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

Sean F. Reardon

520 Galvez Mall, #526
Stanford University
Stanford, CA 94305-3084
(650)-736-8517
sean.reardon@stanford.edu

Stephen W. Raudenbush

1126 East 59th Street
University of Chicago
Chicago IL 60637
(773)-834-1904
sraudenb@uchicago,edu

forthcoming in *Sociological Methods and Research*

This version: December, 2011

Direct correspondence to Sean F. Reardon (sean.reardon@stanford.edu). This work was supported by a grant from the Institute for Education Sciences (R305D090009), and benefitted enormously from lengthy conversations with Howard Bloom, Fatih Unlu, Pei Zhu, and Pamela Morris. An earlier version of this paper was presented at the Annual Meeting of the Society for Research on Educational Effectiveness, Washington, DC, March, 2011. All errors are our own.

ABSTRACT

The increasing availability of data from multi-site randomized trials provides a potential opportunity to use instrumental variables 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 non-trivial and cannot be empirically verified, suggesting that multiple-site, multiple-mediator instrumental variables models must be justified by strong theory.

1. 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 & Robb, 1985b; Imbens & 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 on health, social behavior, education, and economic self-sufficiency of adolescents and adults. Reasoning that the instrument might affect outcomes through mechanisms other than neighborhood poverty, 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 ten[1] instruments ("site-by-randomization interactions") to identify the impact of the two

[1] The five sites generate ten site-by-treatment interactions as instruments because there were three (randomly assigned) treatment conditions per site.

mediators of interest, neighborhood poverty and experimental compliance. Using a similar strategy, Duncan, Morris, and Rodrigues (2011) use data from sixteen 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 US Institute of Education Sciences over the past decade, the majority were multi-site 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 strategy (hereafter MSMM-IV) is to identify the average causal 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 of where a single instrument (which we denote as T) operates through a set of mediators $\mathbf{M} = \{M_1, M_2, \dots, M_P\}$, 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 additional 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 multi-valued.

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 average treatment effects of the mediators, three of which are specific to the MSMM-IV case, and which we discuss in some detail.

2. THE SINGLE-SITE, SINGLE-MEDIATOR CASE

Notation

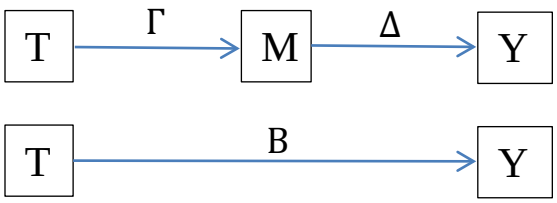
Suppose that each participant in a single-site study is exposed to a treatment T taking on values in the domain $\mathbb{T} \subset \mathbb{R}$. We hypothesize that T may affect some outcome Y through its effect on some mediator M. Thus, in our notation, T is an instrument that will be used to identify 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 multi-valued or continuous.

Note that our terminology and notation differ here from those in standard econometric discussions of instrumental variables. 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 effect(s) of the mediator(s) through which T may operate are of interest. Throughout the remainder of this paper, 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 a the population (often referred to as the "intent-to-treat" effect in the program evaluation literature, or the "reduced form" effect in the econometrics literature) is therefore $\beta = E[\mathrm{B}]$.

Figure 1:



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): <u>*Stable unit treatment value assumptions*</u> (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, $m_i(t_1, t_2, \dots, t_N) = m_i(t_i)$ for all $i \in \{1,2, \dots, N\}$.

(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, $y_i(t_1, t_2, \dots, t_N, m_1, m_2, \dots, m_N) = y_i(t_i, m_i)$ for all $i \in \{1,2, \dots, N\}$.

Given the SUTVA assumptions, we can represent the potential outcome Y for a participant who experiences treatment t and mediator value $m(t)$ as $y\big(t, m(t)\big)$ (we drop the subscript i throughout the remainder of this paper 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): <u>*Exclusion restriction*</u>:

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

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, $y_i(m_1, m_2, \dots, m_N) = y_i(m_i)$ for all $i \in \{1,2, \dots, N\}$.

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): <u>*Person-specific linearity of the mediator M in T*</u>: 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, $y\big(m(t)\big) = y(m = 0) + m(t)\Delta$.

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

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

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

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

The population average intent-to-treat effect (ITT) of interest here is $E(\mathrm{B}) = \beta$. 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 instrumental variables methods rely on the instrument to induce some exogenous variation in the mediator (for at least some individuals), we require γ to be non-zero:

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 instrumental variables estimator is the ratio of the intent-to-treat effect to the average compliance:

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

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 average treatment effect, rather than a weighted average treatment effect, particularly where the weights are some unobservable and instrument-specific set of $\Gamma's$ (Heckman & Robb, 1985a, 1986; Heckman, Urzua, & 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: $Cov(\Gamma, \Delta) = 0$,

in which case (3) identifies the population average treatment effect (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, Imbens and Rubin (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 those who experience the opposite of the intended mediator condition ("defiers"). They then assume that there are no "defiers" in the population—no one for whom exposure to the instrument T causes them to switch from $M = 1$ to $M = 0$:

Assumption (vii.b): <u>*No defiers*</u>: $\Gamma \in \{0,1\}$.

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, Imbens, and

Rubin (1996) termed δ_c the "local average treatment effect" (LATE), also known as the average treatment effect on the compliers, the complier average treatment effect (CATE) or the complier average causal effect (CACE). 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 instrumental variable 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 instrumental variable estimand is a compliance weighted average treatment effect (CWATE): those persons whose mediator is most affected by the instrument will be assigned the greatest weight in the estimand.

3. 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 intent-to-treat effect β and the average compliance γ. We now consider the multi-site, multiple mediator case, where subjects within a multi-site trial are exposed to a treatment T, which may influence Y

[2] In some settings (e.g., Little & 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 LATE becomes the "treatment effect on the treated" (TOT), that is $\delta_C = E(\Delta|\Gamma = 1) = E(\Delta|m = 1) \equiv \delta_{TOT}$.

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 non-zero 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): <u>Stable unit treatment value assumptions (SUTVA)</u>:

(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, $y_i(t_1, t_2, \dots, t_N, \mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_N) = y_i(t_i, \mathbf{m}_i)$ for all $i \in \{1,2, \dots, N\}$.

