Identifying Mechanisms behind Policy Interventions via Causal Mediation Analysis*

Luke Keele† Dustin Tingley‡ Teppei Yamamoto§

This draft: April 20, 2015

Abstract

Causal analysis in program evaluation has primarily focused on the question about whether or not a program, or package of policies, has an impact on the targeted outcome of interest. However, it is often of scientific and practical importance to also explain why such impacts occur. In this paper, we introduce causal mediation analysis, a statistical framework for analyzing causal mechanisms that has become increasingly popular in social and medical sciences in recent years. The framework enables us to show exactly what assumptions are sufficient for identifying causal mediation effects for the mechanisms of interest, derive a general algorithm for estimating such mechanism-specific effects, and formulate a sensitivity analysis for the violation of those identification assumptions. We also discuss an extension of the framework to analyze causal mechanisms in the presence of treatment noncompliance, a common problem in randomized evaluation studies. The methods are illustrated via applications to two intervention studies on pre-school classes and job training workshops.

Key Words: causal mechanisms, noncompliance, instrumental variables, direct and indirect effects, potential outcomes, sensitivity analysis, mediation

*For helpful comments and suggestions we thank Jeff Smith and three anonymous reviewers. An earlier version of this paper was presented at the 2012 Fall Research Conference of the Association for Public Policy Analysis & Management in Baltimore, MD. The methods discussed in this paper can be implemented via an R package `mediation` (Imai et al., 2010b), which is freely available for download at the Comprehensive R Archive Network (http://cran.r-project.org/web/packages/mediation).

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

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

§Assistant Professor, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: teppei@mit.edu

Introduction

In program evaluation, researchers often use randomized interventions to analyze the causal relationships between policies and social outcomes. The typical goal in evaluation studies is to assess the impact of a given policy. Although impact assessment is certainly of primary importance in many substantive contexts, an exclusive focus on the question of *whether* and *how much* has often invited criticisms from scholars both within and outside of the policy community (e.g., Skrabanek, 1994; Heckman & Smith, 1995; Brady & Collier, 2004; Deaton, 2010a,b). Rather, it is often of both scientific and practical interest to explain *why* a policy intervention works (Bloom, 2006, pg.18). Answering such questions will not only enhance the understanding of causal mechanisms behind the policy, but may also enable policymakers to prescribe better policy alternatives.

In this paper, we introduce a statistical framework for the analysis of causal mechanisms that is becoming increasingly popular in many disciplines of social and medical sciences, including epidemiology, psychology, and political science (Greenland & Robins, 1994; Jo, 2008; Imai et al., 2011). This framework, often referred to as *causal mediation analysis* in the recent literature on causal inference, defines a mechanism as a process where a causal variable of interest, i.e., a treatment, influences an outcome through an intermediate variable, which is referred to as a mediator. The goal in such analysis is to decompose the total treatment effect on the outcome into the indirect and direct effects. In this type of analysis, the indirect effect reflects one possible explanation for why the treatment works, and the direct effect represents all other possible explanations.

While the statistical analysis of causal mechanisms has not historically been widespread in economics and public policy, there has recently been increasing awareness of the importance of mechanisms in policy analysis. Indeed, a recent review article highlights how understanding mechanisms in policy analyses plays a "crucial and underappreciated role" (Ludwig et al., 2011, p.20). A recent speech by the President of the William T. Grant foundation noted how "(t)he next generation of policy research in education will advance if it offers more evidence on mechanisms so that the key elements of programs can be supported, and the key problems in programs that fail to reach their goals can be repaired" (Gamoran, 2013). A recent special issue of the Journal of Research on Educational Effectiveness focused on mediation analyses. The lead editorial to this special issue

noted that "such efforts (in mediation analysis) are fundamentally important to knowledge building, hence should be a central part of an evaluation study rather than an optional 'add-on' " (Hong, 2012). In the large literature on neighborhood effects, recent work has called for an increased focus on mechanisms (Galster, 2011; Harding et al., 2011).[1]

The primary goal of the current paper is to provide an outline of recent theoretical advances on causal mediation analysis and discuss their implications for the analysis of mechanisms behind social and policy interventions with empirical illustrations. Below, we discuss three important aspects of investigating causal mechanisms in the specific context of program evaluation. First, we outline the assumptions that are sufficient for identifying a causal mechanism from observed information. A clear understanding of the key assumption at a minimum provides important insights into how researchers should design their studies to increase the credibility of the analysis. The identification result we present is nonparametric, in the sense that it is true regardless of the specific statistical models chosen by the analyst in a given empirical context. This result has led to a flexible estimation algorithm that helps policy analysts since it allows for a range of statistical estimators unavailable in previous approaches to mediation (Imai et al., 2010a).

Second, we discuss how sensitivity analyses can be used to probe the key assumption in causal mediation analysis. Sensitivity analysis is a general framework for investigating the extent to which substantive conclusions rely on key assumptions (e.g., Rosenbaum, 2002b). Sensitivity analysis is essential in causal mediation analysis because, unlike the identification of total treatment effects, identifying direct and indirect effects requires assumptions that are not simply satisfied by randomizing the treatment. This implies that, although studies can be designed to enhance the plausibility of those assumptions, it is fundamentally impossible to guarantee their satisfaction. Thus, sensitivity analysis consists a crucial element of causal mediation analysis by allowing policy analysts to report how strongly their conclusions rely on those assumptions, rather than hiding behind them.

Third, we engage with the problem of treatment noncompliance, an issue that is of central importance in policy analysis but has been understudied in the methodological literature on causal mechanisms. Noncompliance with assigned treatment status is widespread in policy intervention

[1]Recent examples of empirical research focusing on causal mechanisms in policy analysis include Flores & Flores-Lagunes (2009) and Simonsen & Skipper (2006).

studies (Magat et al., 1986; Puma & Burstein, 1994; Hill et al., 2002), and policy analysts are often interested in causal mechanisms behind interventions in the presence of noncompliance. For example, in a recent study reported in this journal, Wolf et al. (2013) investigate the effects of offers to participate in the District of Columbia's Opportunity Scholarship Program on various educational outcomes and speculate about the potential mechanisms driving those effects by highlighting several possibilities (pg.266). The study suffered from the problem of noncompliance because the offers were not always accepted. Ignoring the noncompliance problem and analyzing those mechanisms with standard techniques would have lead to biased inferences. Below, we outline how the intention-to-treat (ITT) effect of the treatment assignment and the average treatment effect on the treated units (ATT) may be decomposed into the direct and indirect effects under the assumptions similar to those commonly made in the instrumental variables literature (Angrist et al., 1996).

To help make abstract concepts concrete, we present original analyses of two well-known policy interventions. In the first application, we analyze data from the Perry Preschool project (Schweinhart & Weikart, 1981). We focus on the causal mechanisms behind the impact of this early education program on high school graduation rates, an outcome that has never been examined in previous research, including a recent study focusing on indirect effects by Heckman & Pinto (2014). In the second application, we analyze data from the JOBS II job training intervention (Vinokur et al., 1995). The JOBS II study is one intervention where a large component of the study was devoted to understanding casual mechanisms and a number of studies have conducted mediation analyses using data from this randomized trial (Imai et al., 2010a; Jo, 2008; Vinokur & Schul, 1997). However, previous analyses have not accounted for the widespread levels of noncompliance that were present in JOBS II. Below, we demonstrate how noncompliance has important implications for a mediation analysis of the data from JOBS II.

The rest of the paper proceeds as follows. In next section, we describe the two empirical examples that showcase the importance of understanding the causal mechanisms present in policy interventions. Then we lay out our statistical approach to causal mediation analysis and illustrates the approach with the first example. Next, we extend our approach to the setting where there is treatment noncompliance, and we analyzes the second example to illustrate the approach. Finally,

we conclude and discusses a variety of practical considerations that our paper gives rise to, including issues of cost and ethical considerations.

Examples of Causal Mechanisms in Program Evaluation

We first introduce the two empirical examples we use as illustrations to motivate the concepts. In the first application, we use data from the Perry Preschool Project randomized trial. The Perry project was a preschool program targeted at disadvantaged African American children during the mid-1960s in Ypsilanti, Michigan. The Perry program was designed to test the effect of preschool classes on a wide range of outcomes. Participants all entered at age 3. The first cohort participated for one year, and the second cohort participated for two years. Following Heckman et al. (2010a) and Heckman et al. (2010b) we ignore dose and measure treatment as a binary indicator.Heckman et al. (2010a) and Heckman et al. (2010b) have shown that the Perry Program affected a diverse set of outcomes including income and criminal behavior later in life. One remarkable aspect of the Perry Program is that it appears to have produced beneficial effects such as higher incomes, better educational outcomes, better health and lower levels of criminality at later ages. A standard analysis of data can only reveal that the Perry program had such impacts on those who participated. These estimates, however, tell us nothing about why the Perry program worked. Did the preschool program change intermediate covariates like cognitive ability that in turn produced these outcomes? A mediation analysis can provide some evidence for why a preschool intervention had lasting effects on outcomes measured many years later. Here, we focus on the question of how much of the Perry program effect on children's high school graduation rate can be attributed to the fact that the treatment increased cognitive ability at an early age. Evidence for a mechanism would suggest that future interventions might accentuate the aspects of the Perry project designed to increase cognitive ability. Here, our goal is to uncover a mechanism that has not been discovered.

In the second application, we use data from the Job Search Intervention Study (JOBS II) (Vinokur et al., 1995; Vinokur & Schul, 1997). JOBS II was a randomized job training intervention for unemployed workers. The program was designed with two goals in mind: to increase reemployment for those that are unemployed and improve the job seeker's mental health. Later analysis found that the JOBS II intervention did in fact increase employment outcomes and improve mental health

(Vinokur et al., 1995). What explains the effects of the program on employment and mental health? The study analysts hypothesized that workshop attendance would lead to increases in employment and mental health by improving confidence in job search ability (Vinokur et al., 1995; Vinokur & Schul, 1997). Because the intervention was specifically designed to improve employment outcomes by enhancing the participants' mental well-being, it is of theoretical interest to analyze whether its overall effect can be attributed to improvement in indicators of mental attitude such as self-confidence. If, on the other hand, the total treatment effect is found to be predominantly due to the direct effect, it may be concluded that the effect of the intervention was primarily through other channels, including the acquisition of more technical job-search skills.Again, like the Perry intervention, the JOBS treatment is multi-faceted. Evidence for a mechanism suggest that future intervention should emphasize elements that improve confidence. Here, our goal is to question conclusions from a previously discovered mechanism.

