1

Identification of Causal Parameters in Randomized Studies with Mediating Variables

Michael E. Sobel
Columbia University

*The material in sections 1-3 was presented at the Arizona State University Preventive Intervention Research Center "Conference on Mediational Models in Prevention Research" in Tempe, Arizona, March 23-24, 1998 in a working paper entitled "Identification of Causal Parameters in Randomized Studies with Mediators". I am especially grateful to Booil Jo, David MacKinnon and Bengt Muthen for helpful comments and advice on that iteration and several others since then. In addition, I thank Jennifer Hill and several anonymous reviewers for helpful remarks and advice.

ABSTRACT

Randomized trials are used to assess the effectiveness of one or more treatments in inducing outcomes of interest. Treatments are typically designed to target key mediating variables that are thought to be causally related to the outcome. Thus, researchers want to know not only if the treatment is effective, but how the mediators affect the outcome. Data from such studies are often analyzed using recursive linear structural equation models, and model coefficients, including the coefficient relating the mediator(s) to the outcome, are endowed with a causal interpretation. However, because only assignment to treatment groups is randomized, not assignment to the mediators, the latter are self selected treatments. In order to believe that the so-called "direct effect" of the mediator on the outcome variable in a structural equation model warrants a causal interpretation one must believe there is no selection bias with respect to the mediator. Holland (1988) studied the case of a single continuous mediator. He criticized the use of structural equation models and showed how to estimate the effect of the mediator on the outcome using treatment assignment as an instrumental variable. However, the assumptions he used to justify the instrumental variable approach are overly strong and substantively implausible. This paper has several goals: 1) to make explicit the assumptions needed to justify equating the parameters of structural equation models with the effects of mediators, 2) to provide weaker and more plausible conditions than those used by Holland under which the instrumental variable estimand may be interpreted as a causal parameter, and 3) to extend the analysis to include the case where subjects do not necessarily take up the treatment to which they are assigned. I also briefly discuss the role of covariates and other possible assumptions to aid in the identification of mediated effects in randomized studies.

1. INTRODUCTION

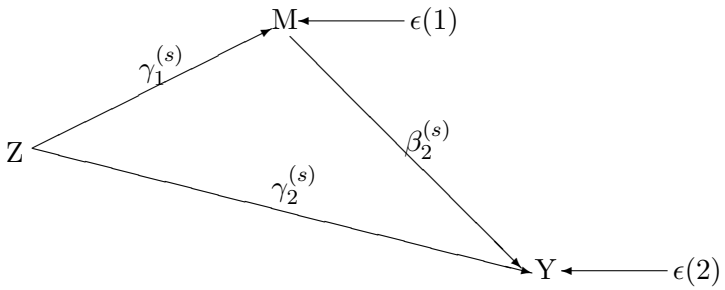
Randomized studies are often used to assess the effectiveness of one or or more treatments in bringing about a desired outcome (set of outcomes), for example, cessation or reduction in smoking (Botvin, Dusenbury, Baker, James-Ortiz, Botvin and Kerner 1992; Donaldson, Graham and Hansen 1994; Hansen and McNeal 1997; MacKinnon, Johnson, Pentz, Dwyer, Hansen, Flay and Wang 1991) among teens. Researchers often design treatments to affect subject's responses on key mediators that are believed to cause the outcome(s). Thus, they want to know if the program affects the targeted mediators and also if the mediators affect the outcome (MacKinnon 1994; MacKinnon and Dwyer 1993). Targeted variables that are not affected by the treatment point to problems in program design and/or delivery, while targeted variables that do not affect the outcomes of interest point to problems with the substantive theory underlying the program design. Both types of knowledge are useful for designing more effective treatments.

[Figure 1 About Here]

Typically, subjects assigned to a treatment group are compared to subjects in a control group and mediation is assessed using recursive linear structural equation models, employing criteria proposed by Kenny and his collaborators (for example, Judd and Kenny 1981a, 1981b, Baron and Kenny 1986). The path diagram corresponding to the model, for the case of a single mediator, is displayed in Figure 1; the parameters in that figure are superscripted with an "s" to denote that these are structural equation parameters. To demonstrate mediation, the following criteria are often proposed: a) the "total effect" $\tau^{(s)} = \gamma_2^{(s)} + \gamma_1^{(s)}\beta_2^{(s)}$ of treatment assignment Z on outcome Y is not 0, b) Z affects the mediator M , that is, the "direct effect" $\gamma_1^{(s)}$ of offering the treatment on the mediator is not 0, c) the "direct effect" $\beta_2^{(s)}$ of M on Y is not 0 in the structural equation model relating the outcome to treatment assignment and the mediator. Some authors do not require criterion a) per se (Baron and Kenney 1986) but if the "direct effect" $\gamma_2^{(s)}$ of the program on the outcome is required to be 0, criterion a) will hold if b) and c) hold. Baron and Kenney (1986) argue that when the parameter $\gamma_2^{(s)}$ is not 0, M only partially mediates the relationship between Z and Y, pointing to the presence of additional mediators not included in the model. MacKinnon and Dwyer (1993) have also argued that the mediated (or "indirect effect") $\gamma_1^{(s)}\beta_2^{(s)}$ of offering the

treatment on the outcome should be significantly different from 0. More generally, in cases where there are multiple mediators, mediation is studied applying the types of considerations above to each mediator in the model. Judd and Kenny (1981a) argue that if randomization is not used to

Figure 1: Path diagram for the 3 variable structural equation model



assign subjects to the treatment or control group, the coefficients above should not be interpreted as effects, unless covariates are included to adjust for selection into treatment groups. However, because it is not possible to know in an observational study whether the adjustments made are adequate, randomized studies are preferred. The most common type of randomized study is the "completely randomized" experiment (Rosenbaum 2002), in which subjects are assigned at random to a treatment or control group. This is the case studied in this paper; however, the analysis herein is easily extended to the case of a randomized block experiment and the case of an observational study where it is necessary to adjust for covariates.

In using structural equation models with data from randomized experiments to study mediation, many researchers assume that randomization enables estimates of model parameters to be given a causal interpretation. Building on Rubin's model for causal inference (Rubin 1974,1977,1978,1980), Holland (1988) argued that this is incorrect. As an example, he constructed a linear causal model for a hypothetical randomized encouragement design with a binary treatment variable Z indexing whether a subject is encouraged or not to study, a continuous mediator M (amount of time studied) and a continuous outcome Y (subject's test performance), and he compared the parameters of the causal model (superscripted with a "c") to the corresponding parameters of the structural equation model depicted in Figure 1. By virtue of randomization,

the effect $\gamma_1^{(c)}$ of Z on M in the causal model equals the "direct effect" $\gamma_1^{(s)}$ in the corresponding structural equation model. However, the effect $\gamma_2^{(c)}$ of Z on Y unmediated by M is not equal to the "direct effect" $\gamma_2^{(s)}$. Nor is the causal effect $\beta_2^{(c)}$ of M on Y equal to the "direct effect" $\beta_2^{(s)}$. Finally, the causal effect of Z on Y, $\tau^{(c)} = \tau^{(s)} = \gamma_2^{(s)} + \gamma_1^{(s)}\beta_2^{(s)}$, also by virtue of randomization.