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

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

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): <u>Person-specific linearity of each mediator in T</u>: 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): <u>Person-specific linearity of Y in $\boldsymbol{M}$</u>: The person-specific effect of each mediator M_p on Y is linear. That is, $Y(\mathbf{m}) = Y(\mathbf{m} = \mathbf{0}) + \sum_{p=1}^{P} m_p \Delta_p$.

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

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): <u>Parallel mediators</u>:

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

Together, the five assumptions above define the person-specific intent-to-treat effect as

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

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 Section IV below). Taking the expectation of (5) over the population within a site *s* yields

$$E(\mathrm{B}|S=s)=\beta_s=E\left[\sum_{1}^{P}\Delta_p\Gamma_p\,\middle|\,S=s\right].$$

(6)

As in the single-site case, we shall need unbiased estimates of the average compliances and intent-to-treat 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: $T \perp Y(t)|s,\ T \perp \mathbf{m}(t)|s,\ \forall\, t \in \mathbb{T}, s \in \{1,\ldots,K\}$.

As in the single-site case, it will next be useful to make either a set of 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 average treatment effect (ATE) or the complier average treatment effect (LATE).

First, if we wish to identify the average treatment effects (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:

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

Alternatively, in the case where both T and M are binary and we wish to identify LATE, we invoke

Assumption (vii.b): <u>*No defiers*</u>: $\Gamma_p \in \{0,1\}$ for all p.

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

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

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 problem of exactly the same form. Specifically, we can write (6) as

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

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 $\sum_{p=1}^P (\delta_{cps} - \delta_{cp})\gamma_{ps}$.

Equations (7) and (8) use the same outcome β_s and the same predictors

$\gamma_{ps}, p = 1,2, \dots P$. However, invoking the no-covariance assumption identifies the coefficients of this model as the ATEs $\delta_{ps}, p = 1,2, \dots, P$ with random error ω_s in (7) , while invoking the no-defiers assumption identifies the coefficients of this model as $\delta_{cps}, p = 1,2, \dots, P$ and the errors as ω_{cs} in (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): <u>Site-by-mediator compliance matrix has sufficient rank</u>. In particular, if **G** is the K x P matrix of the $\gamma_{ps}'s$, we require $rank(\mathbf{G}) = P$. This implies three specific conditions:

(viii.a) The compliance of at least $P - 1$ of the mediators varies across sites. That is,

$Var(\gamma_{ps}) = 0$, for at most one $p \in \{1,2, \dots, P\}$.

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

(viii.c) There is some subset of Q site-specific compliance vectors,

$\boldsymbol{\gamma}_s = \{\gamma_{1s}, \gamma_{2s}, \dots, \gamma_{Ps}\}$, 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 non-zero average impact on the mediator.

Our final assumption requires that the error term ω_s of Equation (7) or ω_{cs} of (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 $E\left[\delta_{qs}|\gamma_{1s}, \gamma_{2s}, \dots, \gamma_{Ps}\right] = E\left[\delta_{qs}\right] = \delta_q$ for all $q \in 1, \dots, P$.

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 $E\left[\delta_{cqs}|\gamma_{1s}, \gamma_{2s}, \dots, \gamma_{Ps}\right] = E\left[\delta_{cqs}\right] = \delta_{cq}$.

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

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

By the same logic, Assumption (ix.b) implies that the expected value of the error term ω_{cs} in (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 cannot be correlated with the site average effect of any mediator M_p on Y.

4. DISCUSSION

Summary of Multiple-Site, Multiple-Mediator IV Assumptions

To summarize, in the case of a multi-site 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 MSSM-IV methods. In order to identify the average treatment effect in the population, the relevant assumptions are

(i) Stable unit treatment value assumptions

(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 complier average treatment effect (LATE) 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.

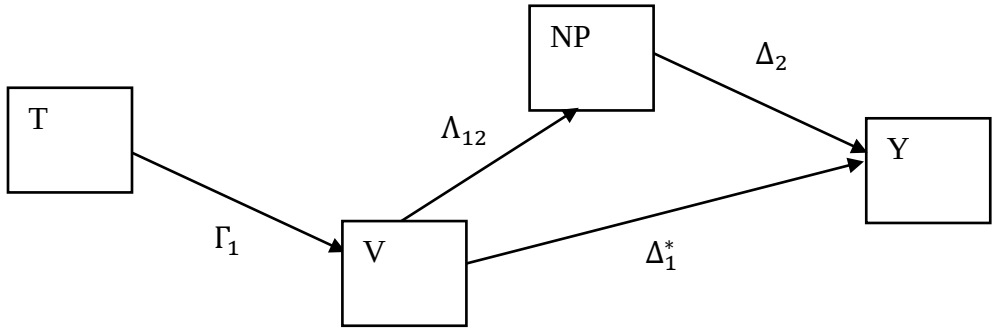
The Parallel Mediators Assumption

The assumption that the mediators impact an outcome in parallel is a non-trivial assumption (see Appendix A for a detailed discussion). Consider the Duncan, Morris, and Rodrigues (2011) study described above. In this study, sixteen implementations of random-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 multiple-site, multiple mediator 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, Liebman, and Katz (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 neighborhood poverty. Because neighborhood poverty could not be influenced except through use of the voucher, the implied structural model is that shown in Figure 2.

Figure 2:



In this model, treatment assignment affects neighborhood poverty (NP) only through use of a voucher (V). Both NP and V may then affect an outcome Y. As detailed in Appendix A, 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 neighborhood poverty experienced by a family if they use the voucher are uncorrelated with the effect of neighborhood poverty 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 neighborhood poverty and average voucher take-up rate; b) the average impact of neighborhood poverty and the average impact of voucher use on neighborhood poverty 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 neighborhood poverty rates. If, for example, sites where the use of a voucher had a large impact on neighborhood poverty (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 neighborhood poverty 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 non-trivial. 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 multi-site study of the impacts of welfare-to-work programs, as in Duncan, Morris, and Rodrigues (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 multi-site study with P mediators and in which each of the nine assumptions is met, a plot in $(P + 1)$-space of the site-average intent-to-treat effects (the β_s's) against the P site-average compliance effects (the γ_{ps}'s) will display a pattern of points scattered (with heteroskedastic 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 non-linearity, allowing one to reject the compliance-effect independence assumption.