Like in many policy interventions, noncompliance with assigned treatment status was a common feature of the JOBS II study. Indeed, a substantial proportion of those assigned to the intervention failed to participate in the job training seminars, while those assigned to the control group were not given access to the treatment. While those assigned to control could have sought out other similar job services, they could not access the JOBS II intervention, and given the novelty of JOBS II, it is unlikely similar services were available. Because the workers in the treatment group selected themselves into either participation or non-participation in job-skills workshops, identification of causal relationships requires additional assumptions. In fact, as we highlight in below, such noncompliance creates more complications for the identification of causal mechanisms than for the analysis of total treatment effects.

Framework for Causal Mechanism Research in Policy Analysis

Following prior work (e.g., Robins & Greenland, 1992; Pearl, 2001; Glynn, 2008; Imai et al., 2010c), we use the potential outcomes framework (e.g., Holland, 1986) to define causal mediation effects. Without reference to specific statistical models, the potential outcomes framework clarifies what assumptions are necessary for valid calculation of causal mediation effects. This framework also enables the formal analysis of a situation that is of specific interest to policy analysts, treatment

noncompliance, the issue we take up later.

Potential Outcomes and Causal Effects

The causal effect of a policy intervention can be defined as the difference between one potential outcome that would be realized if the subject participated in the intervention, and the other potential outcome that would be realized if the subject did not participate. Formally, let T_i be a treatment indicator, which takes on the value of 1 when unit i receives the treatment and 0 otherwise. We here focus on binary treatment for simplicity, but the methods can be extended easily to non-binary treatment (see Imai et al., 2010a). We then use $Y_i(t)$ to denote the potential outcome that would result when unit i is under the treatment status t.[2] The outcome variable is allowed to be any type of random variable (continuous, binary, categorical, etc.). Although there are two potential outcomes for each subject, only the one that corresponds to his or her actual treatment status is observed. Thus, if we use Y_i to denote the observed outcome, we have $Y_i = Y_i(T_i)$ for each i. For example, in the Perry project, $T_i = 1$ if child i is assigned to the preschool program and $T_i = 0$ if not. Here, $Y_i(1)$ represents whether child i graduates from high school if she is in the program and $Y_i(0)$ is the potential high school graduation indicator for the same student not in the program.

Under the potential outcomes framework, the causal effect of T_i on the outcome is typically defined as difference in the two potential outcomes, $\tau_i \equiv Y_i(1) - Y_i(0)$. Of course, this quantity cannot be identified because only either $Y_i(1)$ or $Y_i(0)$ is observable. Thus, researchers often focus on the identification and estimation of the average causal effect, which is defined as $\bar{\tau} \equiv \mathbb{E}(Y_i(1) - Y_i(0))$ where the expectation is taken with respect to the random sampling of units from a target population.[3] In a randomized experiment like the Perry project, T_i is statistically independent of

[2]This notation implicitly assumes the Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1990), which requires that (1) there be no multiple versions of the treatment and (2) there be no interference between units. In particular, the latter implies that potential outcomes for a given unit cannot depend on the treatment assignment of other units. This assumption can be made more plausible by carefully designing the study, for example by not studying individuals from the same household.

[3]This implies that our target causal quantity $\bar{\tau}$ is the population average causal effect, as opposed to the sample average causal effect where the expectation operator is replaced with an average over the units in a given sample. Here and for the rest of the paper, we focus on inference for population-level causal effects which are more often the target quantities in public policy applications.

$(Y_i(1), Y_i(0))$ because the probability of receiving the treatment is unrelated to the characteristics of units; formally, we write $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i$. When this is true, the average causal effect can be identified as the observed difference in mean outcomes between the treatment and control groups, since $\mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(Y_i(1) \mid T_i = 1) - \mathbb{E}(Y_i(0) \mid T_i = 0) = \mathbb{E}(Y_i \mid T_i = 1) - \mathbb{E}(Y_i \mid T_i = 0)$. Therefore, in randomized experiments, the difference-in-means estimator is unbiased for the average causal effect. In the mediation analysis, the average causal effect is referred to as the total effect for reasons that will be clear in the next section.

Causal Mediation Effects

The potential outcomes framework can be extended to define and analyze causal mediation effects. Let $M_i(t)$ denote the potential mediator, the value of the mediator that would be realized under the treatment status t. Similarly to the outcome variable, the mediator is allowed to be any type of random variable. In the Perry project, $M_i(t)$ represents child i's cognitive ability at at ages 6–8 (measured by her IQ score at that time) that would be observed if she had been in the preschool program ($t = 1$) or not ($t = 0$). As before, only the potential mediator that corresponds to the actual treatment for child i can be observed, so that the observed mediator is written as $M_i = M_i(T_i)$. Next, we use $Y_i(t, m)$ to represent the potential outcome that would result if the treatment and mediating variables equaled t and m for i, respectively. For example, in the Perry project, $Y_i(1, 100)$ represents the high school graduation indicator for child i that would be observed if she had been in the preschool program and her cognitive ability equaled the IQ score of 100. Again, we only observe one of the (possibly infinitely many) potential outcomes, and the observed outcome Y_i equals $Y_i(T_i, M_i(T_i))$.

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

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

for $t = 0, 1$. In this definition, the causal mediation effect represents the indirect effects of the treatment on the outcome through the mediating variable (Pearl, 2001; Robins, 2003). The indirect effect essentially answers the following counterfactual question: What change would occur to the outcome if the mediator changed from what would be realized under the treatment condition, i.e., $M_i(1)$, to

what would be observed under the control condition, i.e., $M_i(0)$, while holding the treatment status at t? Although $Y_i(t, M_i(t))$ is observable for units with $T_i = t$, $Y_i(t, M_i(1-t))$ can never be observed for any unit. In the Perry project, $Y_i(1, M_i(1))$ represents high school graduation for child i with the IQ score at age 6–8 after participating in the preschool program, and $Y_i(1, M_i(0))$ represents high school graduation for the same child that participated in the program but had the IQ score as if she had not been in the Perry program. This indirect effect represents a posited mechanism or explanation for why the treatment worked. In our example, the mechanism posits that the reason the Perry intervention (at least partially) worked is because it increased cognitive ability at age 6–8. Similarly, we can define the direct effects of the treatment for each unit as

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

for $t = 0, 1$. In the Perry project, for example, this is the direct effect of the preschool program on child i's high school graduation while holding the mediator, IQ score at age 6–8, at the level that would be realized if she had not been in the program.[4] The direct effect represents all other possible mechanism or explanations for why the treatment worked.

The total effect of the treatment, τ_i, can be decomposed into the indirect and direct effects in the following manner, $\tau_i \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \delta_i(1) + \zeta_i(0) = \delta_i(0) + \zeta_i(1)$.[5] In addition, if direct and causal mediation effects do not vary as functions of treatment status (i.e., $\delta_i = \delta_i(1) = \delta_i(0)$ and $\zeta_i = \zeta_i(1) = \zeta_i(0)$, the assumption often called the no-interaction assumption), then the total effect is the simple sum of the mediation and direct effects, i.e., $\tau_i = \delta_i + \zeta_i$. The total effect is equivalent to the unit-level causal effect of T_i as defined in the previous section.

The causal mediation effect, direct effect and total effect are defined at the unit level, which means that they are not directly identifiable without unrealistic assumptions. The reason is that they are defined with respect to multiple potential outcomes for the same individual and only one of

[4] Pearl (2001) calls $\zeta_i(t)$ the *natural direct effects* to distinguish them from the *controlled direct effects* of the treatment. Imai, Tingley, & Yamamoto (2013) argue that the former better represents the notion of causal mechanisms, whereas the latter represents the causal effects of directly manipulating the mediator.

[5] These two alternative ways of decomposition arise due to the presence of the interaction effect. VanderWeele (2013) proposes a three-way decomposition which isolates the term representing the interaction effect from the sum of the pure direct and indirect effects, $\delta_i(0) + \zeta_i(0)$.

those potential outcomes is observed in reality. We thus focus on the population averages of those effects. First, the *average causal mediation effects* (ACME) can be defined as,

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

for $t = 0, 1$. The ACME can be interpreted similarly to the individual-level mediation effect (equation (1)), except that it now represents the average of those individual effects. Thus in the Perry project, $\bar{\delta}(t)$ represents the portion of the average effect of the preschool program on high school graduation that is transmitted by the change in cognitive ability at ages 6–8 induced by the Perry intervention. Similarly, we can define the average direct effect (ADE) and average total effect as $\bar{\zeta}(t) \equiv \mathbb{E}(Y_i(1, M_i(t)) - Y_i(0, M_i(t)))$ and $\bar{\tau} \equiv \mathbb{E}(Y_i(1, M_i(1)) - Y_i(0, M_i(0))) = \bar{\delta}(0) + \bar{\zeta}(1) = \bar{\delta}(1) + \bar{\zeta}(0)$, respectively. Again, if we make the no-interaction assumption, the average direct effect and average causal mediation effect simply sum to the average (total) causal effect defined in the previous section, i.e., $\bar{\tau} = \bar{\delta} + \bar{\zeta}$.

The definitions of the ACME and ADE make the goal of a causal mediation clear: to take the total effect and decompose it into its indirect and direct components. The indirect component represents a posited explanation for why the treatment works, while the direct component represents all other possible explanations. Interest often focuses on what proportion of the total effect is indirect.

Nonparametric Identification under Sequential Ignorability

Given the counterfactual nature of the ACME and ADE, a key question is what assumptions will allow them to be *nonparametrically identified*. In general, a causal quantity is said to be identified under a certain set of assumptions if it can be estimated with an infinite amount of data. If the set of assumptions for identification does not involve any distributional or functional form assumptions, it is said that the identification is achieved nonparametrically. Only after nonparametric identifiability of a causal parameter is established is it meaningful to consider the questions of statistical inference for the parameter (Manski, 1995, 2007).

As we discussed above, only the randomization of the treatment is required for the nonparametric identification of the average (total) causal effect, $\bar{\tau}$ (as well as the SUTVA; see footnote 2). The ACME and ADE, however, require additional assumptions for identification. Let $X_i \in \mathcal{X}$ be a vector

of the observed pretreatment confounders for unit i where $\mathcal{X}$ denotes the support of the distribution of X_i. Given these observed pretreatment confounders, Imai et al. (2010c) show that the ACME and ADE can be nonparametrically identified under the following condition.