Holland then showed that when a) the unmediated effect of Z on Y, $\gamma_2^{(c)} = 0$, b) the effect of Z on M, $\gamma_1^{(c)} \neq 0$, and c) all the effects in the causal model are identical for all subjects, the effect $\beta_2^{(c)}$ of M on Y is equal to the instrumental variable (IV) estimand (the effect of Z on Y divided by the effect of Z on M).

While Holland's work demonstrates that the routine application of structural equation modeling may lead to causal inferences that are not valid, his results on the IV estimand are of very limited practical value. First, the assumption that all the effects are the same for all subjects is implausible in studies with human subjects. (Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) relaxed this assumption for a binary mediator, and Angrist and Imbens (1995) consider the case of a discrete metrical mediator.) Second, Holland did not give conditions under which the "direct effects" in structural equation models could be endowed, as in the existing literature on mediation, with a causal interpretation. Understanding these conditions is important to any researcher who is contemplating the use of these models to study mediated effects.

This paper examines the identifiability of mediating effects using structural equation models and IV methods. Using potential outcomes, in section 2, I construct a linear causal model analogous to the recursive linear structural equation model in Figure 1. Theorem 1 compares the parameters of the causal model with the analogous parameters of the structural equation model, giving sufficient conditions for structural equation models to yield valid inferences about the effects of mediators on outcomes. This allows researchers to ask if such conditions are substantively plausible in particular applications. Section 3 compares the IV estimand for the effect of the mediator on the outcome with the "direct effect" $\beta_2^{(s)}$ of the mediator on the outcome in the structural equation model. The assumptions used by Holland to justify interpreting the IV estimand as the effect of M on Y are neither stronger nor weaker than the assumptions in Theorem 1. Thus, it is possible for the parameters of a structural equation model to warrant a causal interpretation when the IV estimand does not. Next, I weaken the identifying assumptions

used by Holland and give alternative and more plausible conditions under which the IV estimand identifies the effect of the mediator on the outcome. Importantly (when an "exclusion restriction" also holds), these conditions are weaker than those needed to identify mediated effects using structural equation models. Section 4 presents a brief example illustrating the results. In addition, the case where these conditions do not hold is studied. The effects of treatment assignment Z on M and Y in sections 2 and 3 are average treatment effects only when the treatment received and assigned are identical. But in experiments, for example, a job training program in which the treatment consists of attending a series of workshops, and where subjects assigned to a treatment group may not attend the workshops, as in Jo (2002a,2002b) and Little and Yau (1998), the distinction between treatment assignment and treatment received can be quite important. While the effects of treatment assignment considered in sections 2 and 3 are clearly still of interest (as they provide information about the effectiveness of the entire treatment package), an investigator typicallly wants to also know the effects of treatment itself. Identifying these effects is a difficult problem, as the treatment received is self-selected. Section 5 extends the analysis to this case, thereby incorporating into a more general framework some of the recent statistical work on compliance (for example, Angrist et al. 1996). Section 6 contains a brief discussion of the case where there are multiple mediators and section 7 concludes.

2. IDENTIFICATION OF MEDIATED AND UNMEDIATED EFFECTS USING STRUCTURAL EQUATION MODELS

In section 2.1, I define mediated and unmediated effects, both at the individual and population level. A linear causal model is constructed for comparison with the linear recursive structural equation models featured in the literature on mediation and discussed in section 2.2. Unlike the parameters of the linear causal model, the structural equation parameters are identifiable from the population moments of the observable data. In general (even in randomized studies) the structural and causal parameters are not equal, implying that the structural parameters should not be interpreted as effects. However, under certain (restrictive) conditions, characterized

in Theorem 1 of section 2.3, the causal parameters of interest are equal to the corresponding structural equation parameters. By making the relevant conditions explicit, Theorem 1 enables researchers to ask whether these are reasonable or not in their applications.

2.1 Defining The Causal Parameters of Interest

Let Z denote the treatment group to which a subject is assigned ($Z = 1$ if assigned to the treatment group, 0 otherwise), and let M and Y denote, respectively, a continuous mediator and a continuous outcome. The case where the treatment is targeted at several mediators that may affect the outcome is discussed in section 6.

For each subject i in a population $\mathcal{P}$, let the random variable $M_i(0)$ denote i's value on the mediator when assigned to the control group; similarly, let $M_i(1)$ denote the value when i is assigned to the treatment group. The pair $(M_i(0), M_i(1))$ is called a set of potential outcomes because only one of these can be realized and observed. Similarly, let $Y_i(0, M_i(0))$ $(Y_i(1, M_i(1)))$ denote i's value on the outcome variable when assigned to the control (treatment) group. Note that the response of unit i depends only on the treatment assigned to i and not the treatment to which any other unit is assigned. Rubin (1980) calls this the stable unit treatment value assumption (SUTVA), and it is made throughout this paper.

For unit i, the causal effects of Z on M and Y, respectively, are:

$$M_i(1) - M_i(0), \tag{1}$$

$$Y_i(1, M_i(1)) - Y_i(0, M_i(0)). \tag{2}$$

Averaging these over $\mathcal{P}$ gives the average effects of Z on M and Z on Y:

$$E(M(1) - M(0)) \equiv \gamma_1^{(c)}, \tag{3}$$

$$E(Y(1, M(1)) - Y(0, M(0))) \equiv \tau^{(c)}. \tag{4}$$

Conceptually, the treatment offered affects the outcome by 1) affecting the mediator, which in turn affects the outcome, and 2) affecting other variables that in turn affect the outcome. Because these other variables are not measured, the second type of effect appears as a "direct"

(unmediated) effect of Z on Y. To separate these two types of effects, some additional structure must be imposed. Let Ω_M denote the set of values the mediator can take. For each unit, potential outcomes $Y_i(z, m)$ are now defined for all $m \in \Omega_M$ and $z = 0, 1$; thus $Y_i(z, m)$ is i's outcome when he is assigned to treatment z and receives amount m of the mediator. The causal effects of Z on Y may now be represented as the sum of two types of effects: 1) a "mediated" effect of Z on Y through M and 2) an "unmediated" effect of Z on Y:

$$\begin{aligned} & Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \\ & \{Y_i(1, M_i(1)) - Y_i(0, M_i(1))\} + \{Y_i(0, M_i(1)) - Y_i(0, M_i(0))\} = \\ & \{Y_i(1, M_i(0)) - Y_i(0, M_i(0))\} + \{Y_i(1, M_i(1)) - Y_i(1, M_i(0))\}. \end{aligned} \tag{5}$$

In the two decompositions above, the first component is an unmediated effect and the second is a mediated effect. In general, the values of these two components may depend on which decomposition is used. To obtain a unique decomposition, it is necessary to assume, for example, for all i, m and $m*$,

$$Y_i(1, m) - Y_i(0, m) = Y_i(1, m^*) - Y_i(0, m^*). \tag{6}$$

Holland (1988) refers to this as an "additivity" assumption; Robins (2003) and Ten Have, Elliot, Joffe, Zanutto and Datto (2004), among others, have also considered this assumption.

