Under What Assumptions Do Site-by-Treatment Instruments Identify Average Causal Effects?

Sean F. Reardon and Stephen W. Raudenbush

Sociological Methods & Research 2013 42: 143 originally published online 26 July 2013
DOI: 10.1177/0049124113494575

Article

Sociological Methods & Research
42(2) 143-163

DOI: 10.1177/0049124113494575
smr.sagepub.com

Under What Assumptions Do Site-by-Treatment Instruments Identify Average Causal Effects?

Sean F. Reardon[1] and Stephen W. Raudenbush[2]

Abstract

The increasing availability of data from multisite randomized trials provides a potential opportunity to use instrumental variables (IV) methods to study the effects of multiple hypothesized mediators of the effect of a treatment. We derive nine assumptions needed to identify the effects of multiple mediators when using site-by-treatment interactions to generate multiple instruments. Three of these assumptions are unique to the multiple-site, multiple-mediator case: (1) the assumption that the mediators act in parallel (no mediator affects another mediator); (2) the assumption that the site-average effect of the treatment on each mediator is independent of the site-average effect of each mediator on the outcome; and (3) the assumption that the site-by-compliance matrix has sufficient rank. The first two of these assumptions are nontrivial and cannot be empirically verified, suggesting that multiple-site, multiple-mediator IV models must be justified by strong theory.

[1] Stanford University, Stanford, CA, USA
[2] University of Chicago, Chicago, IL, USA

Corresponding Author:
Sean F. Reardon, 520 Galvez Mall, #526, Stanford University, Stanford, CA 94305, USA.
Email: sean.reardon@stanford.edu

Keywords

instrumental variables, multisite trials, mediation analysis, causal inference

Introduction

In canonical applications of the instrumental variable method, exogenously determined exposure to an instrument induces exposure to a treatment condition which in turn causes a change in a later outcome. A crucial assumption known as the exclusion restriction is that the hypothesized instrument can influence the outcome only through its influence on exposure to the treatment of interest (Heckman and Robb 1985b; Imbens and Angrist 1994). It may be the case, however, that an instrument affects the outcome through multiple treatments, in which case a single instrument will not suffice to identify the causal effects of interest.

To cope with this problem, analysts have recently exploited the fact that a causal process is often replicated across multiple sites, generating the possibility of multiple instruments in the form of site-by-instrument interactions. These multiple instruments can, in principle, enable the investigator to identify the impact of multiple processes regarded as the mediators of the effect of an instrument. Kling, Liebman, and Katz (2007), for example, used random assignment in the Moving to Opportunity (MTO) study as an instrument to estimate the impact of neighborhood poverty (*NP*) on health, social behavior, education, and economic self-sufficiency of adolescents and adults. Reasoning that the instrument might affect outcomes through mechanisms other than *NP*, they control for a second mediator, use of the randomized treatment voucher. To do so, they capitalize on the replication of the MTO experiment in five cities, generating 10 instruments (site-by-randomization interactions)[1] to identify the impact of the two mediators of interest, *NP* and experimental compliance. Using a similar strategy, Duncan, Morris, and Rodrigues (2011) use data from 16 implementations of welfare-to-work experiments to identify the impact of family income, average hours worked, and receipt of welfare as mediators.

Clearly, this strategy for generating multiple instruments has potentially great appeal in research on causal effects in social science. For example, Spybrook (2008) found that, among 75 large-scale experiments funded by the U.S. Institute of Education Sciences over the past decade, the majority were multisite trials in which randomization occurred within sites. In principle, these data could yield a wealth of new knowledge about causal effects in education policy. It is essential, however, that researchers understand the assumptions required to pursue this strategy successfully. To date, we know of no complete account of these assumptions.

Our purpose therefore is to clarify the assumptions that must be met if this "multiple-site, multiple-mediator" instrumental variables (hereafter, MSMM-IV) strategy is to identify the average treatment effects (ATE) in the populations of interest. For simplicity of exposition and corresponding to the applications of MSMM-IV to date, we consider the case where a single instrument (which we denote as T) operates through a set of mediators $\boldsymbol{M} = \{M_1, M_2, \ldots, M_p\}$, that are linearly related to an outcome Y. We conclude that, in addition to the assumptions typically required in the single-site, single-instrument, single-mediator case, three new assumptions are required in the MSMM-IV case.

We begin by delineating the assumptions required for identification in the case of a single instrument and a single mediator within a single-site study. We describe the assumptions needed to identify the "local average treatment effect" (LATE) described by Angrist, Imbens, and Rubin (1996) and the (slightly different) assumptions needed to identify the average treatment effect (ATE) among the population. Additionally, we consider the general case where both the instrument and the mediator may be continuous or multivalued.

Following a discussion of the single-site, single-mediator case, we then turn our attention to the case of primary interest: the MSMM-IV design. We specify a set of nine assumptions required for the MSMM-IV model to identify the ATEs of the mediators, three of which are specific to the MSMM-IV case, and which we discuss in some detail.

The Single-Site, Single-Mediator Case

Notation

Suppose that each participant in a single-site study is exposed to a treatment T taking on values in the domain $\mathbb{T} \subset \mathbb{R}$. We hypothesize that T may affect some outcome Y through its effect on some mediator M. Thus, in our notation, T is an instrument that will be used to identify the effect of some mediator M. We often consider treatments taking on values in the domain $\mathbb{T} = \{0, 1\}$, where $T = 1$ if the participant is assigned to the "treatment" condition or $T = 0$ if she is assigned to the alternative "control" condition. Likewise, we often consider mediators taking on values in the domain $\mathbb{M} = \{0, 1\}$, where $M = 1$ if the individual experiences the mediator condition and $M = 0$ if she does not. More generally, however, both T and M may be multivalued or continuous.