Assumption 1 (Sequential Ignorability (Imai et al., 2010c)) *The following two statements of conditional independence are assumed to hold,*

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

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

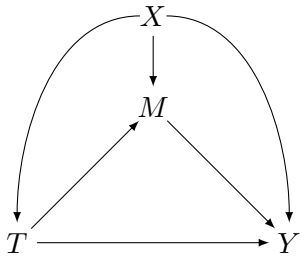
where $0 < \Pr(T_i = t \mid X_i = x)$ *and* $0 < p(M_i = m \mid T_i = t, X_i = x)$ *for* $t = 0, 1$, *and all* $x \in \mathcal{X}$ *and* $m \in \mathcal{M}$.

In the program evaluation literature, Flores & Flores-Lagunes (2009) use a similar identification assumption in the context of an analysis of the Job Corps, but impose an additional functional form assumption. They also ignore the problem of treatment noncompliance, which we discuss later. Flores & Flores-Lagunes (2010) also examine mechanisms but do so using a partial identification approach.

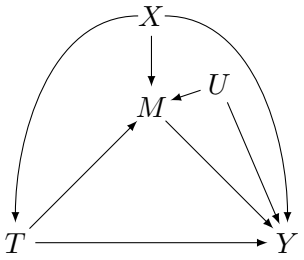
Assumption 1 is called sequential ignorability because two ignorability assumptions are sequentially made (Imai et al., 2011).[6] First, given the observed pretreatment confounders, the treatment assignment is assumed to be ignorable, i.e., statistically independent of potential outcomes and potential mediators. This part of Assumption 1 is guaranteed to be satisfied in a randomized experiment like the Perry project, since the treatment assignment is explicitly randomized by the researchers. If randomization was not used to assign T, then this part of the assumption is much less certain, since the subjects that select into the treatment may be different than those who do not in many ways observable and unobservable.

The second part of Assumption 1, however, requires particular attention. Unlike the first part, the second part may not be satisfied even in an ideal randomized experiment, since randomization of the treatment assignment does not imply that this second part of the assumption holds. For the second

[6]The term "sequential ignorability" was originally used by Robins (2003) and it referred to an assumption that is slightly weaker than Assumption 1 but based on the same substantive intuition of a sequential natural experiment. See Imai et al. (2010c) and Robins & Richardson (2010) for discussions about the technical and conceptual differences among alternative assumptions for mediation analysis.



(a) Example with observed covariate.



(b) Example with unobserved covariate.

Figure 1: Figure 1a is an example with an observed pretreatment covariate, X, that affects the treatment, mediator and outcome. Conditioning on X satisfies the sequential ignorability assumption. In Figure 1b sequential ignorability does not hold even after conditioning on X, since there is an unobserved pretreatment covariate, U, that affects mediator and outcome.

part of the assumption to hold, if there are any pre-treatment covariates that affect both the mediator and the outcome, we must condition on those covariates to identify the indirect and direct effects. The second stage of sequential ignorability is a strong assumption, since there can always be unobserved variables confounding the relationship between the mediator and the outcome even if the treatment is randomized and all observed covariates are controlled for. Furthermore, the conditioning set of covariates must be pre-treatment variables. Indeed, without an additional assumption, we cannot condition on the post-treatment confounders even if such variables are observed by researchers (Avin et al., 2005). The implication is that it is difficult to know for certain whether or not the ignorability of the mediator holds even after researchers collect as many pretreatment confounders as possible. This gives causal mediation analysis the character of observational studies, where confounding between M and Y must be ruled out "on faith" to some extent.

The diagrams in Figure 1 demonstrate two contrasting situations: one where the sequential ignorability assumption holds and another where it does not. In Panel 1a, X is an observed pretreatment covariate that affects T, M, and Y. So long as we condition on X, sequential ignorability will hold and the ACME and ADE can be nonparametrically identified. Randomization of T simply eliminates the arrow from X to T, but we would still need to condition on X to address the M-Y confounding for identification. In Panel 1b, an unobserved pretreatment covariate, U, affects both M and Y. Under such conditions, sequential ignorability does not hold and the ACME and ADE are not identified.

In the Perry project, the second part of sequential ignorability implies that cognitive ability at ages 6–8 must be regarded as "as-if" randomized among the children who have the same treatment status (participation in the preschool program or not) and the same pre-treatment characteristics. To satisfy this second part of the sequential ignorability assumption, we must control for all pre-treatment covariates that may confound the relationship between cognitive ability and high school graduation. The Perry data contain some pretreatment covariates, including the pretreatment level of the mediator which we regard as a key covariate to condition on, but there is always the possibility that this set of covariates is not sufficient. Later, we outline a sensitivity analyses to quantify how robust the empirical findings based on the sequential ignorability assumption are to the violation of that assumption. When having to make nonrefutable assumptions, sensitivity analyses are particularly valuable because they allow the researcher to examine the consequences of violations of the assumption.

One might assume that randomizing both the mediator and the treatment might solve this identification problem. However, randomizing both the treatment and mediator by intervention will not be sufficient for the identification of ACME or ADE in the underlying causal relationships. This is because intervening on the mediator merely fixes its value to an artificial level, instead of making the natural level of the mediator ($M_i(t)$) itself randomized or as-if random. Hence the "causal chain" approach, where in one experiment the treatment is randomized to identify its effect on the mediator and in a second experiment the mediator is randomized to identify its effect on the outcome (Spencer et al., 2005), does not identify the ACME or ADE. Unfortunately, even though the treatment and mediator are each guaranteed to be exogenous in these two experiments, simply combining the two is not sufficient for identification. For further discussion and proofs of these points, see Imai et al. (2011, 2013).

Implications for Design

The sequential ignorability assumption has important implications for the design of policy interventions. Given that randomized experiments rule out unmeasured confounding between the treatment and outcome, pretreatment covariates are often of secondary importance when the goal is to simply estimate the total treatment effects. While covariates may increase the precision of estimated

treatment effects, these estimates are guaranteed to be unbiased without collecting a rich set of pretreatment covariates. However, if a causal mediation analysis will be part of the analysis, collection of pretreatment covariates is of critical importance. A richer set of covariates will help bolster the plausibility of the sequential ignorability assumption.

In particular, baseline measurements of the outcome and mediator are worth collecting. In evaluations where such measurements are possible, the plausibility of sequential ignorability will be much stronger. One example would be in education interventions where the outcome is measured by test scores. Test scores are often fairly stable over time and past scores explain a large amount of the variation in present scores. As an example, Shapka & Keating (2003) study whether single-sex classrooms increase math scores. They explore whether math anxiety acts as a mediator. Here, the outcome is measured using mathematics test scores. In a study of this type, measures of both the mediator and outcome can be collected at baseline. Moreover, the study was conducted over a two year period. Given this time frame there are fewer alternative reasons why either math scores or anxiety should be higher, and past measures of math anxiety and math scores should explain large amount of the variation in measures used in the mediation analysis.

Contrast this study with the original mediation analysis of the Perry preschool program (Heckman et al., 2013). While measures of cognitive ability were collected at baseline, other mediators such as academic motivation were not collected at baseline, and there is no way to collect pretreatment measures of outcomes such as employment status at age 27. In sum, analysts can enhance the plausibility of identification assumptions by considering the possibility of mediation analysis from the beginning of an evaluation study. A clear understanding of the key identification assumption underscores the attention that needs to be paid to the specification requirements in a mediation analysis. Even in a randomized experiment, it is essential to collect information on pretreatment covariates that are likely to affect the mediator and outcome, including the baseline values of those variables whenever feasible. The need for additional data at baseline may also create trade-offs in terms of the resources that are needed for such data collection efforts.

Estimation of Causal Mediation Effects

We now turn to the subject of estimation. First, we outline how LSEM may be used to estimate causal mediation effects when an additional set of assumptions are satisfied. We then review a more general method of estimation that allows for a wide class of nonlinear models.

0.0.1 Relationship to Identification Within the Structural Equation Framework

Here, we briefly demonstrate how mediation analysis using traditional linear structural equation models (LSEM) is encompassed by the potential outcomes framework. For illustration, consider the following set of linear equations,

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

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

Under the popular Baron-Kenny approach to mediation (Baron & Kenny, 1986) researchers would conduct a set of significance tests on the estimated coefficients $\hat{\beta}_2$ and $\hat{\gamma}$, as well as on the effect of the treatment on the outcome variable without controlling for the mediator. This procedure, however, both does not give an actual estimate of the mediation effect but also breaks down when the coefficients on $\hat{\beta}_2$ and $\hat{\gamma}$ are in opposite directions (known as "inconsistent mediation" MacKinnon et al., 2000). In order to get an estimate of the mediation effect, one can use the product of coefficients method which uses $\hat{\beta}_2\hat{\gamma}$ as an estimated mediation effect (MacKinnon et al., 2002).

Imai et al. (2010c) prove that the estimate based on the product of coefficients method can be interpreted as a consistent estimate of the causal mediation effect only under the following conditions: (1) Assumption 1 is satisfied, (2) the effect of the mediator on the outcome does not interact with the treatment status, and (3) the conditional expectations of the potential mediator and outcome are indeed linear and additive as specified in equations (5) and (6) (see also Jo, 2008). Next, we discuss a more general estimation framework that can be used even when conditions (2) and (3) do not hold.

A General Method of Estimation

While we can use LSEMs to estimate causal mediation effects, the linearity assumptions required with LSEMs are often inappropriate. For example, in the Perry program example, the outcome of interest is the binary indicator of whether or not the student graduated from high school. Imai et al. (2010a) develop a general algorithm for computing the ACME and the ADE that can accommodate any statistical model so long as sequential ignorability holds. Here, we provide a brief summary of the two-step algorithm,[7] and refer interested readers to Imai et al. (2010a, in particular Appendix D and E) who provide theoretical justification as well as Monte-Carlo based evidence for its finite sample performance. The algorithm is implemented in the R package, `mediation`.

First, analysts posit and fit regression models for the mediator and outcome of interest. Corresponding to the sequential ignorability assumption, the mediator model should include as predictors the treatment and any relevant pretreatment covariates. Similarly, the outcome should be modeled as a function of the mediator, the treatment, and the pretreatment covariates. The algorithm can accommodate any form of model for the mediator and outcome. For example, the models can be nonlinear (e.g., logit, probit, poisson, etc.) or even nonparametric or semiparametric (e.g. generalized additive models).[8] Based on the fitted mediator model, we then generate two sets of predicted mediator values for each observation in the sample, one under the treatment and the other under the control conditions. In the Perry project example, we would generate predicted levels of IQ scores for the children with and without participation in the program.