In Appendices B and C, we derive expressions for the bias in the 2SLS 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.

5. 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 additional 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 multi-site 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 non-trivial 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.

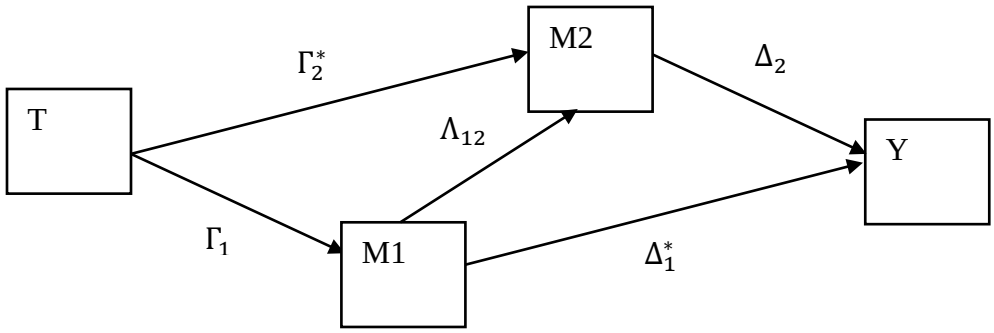
Finally, although the nine assumptions we outline above ensure the consistent estimation of the effects of multiple mediators, they do not ensure unbiased estimation in finite samples. In single-site single-mediator instrumental variables models, finite sample

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 paper, however.

APPENDIX A: RATIONALE FOR THE PARALLEL MEDIATORS ASSUMPTION

For illustration, consider a simple case in which a treatment T affects Y through two mediators, $M1$ and $M2$, one of which affects the other, as illustrated in Figure A1 below.

Figure A1:



Let Γ_1 and Γ_2 be the person-specific effects of T on $M1$ and $M2$, respectively. Note that

$$\Gamma_2 = \Gamma_2^* + \Gamma_1\Lambda_{12}. \quad \text{(A1)}$$

where Γ_2^* is the direct effect of T on $M2$ (the effect not mediated by $M1$), and Λ_{12} is the effect of $M1$ on $M2$. Likewise, let Δ_1 and Δ_2 be the effects of $M1$ and $M2$ on Y, respectively. Note that

$$\Delta_1 = \Delta_1^* + \Lambda_{12}\Delta_2, \quad \text{(A2)}$$

where Δ_1^* is the direct effect of $M1$ on Y (the effect not mediated by $M2$).

Now, the person-specific effect of T on Y is given by

$$\mathrm{B} = \Gamma_1\Delta_1^* + \Gamma_2\Delta_2. \quad \text{(A3)}$$

Typically, we want to estimate $\delta_1 = E[\Delta_1]$ and $\delta_2 = E[\Delta_2]$. Given a multi-site trial, within each site s, we have

$$\beta_s = E[B|s] = E[\Gamma_1\Delta_1^*|s\,] + E[\Gamma_2\Delta_2|s\,]$$
$$= \delta_{1s}^*\gamma_{1s} + \delta_{2s}\gamma_{2s} + Cov_s(\Gamma_1, \Delta_1^*) + Cov_s(\Gamma_2, \Delta_2).$$

(A4)

Let us assume $Cov_s(\Gamma_1, \Delta_1^*) = 0$ and $Cov_s(\Gamma_2, \Delta_2) = 0$. The first of these says that the person-specific compliance of $M1$ is uncorrelated with the direct effect of $M1$ on Y. The second can be written as

$$\begin{aligned} Cov_s(\Gamma_2, \Delta_2) &= Cov_s(\Gamma_2^* + \Gamma_1\Lambda_{12}, \Delta_2) \\ &= Cov_s(\Gamma_2^*, \Delta_2) + Cov_s(\Gamma_1\Lambda_{12}, \Delta_2) \\ &= Cov_s(\Gamma_2^*, \Delta_2) + \gamma_{1s}Cov_s(\Lambda_{12}, \Delta_2) + \lambda_{12s}Cov_s(\Gamma_1, \Delta_2) \\ &\quad + E[(\Gamma_1 - \gamma_{1s})(\Lambda_{12} - \lambda_{12s})(\Delta_2 - \delta_{2s})|S = s] \\ &= 0. \end{aligned}$$

(A5)

This says that the person-specific effect of $M2$ cannot be correlated with any of the paths leading to it (and that the third centered moment of $\{\Gamma_1, \Lambda_{12}, \Delta_2\}$ must be zero, a condition that is met if the three terms are linearly related to one another and if each of them has a non-skew distribution). Thus, if the mediators are not parallel, then assumption (vii.a) must be expanded to include the assumption that, within sites, the direct effect of any mediator cannot be correlated with any upstream pathway leading from the treatment to that mediator.

Given this assumption, we have

$$\begin{aligned} \beta_s &= \delta_{1s}^*\gamma_{1s} + \delta_{2s}\gamma_{2s} \\ &= \delta_1^*\gamma_{1s} + \delta_2\gamma_{2s} + \omega_s, \end{aligned}$$

(A6)

where $\omega_s = (\delta_{1s}^* - \delta_1^*)\gamma_{1s} + (\delta_{2s} - \delta_2)\gamma_{2s}$. As above, we require the assumption that this error term be independent of γ_{1s} and γ_{2s}:

$$E[\omega_s|\gamma_{1s},\gamma_{2s}] = \gamma_{1s}\cdot E[(\delta_{1s}^* - \delta_1^*)|\gamma_{1s},\gamma_{2s}] + \gamma_{2s}\cdot E[(\delta_{2s} - \delta_2)|\gamma_{1s},\gamma_{2s}] = 0.$$

(A7)

A necessary, but not sufficient, condition for this to be true is that

$$Cov(\delta_{1s}^*, \gamma_{1s}) = 0;$$

$$Cov(\delta_{2s}, \gamma_{1s}) = 0;$$

$$Cov(\delta_{1s}^*, \gamma_{2s}) = 0;$$

$$Cov(\delta_{2s}, \gamma_{2s}) = 0.$$

(A8)