Note that our terminology and notation differ here from those in standard econometric discussions of instrumental variables (IV). In the econometric tradition, an instrument Z is used to identify the effect of a treatment T on

an outcome Y. In this tradition, the reduced form effect of Z on Y is often not of substantive interest; rather, Z is of interest to the econometrician largely because it may be "instrumental" in identifying the effect of T on Y. In our terminology, however, assignment to a treatment T (such an intervention or policy condition) is used as an instrument to identify the effect of mediator M on an outcome Y. Our terminology derives from the program evaluation tradition, in which both the reduced-form effect of T on Y and the effects of the mediators through which T may operate are of interest. Throughout the remainder of this article, we shall use T to denote a treatment assignment condition that is used as an instrument, and we shall use M to denote an experienced mediator condition.

Figure 1 summarizes our notation. We refer to the effect of T on M as the "compliance;" the person-specific compliance is denoted Γ; the average compliance in the population is $\gamma = E[\Gamma]$. Similarly, the person-specific effect of the mediator M on the outcome Y is denoted as Δ; the average effect of M on Y in the population (often the estimand of interest) is denoted as $\delta = E[\Delta]$. Finally, we denote the person-specific effect of T on Y as B; the average effect of T on Y in the population (often referred to as the "intent-to-treat" [ITT] effect in the program evaluation literature or the "reduced form" effect in the econometrics literature) is therefore $\beta = E[B]$.

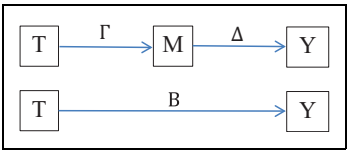

Figure 1. Mediated and reduced-form effects of T on Y.

Identifying Assumptions

In order to define a set of causal estimands of interest, we first require the assumption that an individual's potential outcomes depend only on the treatment condition and mediator condition to which that particular individual is exposed (and not on the treatment and mediator conditions of others), known as the Stable Unit Treatment Value Assumption (SUTVA; Rubin 1986). In the standard potential outcomes framework, we typically require a single SUTVA assumption stating that one individual's potential outcomes do not depend on others' treatment status. In the IV model, however, the presence of three variables of interest—the treatment T, a mediator M, and an outcome Y—necessitates a pair of such assumptions (Angrist et al. 1996), stated formally below.

Assumption (i): SUTVA:

(i.a) Each unit i has one and only one potential value of the mediator M for each treatment condition t: in particular, for a population of size N, $m_i(t_1, t_2, \ldots, t_N) = m_i(t_i)$ for all $i \in \{1, 2, \ldots, N\}$.

(i.b) Each unit i has one and only one potential outcome value of Y for each pair of values of treatment condition t and mediator value m: in particular, for a population of size N, $y_i(t_1, t_2, \ldots, t_N, m_1, m_2, \ldots, m_N) = y_i(t_i, m_i)$ for all $i \in \{1, 2, \ldots, N\}$.

Given the SUTVA assumptions, we can represent the potential outcome Y for a participant who experiences treatment t and mediator value $m(t)$ as $y(t, m(t))$ (we drop the subscript i throughout the remainder of this article except when necessary for clarity).

Our second assumption is that T affects Y only through its impact on the mediator M. This is the standard exclusion restriction assumption:

Assumption (ii): Exclusion restriction:

$$y(t) = y(t, m(t)) = y(m(t)).$$

The exclusion restriction combined with the second SUTVA Assumption (i.b) implies a third SUTVA condition: (i.c) Each unit i has one and only one potential outcome value of Y for each value of the mediator m: in particular, for a population of size N, $y_i(m_1, m_2, \ldots, m_N) = y_i(m_i)$ for all $i \in \{1, 2, \ldots, N\}$.

The SUTVA assumptions are necessary in order to define the causal estimands of interest. If the treatment variable is binary, for example, the first SUTVA Assumption (i.a) implies that we can define the person-specific casual effect of the treatment on M as $\Gamma = m(1) - m(0)$. If, however, the treatment is not binary, it will be useful to assume that the person-specific effect of T on M is linear in T, in which case $\Gamma = m(t) - m(t-1)$:

Assumption (iii): Person-specific linearity of the mediator M in T: the person-specific effect of T on mediator M is linear. That is, $m(t) = m(0) + t\Gamma$.

Likewise, it will be useful to assume that the person-specific effect of M on Y is linear in M. This is a standard, if not unproblematic, assumption in IV models. In this case, the third SUTVA condition (i.c) implies that we can define the person-specific casual effect of the mediator Y as $\Delta = y(m) - y(m-1)$:

Assumption (iv): Person-specific linearity in m: the person-specific effect of the mediator $m(t)$ on Y is linear. That is, $y(m(t)) = y(m = 0) + m(t)\Delta$.

The combination of (ii), (iii), and (iv) implies that the person-specific effect of T on Y is linear in T:

$$\begin{aligned} y(m(t)) &= y(m(0) + t\Gamma) \\ &= y(m = 0) + m(0)\Delta + t\Gamma\Delta. \end{aligned} \tag{1}$$

Thus, defining B as the person-specific effect of T on Y, we can relate the person-specific effects of T on M and of M on Y to the person-specific effect of T on Y by:

$$y(t) - y(t-1) = B = \Gamma\Delta. \tag{2}$$

The population average ITT effect of interest here is $E(B) = \beta$. The parameter β is not directly observable, however, because it is the mean of differences in counterfactual outcomes. If we are justified in assuming that persons are assigned ignorably to treatments $T = t$ for $t \in \mathbb{T}$, as would be true in a randomized experiment, we can estimate β from sample data.