Next, we use the outcome model to impute potential outcomes. First, we obtain the predicted value of the outcome corresponding to the treatment condition ($t = 1$) and the predicted mediator value for the treatment condition we obtained in the previous step. Second, we generate the predicted counterfactual outcome, where the treatment indicator is still set to 1 but the mediator is set to its predicted value under the control, again obtained in the previous step of the algorithm. The ACME, then, is computed by averaging the differences between the predicted outcome under the two values of the mediator across observations in the data. For the Perry project example, this would

[7] Huber (2012) considers an alternative estimation strategy based on inverse probability weighting.

[8] The resulting algorithm, therefore, can be considered either parametric, semi-parametric or nonparametric, depending on the specific models used in the application.

correspond to the average difference in high school graduation rates under the treatment across the levels of IQ scores at ages 6–8 with and without participation in the program.

Finally, we repeat the two simulation steps many times in order to obtain uncertainty estimates. In addition to prediction uncertainty, which is incorporated through those two steps, we also need to take into account sampling variability in order to correctly represent the overall estimation uncertainty for the quantity of interest. This can be achieved in two alternative ways. First, one can simulate model parameters for the mediator and outcome models from their (asymptotically normal) sampling distributions and conduct the two prediction steps for each copy of the simulated model parameters. This approach is based on King et al. (2000). Second, one can simply resample the observations with replacement and apply the two-step procedure to each resample. This nonparametric bootstrap method is more generally applicable, but often slower than the first approach. With estimates of uncertainty, one can use hypothesis tests to understand whether indirect and direct effects are statistically different from zero. For example, in the Perry project analysis, we can test whether the indirect effect of the treatment through cognitive ability is statistically significant.

Sensitivity Analysis

The identification results and estimation procedures we discussed above are only valid under the sequential ignorability assumption. Unfortunately, observed data in an experiment like the Perry project cannot be used to test whether the assumption is satisfied. Even when researchers have theoretical reasons to believe that they have appropriately controlled for confounding variables, such arguments will rarely be dispositive. A powerful approach to address the concern about unobserved confounding that might still remain is to examine how sensitive their results are to the existence of such confounders. As we describe next, a formal sensitivity analysis can be done to quantify how results would change as the sequential ignorability assumption was relaxed. Results that become statistically insignificant, or even change signs, with small violations of the assumption are considered to be sensitive and unreliable.

Imai et al. (2010a,c) develop procedures for conducting such sensitivity analyses under the linear and non-linear structural equations models such as equations (5) and (6). Their analysis is based on the idea that the degree of violation of equation (4), i.e., the second part of the sequential

ignorability assumption, can be represented by the correlation coefficient between the two error terms, ϵ_{i2} and ϵ_{i3}. This is because omitted pretreatment covariates that confound the mediator-outcome relationship will be components of both error terms, resulting in nonzero correlation between ϵ_{i2} and ϵ_{i3}. Formally, let ρ represent this correlation: When $\rho = 0$, the two error terms do not contain any common component, implying that equation (4) is satisfied. Conversely, if $\rho \neq 0$, existence of unobserved confounding is implied and therefore the sequential ignorability assumption is violated. Thus, varying ρ between -1 and 1 and inspecting how the ACME and ADE change enable us to analyze sensitivity against unobserved mediator-outcome confounding.[9] Imai et al. (2010c) show that the ACME and ADE can be consistently estimated for any assumed value of ρ in this range, and that standard errors for those estimates can be obtained via the simulation-based procedures similar to those described above. For example, in the Perry project, we may not have controlled for confounders that affect both cognitive ability at ages 6–8 and high school graduation. A sensitivity analysis would calculate the ρ at which the ACME or ADE is zero (or their confidence intervals contain zero).

The above method uses error correlation as a means of quantifying the severity of unobserved mediator-outcome confounding. This approach, while statistically straightforward, has the important drawback that the sensitivity parameter itself is rather difficult to interpret directly. Here, we present an alternative method for the interpretation of the sensitivity analysis. Imai et al. (2010d) show how to interpret the same sensitivity analysis using the following decomposition of the error terms for equations (5) and (6),

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

for $j = 2, 3$ where U_i is an unobserved pre-treatment confounder that influences both the mediator and the outcome, and λ_j represents an unknown coefficient for each equation. They show that ρ can be written as a function of the coefficients of determination, i.e., R^2s. This allows for the sensitivity analysis to be based on the magnitude of an effect of the omitted variable. Here, the sensitivity anal-

[9] This form of sensitivity analysis is related to methods that Altonji et al. (2005) develop to analyze the effectiveness of Catholic schools. Imbens (2003) also develops a similar sensitivity analysis for the problem of selection on unobservables in the standard program evaluation context.

ysis is based on the proportion of original variance that is explained by the unobserved confounder in the mediator and outcome regressions. These terms are $\widetilde{R}^2_M \equiv \{\text{Var}(\epsilon_{i2}) - \text{Var}(\epsilon'_{i2})\}/\text{Var}(M_i)$ and $\widetilde{R}^2_Y \equiv \{\text{Var}(\epsilon_{i3}) - \text{Var}(\epsilon'_{i3})\}/\text{Var}(Y_i)$, respectively.

The expression for ρ is given by $\text{sgn}(\lambda_2\lambda_3)\widetilde{R}_M\widetilde{R}_Y/\sqrt{(1-R^2_M)(1-R^2_Y)}$ where and R^2_M and R^2_Y are the usual coefficients of determination for the mediator and outcome regressions. Thus, in all cases considered in this section, we can interpret the value of ρ using two alternative coefficients of determination. This implies that, as before, we can analyze the sensitivity of ACME and ADE estimates against unobserved mediator-outcome confounding by varying $\tilde{R}^2_M$ and $\tilde{R}^2_Y$ and reestimating the implied ACME and ADE under the assumed level of unobserved confounding. Again, a result that is strong would be one where the omitted confounder would need to explain a large amount of variation in either the mediator or outcome in order for the substantive results to change. Although mathematically equivalent to the error correlation approach, the variance decomposition approach has the advantage of allowing the mediator and outcome to be separately analyzed.

Sensitivity analysis is not without its limitations. These limitations range from conceptual to more practical ones. Conceptually, the above sensitivity analysis itself presupposes certain causal relationships. First, the causal ordering between the treatment, mediator, outcome and observed covariates assumed by the analyst must be correct in the first place. Second, the treatment is assumed to be ignorable conditional on the pretreatment covariates (equation 3). These conditions, however, can often be made plausible by careful research design (e.g. randomizing the treatment and defining and measuring the mediator and outcome in accordance with the assumed ordering), whereas the mediator-outcome confounding (equation 4) is more difficult to be controlled by the researcher. Third, the above sensitivity analysis can only be used for pretreatment mediator-outcome confounding and does not address posttreatment confounding. For example, if the omitted confounder is itself influenced by the treatment, and then influences the mediator and outcome, this type of sensitivity analysis is no longer appropriate. Alternative procedures have recently been developed to address such situations (e.g. Imai & Yamamoto, 2013; Albert & Nelson, 2011).[10]

[10] A related issue is the choice of conditioning sets. When the treatment is not randomized, and researchers must appeal to the use of control variables to establish the ignorability of the treatment, there arises the issue of what pretreatment covariates to include in the mediator and outcome models. The recent exchange between Pearl (2014)

There are two more practical limitations. First, there is no accepted threshold for which a particular result can be dichotomously judged to be unacceptable, as is the case with similar forms of sensitivity analyses in general. We echo the common recommendation that the degree of sensitivity be assessed via cross-study comparisons (Rosenbaum, 2002a, p.325). It is important to note that such comparisons can only be practiced if sensitivity analyses are routinely conducted and reported in empirical research. Second, the existing sensitivity analysis methods for unobserved mediator-outcome confounding are highly model-specific, in that a different procedure has to be derived for each particular combination of mediator and outcome models. While the existing procedures do cover the most commonly used parametric models, future research could derive methods for other types of models.

Instrumental Variables and Mediation Effects

In program evaluation, researchers often rely on instrumental variables (IV) and related statistical methods to analyze causal relationships. Such techniques are typically used when the causal variable of interest, e.g. actual reception of a policy intervention, cannot be plausibly regarded as ignorable. Since the identification of the ACME and ADE requires ignorability assumptions, it is unsurprising that IVs can play valuable roles in the analysis of causal mechanisms. Here, we provide a brief overview and conceptual clarification for the various existing IV-based methods for analyzing causal mechanisms. We think this clarification is important since in one case IV is ill-suited to mechanisms, but useful in two other contexts.

Indeed, there are at least three distinct ways in which researchers can use IV-based methods for causal mechanisms. The three approaches can best be differentiated by focusing on what variable performs the role analogous to the "instrument" in the standard IV framework. The first, most traditional approach treats the treatment itself (T_i in the above notation) as the instrumental variable and apply a standard IV estimation method for the ACME. This approach originates in Holland (1988) and has recently been further explored by several researchers (Albert, 2008; Jo, 2008; Sobel,

and Imai et al. (2014) reveals that substantial ambiguity is likely to remain in practice with respect to the choice of conditioning sets, which suggests another important dimension for sensitivity analysis (see Imai et al., 2014, for some initial ideas). We, however, emphasize that such considerations are not relevant if the treatment is randomized.

2008). This approach relies on the rather strong assumption that the direct effect is zero. In the jargon of IV methods, this assumption implies that the treatment satisfies the exclusion restrictions with respect to the mediator and outcome, i.e., the treatment can only affect the outcome through its effect on the mediator. Under this assumption and the ignorability of the treatment (i.e. equation 3, the first stage of sequential ignorability), the standard IV methods can be used to obtain valid estimates of the causal mediation effects. The primary advantage of this approach is that it is no longer necessary to assume the absence of unobserved mediator-outcome confounding (equation 4). The obvious drawback, however, is that it assumes *a priori* that there are no alternative causal mechanisms other than the mediator of interest. For example, in the context of the Perry project, this approach will be valid only if the effect of the preschool program on high school graduation is entirely mediated through cognitive ability at ages 6–8. This approach is often invoked under the rubric of principal stratification (Page, 2012), but has been criticized due to the reliance on the exclusion restriction (VanderWeele, 2012).

