervals are based on Algorithm 1 with 1000 Monte Carlo draws. The first column displays the results under the no-interaction assumption, whereas the other columns display the results without this assumption.

|  | No Interaction | With Interaction |  |
|---|---|---|---|
|  |  | Under Treatment | Under Control |
|  |  | $(t = 1)$ | $(t = 0)$ |
| Average Mediation Effect | .004 | .003 | .007 |
| $\bar{\delta}(t)$ | $[-.001,\ .012]$ | $[-.002,\ .020]$ | $[-.001,\ .020]$ |
| Average Direct Effect | 0.057 | .054 | .059 |
| $\bar{\zeta}(t)$ | $[-.008,\ .124]$ | $[-.009,\ .118]$ | $[-.005,\ .121]$ |
| Average Total Effect | 0.061 | .062 |  |
| $\bar{\tau}$ | $[-.006,\ .128]$ | $[-.003,\ .125]$ |  |
| Proportion Mediated | 0.058 | .072 |  |
| $\bar{\upsilon}$ | $[-.104,\ .300]$ | $[-.145,\ .402]$ |  |

Table 5: Estimated Causal Quantities of Interest with a Binary Outcome. Outcome is whether a respondent was working more than 20 hours per week after the training sessions. $95\%$ confidence intervals are based on Algorithm 2 with 1000 resamples. The first column displays the results under the no-interaction assumption, whereas the other columns display the results without this assumption.

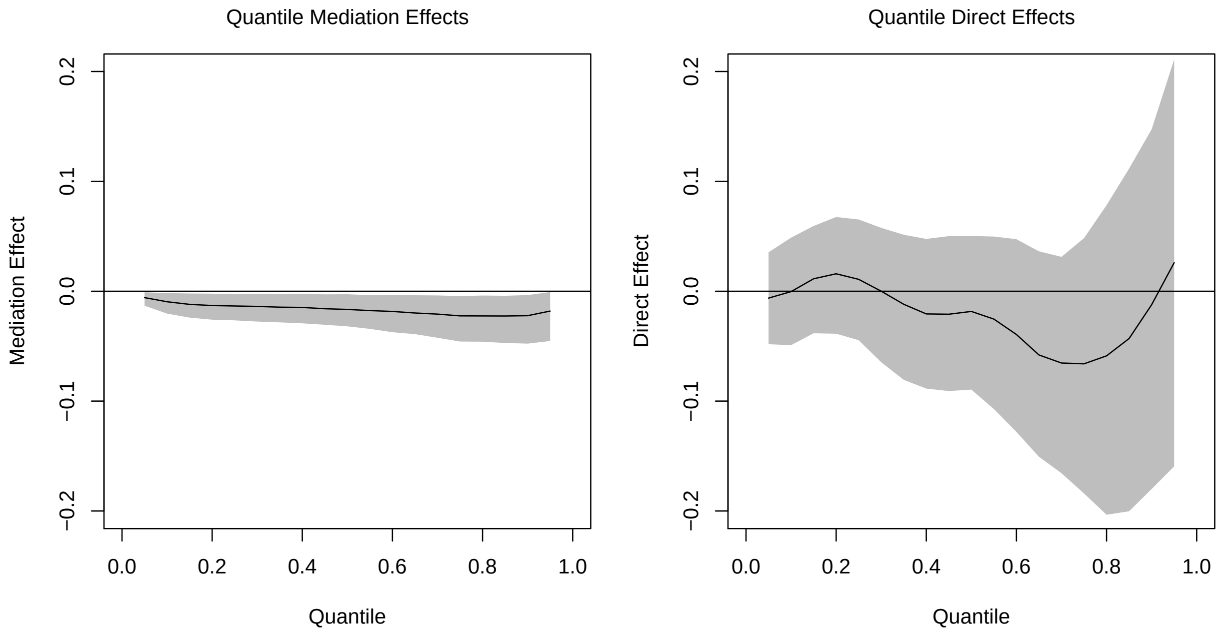

Figure 1: Estimated Quantile Causal Mediation and Direct Effects Using the JOBS II Data. The outcome variable is a measure of depression for each worker, and the mediator represents the level of their job-search self-efficacy. The figure presents the estimated quantile mediation effect (left panel) and the estimated quantile direct effect (right panel), 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.