Assumption (v): Ignorable treatment assignment: $T \perp Y(t), T \perp M(t)$, $t \in \mathbb{T}$.

Likewise, Assumption (v) enables us to estimate $E(\Gamma) = \gamma$ the average causal effect of T on the mediator M (which we refer to as the average compliance) from sample data. Because IV methods rely on the instrument to induce some exogenous variation in the mediator (for at least some individuals), we require γ to be nonzero:

Assumption (vi): Effectiveness of the instrument: $\gamma \neq 0$.

In the simple case in which we have a single instrument and a single mediator, the target of the IV estimator is the ratio of the ITT effect to the average compliance:

$$\frac{\beta}{\gamma} = \frac{E[\Gamma\Delta]}{E[\Gamma]} = \frac{\gamma\delta + \mathrm{Cov}(\Gamma, \Delta)}{\gamma} = \delta + \frac{\mathrm{Cov}(\Gamma, \Delta)}{\gamma}. \tag{3}$$

Equation (3) may be regarded as defining a "compliance-weighted average treatment effect" (CWATE) because each person's treatment effect Δ is weighted by his or her compliance, Γ. This is a rather unsatifying estimand, as we are typically interested in estimating δ, the ATE, rather than a weighted

ATE, particularly where the weights are some unobservable and instrument-specific set of Γ's (Heckman and Robb 1985a, 1986; Heckman, Urzua, and Vytlacil 2006).

There are two different solutions to this problem that yield a well-defined estimand. First, we can simply assume:

Assumption (vii a): No person-specific compliance-effect covariance: $\mathrm{Cov}(\Gamma, \Delta) = 0$,

in which case equation (3) identifies the population ATE as $\delta = \beta/\gamma$. However, this assumption may be implausibly strong in some applications. The assumption says literally that the person-specific impact of M on Y is uncorrelated with that person's inclination to comply. However, if persons have some knowledge of how well they will respond to M, they may select a level of compliance accordingly. For example, a person who correctly expects Δ to be large will be motivated to seek a higher value of M; if assignment to treatment facilitates access to a higher value of M, such a person will comply more than will a person who correctly expects Δ to be zero.

In the case where both T and M are binary, we can adopt an alternative assumption that may be more tenable than (vii.a). In this case, Angrist et al. (1996) note that Γ can take on only three possible values: $\Gamma = 1$ for those for whom the instrument T determines their mediator value ("compliers"); $\Gamma = 0$ for those for whom the instrument does not affect the mediator ("always-takers" and "never-takers"); or $\Gamma = -1$ for whom the instrument causes them to experience the opposite of the intended mediator condition ("defiers"). They then assume that there are no "defiers" in the population:

Assumption (vii.b): No defiers (or "montonicity"): $\Gamma \in \{0, 1\}$.

Under this assumption, we can simplify the expression for the CWATE in equation (3) to

$$\begin{aligned}
\frac{\beta}{\gamma} &= \frac{\Pr(\Gamma = 1) \cdot E[\Gamma \cdot \Delta | \Gamma = 1] + \Pr(\Gamma = 0) \cdot E[\Gamma \cdot \Delta | \Gamma = 0]}{\Pr(\Gamma = 1) \cdot E[\Gamma | \Gamma = 1] + \Pr(\Gamma = 0) \cdot E[\Gamma | \Gamma = 0]} \\
&= \frac{\Pr(\Gamma = 1) \cdot E[\Delta | \Gamma = 1] + \Pr(\Gamma = 0) \cdot 0}{\Pr(\Gamma = 1) \cdot 1 + \Pr(\Gamma = 0) \cdot 0}, \\
&= E(\Delta | \Gamma = 1) \\
&\equiv \delta_c.
\end{aligned} \tag{4}$$

where $\Pr(\Gamma = 1)$ is the proportion of compliers in the population. Angrist et al. (1996) termed δ_c the LATE, also known as the ATE among the compliers, the complier average treatment effect (CATE) or the complier average causal effect. Equation (4) shows that the LATE is a special case of the CWATE when both T and M are binary and the no defiers assumption holds.[2]

Summary of Single-Site, Single-Mediator IV Assumptions

Approaching the IV model from a potential outcomes framework is particularly useful when we allow mediator effects to be heterogeneous. After imposing Assumptions (i)–(vi); SUTVA, exclusion restriction, linearity, instrument effectiveness, and ignorable treatment assignment), this framework reveals the importance of either (vii.a), the no-compliance-effect-covariance assumption, or (vii.b) the no-defiers assumption. If both of these assumptions fail, the IV estimand is a CWATE: those persons whose mediator is most affected by the instrument will be assigned the greatest weight in the estimand.

The IV Model with Multiple Sites and Multiple Mediators

In the single-site, single-mediator case, our challenge was to derive assumptions that define the ATE (δ) or the LATE (δ_c) as a function of the average ITT effect β and the average compliance γ. We now consider the multisite, multiple-mediator case, where subjects within a multisite trial are exposed to a treatment T, which may influence Y through P distinct mediators $M_1, M_2, \ldots, M_P$. We derive a set nine assumptions required to identify the effects of these mediators. The key insight that enables us to identify these effects is that site-specific values of β become outcomes in a regression where multiple site-specific compliances are predictors.