The first two of these expressions indicate that the site-average compliance of mediator 1 is uncorrelated with the site average direct effects of both mediators 1 and 2. The third and fourth expressions can be written as

$$\begin{aligned} Cov(\delta_{1s}^*, \gamma_{2s}) &= Cov(\delta_{1s}^*, \gamma_{2s}^* + \gamma_{1s}\lambda_{12s}) \\ &= Cov(\delta_{1s}^*, \gamma_{2s}^*) + \gamma_1 Cov(\delta_{1s}^*, \lambda_{12s}) + \lambda_{12} Cov(\delta_{1s}^*, \gamma_{1s}) \\ &+ E[(\delta_{1s}^* - \delta_1^*)(\gamma_{1s} - \gamma_1)(\lambda_{12s} - \lambda_{12})], \end{aligned}$$

and

$$\begin{aligned} Cov(\delta_{2s}, \gamma_{2s}) &= Cov(\delta_{2s}, \gamma_{2s}^* + \gamma_{1s}\lambda_{12s}) \\ &= Cov(\delta_{2s}, \gamma_{2s}^*) + \gamma_1 Cov(\delta_{2s}, \lambda_{12s}) + \lambda_{12} Cov(\delta_{2s}, \gamma_{1s}) \\ &+ E[(\delta_{2s} - \delta_2)(\gamma_{1s} - \gamma_1)(\lambda_{12s} - \lambda_{12})]. \end{aligned}$$

(A9)

Thus, we require that the site-average direct effects of each mediator be independent of the site-average compliance of each mediator and independent of the site-average effect of mediator 1 on mediator 2 (and that the third centered moment of $\{\gamma_{1s}, \lambda_{12s}, \delta_{2s}\}$ must be zero). In particular, we require $Cov(\delta_{1s}^*, \lambda_{12s}) = Cov(\delta_{2s}, \lambda_{12s}) = 0$.

Given these assumptions, and the ignorable treatment assignment and sufficient rank assumptions (assumptions viii and ix), we can identify δ_1^* and δ_2 from the regression model

$$\beta_s = \delta_1^* \gamma_{1s} + \delta_2 \gamma_{2s} + \omega_s, \tag{A10}$$

because the β_s's, γ_{1s}'s, and γ_{2s}'s are directly estimable from the observed data. Importantly, however, the assumptions are not sufficient to identify δ_1, the total effect of $M1$. Our assumptions imply that $\delta_1 = \delta_1^* + \lambda_{12}\delta_2$, but because our assumptions are, in general, insufficient to identify λ_{12}, we therefore cannot identify δ_1. To identify λ_{12}, we would require a further assumption regarding the independence of γ_{1s} and γ_{2s}^*.[3] In general then, if we replace the parallel mediators assumption with a stronger set of assumptions about the independence of the person-specific and site-specific direct effects of each mediator with everything upstream from that mediator, we still can only identify the direct effect of each mediator (that part of the effect that does not operate through any other mediator in the model).

[3] To see this, consider the lefthand part of Figure A1. If we consider $M2$ as the outcome, then T affects $M2$ both directly and through $M1$. Now construct a second mediator M^* that is in the direct pathway between T and $M2$. Let $M^* = T$ for all individuals, implying that Γ^*, the person-specific effect of T on M^*, is equal to 1 for all individuals, and that Δ^*, the person-specific effect of M^* on $M2$ is equal to Γ_2^* for all individuals. Now we have a case of parallel mediators—T affects $M2$ through two parallel mediators $M1$ and M^*. Assumption (vii) implies that γ_{1s} is independent of δ_s^*, but this is the same as assuming $\gamma_{1s} \perp \gamma_{2s}^*$. Thus, to identify λ_{12}, we require the additional assumption that the direct effects of T on both mediators are independent. Note that nowhere else have we assumed that compliances are uncorrelated; this is a strong, and generally untenable, assumption.

APPENDIX B: MEAN AND VARIANCE OF MSMM-IV ESTIMATORS

In this appendix, we first show that using two-stage least squares (2SLS) to estimate the MSMM-IV model with site fixed effects and site-by-treatment interactions as instruments is equivalent to fitting a site-level weighted least squares regression model where the estimated ITT effect in a given site is the outcome and where the site-specific first-stage effects (the 'compliances') are predictors. We next show that this estimator is unbiased under the assumptions we outline in the paper. Finally, we derive expressions for the sampling variance of the 2SLS MSMM-IV model under conditions of both homogeneity and heterogeneity of the mediator effects.

Notation

We have persons $i = 1, \dots, n_s$ nested within sites $s = 1, \dots, K$. Let $N = \sum_{s=1}^{K} n_s$. Person i in site s is assigned to treatment condition T_{is} (which is measured in an interval-scaled metric) and is observed to have a vector of P continuous mediators $\mathbf{M}_{is} = (M_{1is}, M_{2is}, \dots, M_{Pis})'$, and outcome Y_{is}. Under the nine assumptions outlined above, T_{is} is an instrument that identifies the vector of effects $\boldsymbol{\delta} = (\delta_1, \delta_2, .., \delta_P)'$ of mediators $M_1, M_2, \dots, M_P$ on Y.

Let $\mathbf{Y}$ be the N x 1 vector of observed outcomes. Let $\mathbf{1}$ be the N x 1 vector with elements equal to unity and let η be a scalar. Let $\mathbf{M}$ be the N x P matrix of observed mediators. Let $\mathbf{T}$ be the N x N matrix with the values of T on the diagonals. Finally, let $\mathbf{S}$ be the N x K matrix with element $s_{is} = 1$ if person i is in site k and $s_{ik} = 0$ otherwise.