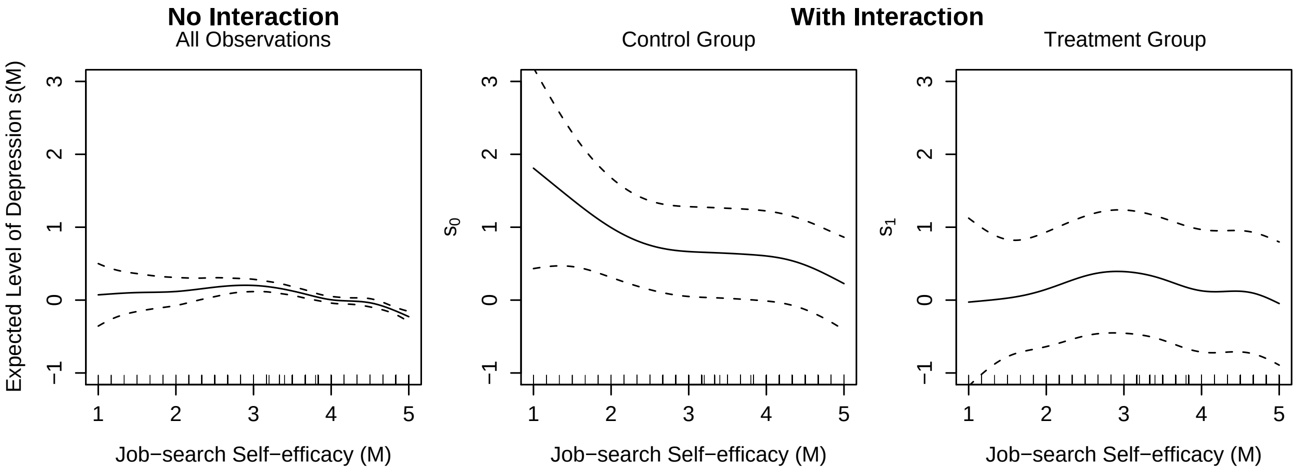

Figure 2: Generalized Additive Models and Estimated Non-linear Relationships between the Mediator (job-search self-efficacy) and the Outcome (depression level). The left panel assumes no interaction between the treatment and the mediator, whereas the other two panels allow for the interaction. The solid lines represent the estimated nonlinear relationships between the mediator (the horizontal axis) and the expected outcome (the vertical axis). $95\%$ confidence intervals (dashed lines) are based on non-parametric 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.