It is important to realize that the additivity assumption is not innocuous and empirical researchers who wish to conduct mediational analyses should attempt to address the reasonableness of this assumption before proceeding further. Unfortunately, unlike the case where Y is regressed on Z and M, assumption (6) is not testable. One instance where the additivity assumption holds is when the treatment does not have any effect on the outcome except through the mediator, i.e., the unmediated effect is 0. This is the "so-called" exclusion restriction (subsequently discussed), and it is likely to hold in a double blinded study.

Under additivity, the average unmediated effect of Z on Y can be written as

$$E(Y(1, m) - Y(0, m)) \equiv \gamma_2^{(c)}. \tag{7}$$

Using additivity and averaging over (5) gives

$$\tau^{(c)} = \gamma_2^{(c)} + E(Y(1, M(1)) - Y(1, M(0))), \tag{8}$$

the sum of the average unmediated effect of Z on Y and the average mediated effect of Z on Y.

While it is possible to separate the average effect of Z on Y into mediated and unmediated components using only the additivity assumption, further decomposition of the mediated effect into components due to the effect of Z on M and M on Y requires imposing additional structure. Following the conventional literature on mediation, in which linear structural equation models are used to express the mediated effect as the product of the effect of Z on M with the effect of M on Y, I construct a causal model in which the relationship between the outcome and the mediator is linear. The additivity assumption is weaker than the linearity assumption, that is, if the linear causal model holds, additivity holds. But even if the linearity assumption holds, additional assumptions are needed in order to express the mediated effect as the product of the effect of Z on M with the effect of M on Y. If these types of assumptions are not substantively warranted, a researcher would be well advised to abandon the attempt to further decompose the unmediated effect $E(Y(1, M(1)) - Y(1, M(0)))$.

With the foregoing caveat in mind, I now use potential outcomes to express the causal relationships between Z,M and Y as the system:

$$M_i(z) = \alpha_1^{(c)} + \gamma_1^{(c)} z + \varepsilon_i(z), \tag{9}$$

$$Y_i(z, m) = \alpha_2^{(c)} + \gamma_2^{(c)} z + \beta_2^{(c)} m + \varepsilon_i(z, m), \tag{10}$$

where $E(\varepsilon(z)) = 0$ for $z = 0, 1$ and $E(\varepsilon(z, m)) = 0$ for all values of the pair (z, m). This implies $\gamma_1^{(c)}$ and $\gamma_2^{(c)}$ are as previously described and the effect of M on Y, at level m versus m^*, is

$$\beta_2^{(c)}(m - m^*) = E(Y(z, m) - Y(z, m^*)). \tag{11}$$

2.2 Structural Equation Parameters

Researchers using recursive structural equation models do not consider the causal model (9) and (10). They consider the linear system in the observed outcomes:

$$M_i(Z_i) = \alpha_1^{(s)} + \gamma_1^{(s)} Z_i + \epsilon_i(1), \tag{12}$$

$$Y_i(Z_i, M_i(Z_i)) = \alpha_2^{(s)} + \gamma_2^{(s)} Z_i + \beta_2^{(s)} M_i(Z_i) + \epsilon_i(2), \tag{13}$$

where $Z_i = 1$ if subject i is assigned to the treatment group, 0 otherwise, and $M_i(Z_i)$ is the observed value of the mediator. The parameters of (12) and (13) are identified through the definitions $E(\epsilon(1) \mid Z = z) = 0$ and $E(\epsilon(2) \mid Z = z, M(Z) = m) = 0$. Thus,

$$\gamma_1^{(s)} = E(M(1) \mid Z = 1) - E(M(0) \mid Z = 0), \tag{14}$$

$$\gamma_2^{(s)} = E(Y(Z, M(Z)) \mid M(Z) = m, Z = 1) - E(Y(Z, M(Z)) \mid M(Z) = m, Z = 0), \tag{15}$$

$$\beta_2^{(s)}(m - m^*) = E(Y(Z, M(Z)) \mid M(Z) = m, Z = z) - E(Y(Z, M(Z)) \mid M(Z) = m^*, Z = z). \tag{16}$$

The "total effect" of Z on Y is typically defined as: $\tau^{(s)} \equiv \gamma_2^{(s)} + \beta_2^{(s)}\gamma_1^{(s)}$; from (13) it is easy to see

$$\tau^{(s)} = E(Y(1, M(1)) \mid Z = 1) - E(Y(0, M(0)) \mid Z = 0). \tag{17}$$

2.3 Comparison of Causal Parameters and Structural Equation Parameters in Non-Randomized and Randomized Studies

In general, the value of the structural equation parameter $\gamma_1^{(s)}$ is not equal to the value of the causal parameter $\gamma_1^{(c)}$. Similar remarks apply to the parameters $\gamma_2^{(s)}$ and $\gamma_2^{(c)}$ and the parameters $\beta_2^{(s)}$ and $\beta_2^{(c)}$. Nor is $\tau^{(c)}$ in (4) generally equal to $\tau^{(s)}$ in (17).

However, in randomized studies, $\gamma_1^{(s)} = \gamma_1^{(c)}$ and $\tau^{(c)} = \tau^{(s)}$ (Holland 1988). This is because in a randomized study, it is reasonable to assume that treatment assignment is "ignorable" (Rosenbaum and Rubin 1983), that is, treatments are assigned independently of the potential outcomes:

$$M(0), M(1), Y(0, M(0)), Y(1, M(1)) \underline{\perp\!\!\!\perp} Z, \tag{18}$$

where the symbol "$\underline{\perp\!\!\!\perp}$" denotes statistical independence. (Note that independence of the potential outcomes and treatment assignment does not imply observed outcomes are independent of treatment assignment, e.g., $M = M(0) + (M(1) - M(0))Z$.)

Many researchers also impart a causal interpretation to the parameters $\gamma_2^{(s)}$ and $\beta_2^{(s)}$. In general, this interpretation is unwarranted, even when (18) holds. To see this, consider the

conditional expectation $E(Y(Z, M(Z)) \mid Z = z, M(Z) = m) = E(Y(z, m) \mid Z = z, M(Z) = m)$. Under (18), this reduces to $E(Y(z, m) \mid M(Z) = m)$; combining this result with (16) gives

$$\beta_2^{(s)}(m - m^*) = E(Y(z, m) \mid M(Z) = m) - E(Y(z, m^*) \mid M(Z) = m^*). \tag{19}$$

The first term of (19) is the average value of $Y(z, m)$ in the subpopulation of subjects with $M(z) = m$, while the second term is the average value of $Y(z, m^*)$ in the subpopulation of subjects with $M(z) = m^*$. Thus, the "direct effect" $\beta_2^{(s)}$ of M on Y in the structural equation model compares the values of the outcome in two different sub-populations of units. In contrast, the causal effect $\beta_2^{(c)}$ compares the same subjects under different conditions. Similar remarks apply to the "direct effect" $\gamma_2^{(s)}$ of Z on Y in the structural equation model (12)-(13) and the causal effect $\gamma_2^{(c)}$ of Z on Y in the causal model (9)-(10)

However, comparison of (16) with (11) and (15) with (7) leads trivially to a sufficient condition for equality of the remaining causal and structural parameters:

Theorem 1. Assume the causal model (9)-(10) and the ignorability assumption (18) hold and that also

$$Y(z, m) \underline{\perp\!\!\!\perp} M(z) \tag{20}$$

for $z = 0, 1$ and for all m. Then $\gamma_2^{(s)} = \gamma_2^{(c)}$, $\beta_2^{(s)} = \beta_2^{(c)}$.