The Two-stage Least Squares Estimator

The 2SLS model is

$$\mathbf{Y} = \mathbf{S\eta} + E(\mathbf{M}|\mathbf{T})\boldsymbol{\delta} + \mathbf{u}, \quad \mathbf{u} \sim (\mathbf{0}, \sigma_u^2\mathbf{I}),$$

(B1)

where $\boldsymbol{\eta}$ is a K x 1 vector of site-specific intercepts, and where the conventional "first stage" gives us $E(\mathbf{M}|\mathbf{T}) = \mathbf{S\mu} + \mathbf{TS\gamma}$, where $\boldsymbol{\mu}$ is the K x P matrix of site-specific intercepts from the first stage equations and $\boldsymbol{\gamma}$ is the K x P matrix of compliance parameters from the first stage equations (i.e., γ_{sp} is the average effect of T on M_p in site s).[4] Thus, (B1) is equivalent to

$$\mathbf{Y} = \mathbf{S\eta} + \mathbf{S}\boldsymbol{\mu}\boldsymbol{\delta} + \mathbf{TS\gamma\delta} + \mathbf{u}.$$

(B2)

The fixed effects estimator of $\boldsymbol{\delta}$ can be obtained by centering the elements of (B2) around their site means, yielding

$$\mathbf{Y}^* = \mathbf{T}^*\mathbf{S\gamma\delta} + \mathbf{u}^*,$$

(B3)

where $\mathbf{Y}^*$ is the N x 1 vector with elements $Y_{is}^* = Y_{is} - \bar{Y}_{\cdot s}$; $\mathbf{T}^*$ is the N x N matrix with diagonal elements $T_{is}^* = T_{is} - \bar{T}_{\cdot s}$; and $\mathbf{u}^*$ is the N x 1 vector with elements $u_{is}^* = u_{is} - \bar{u}_{\cdot s}$. Now, the OLS estimator for (B3) will be

$$\hat{\boldsymbol{\delta}} = (\boldsymbol{\gamma}'\mathbf{S}'\mathbf{T}^{*\prime}\mathbf{T}^*\mathbf{S\gamma})^{-1}(\boldsymbol{\gamma}'\mathbf{S}'\mathbf{T}^{*\prime}\mathbf{Y}^*).$$

(B4)

[4] Note that (B1) assumes that errors are i.i.d.; this is a standard (though potentially problematic) assumption in 2SLS models. In particular, if the δ_{ps}'s vary across sites, the i.i.d. assumption is likely to be invalid. Note that the i.i.d. assumption is an assumption of a specific IV estimator, rather than an identifying assumption of the MSMM-IV method in general.

Now we note that $\mathbf{W} = \mathbf{S}'\mathbf{T}^{*\prime}\mathbf{T}^{*}\mathbf{S}$ will be the diagonal K x K weight matrix with diagonal elements equal to $w_s = n_s\sigma^2_{Ts}$, where σ^2_{Ts} is the variance of T in site s. So we have

$$\widehat{\boldsymbol{\delta}} = (\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^{-1}(\boldsymbol{\gamma}'\mathbf{S}'\mathbf{T}^{*\prime}\mathbf{Y}^{*}). \tag{B5}$$

Now note that if we fit the reduced form fixed effects model $\mathbf{Y} = \mathbf{S}\boldsymbol{\theta} + \mathbf{TS}\boldsymbol{\beta}$, where $\boldsymbol{\theta}$ is a K x 1 vector of site-specific intercepts and $\boldsymbol{\beta}$ is the K x 1 vector of ITT effects (i.e., β_s is the average effect of T on Y in site s), using the same centering strategy as above, we get

$$\widehat{\boldsymbol{\beta}} = \mathbf{W}^{-1}(\mathbf{S}'\mathbf{T}^{*\prime}\mathbf{Y}^{*})$$

$$\mathbf{W}\widehat{\boldsymbol{\beta}} = \mathbf{S}'\mathbf{T}^{*\prime}\mathbf{Y}^{*}. \tag{B6}$$

Substituting (B6) into (B5) yields

$$\widehat{\boldsymbol{\delta}} = (\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^{-1}\left(\boldsymbol{\gamma}'\mathbf{W}\widehat{\boldsymbol{\beta}}\right). \tag{B7}$$

We can therefore reformulate the 2SLS regression model in (B2) as a site-level weighted least squares regression of the estimated ITT effects on the site-specific compliances, where the weights are w_s:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\gamma}\boldsymbol{\delta} + \boldsymbol{\omega}, \qquad \boldsymbol{\omega} \sim (\mathbf{0}, \sigma^2\mathbf{W}^{-1}). \tag{B8}$$

Equation (B8) implicitly assumes that $\boldsymbol{\delta}$ is homogenous across sites. More generally, however, the effect of the mediators may vary among sites. In order to compute the bias and variance of the MSMM-IV 2SLS estimator, we consider the general case where the effect of each mediator may be heterogenous. First we define $\boldsymbol{\gamma}_s$ as the s^{th} row of $\boldsymbol{\gamma}$ (that is

$\boldsymbol{\gamma}_s$ is the 1 x P vector of compliances in site s) and we define $\boldsymbol{\delta}_s$ is the P x 1 vector of effects of the mediators in site s. Note that we can write the estimated ITT effect in site s as

$$\begin{aligned}
\hat{\beta}_s &= E(\mathrm{B}|s) + e_\mathrm{s} \\
&= \sum_{r=1}^{P} E(\Gamma_r \cdot \Delta_r|s) + e_\mathrm{s} \\
&= \sum_{r=1}^{P} \gamma_{rs}\delta_{rs} + cov(\Gamma_r, \Delta_r)|s + e_\mathrm{s} \\
&= \sum_{r=1}^{P} [\gamma_{rs}\delta_r + \gamma_{rs}(\delta_{rs} - \delta_r) + cov(\Gamma_r, \Delta_r)|s] + e_\mathrm{s} \\
&= \boldsymbol{\gamma}_s\boldsymbol{\delta} + \boldsymbol{\gamma}_s(\boldsymbol{\delta}_s - \boldsymbol{\delta}) + C_s + e_\mathrm{s} \\
&= \boldsymbol{\gamma}_s\boldsymbol{\delta} + \boldsymbol{\gamma}_s\mathbf{b}_s + C_s \; + e_\mathrm{s},
\end{aligned} \tag{B9}$$

where $e_s = \hat{\beta}_s - \beta_s \sim N(0, \sigma^2/w_s)$; $\mathbf{b}_s = \boldsymbol{\delta}_s - \boldsymbol{\delta} \sim N(\mathbf{0}, \boldsymbol{\tau})$, $\boldsymbol{\tau}$ being a P x P covariance matrix, and where $C_s = \sum_{r=1}^{P} cov(\Gamma_r, \Delta_r)|s$. We can then write (B9) as