Six of our assumptions are straightforward extensions of the assumptions derived above in the single-site case, single-mediator case. These include SUTVA, the exclusion restriction, the two linearity assumptions, the assumption of ignorable assignment to T, and either a no compliance-effect covariance assumption (to identify ATE) or a "no defiers" assumption in the binary treatment, binary mediator case (to identify LATE). The assumption of nonzero average compliance that was needed in the single-site case is generalized to the assumption that there exists a full column rank site-by-compliance matrix, literally a design matrix within a multiple regression framework. Standard requirements of regression then generate two additional assumptions: an assumption that one mediator does not affect another

and an assumption of independence among the site-level compliances and site-level causal effects. These assumptions are described below.

We first assume that both SUTVA assumptions hold (i.a and i.b) with respect to the vector of P mediators:

Assumption (i): SUTVA:

(i.a) Each unit i has one and only one potential value of the vector of mediators $\mathbf{m}_i = \{m_{1i}, m_{2i}, \ldots, m_{Pi}\}$ for each treatment condition t: in particular, for a population of size N, $\mathbf{m}_i(t_1, t_2, \ldots, t_N) = \mathbf{m}_i(t_i)$ for all $i \in \{1, 2, \ldots, N\}$.

(i.b) Each unit i has one and only one potential outcome value of Y for each treatment condition t and each vector of mediator values $\mathbf{m}_i$: in particular, for a population of size N, $y_i(t_1, t_2, \ldots, t_N, \mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_N) = y_i(t_i, \mathbf{m}_i)$ for all $i \in \{1, 2, \ldots, N\}$.

We next assume that assignment to T influences Y only through the list of P distinct and observable mediators $M_1, M_2, \ldots, M_P$. Specifically, each participant has potential mediator values $m_1(t), m_2(t), \ldots, m_P(t)$ for $t \in \mathrm{T}$. The exclusion restriction now requires that T affects Y only through its effects on one or more of the mediators. That is:

Assumption (ii): Exclusion restriction: The treatment T affects Y only through its impact on the set of P mediators, $\boldsymbol{M} = \{M_1, M_2, \ldots, M_P\}$. That is, $Y(t) = Y(t, \mathbf{m}(t)) = Y(\mathbf{m}(t))$.

As above, we also assume person-specific linearity of each M in T (iii) and person-specific linearity of Y in each of the mediators (iv). Specifically, we assume that the outcome Y is a linear function of the mediators and that there are no interactions among the mediators.

Assumption (iii): Person-specific linearity of each mediator in T: the person-specific effect of T on each mediator M_p is linear. That is, $m_p(t) = m_p(0) + t\Gamma_p$ for each p.

Assumption (iv): Person-specific linearity of Y in $\mathbf{M}$: the person-specific effect of each mediator M_p on Y is linear. That is, $Y(\mathbf{m}) = Y(\mathbf{m} = 0) + \sum_{p=1}^{P} m_p \Delta_p$.

These imply, respectively, that the person-specific causal effect of T on M_p is $\Gamma_p = m_p(t) - m_p(t-1)$, and that that the person-specific causal effect

of M_p on Y is $\Delta_p = y(m_p) - y(m_p - 1)$, for all $p \in 1, 2, \ldots, P$. As above, the person-specific causal effect of T on Y is $B = y(t) - y(t-1)$. The observed outcome is $y(t) = y(0) + tB$.

We next assume that assignment to T does not influence a given mediator M_p through any other mediator M_q. That is, the mediators do not influence one another. This is required so that the estimation of the effects of a given mediator M_q on Y are not confounded with the effects of another mediator M_p.

Assumption (v): Parallel mediators:

$$m_p(t, m_1, \ldots, m_{p-1}, m_{p+1}, \ldots, m_P) = m_p(t) \text{ for all } p \in 1, 2, \ldots, P.$$

Together, the five assumptions above define the person-specific ITT effect as:

$$\begin{aligned} \mathrm{B} &= y(t) - y(t-1) \\ &= y(m_1(t), m_2(t), \ldots, m_P(t)) - y(m_1(t-1), m_2(t-1), \ldots, m_P(t-1)) \\ &= \sum_1^P \Delta_p \Gamma_p. \end{aligned} \tag{5}$$

Equation (5) says that the person-specific effect of T on Y can be written as the sum of the products of the person-specific effects of T on each mediator and the person-specific effects of that mediator on the Y (we discuss the implications of a failure of the parallel mediator assumption in the Discussion section below). Taking the expectation of equation (5) over the population within a site s yields:

$$E(B|S = s) = \beta_s = E\left[\sum_1^P \Delta_p \Gamma_p | S = s\right]. \tag{6}$$

As in the single-site case, we shall need unbiased estimates of the average compliances and ITT effects within each site. Letting K denote the number of sites, we invoke:

Assumption (vi): Ignorable within-site treatment assignment: The assignment of the instrument T must be independent of the potential outcomes within each site: $T \perp Y(t)|s, T \perp \mathbf{m}(t)|s, \forall t \in \mathrm{T}, s \in \{1, \ldots, K\}$.

As in the single-site case, it will next be useful to make either a set of no-compliance-effect covariance assumptions, analogous to (vii.a), or a set of "no defiers" assumptions, analogous to (vii.b). The assumptions made here determine whether the model identifies the ATE or the CATE.