Proof. From (18), $E(Y(Z, M(Z)) \mid Z = z, M(Z) = m) = E(Y(z, m) \mid Z = z, M(Z) = m) = E(Y(z, m) \mid M(Z) = m)$, which reduces to $E(Y(z, m))$ when (20) holds. Substituting this result into (15) gives $\gamma_2^{(s)} = \gamma_2^{(c)}$. Similarly, $\beta_2^{(s)} = \beta_2^{(c)}$.
Similar results have been obtained by Tenhave et al. (2005) and Egleston et al. (2006).

Assumption (20) states that if $M(z)$ is ignorable with respect to $Y(z, m)$, for $z = 0, 1$ and for all m, structural equation models can be used to study mediation. This would be the case if values of $M(z)$ were randomly assigned to subjects. But $M(z)$ is self-selected, and although this does not imply (20) does not hold, in many instances, this assumption may not be reasonable. For example, consider Holland's hypothetical encouragement study. If the amount of time studied varies systematically by intelligence, the assumption $Y(0, m) \underline{\perp\!\!\!\perp} M(0)$ will not be reasonable. For

example, suppose there are two types of students. The smart students study $m_s(0)$ hours when they are not encouraged and the others study $m_d(0) \neq m_s(0)$ hours. Because the smart students will outperform the others at any value of m, $Y(0,m)$ will not be independent of $M(0)$.

Theorem 1 is very useful. Researchers who have used structural equation models to study mediation have implicitly made the additional assumption (20). Making this explicit allows researchers to reexamine the validity of previous work and ask if it is reasonable or not to assume (20) in a particular application under consideration.

It is also important to note that although Theorem 1 only reports sufficient conditions for equality of corresponding structural and causal parameters, the condition (20) cannot be relaxed in a useful way: although (20) can be replaced by the assumption of mean independence $E(Y(z,m) \mid M(z) = m) = E(Y(z,m))$ for $z = 0, 1$, it is not easy to think of applications where mean independence holds, but (20) does not.

3. IDENTIFICATION OF CAUSAL EFFECTS USING AN INSTRUMENTAL VARIABLE

Section 3.1 takes up the case considered by Holland, where causal effects are assumed to be homogeneous across units. Section 3.2 relaxes the homogeneity assumption and gives identification conditions under which the IV estimand admits a causal interpretatoin.

3.1. Homogeneous Unit Effects and the IV estimand

Holland (1988) criticized the use of structural equation models and used the IV estimand $\tau^{(s)}/\gamma_1^{(s)}$ to identify the causal effect $\beta_2^{(c)}$. In section 2, the effects $\gamma_1^{(c)}$, $\gamma_2^{(c)}$, and $\beta_2^{(c)}(m - m^*)$ were defined as averages of the heterogeneous unit effects. Holland assumes the unit effects are the same for all subjects, that is, for all $i \in \mathcal{P}$,

$$M_i(1) - M_i(0) = \gamma_1^{(c)}, \tag{21}$$

$$Y_i(1,m) - Y_i(0,m) = \gamma_2^{(c)}, \tag{22}$$

$$Y_i(z,m) - Y_i(z,m^*) = \beta_2^{(c)}(m - m^*). \tag{23}$$

In terms of the causal model (9) and (10), assumption (21) is equivalent to assuming that for all i, a) $\varepsilon_i(0) = \varepsilon_i(1)$, and assumptions (22) and (23) are equivalent to assuming for all i b) $\varepsilon_i(z,m) = \varepsilon_i(z^*,m^*)$ for all z, z^*, m, m^*. The next result is an immediate consequence of (21)-(23).

Theorem 2. (Holland 1988). If the constant effects assumption (21)- (23) holds, $\tau^{(c)} = \gamma_2^{(c)} + \beta_2^{(c)}\gamma_1^{(c)}$.

Proof. Substituting (22) and (23) into the decomposition (5) yields $Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \gamma_2^{(c)} + \beta_2^{(c)}(M_i(1) - M_i(0))$. The result then follows using (21). But (21) is unnecessary, as the result also follows from averaging both sides of the decomposition above.

Because $\tau^{(c)} = \tau^{(s)}$ and $\gamma_1^{(c)} = \gamma_1^{(s)}$ when the ignorability assumption (18) holds, $\tau^{(s)} = \gamma_2^{(c)} + \beta_2^{(c)}\gamma_1^{(s)}$, yielding one linear equation in the two unknowns $\gamma_2^{(c)}$ and $\beta_2^{(c)}$. In applications where it is reasonable to assume that the treatment effect is transmitted solely through the mediator, the exclusion restriction

$$Y_i(0,m) = Y_i(1,m) \tag{24}$$

for all $i \in \mathcal{P}$ and $m \in \Omega_M$ holds, implying the direct effect $\gamma_2^{(c)} = 0$. This leads to the following result, also in Holland (1988).

Theorem 3. Under the ignorability assumption (18), the constant effects assumption (21)- (23), the exclusion restriction (24) and the assumption that the average effect of Z on M $\gamma_1^{(c)} \neq 0$, the IV estimand $\tau^{(s)}/\gamma_1^{(s)} = \beta_2^{(c)}$.

Unfortunately, the exclusion restriction is not always plausible and (assuming the other assumptions in Theorem 3 hold) the IV estimand is biased by the amount $\gamma_2^{(c)}/\gamma_1^{(s)}$ when the restriction does not hold. However, as the denominator of the bias term is identifiable, assumptions on the sign and/or magnitude of $\gamma_2^{(c)}$ can be used to bound the causal effect $\beta_2^{(c)}$.

In addition, the constant effects assumption (21)- (23) is not (mathematically) weaker than assumption (20). Similarly, assumption (20) is not weaker than the constant effects assumption. Examples where the IV estimand admits a causal interpretation but the structural equation parameters $\gamma_2^{(s)}$ and $\beta_2^{(s)}$ do not are readily constructed. In applications with human subjects, the assumptions needed to identify causal effects using structural equation models will often be implausible; the constant effects assumption will almost always be implausible. Although Theorems 2 and 3 can be weakened by removing (21) and assuming only that the effects (22) and (23) are constant within known covariate classess, this is only useful if all the sources of heterogenous causal effects are known. This is also implausible.

3.2 Heterogeneous Causal Effects, the IV estimand and the Average Mediated Effect

As previously noted, the identification condition (20) used to equate causal and structural equation parameters can only be weakened slightly. Thus, it is useful to consider weakening the assumption of constant effects. This is the content of Theorems 4 and 5. Theorem 6 shows that the new identification condition in Theorem 4 is also weaker than (20). That is, provided the exclusion restriction holds, the IV estimand can be used in some instances where the identification conditions required for using structural equation models are not met. Theorem 7 explores the substantive plausibility of using the IV estimand with heterogeneous causal effects.