The second approach, proposed by Imai et al. (2013), uses an IV in order to cope with the possible existence of unobserved confounding between the mediator and outcome (i.e. violation of equation 4). This approach presupposes the situation where researchers can partially manipulate the mediating variable by random encouragement. It can then be shown that, if the encouragement is applied to a randomly selected subset of the sample, and the encouragement satisfies the standard set of IV assumptions (exclusion restrictions and monotonicity), then the ACME and ADE can be nonparametrically bounded for a meaningful subgroup of the population defined by their compliance to the encouragement. Since such direct manipulation of mediating variables is relatively uncommon (though certainly not impossible) in program evaluation, we omit further details and refer interested readers to the aforementioned article.

A third approach developed by Yamamoto (2013) uses the IV framework for causal mediation analysis in yet another way. Unlike the above two methods, this approach is designed to address the nonignorability of the treatment variable (i.e. violation of equation 3) due to treatment non-compliance, a common problem in randomized evaluation studies. Indeed, as mentioned in above, the JOBS II study involved a substantial number of participants who were assigned to the job train-

ing workshops but did not comply with their assigned treatment. Thus, the identification results and estimation methods discussed thus far cannot be applied to the JOBS II example. Given the prevalence of treatment noncompliance in program evaluation, we discuss this approach in detail in Sections 0.0.1 and 0.0.1.

Mediation Effects in the Perry Preschool Program

We now present a causal mediation analysis for the Perry program study. Our focus is to illustrate how interpretation in a mediation analysis differs from a standard analysis of total treatment effects. As described previously, we study whether the Perry program increased their likelihood of graduating from high school by improving childrens' cognitive ability at early ages. Our mediator of interest is therefore cognitive skills (as measured by IQ scores at ages 6–8)[11] and the outcome is the indicator of high school graduation.

Children in the Perry Preschool Project were randomized to either two-years of specialized preschool classes that lasted 2.5 hours for five days a week or were excluded from the specialized preschool classes. Treated students were also visited by teachers at home for 1.5 hour sessions designed to engage parents in the development of their children (Schweinhart & Weikart, 1981). Overall there were 123 participants. The experiment suffered very little from usual complications such as attrition and noncompliance. All outcomes are observed for high school graduation, and all participants complied with the assigned treatment (Schweinhart & Weikart, 1981; Weikart et al., 1978). Because admission to the program was randomized and compliance was perfect, we can safely assume that the first stage of sequential ignorability (equation 3) is satisfied in the Perry study.

Another key feature of the Perry program data is that they contain a number of pretreatment covariates, including the mother's level of education, whether the mother works or not, whether the father was present in the home, the mother's age, whether the father did unskilled work, the density of people living in the child's home, the child's sex, and baseline levels of cognitive skills. As discussed above, the second stage of sequential ignorability (equation 4) crucially depends on the quality of this pretreatment data. That is, if there are unobserved pretreatment covariates that affect both cognitive ability and high school graduation, this assumption will be violated and the

[11] Cognitive ability is just one of three mediators analyzed in Heckman & Pinto (2014).

ACME and ADE will not be identified.

How plausible, then, is the second stage of sequential ignorability in the Perry study? The rich set of pretreatment covariates is a big plus. In particular, we do have a measure for the mediator at baseline, and it is logically impossible to consider the baseline measurement of the outcome. However, the possibility of unobserved confounding still remains. For example, consider depressive symptoms at baseline, which were not measured. Clinical depression may both reduce cognitive ability as measured by IQ test and reduce the likelihood of graduating from high school, especially if it goes untreated. Ruling out the presence of possible unobserved confounding completely is unfortunately impossible. Nevertheless, as we discussed below, we can address the possibility of unobserved pretreatment confounding via a sensitivity analysis.

We first estimate the ACME and ADE assuming sequential ignorability. Because the outcome is a binary indicator, we model it by a logistic regression model with the mediator, treatment, and the full set of pretreatment covariates listed above. For the mediator, we use a normal linear regression model including the treatment and the same set of pretreatment covariates. We then apply the general estimation procedure described above, which easily accommodates the combination of these two different types of statistical models.[12]

Table 1 shows the estimated ACME, ADE and average total effect. The average total effect (bottom row), which is equivalent to the usual average treatment effect, is estimated to be 0.224 with the 95% confidence interval ranging between 0.044 and 0.408. Thus, the Perry program increased the percent chance of high school graduation by just over 22 points. This estimate strongly suggests that the Perry program increased the graduation rate by a significant margin, both statistically and substantively. In an analysis of the causal mechanism, however, the primary goal is to decompose this effect into direct and indirect effects. To reiterate, the indirect effect (ACME) is the portion of the average total effect that is transmitted through higher cognitive ability, and the direct effect (ADE) is the remaining portion of the Perry program effect attributable to all other possible causal

[12]We omit an interaction term between the treatment and the mediator variable from the specifications of our models, as we found no evidence for such an interaction. Inclusion of this interaction would allow the ACME and ADE to differ depending on the baseline condition. For a focused discussion about how to interpret treatment/mediator interactions, see Muller et al. (2005).

Table 1: Estimated Causal Quantities of Interest for Perry Preschool Project.

		Graduate High School
Average Causal Mediation Effects	$\bar{\delta}$	0.069 [0.002, 0.154]
Average Direct Effects	$\bar{\zeta}$	0.169 [−0.011, 0.334]
Average Total Effect	$\bar{\tau}$	0.224 [0.044, 0.408]

Note: $N = 123$. Outcome is whether a student graduated from high school and the mediator is cognitive ability as measured by an average of IQ scores across ages 6–8. In square brackets are 95% bootstrap percentile confidence intervals. The model for the outcome is a logistic regression, and the model for the mediator is a linear regression model. Both models are specified with a number of covariates. Estimates are on a probability scale.

mechanisms. Here, we find that a substantial portion of the average total effect is due to changes in cognitive ability at ages 6–8. That is, the ACME for the cognitive ability (top row) is estimated to be approximately 0.069, with the 95% confidence interval ranging from 0.001 to 0.154 points. This implies that treatment-induced changes in cognitive ability account for about 29% of the total effect. On the other hand, the estimated Perry program ADE, which represents all other possible mechanisms, is 0.169, with a 95% confidence interval of -0.011 to 0.334. Overall, the analysis suggests that the Perry program increases high school graduation rates and some of that change is due to an increase in cognitive ability at ages 6–8. The mediation results suggest that components of the intervention that increase cognitive ability are important.

The analysis thus far rests on the strong assumption that there is not a common unobserved confounder that affects both cognitive ability and high school graduation. As discussed above, this part of the sequential ignorability assumption is required for identification of the ACME and ADE but is not guaranteed to hold even in a randomized intervention like the Perry Preschool project. Indeed, it is not unreasonable to think this assumption may have been violated in the Perry program study. As we noted above, depression is one possible confounder that is not measure at baseline but

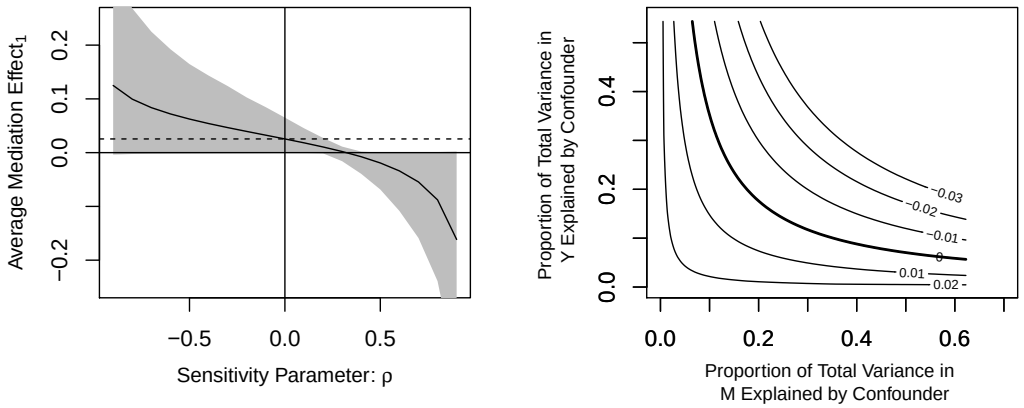


Figure 2: Sensitivity analysis for the Perry Preschool Project study, with high school graduation status as outcome. In the left panel, the correlation between the error terms in the mediator and outcome regression models (ρ) is plotted against the true ACME. The estimated ACME (assuming sequential ignorability) is the dashed line and 95% confidence intervals represented by the shaded regions. The right panel plots the true ACME as a function of the proportion of the total mediator variance (horizontal axis) and the total outcome variance (vertical axis) explained by an unobserved confounder. In this graph the mediator and outcome variables are assumed to be affected in the same directions by the confounder. Note that the contour lines terminate at the maximum allowable values of the sensitivity parameters implied by the observed information.

could affect both cognitive ability and high school graduation. Therefore, a sensitivity analysis is necessary in order to understand whether our conclusion is highly contingent on the assumption of no unobserved mediator-outcome confounding.

We now apply the sensitivity analysis discussed above. First, we conduct the analysis based on the ρ parameter. Recall that ρ represents the correlation between the error terms of the mediation and outcome models. When the second part of the sequential ignorability assumption holds, ρ is zero. Therefore, nonzero values of ρ represent violations of the key identifying assumption. In the sensitivity analysis, we can compute the indirect effect as a function of ρ. If the indirect effect is zero for small values of ρ that indicates that a minor violation of the sequential ignorability assumption would reverse the conclusions in the study. The result is shown in the left panel of Figure 2. We find that, for this outcome, the estimated ACME equals zero when ρ equals 0.3. However, given sampling uncertainty the confidence intervals for ρ always include zero. Thus if there were a modest violation of the sequential ignorability assumption, the true ACME could be zero.

We can also express the degree of sensitivity in terms of the $\tilde{R}^2$ parameters, i.e., how much of

the observed variations in the mediator and outcome variables are each explained by a hypothesized omitted confounder. In the right panel of Figure 2, the true ACME is plotted as contour lines against the two sensitivity parameters. On the horizontal axis is $\tilde{R}^2_M$, the proportion of the variance in the mediator, and on the vertical axis is $\tilde{R}^2_Y$, the proportion of the variance for the outcome, that are each explained by the unobserved confounder. In this example, we let the unobserved confounder affect the mediator and outcome in the same direction, though analysts can just as easily explore the alternative case. The dark line in the plot represents the combination of the values of $\tilde{R}^2_M$ and $\tilde{R}^2_Y$ for which the ACME would be zero. Note that, as is evident in the figure, these two sensitivity parameters are each bounded above by one minus the overall R^2 of the observed models, which represents the proportion of the variance that is not yet explained by the observed predictors in each model. Here, we find that the true ACME changes sign if the product of these proportions are greater than 0.037 and the confounder affects both cognitive ability and high school graduation in the same direction. For example, suppose that clinical depression was the unmeasured pretreatment confounder, which would most likely decrease both cognitive ability and high school graduation rate. Then, the true ACME would be zero or negative if depression explained about 20 percent of the variances in both of these variables. This level of sensitivity, again, is largely comparable to existing empirical studies (Imai et al., 2010c,a, 2011). In sum, our sensitivity analysis suggests that the positive mediation effect of cognitive ability for the effect of the Perry program on high school graduation is moderately robust to the possible unobserved pretreatment confounding.