First, if we wish to identify the ATEs of the mediators, we may make the assumption that there is no within-site covariance between Δ_p and Γ_p for each mediator p:

Assumption (vii.a): No within-site compliance-effect covariance:

$$Cov_s(\Gamma_p, \Delta_p) = \left[Cov(\Gamma_p, \Delta_p)|S = s\right] = 0, \text{ for all } p \text{ and } s.$$

Alternatively, in the case where T and each of the mediators M_1, M_2, ..., M_p are binary and we wish to identify LATE, we invoke:

Assumption (vii.b): No defiers (or "montonicity"): $\Gamma_p \in \{0, 1\}$ for all p.

Either of these two assumptions, in combination with Assumptions (i–vi) generates a multiple regression equation in which an estimable site-average ITT effect β_s is the outcome and estimable site-average compliances γ_{ps}, $p = 1, 2, \ldots, P$ are predictors. To see this, consider first the case of ATE where we invoke Assumption (vii.a). Under this assumption, equation (6) is:

$$\begin{aligned}
\beta_s &= E\left[\sum_1^P \Delta_p \Gamma_p | S = s\right] \\
&= \sum_1^P \delta_{ps}\gamma_{ps} + \sum_1^P Cov_s(\Delta_p, \Gamma_p) \\
&= \sum_1^P \delta_{ps}\gamma_{ps} \\
&= \sum_1^P \delta_p\gamma_{ps} + \sum_1^P (\delta_{ps} - \delta_p)\gamma_{ps} \\
&= \sum_1^P \delta_p\gamma_{ps} + \omega_s,
\end{aligned} \tag{7}$$

where δ_{ps} and γ_{ps} are the average effect of M_p on Y in site s and the average effect of T on M_p in site s, respectively; where δ_p is the average, across sites, of the δ_{ps}'s; and where the error term is $\omega_s = \sum_1^P (\delta_{ps} - \delta_p)\gamma_{ps}$.

If, in contrast, we have a binary M and seek to estimate LATE, we invoke Assumption (vii.b), generating a multiple regression equation of exactly the same form. Specifically, we can write equation (6) as:

$$\begin{aligned}\beta_s &= E\left[\sum_1^P \Delta_p \Gamma_p | S = s\right] \\ &= E\left[\sum_1^P \left(\Delta_p | \Gamma_p = 1\right) \cdot Pr\left(\Gamma_p = 1\right) | S = s\right] \\ &= \sum_1^P E\left[\Delta_p | \Gamma_p = 1, S = s\right] \cdot \gamma_{ps} \\ &= \sum_1^P \delta_{cps}\gamma_{ps} \\ &= \sum_{p=1}^P \delta_{cp}\gamma_{ps} + \sum_{p=1}^P \left(\delta_{cps} - \delta_{cp}\right)\gamma_{ps} \\ &= \sum_{p=1}^P \delta_{cp}\gamma_{ps} + \omega_{cs},\end{aligned} \tag{8}$$

where δ_{cps} is the complier average effect of M_p on Y in site s (the LATE for mediator p in site s); δ_{cp} is the complier average effect of M_p on Y in the population; γ_{ps} is the average effect of T on M_p in site s (which, under the no-defiers assumption, is equal to the proportion of the population in site s who are compliers with respect to mediator p); and ω_{cs} is an error term equal to $\sum_{p=1}^P \left(\delta_{cps} - \delta_{cp}\right)\gamma_{ps}$.

Equations (7) and (8) use the same outcome β_s and the same predictors $\gamma_{ps}, p = 1, 2, \ldots, P$. However, invoking the no-covariance assumption identifies the coefficients of this model as the ATEs $\delta_{ps}, p = 1, 2, \ldots, P$ with random error ω_s in equation (7), while invoking the no-defiers assumption identifies the coefficients of this model as $\delta_{cps}, p = 1, 2, \ldots, P$ and the errors as ω_{cs} in equation (8). To identify either of these models thus requires additional standard assumptions for regression, namely that the design matrix be of full rank and that the model errors be ignorable. Thus, in either case, we assume:

Assumption (viii): Site-by-mediator compliance matrix has sufficient rank. In particular, if $\boldsymbol{G}$ is the $K \times P$ matrix of the $\gamma_{ps}'s$, we require $rank(\boldsymbol{G}) = P$. This implies three specific conditions:

(viii.a) The compliance of at least $P - 1$of the mediators varies across sites. That is, $\mathrm{Var}(\gamma_{ps}) = 0$ for at most one $p \in \{1, 2, \ldots, P\}$.

(viii.b) There are at least as many sites as mediators: $K \geq P$.

(viii.c) There is some subset of Q site-specific compliance vectors, $\gamma_s = \{\gamma_{1s}, \gamma_{2s}, \ldots, \gamma_{Ps}\}$, where $K \geq Q \geq P$, that are linearly independent.

The sufficient rank assumption is a generalization of the familiar instrument effectiveness assumption (Assumption [vi] in the first section). Note that when there is a single mediator ($P = 1$), the site-by-mediator compliance matrix will have rank 1, so long as $\gamma_{1s} \neq 0$ for at least one site s (the average compliance across sites may be zero, as long as it is not zero in every site). Thus, when there is a single site and a single mediator, the sufficient rank assumption is identical to the usual condition that the treatment has a nonzero average impact on the mediator.

Our final assumption requires that the error term ω_s of equation (7) or ω_{cs} of equation (8) be ignorable. In order to identify the ATEs, we assume:

Assumption (ix.a): Between-site compliance-effect independence: The site-average compliance of each mediator is independent of the site-average effect of each mediator. That is, $E[\delta_{qs}|\gamma_{1s}, \gamma_{2s}, \ldots, \gamma_{Ps}] = E[\delta_{qs}] = \delta_q$ for all $q \in 1, \ldots, P$.