Theorem 4. Under the causal model (9)-(10), $\tau^{(c)} = \gamma_2^{(c)} + \beta_2^{(c)}\gamma_1^{(c)}$ if and only if

$$E(\varepsilon(1, M(1)) - \varepsilon(0, M(0))) = 0. \tag{25}$$

Proof. Under (9)-(10),

$$Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \gamma_2^{(c)} + \beta_2^{(c)}(M_i(1) - M_i(0)) + \varepsilon_i(1, M_i(1)) - \varepsilon_i(0, M_i(0)). \tag{26}$$

Averaging both sides gives the result.

To understand the relationship between this result and the assumption of constant effects, recall that (22), (23) and (10) jointly imply, for all i, $\varepsilon_i(z,m) = \varepsilon_i(z^*,m^*)$ for all z, z^*, m, m^*, whence $\varepsilon_i(1,M(1)) - \varepsilon_i(0,M(0)) = 0$ for all i. So theorem 4 replaces this strong implication of the constant effects assumption with the weaker assumption that the difference in the potential errors $\varepsilon_i(1,M(1))$ and $\varepsilon_i(0,M(0))$ is 0 on average.

Replacing the assumption of constant effects in theorem 3 with the weaker assumption (25) now gives:

Theorem 5. Under the asssumption (18) that treatment assignment is ignorable, the causal model (9)-(10), assumption (25), the exclusion restriction (24) and the assumption that the average effect of the treatment on the mediator $\gamma_1^{(c)} \neq 0$, the IV estimand $\tau^{(s)}/\gamma_1^{(s)} = \beta_2^{(c)}$.

An important point is that assumption (25) relaxes (is weaker than) the assumption (20) used to justify structural equation modeling of mediated effects:

Theorem 6. If the ignorability assumption (20) holds, (25) holds as well. The converse does not hold.

Proof. Assumption (20) is equivalent to the assumption

$$\varepsilon(z,m) \underline{\perp\!\!\!\perp} M(z) \tag{27}$$

for $z = 0, 1$ and $m \in \Omega_M$. The expectations $E(\varepsilon(z,M(z)))$ can be reexpressed as:

$$EE(\varepsilon(z,M(z)) \mid M(z) = m) = EE(\varepsilon(z,m) \mid M(z) = m). \tag{28}$$

By (27), $E(\varepsilon(z,m) \mid M(z) = m)$ does not depend on $M(z)$, hence has value 0. However, (25) does not imply (20): for example, take $\varepsilon_i(z,m) = \kappa(M_i(z) - E(M(z))$ for $z = 0, 1$.

Although (25) relaxes both the assumption of constant effects and the ignorability assumption

(20), the question remains as to whether the conditions in Theorem 5 are plausible in empirical work. Theorem 7 gives a sufficient condition for (25) to hold. Although stronger than (25), this condition is important because it is easy to think about substantively and it also represents the most plausible way in which (25) is likely to come about.

Whereas assumption (20) in Theorem 1 states that there is no selection on the mediators with respect to the potential outcomes $Y(z,m)$ (or equivalently, on the potential errors $\varepsilon(z,m)$), the conditions in theorem 7 are substantively weaker, requiring only that there be no selection on the difference between the potential errors. As discussed below, whereas the ignorability assumption (20) is not plausible in the hypothetical encouragement study considered by Holland (1988), assumption (25) is much more reasonable.

Theorem 7. For all m and m^*, suppose

$$\varepsilon(1,m) - \varepsilon(0,m^*) \underline{\perp\!\!\!\perp} M(0), M(1). \tag{29}$$

Then $E(\varepsilon(1,M(1)) - \varepsilon(0,M(0)) = 0$.

Proof

$$E(\varepsilon(1,M(1)) - \varepsilon(0,M(0))) = EE(\varepsilon(1,M(1)) - \varepsilon(0,M(0)) \mid M(0)=m^*, M(1)=m) =$$
$$EE(\varepsilon(1,m) - \varepsilon(0,m^*) \mid M(0)=m^*, M(1)=m) = E(\varepsilon(1,m) - \varepsilon(0,m^*)) = 0, \tag{30}$$

where the last equality follows from the independence hypothesis (29) and the fact that the errors in the causal model have mean 0.

(Technically the independence assumption (29) is not weaker than the marginal ignorability assumption (27) or the equivalent assumption (20). However, from a substantive point of view, the joint ignorability assumption $\varepsilon(0,m),\varepsilon(1,m) \underline{\perp\!\!\!\perp} M(0), M(1)$ is not much stronger than (27), and (29) is mathematically weaker than this.)

To address the plausibility of the conditions in Theorem 7, recall the previous discussion of the encouragement design. The smart students study $m_s(0)$ hours when not encouraged while the others study $m_d(0) \neq m_s(0)$ hours when not encouraged. But had the smart students chosen to

study $m_d(0)$ hours when not encouraged, their performance would still have exceeded that of the others. Thus $E(Y(0, m_d(0)) \mid M(0) = m_d(0)) < E(Y(0, m_d(0)))$, that is, the less able students fall below the mean $E(Y(0, m_d(0)))$; equivalently $E(\varepsilon(0, m_d(0)) \mid M(0) = m_d(0)) < 0$. Similarly, $E(Y(1, m_d(1) \mid M(0) = m_d(1)) < E(Y(1, m_d(1))$, implying $E(\varepsilon(1, m_d(1) \mid M(0) = m_d(1)) < 0$. Under both conditions, the less able students fall below the mean, violating (27) (and (20). However, if on average, the less able students fall below the mean by the same amount under the two conditions, theorem 7 will hold.

Finally, the need for assumption (25) stems from the use of the causal linear regression (10) to separate the average mediated effect of Z on Y into components. Of course other parametric forms might be used, but it would still be necessary to assume (25) or something similar. When such assumptions are not plausible, it is still possible to conduct a mediational analysis. Theorem 8 takes a somewhat different tack, giving useful sufficient conditions for non-parametrically identifying the average mediated effect:

Theorem 8. Suppose treatment assignment is ignorable (assumption (18)) and the exclusion restriction (24) holds. Then $\tau^{(s)}$ is the average mediated causal effect of Z on Y.

Proof Under (18) $\tau^{(c)} = \tau^{(s)}$, and under additivity, $\tau^{(c)}$ is given by (8). Under the exclusion restriction (24), $\gamma_2^{(c)}$ in (8) is 0, whence the result.

Though simple, Theorem 8 is very important. The identification conditions in theorem 5, while weaker than those in previous research, are considerably stronger than those in theorem 8. Under the hypotheses in Theorem 8, it is still possible to answer two important questions of interest in a mediational analysis: 1) does the treatment package affect the mediator, and 2) what is the value of the mediated effect? The first question is answered by considering the (identified) effect of Z on M. The second is answered by considering the average mediated effect.

4. EXAMPLE

I use the hypothetical encouragement study considered by Holland to numerically compare the structural equation parameters of section 2.2 with the causal parameters of section 2.1 when the latter are identified using the IV estimand in section 3. I assume Z is ignorable (see 18), implying that the average effect of encouragement on amount of time studied M $\gamma_1^{(c)} = \gamma_1^{(s)}$, and the average effect of encouragement on test scores Y $\tau^{(c)} = \tau^{(s)}$. Throughout, I assume $\gamma_1^{(s)} = 3$ hours and $\tau^{(s)} = 18$ points.