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_K \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \vdots \\ \boldsymbol{\gamma}_K \end{pmatrix} \boldsymbol{\delta} + \begin{pmatrix} \boldsymbol{\gamma}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\gamma}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\gamma}_K \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_K \end{pmatrix} + \begin{pmatrix} C_1 \\ C_2 \\ \vdots \\ C_K \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_K \end{pmatrix}, \tag{B10}$$

Or, more compactly, as

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\gamma}\boldsymbol{\delta} + \mathbf{Z}\mathbf{b} + \mathbf{C} + \mathbf{e}, \tag{B11}$$

Where $\mathbf{C}$ and $\mathbf{e}$ are the K x 1 vectors of the C_s's and e_s's and $\mathbf{Z}$ is the diagonal matrix containing the $\boldsymbol{\gamma}_s$ vectors. Substituting (B11) into (B7) yields

$$\begin{aligned}\widehat{\boldsymbol{\delta}} &= (\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^{-1}\left(\boldsymbol{\gamma}'\mathbf{W}\widehat{\boldsymbol{\beta}}\right) \\ &= (\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^{-1}\boldsymbol{\gamma}'\mathbf{W}(\boldsymbol{\gamma}\boldsymbol{\delta} + \mathbf{Z}\mathbf{b} + \mathbf{C} + \mathbf{e}) \\ &= \boldsymbol{\delta} + (\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^{-1}\boldsymbol{\gamma}'\mathbf{W}(\mathbf{Z}\mathbf{b} + \mathbf{C} + \mathbf{e}).\end{aligned} \tag{B12}$$

Bias of the 2SLS Estimator

To find the bias in the 2SLS estimator, we take the conditional expectation of (B12), given $\boldsymbol{\gamma}$:

$$E\left(\widehat{\boldsymbol{\delta}}\middle|\boldsymbol{\gamma}\right) = \boldsymbol{\delta} + (\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^{-1}\boldsymbol{\gamma}'\mathbf{W}[\mathbf{Z}E(\mathbf{b}|\boldsymbol{\gamma}) + E(\mathbf{C}|\boldsymbol{\gamma}) + E(\mathbf{e}|\boldsymbol{\gamma})]. \tag{B13}$$

Under the assumption of ignorable assignment of T, $E(\mathbf{e}|\boldsymbol{\gamma}) = E(\mathbf{e}) = \mathbf{0}$. Under the no within-site compliance-effect covariance assumption, $\mathbf{C} = \mathbf{0}$, so $E(\mathbf{C}|\boldsymbol{\gamma}) = E(\mathbf{C}) = \mathbf{0}$. Finally, under the between-site compliance-effect independence assumption, $E(\mathbf{b}|\boldsymbol{\gamma}) = E(\mathbf{b}) = \mathbf{0}$. Therefore, $E\left(\widehat{\boldsymbol{\delta}}\middle|\boldsymbol{\gamma}\right) = \boldsymbol{\delta}$ and the estimator is unbiased.

Variance of the 2SLS Estimator

Noting that $Var\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{Z}\boldsymbol{\tau}\mathbf{Z}' + \sigma^2\mathbf{W}^{-1}$, we can write the variance of the 2SLS MSMM-IV estimator as

$$\begin{aligned}Var\left(\widehat{\boldsymbol{\delta}}\middle|\boldsymbol{\gamma}\right) &= (\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^{-1}\boldsymbol{\gamma}'\mathbf{W}\left[Var\left(\widehat{\boldsymbol{\beta}}\right)\right]\mathbf{W}'\boldsymbol{\gamma}(\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^{-\mathbf{1}} \\ &= (\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^{-1}\boldsymbol{\gamma}'\mathbf{W}[\mathbf{Z}\boldsymbol{\tau}\mathbf{Z}']\mathbf{W}'\boldsymbol{\gamma}(\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^{-\mathbf{1}} + \sigma^2(\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^{-\mathbf{1}}.\end{aligned} \tag{B14}$$

Note that if the effects are homogenous, that is, if $\boldsymbol{\tau} = \mathbf{0}$, then (B14) becomes simply

$$Var\left(\widehat{\boldsymbol{\delta}}|\boldsymbol{\gamma}\right) = \sigma^2(\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^{-\mathbf{1}}. \tag{B15}$$

APPENDIX C: EXPRESSIONS FOR THE BETWEEN-SITE COMPLIANCE-EFFECT COVARIANCE BIAS WHEN $P = 1$ OR $P = 2$.

Here we derive expressions for the bias due to non-independence of the site-specific compliance and effect parameters in the 2SLS MSMM-IV estimator when there are one or two mediators. In order to simplify these expressions somewhat, and express them in terms of the means, variances, and covariances of the site-specific compliance and effect parameters, we require several simplifying assumptions.

First, we assume w_s is constant across sites (sample sizes and treatment variance are constant across sites). We also assume there is no within-site compliance-effect covariance (i.e., $cov(\Gamma_1, \Delta_1)|s = \ cov(\Gamma_2, \Delta_2)|s = 0$ for all $s \in 1, \dots, K$). Next we assume that the γ_s's and δ_s's are linearly related to one another (i.e., we allow them to be correlated, but constrain them to have a linear relationships such that there are constants a_p and $b_b$such that $\gamma_{ps} = a + b\delta_{ps} + e_{ps}, E\left(e_{ps}|\delta_{ps}\right) = E\left(e_{ps}\right) = 0$, for all $p \in 1, \dots, P$). Finally, we assume that the γ_s's and δ_s's have non-skew distributions (i.e, that $\sum_{s=1}^{K}\left(\gamma_{ps} - \gamma_p\right)^3 = \sum_{s=1}^{K}\left(\delta_{ps} - \delta_p\right)^3 = 0$). Under these assumptions, (B12) becomes

$$E\left(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}|\boldsymbol{\gamma}\right) = (\boldsymbol{\gamma}'\boldsymbol{\gamma})^{-1}\boldsymbol{\gamma}'\mathbf{Z}\mathbf{b}.$$

$$= (\boldsymbol{\gamma}'\boldsymbol{\gamma})^{-1}\sum_{s=1}^{K}\boldsymbol{\gamma}_s'\boldsymbol{\gamma}_s\left(\boldsymbol{\delta}_s - \boldsymbol{\delta}\right)$$