Likewise, to identify the LATEs, we assume:

Assumption (ix.b): Between-site compliance-effect independence: The site-average compliance of each mediator is independent of the site complier average effect of each mediator. That is, $E[\delta_{cqs}|\gamma_{1s}, \gamma_{2s}, \ldots, \gamma_{Ps}] = E[\delta_{cqs}] = \delta_{cq}$.

Under Assumption (ix.a), we can write the expected value of the error ω_s in equation (7) as:

$$\begin{aligned} E[\omega_s|\gamma_{1s}, \gamma_{2s}, \ldots, \gamma_{Ps}] &= E\left[\sum_{q=1}^{P}(\delta_{qs} - \delta_q)\gamma_{qs}|\gamma_{1s}, \gamma_{2s}, \ldots, \gamma_{Ps}\right] \\ &= \sum_{q=1}^{P}\gamma_{qs} \cdot E[(\delta_{qs} - \delta_q)|\gamma_{1s}, \gamma_{2s}, \ldots, \gamma_{Ps}] \\ &= \sum_{q=1}^{P}\gamma_{qs} \cdot E[(\delta_{qs} - \delta_q)] \\ &= 0. \end{aligned} \tag{9}$$

By the same logic, Assumption (ix.b) implies that the expected value of the error term ω_{cs} in equation (8) is zero.

Note that Assumptions (ix.a) and (ix.b) are each stronger than an assumption of no between-site compliance-effect covariance (the latter requires only no linear association between compliance and effect; the former requires no association whatsoever). Moreover, note that Assumptions (ix.a) and (ix.b) require not only that there be no compliance-effect association for a given mediator but also that there be no cross-mediator compliance-effect association. That is, the site-average effect of T on a given mediator M_q must be statistically independent of the site-average effect of any mediator M_p on Y.

Discussion

Summary of Multiple-Site, Multiple-Mediator IV Assumptions

To summarize, in the case of a multisite study in which a treatment T may affect the outcome Y through multiple mediators, we require a number of assumptions in order to identify the average causal effects of the mediators using MSMM-IV methods. In order to identify the ATE in the population, the relevant assumptions are as follows:

(i) SUTVAs;
(ii) Exclusion restriction;
(iii) Person-specific linearity of the mediators with respect to the treatment;
(iv) Person-specific linearity of the outcome with respect to the mediators;
(v) Parallel mediators;
(vi) Within-site ignorable treatment assignment;
(vii.a) Zero within-site compliance-effect covariance for each mediator;
(viii) Compliance matrix has sufficient rank;
(ix.a) Between-site cross-mediator compliance-effect independence.

In order to identify the CATE in the case of a binary treatment and binary mediators, Assumption (vii.a) is replaced by Assumption (vii.b), no defiers for any mediator; and Assumption (ix.a) is replace by (ix.b), between-site independence of the compliance and complier average effects.

Note that six of these assumptions—SUTVA, the exclusion restriction, the two linearity assumptions, ignorable treatment assignment, and either the zero within-site compliance-effect covariance assumption or the no

defiers assumption—are identical to those required for the single-site, single-instrument, single-mediator case (though often the two linearity assumptions are ignored because they are met trivially when the instrument and mediators are binary). Assumptions (v), (viii), and (ix) are specific to the multiple-site, multiple-mediator case (though the sufficient rank Assumption [viii] is equivalent to the instrument effectiveness assumption when there is a single site and single mediator, as we note above). We discuss these three assumptions in more detail below.

The Parallel Mediators Assumption

The assumption that the mediators impact an outcome in parallel is a nontrivial assumption (see Appendix A, which can be found at http://smr.sagepub.com/supplemental/, for a detailed discussion). Consider the Duncan et al. (2011) study described above. In this study, 16 implementations of random

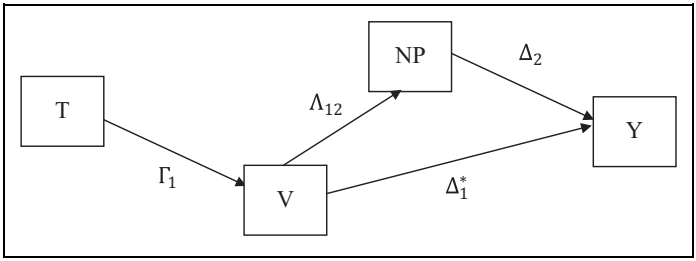

Figure 2. Hypothesized treatment and mediator effects in the MTO study.

assignment welfare-to-work experiments were used to estimate the impact of three hypothesized mediators of the programs: income, hours worked, and welfare receipt. The MSMM-IV models used assume that none of these mediators affects the others. However, this is an implausible assumption, given that both hours worked and welfare receipt are clearly linked to income.

The MTO study analyzed in Kling et al. (2007) provides an opportunity to consider the parallel mediators assumption in concrete terms. In this study, random assignment to a voucher was hypothesized to affect outcomes via two potential mediators—use of the voucher and *NP*. Because *NP* could not be influenced except through use of the voucher, the implied structural model is that shown in Figure 2.