Suppose now that the assumptions in Theorem 1 justifying structural equation modeling hold, with $\gamma_2^{(s)} = \gamma_2^{(c)} = 3$. In that case, each additional hour of study time results in an average improvement of $\beta_2^{(s)} = 5$ points. Because the ignorability assumption (20) in Theorem 1 implies (25) must also hold, $\tau^{(c)} = \gamma_2^{(c)} + \beta_2^{(c)}\gamma_1^{(c)}$, as in Theorem 4. But if, in order to identify β_2^c using the IV estimand, it is incorrectly assumed that $\gamma_2^{(c)} = 0$, the IV estimand 18/3 is biased.

It is also worth noting that conclusions identical to those above are reached in the case where Holland's assumption of constant effects holds. That is because the weaker assumption in (25) is sufficient for the decomposition in Theorem 4 to hold. Thus, Holland's assumptions are not considered further.

Now suppose instead that the assumptions in Theorem 5 justifying the use of the IV estimand hold, but that the ignorability assumption (20) does not. In this case, $\gamma_2^{(c)} = 0$ and $\beta_2^{(c)} = 6$, whereas $\beta_2^{(s)} = \beta_2^{(c)} - \gamma_2^{(s)}/\gamma_1^{(c)} = 6 - 3/3 = 5$.

The general case where the assumptions in Theorem 5 fail is also easy to consider. If (25) does not hold, neither does (20). The bias (BIAS(IV)) of the IV estimand is:

$$\frac{\tau_1^{(c)}}{\gamma_1^{(c)}} - \beta_2^{(c)} = \frac{\gamma_2^{(c)} + E(\varepsilon(1, M(1)) - \varepsilon(0, M(0)))}{\gamma_1^{(c)}}, \tag{31}$$

whereas the bias of $\beta_2^{(s)}$ = BIAS(IV) - $\gamma_2^{(s)}/\gamma_1^{(c)}$. Thus, in the case where the assumptions needed to justify either structural equation modeling or the the use of the IV estimand fail, neither method dominates the other.

Finally, it is worth considering the special case in the preceding paragraph where the exclusion restriction (24) holds, but assumption (25) does not. Here the most sensible strategy is apparently

to forego attempting to further decompose and to fall back on Theorem 8 (which will hold even if the causal model is incorrect). See also the remarks following Theorem 8.

5. LOCAL AVERAGE TREATMENT EFFECTS

In general, the effect of treatment assignment Z on an outcome Y should not be interpreted as the effect of treatment itself unless all subjects comply with their assignments (Bloom, 1984). Thus, it is important to extend the previous analysis to the case of imperfect compliance. To that end, I extend the work of Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996), who consider the interpretation of the IV estimand when Z is used as an instrumental variable for a binary mediator D, indexing treatment received,with potential outcomes $D_i(z)$ for $z = 0, 1$. Theorem 9 presents the main result in Angrist et. al (1996); to facilitate comparison with later material, I consider the intermediate outcome variable, with potential outcomes $M_i(z, d)$, for $z = 0, 1$ and $d = 0, 1$.

Theorem 9. (Angrist et al. 1996) Suppose randomization is used to assign subjects, the exclusion restriction $M_i(0, d) = M_i(1, d)$ holds for $d = 0, 1$ and all i, $D_i(1) \geq D_i(0)$ (monotonicity) and the instrument affects the treatment received, that is, $E(D(1) - D(0)) > 0$. Then the IV estimand

$$\frac{E(M(1, D(1)) - M(0, D(0)))}{E(D(1) - D(0))} = E(M(z, 1) - M(z^*, 0) \mid D(0) = 0, D(1) = 1), \tag{32}$$

where $z = 0, 1$, $z^* = 0, 1$, is the average effect of treatment D (not treatment assigment Z) on M in the subpopulation of compliers (subjects with $D_i(0) = 0, D_i(1) = 1$). Note that this result also holds when the ignorability assumption (18) is made in place of the randomization assumption. Angrist et al. (1996) call this effect a local average treatment effect (LATE). Comparison of the ITT $E(M(1, D(1)) - M(0, D(0)))$ with the LATE (32) is potentially very informative. The former measures the overall success of the treatment package in affecting the mediator. It depends on the substantive content of the package and the compliance rate. If (32) is substantial, but the ITT is not, and the investigator thinks the effect of the treatment package on the mediator for

the non-compliers is also substantial, this suggests the need to redesign the treatment package to increase compliance, as opposed to redesigning the substantive content of the package.

Similarly, let $Y_i(z, d, m)$ denote the outcome when unit i is assigned to treatment z, receives treatment d and has value m of the mediator. Under the conditions in Theorem 9 and the exclusion restriction $Y_i(0, d, m) = Y_i(1, d, m)$ for all d, m and i, the IV estimand

$$\frac{E(Y(1, D(1), M(1, D(1))) - Y(0, D(0), M(0, D(0)))}{E(D(1) - D(0))} \tag{33}$$

is the LATE for the effect of D on Y, and comparison of this with the ITT (4) may also be informative. However, the treatment effects above do not attempt to account for the role of the mediator M in bringing about the outcome. To do so, I decompose the LATE for the effect of D on Y.

5.1 Complier Average Mediated Effects

Parallelling the earlier material, where average unmediated and mediated effects of Z on Y were defined, average unmediated and mediated effects of D on Y are defined. An extension of the exclusion restriction leads to identification of the average mediated complier effect using Z as an instrumental variable. Next, using a linear model for the compliers, conditions under which the IV estimand is the causal effect of M on Y in the subpopulation of compliers are given.

For $z = 0, 1$, the unit effects of D on Y, $Y_i(z, 1, M_i(z, 1)) - Y_i(z, 0, M_i(z, 0))$, may be written as:

$$\{Y_i(z, 1, M_i(z, 1)) - Y_i(z, 0, M_i(z, 1))\} + \{Y_i(z, 0, M_i(z, 1)) - Y_i(z, 0, M_i(z, 0))\}. \tag{34}$$

Under the additivity assumption $Y_i(z, 1, m) - Y_i(z, 0, m) = Y_i(z, 1, m^*) - Y_i(z, 0, m^*)$ for all i, m and m^*, and the exclusion restriction $Y_i(1, d, m) = Y_i(0, d, m)$ for all d and m, the first component of (34) is the unmediated effect of D on Y for unit i, while the second component is the effect of D on Y mediated by M. In this case, the ITT may be written as:

$$\begin{aligned} &E(Y(1, D(1), M(1, D(1))) - (Y(1, D(0), M(1, D(1))) + \\ &E(Y(0, D(0), M(1, D(1))) - (Y(0, D(0), M(0, D(0))), \end{aligned} \tag{35}$$

the sum of the average unmediated effect of D on Y and the average mediated effect of D on Y. If the effect of D on Y is transmitted solely through M, the first component of (35) is 0. The exclusion restrictions previously considered do not imply this. However, this is implied by the stronger exclusion restriction

$$Y_i(0,0,m) = Y_i(0,1,m) = Y_i(1,0,m) = Y_i(1,1,m) \tag{36}$$

for all $i \in \mathcal{P}$ and $m \in \Omega_M$. This leads immediately to the following analogue to Theorem 8:

Theorem 10. Under the hypotheses in theorem 9 and the exclusion restriction (36), the IV estimand

$$\frac{E(Y(1,D(1),M(1,D(1))) - Y(0,D(0),M(0,D(0))))}{E(D(1)-D(0))} \tag{37}$$

is the complier average mediated effect of D on Y.