Causal Mediation Analysis with Noncompliance

In the discussion so far, we have assumed that all subjects comply with the assigned treatment status. However, many randomized evaluation studies suffer from treatment noncompliance. For example, in the JOBS II study, 39% of the workers who were assigned to the treatment group did not actually participate in the job-skills workshops. Noncompliant subjects present a substantial challenge to randomized studies because those who actually take the treatment are no longer a randomly selected group of subjects; the compliers and non-compliers may systematically differ in their unobserved characteristics. A naïve comparison of average employment outcomes between the actual participants in the workshops and those who did not participate will therefore lead to a biased

estimate of the average causal effect of the treatment.

In the presence of treatment noncompliance, the methods from above are no longer valid because the actual treatment status is no longer ignorable. That is, equation (3) in Assumption 1 is violated. Hence, it is crucial to understand the basis under which causal mechanisms can be studied when there is noncompliance, which often occurs in policy interventions. Given the interest in studying mechanisms when noncompliance exists, it is important that we know exactly what assumptions are necessary and what quantities can be estimated from the data.

Alternative Causal Mediation Effects with Noncompliance

We now modify our framework to incorporate treatment noncompliance in causal mediation analysis. In addition to the actual treatment received by the workers (which we continue to denote by T_i), we consider the assigned treatment status Z_i, which equals 1 if worker i is assigned to (but does not necessarily take) the treatment and 0 otherwise. Then, under the assumption that the treatment assignment itself does not directly affect the mediator (exclusion restriction; see Appendix A.1), we can rewrite the potential mediator in terms of the treatment assignment alone, $M_i(z)$, where the dependence on the actual treatment is kept implicit. Likewise, if we assume that the treatment assignment can only affect the outcome through the actual treatment, the potential outcome can be written as $Y_i(z, m)$. In this alternative representation, the observed mediator and outcome can then be expressed as $M_i = M_i(Z_i)$ and $Y_i = Y_i(Z_i, M_i(Z_i))$, respectively.

What causal quantities might we be interested in, when treatment noncompliance exists and our substantive goal is to analyze the causal mechanism represented by the mediator? The quantities we examined earlier in the paper, the ACME and ADE, are difficult to identify without strong assumptions because the observed actual treatment is unlikely to be ignorable. We instead focus on two alternative sets of mechanism-related causal quantities that can be identified under more plausible assumptions.

First, consider the intention-to-treat (ITT) effect, the average effect of treatment assignment itself on the outcome of interest. This effect is the usual estimand in the "reduced-form" analysis of randomized evaluation studies with noncompliance (e.g. Angrist et al., 1996) and can be written in our current modified notation as $\bar{\tau}_{ITT} \equiv \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))]$. Our first set of mechanism-

related quantities decompose this effect. That is, the *mediated* and *unmediated ITT effects* are defined as

$$\bar{\lambda}(z) \equiv \mathbb{E}[Y_i(z, M_i(1)) - Y_i(z, M_i(0))], \tag{7}$$

$$\bar{\mu}(z) \equiv \mathbb{E}[Y_i(1, M_i(z)) - Y_i(0, M_i(z))], \tag{8}$$

for $z \in \{0, 1\}$ respectively. These quantities are identical to the ACME and ADE defined in above except that they are defined with respect to treatment assignment, not actual treatment. That is, the mediated ITT effect is the portion of the average effect of the treatment assignment itself on the outcome that goes through changes in the mediator values, regardless of the actual treatment. In the JOBS II study, $\bar{\lambda}(z)$ represents the average change in the employment in response to the change in self-efficacy induced by assignment to job-skills workshops (regardless of actual participation), holding the actual participation variable at the value workers would naturally choose under one of the assignment conditions. Similarly, the unmediated ITT effect, $\bar{\mu}(z)$, represents the portion of the average effect of the assignment on the outcome that does not go through the mediator. It can be shown that the mediated and unmediated ITT effects sum up to the total ITT effect, $\bar{\tau}_{ITT}$.

Second, we consider decomposing an alternative total effect, the average treatment effect on the treated (ATT). This quantity represents the (total) causal effect of the actual treatment on the outcome among the subjects who actually received the treatment. Under the assumption that (as was true in the JOBS II study) no worker assigned to the control group can actually take the treatment (one-sided noncompliance; see Appendix A.1), this quantity can be written as $\tilde{\tau} \equiv \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0)) \mid T_i = 1]$. Now we define the *average causal mediation effect on the treated* (ACMET) and *average natural direct effect on the treated* (ANDET) respectively as,

$$\tilde{\delta}(z) \equiv \mathbb{E}[Y_i(z, M_i(1)) - Y_i(z, M_i(0)) \mid T_i = 1], \tag{9}$$

$$\tilde{\zeta}(z) \equiv \mathbb{E}[Y_i(1, M_i(z)) - Y_i(0, M_i(z)) \mid T_i = 1], \tag{10}$$

for $z \in \{0, 1\}$. These quantities are equivalent to the ACME and ADE, except that they refer to the average indirect and direct effects among those who are actually treated.[13] In the JOBS II study,

[13] Because $\Pr(Z_i = 1 \mid T_i = 1) = 1$ under one-sided noncompliance, $\tilde{\delta}(z)$ and $\tilde{\zeta}(z)$ represent both the decomposed effects of the treatment assignment and the actual treatment.

these effects correspond to the effects of participation in the job-skills workshops on employment probability mediated and unmediated through self-efficacy among the workers who actually participated in the workshops. Again, it can be mathematically shown that the sum of these two effects is equal to the (total) ATT.

Nonparametric Identification under Local Sequential Ignorability

When can we identify the alternative mediation effects defined in the previous section? Using the more general result of Yamamoto (2013), we can show that the following assumption is sufficient:

Assumption 2 (Local Sequential Ignorability among the Treated)

$$\big\{Y_i(t,m), M_i(t'), T_i(z)\big\} \quad \perp\!\!\!\perp \quad Z_i \;\mid\; X_i, \tag{11}$$

$$Y_i(t',m) \quad \perp\!\!\!\perp \quad M_i \;\mid\; T_i = 1, X_i, \tag{12}$$

for all $z, t, t' \in \{0,1\}$ *and* $m \in \mathcal{M}$*, where* $T_i(z)$ *denotes the potential treatment given assignment to* z*.*

Details are provided in Appendix A.1. Assumption 2 is similar to Assumption 1 but differs from the latter in several important respects. First, equation (11) is satisfied if the treatment assignment Z_i, instead of the actual treatment, is either randomized or can be regarded as if randomized conditional on pretreatment covariates X_i. Since the assignment to job-skills workshops was randomly made in the JOBS II study, equation (11) is guaranteed to hold in our JOBS II dataset. Second, equation (12) is typically more plausible than equation (4) because it assumes the independence of the potential outcomes and the observed mediator only among the treated workers. In the JOBS II study, equation (12) will be satisfied if the observed levels of self-efficacy among the actual participants of the job-skills workshops can be regarded as close to random after controlling for the observed pretreatment covariates that may systematically affect both self-efficacy and employment.

A General Estimation Procedure

Once the nonparametric identification of these alternative mediation effects are achieved under Assumption 2, they can be consistently estimated using the flexible procedure proposed by Yamamoto (2013). The procedure is similar to the general algorithm for the perfect compliance case discussed

before, in that it accommodates various types of parametric and semi-parametric models. Specifically, the estimation procedure entails three regression-like models for the outcome, mediator, and actual treatment.

First, analysts should posit and fit a regression model for the outcome. The model, which we denote by $S(m,t,z,x) \equiv \mathbb{E}[Y_i|M_i = m, T_i = t, Z_i = z, X_i = x]$, should include the mediator, actual treatment, assigned treatment and pretreatment covariates as predictors, and can be fitted via standard estimators such as least squares and maximum likelihood estimators. Second, the analysts should model the conditional density of the mediator, using the actual treatment, assigned treatment and pretreatment covariates as predictors. The model, denoted by $G(m,t,z,x) \equiv p(M_i = m \mid T_i = t, Z_i = z, X_i = z)$, can again be estimated via standard procedures. Finally, the conditional probability of the actual treatment should similarly be modelled as a function of the assigned treatment and covariates. We denote this model by $Q(t,z,x) \equiv \Pr(T_i = t|Z_i = z, X_i = x)$.

The mediated and unmediated ITTs, ACMET, and ANDET can then be estimated by combining these estimates of the conditional expectations and densities. The exact formulas that generally apply for any types of models are given by Yamamoto (2013) and implemented by the `ivmediate` function in the R package `mediation` (Imai et al., 2010b); here, we provide an illustration for the case of a binary mediator and no pretreatment covariate, focusing on the ACMET for the treatment baseline. Using the fitted models $\hat{S}(m,t,z)$, $\hat{G}(m,t,z)$ and $\hat{Q}(t,z)$, this mediation effect can be estimated by the following expression,

$$\hat{\delta}(1) \;=\; \sum_{m=0}^{1} \hat{S}(m,1,1) \left\{ \hat{G}(m,1,1) + \frac{\hat{Q}(0,1)}{\hat{Q}(1,1)} \hat{G}(m,0,1) - \frac{\hat{Q}(0,0)}{\hat{Q}(1,1)} \hat{G}(m,0,0) \right\}. \tag{13}$$

Each of the quantities in the above equation are predicted quantities from the three fitted models with treatment assignment and status set to the appropriate values. For example, $\hat{Q}(1,1)$ is the predicted values from this model: $\Pr(T_i = t|Z_i = z, X_i = x)$ with t and z set to zero.

Valid uncertainty estimates for these quantities can be obtained via the bootstrap. One such procedure, implemented in `ivmediate`, consists of randomly resampling n observations from the sample of size n with replacement, calculating the estimates of mediation effects such as equation (13) for each of the resamples, and using the empirical quantiles of the resulting distributions as confidence intervals. Yamamoto (2013) shows evidence based on a series of Monte Carlo simulations suggesting

that this procedure works well for a reasonably large sample and if compliance rate is not too low.