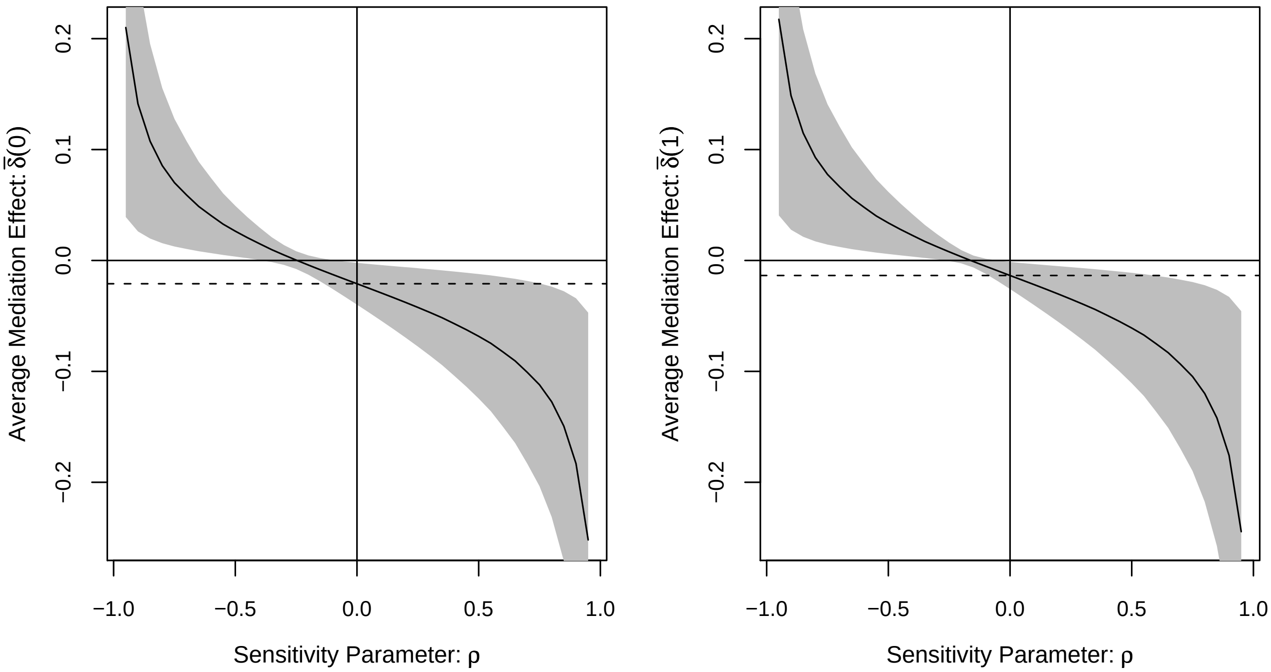

Figure 3: Sensitivity Analysis with Continuous Outcome and Mediator. The right (left) panel is for the estimated average mediation effect under the treatment (control). The dashed line represents the estimated mediation effect for $\rho = 0$. The grey areas represent the 95% confidence interval for the mediation effects at each value of $\rho$. The solid line represents the estimated average mediation effect at different values of $\rho$.

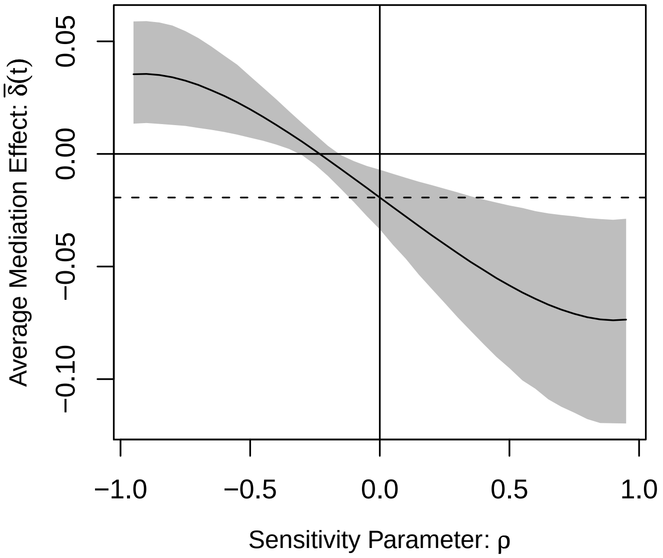

Figure 4: Sensitivity Analysis with Continuous Outcome and Binary Mediator. The dashed line represents the estimated mediation effect. The grey areas represent the 95% confidence interval for the mediation effects at each value of $\rho$.

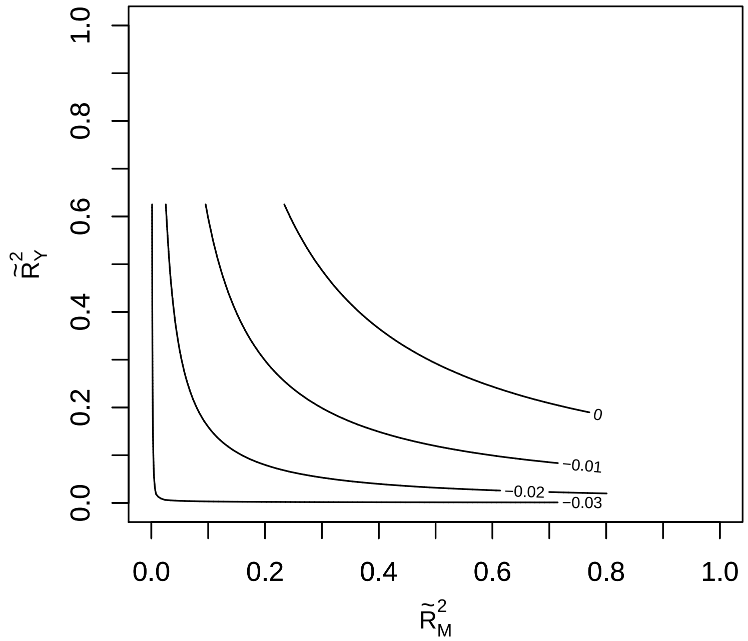

Figure 5: Sensitivity Analysis with Continuous Outcome and Binary Mediator. The plot contains contour lines that represent estimated average mediation effect corresponding to unobserved pre-treatment confounders of various magnitudes. These magnitudes are measured by the coefficients of determination, $\widetilde{R}_M^2$ and $\widetilde{R}_Y^2$, which represent the proportion of original variance explained by the unobserved confounder for the mediator and outcome, respectively. Here we assume $\mathrm{sgn}(\lambda_2\lambda_3) = -1$.