In this model, treatment assignment affects *NP* only through use of a voucher (*V*). Both *NP* and *V* may then affect an outcome *Y*. As detailed in

Appendix A, which can be found at http://smr.sagepub.com/supplemental/, identification of $\delta_2 = E[\Delta_2]$ requires two key sets of additional assumptions. First, within each MTO site s, both a family's likelihood of using the voucher if offered it and the change in *NP* experienced by a family if they use the voucher are uncorrelated with the effect of *NP* on that family. Families for whom a move to low-poverty neighborhoods would be particularly beneficial are no more likely to use the voucher and move to low-poverty neighborhoods than are families for whom such a move would be less beneficial. Second, across MTO sites, there are no correlations between (a) the average impact of *NP* and average voucher take-up rate; (b) the average impact of *NP* and the average impact of voucher use on *NP* rates; (c) the average impact of using of a voucher and the average voucher take-up rate; or (d) the average impact of using of a voucher and the average impact of voucher use on *NP* rates. If, for example, sites where the use of a voucher had a large impact on *NP* (because it was relatively easy for families to move far from their original neighborhood) were also sites where use of a voucher moved families far from family and friendship networks that have a positive effect on outcomes, then the assumption of the independence of the direct effect of the voucher (through network supports in this example) and the effect of one mediator on another would be violated. Note that, in the MTO example, it would be possible to identify the total effect of the first mediator (use of the voucher), because there is no pathway from T to Y that does not go through V. Identifying the effect of *NP* and the direct effect of V on Y, however, requires additional assumptions about the independence of these effects and the effect of V on *NP*. Given the correlation of *NP* and other factors likely to influence the outcomes of interest in the MTO study, such assumptions may not be warranted.

The Site-average Compliance-Effect Independence Assumption

The assumption that the site-average compliances are independent of the site-average effects is nontrivial. Because site-average compliance effects are not randomly assigned to sites, they may not be independent of the site-average mediator effects. Consider a simple example. Suppose we have a multisite study of the impacts of welfare-to-work programs, as in Duncan et al. (2011), where the programs are hypothesized to affect child outcomes by affecting mothers' hours worked, income, and welfare receipt. Suppose that entry-level wages and the cost of living are higher in some sites than others. In this case, randomized assignment to a training program may induce a greater increase in hours worked and income (higher compliance) in high-wage sites than in low-wage sites (because the wage benefits of

work are greater); however, the effect of increased income on child achievement may be lower in high-wage sites than in low-wage sites, because the cost of child care, preschool, and school quality is higher. Such a pattern would induce a negative correlation between the work and income effects of the program and the effects of income on children, violating the assumption of site-average compliance-effect independence.

Although the compliance-effect independence assumption is not empirically verifiable, it may be falsifiable, given sufficient data. Equation (9) implies that, in a multisite study with P mediators and in which each of the nine assumptions is met, a plot in $(P + 1)$ space of the site-average ITT effects (the β_s's) against the P site-average compliance effects (the γ_{ps}'s) will display a pattern of points scattered (with heteroscedastic variance) around a hyperplane passing through the origin with partial slopes $\frac{\partial \beta}{\partial \gamma_p} = \delta_p$, for all p. A violation of the site-average compliance-effect independence assumption, however, implies that $E(\omega_s|\gamma_1, \ldots, \gamma_P) \neq 0$ for some value(s) of $\gamma_1, \ldots, \gamma_P$. As a result, the surface described by $E(\beta_s|\gamma_1, \ldots, \gamma_P)$ will be nonlinear. With sufficient data (a sufficient number of sites and sufficiently precise estimation of the β_s's and γ_s's for each site), one might have adequate statistical power to reliably detect such nonlinearity, allowing one to reject the compliance-effect independence assumption.

In Online Appendices B and C, which can be found at http://smr.sagepub.com/supplemental/, we derive expressions for the bias in the two-stage least squares MSMM-IV estimator when the site-average compliance-effect independence assumption fails.

The Sufficient Rank Assumption

The sufficient rank assumption is relatively straightforward. In order to identify the effects of P mediators using an MSMM-IV model, we require at least as many sites as mediators; we require that the effect of treatment assignment on the mediators varies across sites (for at least $P - 1$ of the mediators); and we require that there are at least P sites among which these effects are linearly independent. In many practical applications, these assumptions are likely to be met. The average effect of treatment assignment on a mediator is likely to vary across sites for a variety of reasons, including differential implementation, heterogeneity of populations, and differences among sites in baseline conditions or capacity. Moreover, unless the mediators are conceptually very similar, the effects of treatment assignment on the mediators are unlikely to be perfectly collinear.

Nonetheless, in practical applications, the effects of treatment assignment on the mediators are likely to be somewhat correlated (though not perfectly) across sites. This may occur because in sites where a treatment is well implemented, the treatment may affect all mediators more than in sites where it is poorly implemented. Or it may occur because the mediators are correlated in the world, leading to a correlation of compliances. For example, because income is correlated with hours worked, sites in which a treatment—such as a welfare-to-work experiment—induces large changes in hours worked will tend to also be sites in which the same treatment induces large changes in income.

Although such correlations among the γ_s's do not pose an identification problem for the MSMM-IV model (we require no assumption regarding the independence of the site-average compliances), they may pose a problem for estimation. Because the identification of the effects of the mediators depends on the separability of the site-average compliances, statistical power will be greatest—all else being equal—when compliances are not positively correlated.