Mediation Effects in the JOBS II Study

Now we apply the method in the previous section to the JOBS II dataset for illustration. As we discussed before, the study's analysts were interested in how much of the causal effects of participation in job-skills workshops on depressive symptoms and employment were due to participants' increased confidence in their ability to search for a job. In the Job Search Intervention Study (JOBS II) program a pre-screening questionnaire was given to 1,801 unemployed workers, after which treatment and control groups were randomly assigned. Job-skills workshops were provided to the treatment group which covered job-search skills as well as techniques for coping with difficulties in finding a job. Individuals in the control group were given a booklet that gave them tips on finding a job. Two key outcome variables were measured: the Hopkins Symptom Checklist which measures depressive symptoms (continuous), and an indicator for whether employment had been obtained (binary).

Here, we focus on the ACMET and ANDET of the workshop attendance on the depression and employment outcomes with respect to the self-efficacy mediator, which respectively represent the portions of the total average effect of the workshop attendance among the actual participants in the workshops that can and cannot be attributed to their increased sense of self-efficacy. We estimate these causal effects of interest based on a series of regression models which include a large set of pretreatment covariates (participants' sex, age, occupation, marital status, race, educational attainment, pre-intervention income, and pre-intervention level of depressive symptoms) to make Assumption 2 more plausible. The sample for our analysis ($N = 1050$) includes all observations for which all key variables were measured without missingness. Of these observations, 441 actually participated in the job-skills workshops, and our estimates apply to those 441 actually treated observations. Results are reported in Table 2.

We begin with a discussion of the results for the depression outcome (left column). As discussed in before, these estimates are obtained by first fitting three models for the outcome, mediator, and treatment compliance, and then combining them into the ACMET and ANDET estimates. Here, we use linear regressions for all three models. The estimate of the average treatment effect on the treated ($\tilde{\tau}$, bottom row) represents the total effect of workshop participation. Here, we observe a

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

		Depression	Employment Status
Average Causal Mediation Effects on the Treated (ACMET)	$\tilde{\delta}(1)$	$-.034$	.001
		$[-.071, -.005]$	$[-.011, .012]$
	$\tilde{\delta}(0)$	$-.044$	.002
		$[-.103, -.006]$	$[-.028, .021]$
Average Natural Direct Effects on the Treated (ANDET)	$\tilde{\zeta}(1)$	$-.009$	.102
		$[-.128, .117]$	$[.012, .192]$
	$\tilde{\zeta}(0)$	$-.019$	.104
		$[-.140, .107]$	$[.017, .187]$
Average Treatment Effect on the Treated (ATT)	$\tilde{\tau}$	$-.053$	.104
		$[-.174, .074]$	$[.018, .186]$

Note: $N = 1050$. Mediator is a continuous measure of job-search self-efficacy measured in the post-intervention interviews. Depression outcome is a continuous measure of depressive symptoms. Employment status outcome is whether a respondent was working more than 20 hours per week after the training sessions. In square brackets are 95% bootstrap percentile confidence intervals. Models for the outcome and mediator were specified with a number of covariates including measures of depressive symptoms measured prior to treatment.

slight decrease in depressive symptoms (about $-.053$ points on the scale of 1 to 5). The estimate does not reach the conventional levels of statistical significance, with the 95% confidence interval of $[-.174, .074]$. The ACMET ($\tilde{\delta}(1)$ and $\tilde{\delta}(0)$, top two rows), however, is negative both under the treatment and control baselines ($-.034$ and $-.044$, respectively) with the 95% confidence interval not overlapping with zero ($[-.071, -.005]$ and $[-.103, -.006]$). This suggest that the workshop attendance slightly but significantly decreased the depressive symptoms among the actual participants by increasing the participants' sense of self-efficacy in job-search process. The ANDET ($\tilde{\zeta}(1)$ and $\tilde{\zeta}(0)$, middle two rows), on the other hand, is even smaller in magnitude ($-.009$ and $-.019$) and statistically insignificant ($[-.128, .117]$ and $[-.140, .107]$), implying that the treatment effect mostly goes through the self-efficacy mechanism among the workshop participants.

Turning to the employment outcome (right column), we use logistic regression to model this variable because it takes on binary values (employed or unemployed). As in the case where treatment compliance is perfect (Sections and 0.0.1), the estimation method used here can accommodate a

large variety of outcome and mediator models. Here, we observe that the treatment increased the probability of obtaining a job among the actual workshop participants by 10.4 percentage points, with the 95% confidence interval of $[.018, .186]$. The estimates of the ACMET and ANDET, however, implies that this statistically significant increase in the employment probability cannot be attributed to the self-efficacy mechanism. The ACMET is very close to zero for both the treatment and control baselines, while the ANDET is estimated to be almost as large as the total effect on the treated for both baseline conditions, with the 95% confidence intervals not overlapping with zero. This suggests that the components of the JOBS II intervention designed to activate self-efficacy were of lesser importance.

Concluding Remarks on Causal Mediation Analysis

In program evaluation, analysts tend to focus solely on the study of policy impact. There is good reason for this since, with randomization, we can estimate average treatment effects under relatively weak assumptions. Policymakers may, however, demand deeper explanations for why interventions matter. Analysts may be able to use causal mechanisms to provide such explanations.

Here, we have outlined the assumptions and methods needed for going beyond average treatment effects to the estimation of causal mechanisms. Researchers often attempt to estimate causal mechanisms without fully understanding the assumptions needed. The key assumption, sequential ignorability, cannot be made plausible without careful attention to study design, especially in terms of collecting a full set of possible pretreatment covariates that might confound the indirect effect. The sensitivity analysis discussed in this paper allows researchers to formally evaluate the robustness of their conclusions to the potential violations of those assumptions. Strong assumptions such as sequential ignorability deserve great care and require a combination of innovative statistical methods and research designs. We also engaged with the issue of treatment noncompliance, a problem that may be of particular importance in policy analysis. We showed that alternative assumptions are necessary to identify the role of a mechanism and that a simple, flexible estimation procedure can be used under those assumptions.

Recent work has explored how analysts can use creative experimental designs to shed light on causal mechanisms. The two examples in this paper both involved a single randomization of the

treatment. The problem with the single experiment design, however, is that we cannot be sure that the observed mediator is ignorable conditional on the treatment and pretreatment covariates. As noted in Howard Bloom's acceptance remarks to the Peter Rossi award, "The three keys to success are 'design, design, design'... No form of statistical analysis can fully rescue a weak research design" (Bloom, 2010). Above we lay out the importance of research designs that collect relevant confounding variables in designs where only the treatment is randomized. Pushing the importance of design further, Imai et al. (2013) propose several different experimental designs and derive their identification power under a minimal set of assumptions. These alternative designs can often provide informative bounds on mediation effects under assumptions that may be more plausible than those required with a single experiment. As such, policy analysts have a number of tools, both statistical and design-based, available when they are interested in moving beyond standard impact assessment.

We conclude with a discussion of an important practical aspect of causal mediation analysis in the field of policy analysis. The need to collect extensive sets of pretreatment covariates suggests increase in cost, compared to traditional intervention studies. A similar consideration arises in measuring mediating variables, since it often means that policy researchers will need to revisit the subjects in their study sample multiple times to collect these measures prior to the ultimate outcomes. And of course, some mediators may be more or less easily measured. Given the likely increase in cost for mediation studies, the role of federal, state and local government funders will be crucial. In the end, we consider it of fundamental importance to answer questions of how and why experimental manipulations work in a policy setting. Equipped with the appropriate statistical tools, like those outlined in this paper, policy analysts can accumulate important knowledge that speaks to pressing public policy concerns.

References

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

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

Altonji, J.G., Elder, T.E., & Taber, C.R. (2005), Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools, Journal of Political Economy, 113, 151–184

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

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

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

Bloom, H.S. (2006), The core analytics of randomized experiments for social research, MRDC working papers on research methodology

Bloom, H.S. (2010), Nine lessons about doing evaluation research: Remarks on accepting the peter h. rossi award

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

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

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

Flores, C.A. & Flores-Lagunes, A. (2009), Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness. IZA Discussion Paper No. 4237

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

Galster, G. (2011), The mechanism(s) of neighbourhood effects: Theory, evidence, and policy implications, in H. van Ham, D. Manley, N. Bailey, L. Simpson, & D. Maclennan (Eds.) Neighbourhood Effects Research: New Perspectives, New York: Springer, pp. 23–56

Gamoran, A. (2013), Educational inequality in the wake of no child left behind, Association for Public Policy and Management

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

Greenland, S. & Robins, J.M. (1994), Ecologic studies: Biases, misconceptions, and counterexamples, American Journal of Epidemiology, 139, 747–760

Harding, D.J., Gennetian, L., Winship, C., Sanbonmatsu, L., & Kling, J. (2011), Unpacking neighborhood influences on education outcomes: Setting the stage for future research, in G. Duncan & R. Murnane (Eds.) Whither Opportunity: Rising Inequality, Schools, and Children's Life Chances, New York: Russell Sage, pp. 277–296

Heckman, J., Moon, S.H., Pinto, R., Savelyev, P., & Yavitz, A. (2010a), Analyzing social experiments as implemented: A reexamination of the evidence from the highscope analyzing social experiments as implemented: A reexamintation from the highscope perry preschool program, Quantitative Economics, 1, 1–46

Heckman, J. & Pinto, R. (2014), Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs, Econometric Reviews, Forthcoming

Heckman, J., Pinto, R., & Savelyev, P. (2013), Understand the mechanisms through which an influential early childhood program boosted adult outcomes, American Economic Review, 103, 2052–2086

Heckman, J.J., Moon, S.H., Pinto, R., Savelyev, P.A., & Yavitz, A. (2010b), The rate of return to the highscope perry preschool program, Journal of Public Economics, 94, 114–128

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

Hill, J., Waldfogel, J., & Brooks-Gunn, J. (2002), Differential effects of high-quality child care,

Journal of Policy Analysis and Management, 21, 601–627. URL http://dx.doi.org/10.1002/pam.10077

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

Holland, P.W. (1988), Causal inference, path analysis, and recursive structural equations models, Sociological Methodology, 18, 449–84

Hong, G. (2012), Editorial comments, Journal of Educational Effectiveness, 5, 213–214