$$= (\boldsymbol{\gamma}'\boldsymbol{\gamma})^{-1} \sum_{s=1}^{K} (\boldsymbol{\gamma}_s' - \bar{\boldsymbol{\gamma}}')(\boldsymbol{\gamma}_s - \bar{\boldsymbol{\gamma}})(\boldsymbol{\delta}_s - \boldsymbol{\delta}) + (\bar{\boldsymbol{\gamma}}'\boldsymbol{\gamma}_s + \boldsymbol{\gamma}_s'\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}'\bar{\boldsymbol{\gamma}})(\boldsymbol{\delta}_s - \boldsymbol{\delta})$$

$$= (\boldsymbol{\gamma}'\boldsymbol{\gamma})^{-1} \sum_{s=1}^{K} (\boldsymbol{\gamma}_s' - \bar{\boldsymbol{\gamma}}')(\boldsymbol{\gamma}_s - \bar{\boldsymbol{\gamma}})(\boldsymbol{\delta}_s - \boldsymbol{\delta})$$

$$+ (\bar{\boldsymbol{\gamma}}'(\boldsymbol{\gamma}_s - \bar{\boldsymbol{\gamma}}) + (\boldsymbol{\gamma}_s' - \bar{\boldsymbol{\gamma}}')\bar{\boldsymbol{\gamma}} + \bar{\boldsymbol{\gamma}}'\bar{\boldsymbol{\gamma}})(\boldsymbol{\delta}_s - \boldsymbol{\delta}),$$

(C1)

where $\bar{\boldsymbol{\gamma}}$ is the 1 x P matrix containing the averages of the γ_{ps}'s across sites. Under the linearity and non-skew assumptions above, $\sum_{s=1}^{K}(\boldsymbol{\gamma}_s' - \bar{\boldsymbol{\gamma}}')\,(\boldsymbol{\gamma}_s - \bar{\boldsymbol{\gamma}})(\boldsymbol{\delta}_s - \boldsymbol{\delta}) = \mathbf{0}$. Likewise, it is straightforward to show that $\sum_{s=1}^{K} \bar{\boldsymbol{\gamma}}'\bar{\boldsymbol{\gamma}}(\boldsymbol{\delta}_s - \boldsymbol{\delta}) = \mathbf{0}$. After applying these assumptions, (C1) is now

$$E\left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\middle|\boldsymbol{\gamma}\right) = (\boldsymbol{\gamma}'\boldsymbol{\gamma})^{-1} \sum_{s=1}^{K} (\bar{\boldsymbol{\gamma}}'(\boldsymbol{\gamma}_s - \bar{\boldsymbol{\gamma}}) + (\boldsymbol{\gamma}_s' - \bar{\boldsymbol{\gamma}}')\bar{\boldsymbol{\gamma}})(\boldsymbol{\delta}_s - \boldsymbol{\delta}).$$

(C2)

Bias in the $P = 1$ Case

When $P = 1$, (C2) becomes

$$E\left(\hat{\delta} - \delta\middle|\boldsymbol{\gamma}\right) = \left(\gamma_1^2 + var(\gamma_{1s})\right)^{-1} \sum_{s=1}^{K} 2\gamma_1(\gamma_{1s} - \gamma_1)(\delta_{1s} - \delta_1)$$

$$= \frac{2\gamma_1 Cov(\gamma_{1s}, \delta_{1s})}{\gamma_1^2 + var(\gamma_{1s})}.$$

(C3)

Note that (C3) can be rewritten as

$$E\left(\hat{\delta} - \delta\middle|\boldsymbol{\gamma}\right) = 2\rho_{\gamma\delta}\sigma_\delta \frac{CV(\gamma)}{CV(\gamma)^2 + 1},$$

(C4)

where $\rho_{\gamma\delta}$ is the correlation between γ_1 and δ_1; σ_δ is the standard deviation of δ_1 across sites; and $CV(\gamma) = \frac{\sigma_\gamma}{\gamma}$ is the coefficient of variation of γ. For given values of $\rho_{\gamma\delta}$ and σ_δ, the bias is maximized when $CV(\gamma) = 1$. The bias decreases to 0 as $CV(\gamma) \Rightarrow 1$ and as $CV(\gamma) \Rightarrow \infty$. Thus, under some simplifying assumptions about the joint distribution of γ and δ (linear association, non-skew distributions), the asymptotic bias in the mulitiple-site IV estimator can be written as a relatively simple function of the variances and covariances of γ and δ. It may be possible to bound the bias term using information about the plausible distributions of the γ's and δ's obtained from other analyses.

Bias in the $P = 2$ Case

When $P = 2$, (C2) becomes

$$
\begin{aligned}
E\left(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta} \middle| \boldsymbol{\gamma}\right) &= (\boldsymbol{\gamma}'\boldsymbol{\gamma})^{-1} \sum_{s=1}^{K} (\bar{\boldsymbol{\gamma}}'(\boldsymbol{\gamma}_s - \bar{\boldsymbol{\gamma}}) + (\boldsymbol{\gamma}_s' - \bar{\boldsymbol{\gamma}}')\bar{\boldsymbol{\gamma}})(\boldsymbol{\delta}_s - \boldsymbol{\delta}) \\
&= K(\boldsymbol{\gamma}'\boldsymbol{\gamma})^{-1} \begin{bmatrix} 2\gamma_1 cov(\gamma_{1s}, \delta_{1s}) + \gamma_1 cov(\gamma_{2s}, \delta_{2s}) + \gamma_2 cov(\gamma_{1s}, \delta_{2s}) \\ \gamma_2 cov(\gamma_{1s}, \delta_{1s}) + 2\gamma_2 cov(\gamma_{2s}, \delta_{2s}) + \gamma_1 cov(\gamma_{2s}, \delta_{1s}) \end{bmatrix} \\
&= K^2 \begin{bmatrix} \gamma_1^2 + var(\gamma_{1s}) & \gamma_1\gamma_2 + cov(\gamma_{1s}, \gamma_{2s}) \\ \gamma_1\gamma_2 + cov(\gamma_{1s}, \gamma_{2s}) & \gamma_2^2 + var(\gamma_{2s}) \end{bmatrix}^{-1} \\
&\quad \cdot \begin{bmatrix} 2\gamma_1 cov(\gamma_{1s}, \delta_{1s}) + \gamma_1 cov(\gamma_{2s}, \delta_{2s}) + \gamma_2 cov(\gamma_{1s}, \delta_{2s}) \\ \gamma_2 cov(\gamma_{1s}, \delta_{1s}) + 2\gamma_2 cov(\gamma_{2s}, \delta_{2s}) + \gamma_1 cov(\gamma_{2s}, \delta_{1s}) \end{bmatrix} \\
&= \frac{1}{D} \begin{bmatrix} \gamma_2^2 + var(\gamma_{2s}) & -\gamma_1\gamma_2 - cov(\gamma_{1s}, \gamma_{2s}) \\ -\gamma_1\gamma_2 - cov(\gamma_{1s}, \gamma_{2s}) & \gamma_1^2 + var(\gamma_{1s}) \end{bmatrix} \\
&\quad \cdot \begin{bmatrix} 2\gamma_1 cov(\gamma_{1s}, \delta_{1s}) + \gamma_1 cov(\gamma_{2s}, \delta_{2s}) + \gamma_2 cov(\gamma_{1s}, \delta_{2s}) \\ \gamma_2 cov(\gamma_{1s}, \delta_{1s}) + 2\gamma_2 cov(\gamma_{2s}, \delta_{2s}) + \gamma_1 cov(\gamma_{2s}, \delta_{1s}) \end{bmatrix},
\end{aligned}
\tag{C5}
$$