Conclusion

If each of the nine assumptions described above is met, the effects of each mediator are, in principle, identifiable from observed data. Such models provide a possible approach to estimating the effects of the mediators of treatment effects when such mediators cannot themselves be easily assigned at random. The assumptions necessary for consistent identification in MSMM-IV models are not, however, trivial. In addition to the usual IV assumptions, such models require several assumptions. The parallel mediator and site-average compliance-effect independence assumptions, in particular, are relatively strong, and cannot be empirically verified (though with large samples the compliance-effect independence assumption may be falsifiable). Justification of such models must rely, therefore, on sufficiently strong theory or prior evidence to warrant these assumptions.

Although we have framed our discussion in the context of a multisite randomized trial, where "sites" are specific locations (different cities in the MTO example, different studies and cities in the welfare-to-work example), the same logic would apply to any study in which randomization occurs within identifiable subgroups of individuals. Thus, one could stratify the sample of a large randomized trial by sex, age, and race, and treat each sex-by-age-by-race cell as a "site" in order to create multiple "site"-by-treatment interactions as instruments. This would, in principle, allow one to identify the effects of multiple mediators within a single (large) randomized trial, but only under the set of assumptions we describe above.

Alternately, one could estimate a set of propensity scores, indicating each individual's "propensity to comply" with each mediator, and then stratify the sample by vectors of these propensity scores. Using such strata as "sites" in an MSMM-IV model would have two advantages: It would ensure there is no or little within-site compliance-effect covariance (because compliance would be near constant within compliance strata); and it may allow one to create strata among which the site-average compliances are uncorrelated, which may increase the precision of the estimates. Estimating "propensity to comply," however, is itself a nontrivial enterprise, relying on an additional set of rather strong assumptions (which we do not address here).

Several important issues remain to be addressed in order to fully understand the use of MSMM-IV models. First, although failure of the assumptions will lead to inconsistent estimates, it is not clear how severe the bias resulting from plausible failures of the parallel mediators and compliance-effect independence assumptions will be. Second, we have not discussed the properties of specific estimators of MSMM-IV models or the computation of standard errors from such models. Both issues merit further investigation.

Finally, although the nine assumptions we outline above ensure the consistent estimation of the effects of multiple mediators, they do not ensure unbiased estimation in finite samples. In single-site single-mediator IV models, finite sample bias is a concern when the average compliance is small relative to its sampling variance. In multiple-site, multiple-mediator models, finite sample bias is more complex. In general, however, finite sample bias is likely to be a concern when both the average compliance (across sites) is small and the variance of the site-average compliances is small, relative to the sampling variation of the site-average compliances. A full discussion of finite sample bias is beyond the scope of this article, however.

Authors' Note

An earlier version of this article was presented at the Annual Meeting of the Society for Research on Educational Effectiveness, Washington, DC, March 2011. All errors are our own.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a grant from the Institute for Education Sciences (R305D090009), and benefited enormously from lengthy conversations with Howard Bloom, Fatih Unlu, Pei Zhu, and Pamela Morris.

Notes

1. The five cities generate 10 site-by-treatment interactions as instruments because there were three (randomly assigned) treatment conditions per site.
2. In some settings (e.g., Little and Yau 1998), participants assigned to the control cannot gain access to the mediator, that is $\Pr(m(0) = 1) = 0$. In this case, there are no "always-takers." We then see that local average treatment effect becomes the "treatment effect on the treated," that is, $\delta_C = E(\Delta|\Gamma = 1) = E(\Delta|m = 1) \equiv \delta_{TOT}$.

References

Angrist, J. D., G. W. Imbens, and D. B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91: 444-55.

Duncan, G. J., P. A. Morris, and C. Rodrigues. 2011. "Does Money Really Matter? Estimating Impacts of Family Income on Young Children's Achievement with Data from Random-assignment Experiments." *Developmental Psychology* 47: 1263-79.

Heckman, J. J. and R. Robb. 1985a. "Alternative Methods for Evaluating the Impact of Interventions." Pp. 156-245 in *Longitudinal Analysis of Labor Market Data*, edited by J. J. Heckman and B. Singer, Vol. 10. New York: Cambridge University Press.

Heckman, J. J. and R. Robb. 1985b. "Using Longitudinal Data to Estimate Age, Period and Cohort Effects in Earnings Equations." Pp. 137-150 in *Cohort Analysis in Social Research beyond the Identification Problem*, edited by W. M. Mason and S. E. Feinberg. New York: Springer-Verlag.

Heckman, J. J. and R. Robb. 1986. "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes." Pp. 63-107 in *Drawing Inferences from Self-selected Samples*, edited by H. Wainer. New York: Springer-Verlag.

Heckman, J. J., S. Urzua, and E. Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Review of Economics and Statistics* 88:389-432.

Imbens, G. W. and J. D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62:467-75.

Kling, J. R., J. B. Liebman, and L. F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75:83-119.

Little, R. J. and L. H. Y. Yau. 1998. "Statistical Techniques for Analyzing Data from Prevention Trials: Treatment of No-shows using Rubin's Causal Model." *Psychological Methods* 3:147-59.

Rubin, D. B. 1986. "Comment: Which Ifs Have Causal Answers." *Journal of the American Statistical Association* 81:961-62.

Spybrook, J. 2008. "Are Power Analyses Reported with Adequate Detail: Findings from the First Wave of Group Randomized Trials Funded by the Institute of Education Sciences." *Journal of Research on Educational Effectiveness* 1:15-235.

Author Biographies

Sean F. Reardon is Professor of Education and (by courtesy) Sociology at Stanford University.

Stephen W. Raudenbush is the Lewis-Sebring Distinguished Service Professor in the Department of Sociology and the Harris School of Public Policy Studies, and Chairman of the Committee on Education, at the University of Chicago.