Huber, M. (2012), Identifying causal mechanisms in experiments (primarily) based on inverse probability weighting. Unpublished Manuscript

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

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2010b), Advances in Social Science Research Using R (ed. H. D. Vinod), chap. Causal Mediation Analysis Using R, Lecture Notes in Statistics, New York: Springer, pp. 129–154

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011), Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies, American Political Science Review, 105, 765–789

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2014), Commentary: Practical implications of theoretical results for causal mediation analysis, Psychological Methods

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

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

Imai, K., Tingley, D., & Yamamoto, T. (2013), Experimental designs for identifying causal mechanisms (with discussions), Journal of the Royal Statistical Society, Series A (Statistics in Society), 176, 5–51

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

Imbens, G.W. (2003), Sensitivity to exogeneity assumptions in program evaluation, American Economic Review, 93, 126–132

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

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

Ludwig, J., Kling, J.R., & Mullainathan, S. (2011), Mechanism experiments and policy evaluations, Journal of Economic Perspectives, 25, 17–38

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

MacKinnon, D.P., Krull, J.L., & Lockwood, C.M. (2000), Equivalence of the mediation, confounding and suppression effect, Prevention Science, 1, 173–181

Magat, W.A., Payne, J.W., & Brucato, P.F. (1986), How important is information format? an experimental study of home energy audit programs, Journal of Policy Analysis and Management, 6, 20–34

Manski, C.F. (1995), Identification Problems in the Social Sciences, Harvard University Press

Manski, C.F. (2007), Identification for Prediction and Decision, Cambridge, MA: Harvard University Press

Muller, D., Judd, C.M., & Yzerbyt, V.Y. (2005), When moderation is mediated and mediation is moderated., Journal of personality and social psychology, 89, 852

Page, L.C. (2012), Principal stratification as a framework for investigating mediational processes in experimental settings, Journal of Research on Educational Effectiveness, 5, 215–244

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

Pearl, J. (2014), Interpretation and identification of causal mediation, Psychological Methods

Puma, M.J. & Burstein, N.R. (1994), The national evaluation of the food stamp employment and training program, Journal of Policy Analysis and Management, 13, 311–330

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), Oxford: Oxford University Press, pp. 70–81

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

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

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

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

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

Schweinhart, L.J. & Weikart, D.P. (1981), Effects of the perry preschool program on youths through age 15, Journal of Early Intervention, 4, 29–39

Shapka, J.D. & Keating, D.P. (2003), Effects of a girls-only curriculum during adolescence: Performance, persistence, and engagement in mathematics and science, American Educational Research Journal, 40, 929–960

Simonsen, M. & Skipper, L. (2006), The costs of motherhood: An analysis using matching estimators, Journal of Applied Econometrics, 21, 919–934

Skrabanek, P. (1994), The emptiness of the black box, Epidemiology, 5, 5553–5555

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

Spencer, S., Zanna, M., & 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, 845–851

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2013), mediation: R package for causal mediation analysis, available at the Comprehensive R Archive Network (CRAN). `http://CRAN.R-project.org/package=mediation`

Tingley, D., Yamamoto, T., Keele, L.J., & Imai, K. (2014), mediation: R package for causal mediation analysis, Journal of Statistical Software, Forthcoming

VanderWeele, T.J. (2012), Comments: Should principal stratification be used to study mediational processes?, Journal of Research on Educational Effectiveness, 5, 245–249

VanderWeele, T.J. (2013), A three-way decomposition of a total effect into direct, indirect, and interactive effects, Epidemiology, 24, 224–232

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

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

Weikart, D.P., Bond, J.T., & McNeil, J.T. (1978), The Ypsilanti Perry Preschool Project: Preschool years and longitudinal results through fourth grade, High/Scope Educational Research Foundation

Wolf, P.J., Kisida, B., Gutmann, B., Puma, M., Eissa, N., & Rizzo, L. (2013), School vouchers and student outcomes: Experimental evidence from washington, dc, Journal of Policy Analysis and Management, 32, 246–270

Yamamoto, T. (2013), Identification and estimation of causal mediation effects with treatment noncompliance. Unpublished manuscript

Appendices

A.1 Mathematical Details for the Noncompliance Case

In this appendix, we provide a formal representation of the two assumptions discussed in Section 0.0.1 and provide a proof of the nonparametric identification result for the mediated and unmediated ITT, ACMET and ANDET.

The two assumptions, exclusion restrictions and one-sided compliance, are commonly made in the analysis of randomized experiments with treatment noncompliance (Angrist et al., 1996). Using the notation introduced in Section 0.0.1, the assumptions can be formally represented as follows.

Assumption 3 (Exclusion Restrictions)

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

Assumption 4 (One-sided Noncompliance)

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

Now, we show that the more general result of Yamamoto (2013) implies the nonparametric identification of the mediated and unmediated ITT effects, ACMET and ANDET under Assumptions 2, 3 and 4. In fact, the result is immediate by noting that Assumption 4 implies the monotonicity assumption in Yamamoto (2013) and that the ACMET, ANDET and Assumption 2 are equivalent to the local average causal mediation effect, local average natural direct effect and the local sequential ignorability assumption in Yamamoto (2013) under Assumption 4, respectively.

The expressions for the identified effects can also be obtained as special cases of the results by Yamamoto (2013). For example, the ACMET for the treatment baseline condition is given by,

$$\begin{aligned}
\tilde{\delta}(1) \;=\; & \int\int \mathbb{E}[Y_i \mid M_i = m, T_i = Z_i = 1, X_i = x] \\
& \times \Bigg\{ p(m \mid T_i = Z_i = 1, X_i = x) + \frac{\Pr(T_i = 0 \mid Z_i = 1, X_i = x)}{\Pr(T_i = 1 \mid Z_i = 1, X_i = x)} p(m \mid T_i = 0, Z_i = 1, X_i = x) \\
& \qquad - \frac{\Pr(T_i = 0 \mid Z_i = 0, X_i = x)}{\Pr(T_i = 1 \mid Z_i = 1, X_i = x)} p(m \mid T_i = Z_i = 0, X_i = x) \Bigg\} \; dm \; dF(x), \qquad (14)
\end{aligned}$$

where $p(m \mid \cdot)$ represents the conditional density of the mediator. Note that this expression differs from the intuitively appealing estimator analogous to the usual Wald estimator for the local average treatment effect (Angrist et al., 1996). That is, one might be tempted to first estimate the mediated ITT effects by simply "ignoring" the actual treatment and applying the estimation procedure in Section to the assigned treatment, mediator and outcome, and then dividing the resulting quantity by the estimated compliance probability to obtain an estimate of ACMET. Unfortunately, this naïve approach leads to a biased estimate even under Assumptions 3, 4 and 2. The reason is that the actual treatment plays the role of a posttreatment mediator-outcome confounder, which renders the mediated ITT effects unidentified without additional assumptions about how $T_i(1)$ and $T_i(0)$ are jointly distributed (see Yamamoto, 2013, for more detailed discussion).

A.2 Software Details

In this section, we illustrate the use of the R package `mediation` (Tingley et al., 2013) for the application of the methods discussed in the main text. Specifically, we show the steps required to reproduce the empirical results in Sections 0.0.1 and 0.0.1. See Tingley et al. (2014) for a full overview of mediation analysis in `R` with the `mediation` package.

First, we show the steps for producing the results in Table 1 and Figure 2. The data from the Perry Preschool program requires a license, so we are unable to distribute the data with a replication file. The code is, however, available from the authors and partially reproduced below.

```
# First, load the mediation package
library(mediation)

# Fit model for mediator as a function of treatment and baseline covariates.
d <- lm(cogn ~ treat + female + fhome + medu + mwork + fskilled + mage
               + binet + density, data=perry)

# Fit outcome model as a function of treatment, mediator, and baseline covariates.
# Note that we omit an interaction between treatment and the mediator.
e <- glm(hs ~ treat + cogn + female + fhome + medu + mwork + fskilled + mage
              + binet + density, data=perry, family=binomial("probit"))

# Estimation with inference via the nonparametric bootstrap
# The two model objects above are passed to the mediate function.
binary.boot <- mediate(d, e, boot=TRUE, sims=5000, treat="treat", mediator="cogn")

# We now summarize the results which are reported in Table 1
```

```
summary(binary.boot)

# Next, we pass the output from the mediate function to the medsens function.
# The medsens function then performs the sensitivity analysis.
sens.binary <- medsens(binary.boot, rho.by=.1, eps=.01, effect.type="indirect")

# Use summary function to display results
summary(sens.binary)

# Plot results from sensitivity analysis
plot(sens.binary, main="", ylim=c(-.25,.25), ask=FALSE)

plot(sens.binary, sens.par="R2", sign.prod=1, r.type=2,
     ylim=c(0,0.4), xlim=c(0,0.7), ylab = "", xlab = "", main="")
title(ylab="Proportion of Total Variance in \n Y Explained by Confounder",
      line=2.5, cex.lab=.85)
title(xlab="Proportion of Total Variance in \n M Explained by Confounder",
      line=3, cex.lab=.85)
```

Next, we provide the code for producing the results in Table 2 using the JOBS II data. The full code and data set are available from the authors as part of the replication materials.

```
# Variable labels for the pretreatment covariates
Xnames <- c("sex", "age", "occp", "marital", "nonwhite",
            "educ", "income", "depress_base")

# Fit models for the treatment, mediator and outcomes
fit.T <- lm(formula(paste(c("comply~treat", Xnames), collapse="+")),
            data=data)
fit.M <- lm(formula(paste(c("job_seek~comply+treat", Xnames), collapse="+")),
            data=data)
fit.Y1 <- lm(formula(paste(c("depress2~job_seek*(comply+treat)", Xnames),
                           collapse="+")), data=data)
fit.Y2 <- glm(formula(paste(c("work~job_seek*(comply+treat)", Xnames),
                            collapse="+")), data=data, family=binomial)

# Now estimate the mediation effects
out1 <- ivmediate(fit.T, fit.M, fit.Y1, sims = 2000, boot = TRUE,
                  enc = "treat", treat = "comply", mediator = "job_seek",
                  conf.level = c(.90,.95), multicore = TRUE, mc.cores=20)
summary(out1, conf.level=.95)

out2 <- ivmediate(fit.T, fit.M, fit.Y2, sims = 2000, boot = TRUE,
                  enc = "treat", treat = "comply", mediator = "job_seek",
                  conf.level = c(.90,.95), multicore = TRUE, mc.cores=20)
summary(out2, conf.level=.95)
```