As before, without imposing additional structure, (37) cannot be separated into the effect of D on M and the effect of M on Y. Nevertheless, (32) and (37) can be used to address two key questions: among the compliers 1) does the treatment affect the mediator, and 2) does the treatment affect the outcome via the mediator?

In keeping with the previous material, a linear causal model (for the compliers) is used to separate the IV estimand (37) into components:

$$Y_i(z,d,m) = \nu_c^{(c)} + \eta_c^{(c)} z + \kappa_c^{(c)} d + \lambda_c^{(c)} m + \delta_i(z,d,m), \tag{38}$$

where $E(\delta(z,d,m) \mid D(0)=0, D(1)=1) = 0$ and the subscript c is used to index the value of the corresponding parameter in the subpopulation of compliers. The exclusion restriction (36) is incorporated by assuming $\eta_c^{(c)} = 0$, $\kappa_c^{(c)} = 0$, $\delta_i(0,0,m) = \delta_i(0,1,m) = \delta_i(1,0,m) = \delta_i(1,1,m)$.

Using (38), (37) may be reexpressed as:

$$\frac{\begin{aligned}\lambda_c^{(c)} E(M(1,D(1)) - M(0,D(0)) \mid D(0)=0, D(1)=1) + \\ E(\delta(1,D(1),M(1,D(1))) - \delta(0,D(0),M(0,D(0)))) \mid D(0)=0, D(1)=1)\end{aligned}}{\Pr(D(0)=0, D(1)=1)}. \tag{39}$$

From (39) it is easy to see that a condition analogous to (25) must hold in order for $\lambda_c^{(c)}$ to equal the average complier effect of a one unit increase in M. Putting together these results and

substituting the IV estimand (32) for the complier average effect of D on M gives:

Theorem 11. Under the hypotheses in theorem 9, the hypothesis $E(M(1, D(1)) - M(0, D(0)) \neq 0$, the causal model (38), the exclusion restriction (36), and the hypothesis $E(\delta(1, D(1), M(1, D(1)) - \delta(0, D(0), M(0, D(0)) \mid D(0) = 0, D(1) = 1) = 0$, the ratio of the IV estimands (37) and (32)

$$\frac{E(Y(1, D(1), M(1, D(1))) - Y(0, D(0), M(0, D(0))))}{E(M(1, D(1)) - M(0, D(0)))} = \lambda_c^{(c)}, \tag{40}$$

is the complier average effect on Y of a one unit increase in M.

An interesting and important feature of theorem 11 is that the complier effect of M on Y is identified without using any information on the treatments subjects have taken up. It is nevertheless useful to estimate the proportion of the population consisting of compliers; if this is small and the average causal effects for compliers and non-compliers are not equal, the study results may be of limited utility for making policy. On the other hand, if the compliers represent the majority of the population, the results will be of greater policy relevance (even if the complier and non-complier effects are not equal).

Finally, when subjects assigned to the control group cannot take up treatment, the compliers consist of all the treated subjects . In this case, the complier effects herein may be interpreted as treatment effects on the subpopulation that is treated.

6. MULTIPLE MEDIATORS

To keep matters as straightforward as possible, only the case of a single mediator M has been considered. Yet study designers often target more than one mediator and wish to simultaneously consider the effects of multiple mediators on the outcome.

The material on structural equation models in section 2 extends straightforwardly to the case of a vector valued mediator $\underline{M}$. In particular, Theorem 1 is generalized to this case by comparing a linear structural equation model with multiple mediators to a linear causal model for potential outcomes $Y_i(z, \underline{m})$ and replacing the condition (20) with the condition $Y(z, \underline{m}) \perp\!\!\!\perp \underline{M}_z$ for $z = 0, 1$ and for all $\underline{m}$.

Theorem 8 also admits a straightforward generalization to this case. Assuming that treatment can affect the outcome Y only through a set of mediators $\underline{M}$ and that all these mediators have been correctly identified by the investigator, τ^s is then the mediated effect of Z on Y through $\underline{M}$. Alternatively, in a randomized double blinded study with treatment targeted at $\underline{M}$, this result should also hold. Clearly Theorem 10 extends as well.

It should also be clear that with only a treatment group and a control group, it will not be possible to identify the effect of any single mediator nor the separate effects of multiple mediators using standard IV methods.

If a single mediator model is assumed and multiple mediators have been targeted in the study, the exclusion restriction will almost surely be violated. Nor are the separate effects of multiple mediators (when these are separable) identified using standard IV methods. To see the latter point as simply as possible, suppose there are two mediators M and P, and the model for the potential outcomes is:

$$Y_i(z,m,p) = \alpha^{(c)} + \beta_m^{(c)} m + \beta_p^{(c)} p + \varepsilon_i(z,m,p), \tag{41}$$

where $E(\varepsilon(z,m,p)) = 0$ and $\varepsilon_i(0,m,p) = \varepsilon_i(1,m,p)$; note the incorporation of the exclusion restriction into the model. Under the assumption that treatment assignment is ignorable, the average effect of treatment assignment on Y is:

$$\begin{aligned} &\beta_m^{(c)} E(M(1) - M(0)) + \beta_p^{(c)} E(P(1) - P(0)) + \\ &\qquad E(\varepsilon(1, M(1), P(1)) - \varepsilon(0, M(0), P(0))). \end{aligned} \tag{42}$$

Supposing also that treatment assignment affects both mediators and that (25) holds, this gives one equation in the two unknowns $\beta_m^{(c)}$ and $\beta_p^{(c)}$. To identify both parameters, either one of the values must be known or additional information must be incorporated into the problem. One way to identify both parameters is to add an additional treatment group in order to pick up a second instrument (Genetian, Morris, Johannes and Bloom 2005). For example, in the first treatment group, the intervention might be targeted only at M and in the second treatment group, the intervention might be targeted at both M and P. Under suitable assumptions, the effect of M can be ascertained by comparing the first treatment group with the control group, and the effect of P can be obtained by comparing the two treatment groups. Other methods

of obtaining identification could also be considered, for example, adding multiple outcomes and imposing cross-equation restrictions. Using information on covariates is another possible means. Similar remarks would apply were compliance explicitly taken into account. A more general treatment of these issues is beyond the scope of this paper and will be given in a future paper.

6. DISCUSSION

This paper examines conditions under which structural equation models and instrumental variable methods can be used to identify causal effects of a mediating variable on an outcome. Even in a randomized experiment, the coefficient on the mediating variable in a structural equation model does not identify a causal parameter, unless additional and often implausible assumptions, explicated herein, are made. Instrumental variable methods are considered and conditions under which the IV estimand identifies the average effect of the mediator are given. The critical assumption (25) that is needed to identify the effect of the mediator using IV in models with heterogeneous unit effects is weaker than the assumption (20) that is needed to identify this effect using structural equation models. Section 5 synthesizes the psychological literature on mediation and recent statistical and econometric literature on compliance, showing that the IV estimand (40) identifies the complier average response to a one unit increase in the mediator.

In this paper, potential outcomes of the form $Y(z, m)$ (and later $Y(z, d, m)$) were considered. Some authors argue that such outcomes are ill defined, as only Z is randomized. Principal stratification can be used to take an alternative approach approach to mediation (Frangakis and Rubin 2002, Jo 2006). In this approach, which generalizes the recent statistical literature on compliance, only the potential outcomes $Y(z, M(z))$ for $z = 0, 1$ are considered, and causal comparisons are made only within latent subclasses defined jointly by values of $M(0)$ and $M(1)$. An extensive discussion of the relative merits of these two approaches to mediation is beyond the scope of this paper. The position taken herein is that both can be useful, as evidenced by section 5, in which the two approaches are blended.

The extension of the results for a single mediator to the case of multiple mediators was also considered. The results in this paper also extend readily to the case where the mediator and/or

outcome variables are latent. In observational studies or conditionally randomized experiments where the ignorability assumption (18) holds after conditioning on pretreatment covariates, the previous results also hold after conditioning on these covariates. In this case, effects that are conditional on the values of the covariates are obtained, and averaging over the distribution of the covariates then gives back the average effects. Kraemer, Wilson, Fairburn and Agras (2002) call such covariates moderators. Even in a randomized study, a researcher will often want to know how the effects of interest are moderated.

Following Angrist et al. (1996), the exclusion restrictions herein were assumed to hold for all subjects. Jo (2002a, 2002b) weakens this assumption and also shows how covariate information may be used in a different way to identify average effects of treatment assignment Z in the subpopulation of compliers. Extensions of the results herein to handle this case would also be useful.

1 REFERENCES

Angrist, J.D., & G.W. Imbens. (1995). "Two Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association,* 90, 431-42.

Angrist, J.D., Imbens, G.W., & D.B. Rubin. (1996). "Identification of Causal Effects Using Instrumental Variables" (with discussion). *Journal of the American Statistical Association,* 91, 444-72.

Baron, R.M., & D.A. Kenny. (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. (1984). "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review,* 8, 225-246.

Botvin, G.J, Dusenbury, L.L., Baker, E., James-Otiz, S., Botvin, E.M., & J. Kerner. (1992). "Smoking Prevention Among Urban Minority Youth: Assessing Effects on Outcome and Mediating Variables." *Health Psychology*, 11, 290-299.

Donaldson, S.I., Graham,J.W., & W.B. Hansen. (1994). "Testing the Generalizability of Intervening Mechanism Theories: Understanding the Effects of Adolescent Drug Use Prevention Interventions." *Journal of Behavioral Medicine,* 17, 1-21.

Egleston, B., Scharfstein, D., Munoz, B., & S. West. (2006). "Investigating Mediation When Counterfactuals Are Not Metaphysical: Does Sunlight UVB Exposure Mediate the Effect of Eyeglasses on Cataracts?" Unpublished manuscript, Johns Hopkins University.

Frangakis, C.E., & D.B. Rubin. (2002). "Principal Stratification in Causal Inference." *Biometrics* 58:21-29.

Genetian, L.A., Morris, P.A., Johannes, M., & H.S. Bloom. (2005). "Constructing Instrumental Variables from Experimental Data to Explore How Treatments Produce Effects." Pp. 75-114 in *Learning More from Social Experiments*, edited by H.S. Bloom.

Hansen, W.B., & R.B. McNeal. (1997). "How D.A.R.E. Works: An Examination of Program Effects on Mediating Variables." *Health Education Behavior,* 24, 165-176.

Holland, P.W. (1988). "Causal Inference, Path Analysis, and Recursive Structural Equation

Models." (with discussion). Pp. 449-493 in *Sociological Methodology* 1988, edited by C. C. Clogg. Washington, D.C: American Sociological Association.

Imbens, G.W., & J.D. Angrist. (1994). "Identification and Estimation of Local Average Treatment Effects." *Econometrica,* 62, 467-475.

Jo, B. (2002a). "Estimation of Intervention Effects with Noncompliance: Alternative Model Specifications" (with discussion). *Journal of Educational and Behavioral Statistics,* 27, 385-420.

Jo, B. (2002b). "Model Misspecification Sensitivity Analysis in Estimating Causal Effects of Interventions with Non-Compliance." *Statistics in Medicine,* 21, 3161-3181.

Jo, B. (2006). "Causal Inference in Randomized Trials with Mediational Processes." Unpublished manuscript, Stanford University.

Judd,C.M., & D.A. Kenny. (1981a). "Process Analysis: Estimating Mediation in Treatment Evaluations." *Evaluation Review,* 5, 602-619.

———1981b. *Estimating the Effects of Social Interventions.* New York: Cambridge University Press.

Kraemer, H.C., Wilson, G.T., Fairburn, C.G., & W. Stewart Agras. (2002). "Mediators and Moderators of Treatment Effects in Randomized Clinical Trials." *Archives of General Psychiatry,* 59, 877-883.

Little, R.J., & L.H.Y. Yau. (1998). "Statistical Techniques for Analyzing Data from Prevention Trials: Treatment of No-Shows Using Rubin's Causal Model". *Psychological Methods,* 3, 147-159.

MacKinnon, D.P. (1994). "Analysis of Mediating Variables in Prevention and Intervention Research." Pp. 127-153 in A. Cazares & L.A. Beatty, eds. *Scientific Methods for Prevention Intervention Research*, NIDA Research Monograph 139. DHHS Pub. No. 94-3631. Washington, DC: U.S. Government Printing Office.

MacKinnon, D.P., Johnson, C.A., Pentz, M.A., Dwyer, J.H., Hansen, W.B., Flay, B.R., & E. Yu-I Wang. (1991). "Mediating Mechanisms in a School-Based Drug Prevention Program: First-Year Effects of the Midwestern Prevention Project." *Health Psychology,* 1, 164-172.

MacKinnon, D.P., & J.H.Dwyer. (1993). "Estimating Mediated Effects in Prevention Stud-

ies." *Evaluation Review,* 17,141-158.

Robins, J.M. (2003). "Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects." Pp. 70-81 in P. Green, N.L Hjort & S. Richardson, eds, *Highly Structured Stochastic Systems*. New York: Oxford University Press.

Rosenbaum, Paul R. (2002). *Observational Studies.* 2d ed. New York: Springer.

Rosenbaum, P.R., & D.B. Rubin. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70, 41-55.

Rubin, D. B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, 66, 688-701.

———(1977). "Assignment to Treatment Groups on the Basis of a Covariate." *Journal of Educational Statistics,* 2, 1-26.

———(1978). "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics,* 6, 34-58.

———(1980). "Comment on 'Randomization Analysis of Experimental Data: The Fisher Randomization Test,' by D. Basu." *Journal of the American Statistical Association,* 75, 591-93.

Ten Have, T.R., Elliott, M.R., Joffe, M., Zanutto, E., & Datto, C. 2004. "Causal Models for Randomized Physician Encouragement Trials in Treating Primary Care Depression." *Journal of the American Statistical Association,* 99, 16-25.

Ten Have, T.R., Joffe, M., Lynch, K., Brown, G., & Maito, S. (2005) "Causal Mediation Analyses with Structural Mean Models." University of Pennsylvania Biostatistics Working Paper.