where

$$D = \left(\gamma_1^2 + var(\gamma_{1s})\right)\left(\gamma_2^2 + var(\gamma_{2s})\right) - \left(\gamma_1\gamma_2 + cov(\gamma_{1s}, \gamma_{2s})\right)^2.$$

(C6)

After a bit more matrix algebra and rearrangement, we have

$$E\left[\hat{\delta}_1 - \delta_1|\gamma\right] = A_{11}Cov(\gamma_{1s}, \delta_{1s}) + A_{12}Cov(\gamma_{1s}, \delta_{2s}) + A_{21}Cov(\gamma_{2s}, \delta_{1s}) + A_{22}Cov(\gamma_{2s}, \delta_{2s})$$

$$E\left[\hat{\delta}_2 - \delta_2|\gamma\right] = B_{11}Cov(\gamma_{1s}, \delta_{1s}) + B_{12}Cov(\gamma_{1s}, \delta_{2s}) + B_{21}Cov(\gamma_{2s}, \delta_{1s}) + B_{22}Cov(\gamma_{2s}, \delta_{2s}),$$

(C7)

where

$$A_{11} = \frac{1}{D}[\gamma_1\gamma_2^2 + 2\gamma_1 Var(\gamma_{2s}) - \gamma_2 Cov(\gamma_{1s}, \gamma_{2s})]$$

$$A_{22} = \frac{-1}{D}[\gamma_1\gamma_2^2 + \gamma_1 Var(\gamma_{2s}) + 2\gamma_2 Cov(\gamma_{1s}, \gamma_{2s})]$$

$$A_{12} = \frac{1}{D}[\gamma_2^3 + \gamma_2 Var(\gamma_{2s})]$$

$$A_{21} = \frac{-1}{D}[\gamma_1^2\gamma_2 + \gamma_1 Cov(\gamma_{1s}, \gamma_{2s})]$$

$$B_{11} = \frac{-1}{D}[\gamma_1^2\gamma_2 + \gamma_2 Var(\gamma_{1s}) + 2\gamma_1 Cov(\gamma_{1s}, \gamma_{2s})]$$

$$B_{22} = \frac{1}{D}[\gamma_1^2\gamma_2 + 2\gamma_2 Var(\gamma_{1s}) - \gamma_1 Cov(\gamma_{1s}, \gamma_{2s})]$$

$$B_{12} = \frac{-1}{D}[\gamma_1\gamma_2^2 + \gamma_2 Cov(\gamma_{1s}, \gamma_{2s})]$$

$$B_{21} = \frac{1}{D}[\gamma_1^3 + \gamma_1 Var(\gamma_{1s})].$$

(C8)

The key thing to note here is that the bias in $\hat{\delta}_1$ depends not only on the covariance between γ_{1s} and δ_{1s}, but also on $Cov(\gamma_{1s}, \delta_{2s})$, $Cov(\gamma_{2s}, \delta_{1s})$, and $Cov(\gamma_{2s}, \delta_{2s})$. Similarly,

the bias in $\hat{\delta}_2$ depends not only on the covariance between γ_{2s} and δ_{2s}, but also on $Cov(\gamma_{1s}, \delta_{2s})$, $Cov(\gamma_{2s}, \delta_{1s})$, and $Cov(\gamma_{1s}, \delta_{1s})$. Moreover, the biases are very complex functions of these covariances, so it will not be easy to predict their magnitude or direction in practical applications.

REFERENCES

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association, 91*(434), 444-455.

Duncan, G. J., Morris, P. A., & Rodrigues, C. (2011). Does Money Really Matter? Estimating Impacts of Family Income on Young Children's Achievement with Data from Random-Assignment Experiments. *Developmental Psychology, 47*(5), 1263-1279.

Heckman, J. J., & Robb, R. (1985a). Alternative Methods for Evaluating the Impact of Interventions. In J. J. Heckman & B. Singer (Eds.), *Longitudinal Analysis of Labor Market Data* (Vol. 10, pp. 156-245). New York: Cambridge University Press.

Heckman, J. J., & Robb, R. (1985b). Using Longitudinal Data to Estimate Age, Period and Cohort Effects in Earnings Equations. In W. M. Mason & S. E. Feinberg (Eds.), *Cohort Analysis in Social Research Beyond the Identification Problem*. New York: Springer-Verlag.

Heckman, J. J., & Robb, R. (1986). Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes. In H. Wainer (Ed.), *Drawing Inferences from Self-Selected Samples* (pp. 63-107). New York: Springer-Verlag.

Heckman, J. J., Urzua, S., & Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics, 88*(3), 389-432.

Imbens, G. W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica, 62*(2), 467-475.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental Analysis of Neighborhood Effects. *Econometrica, 75*(1), 83-119.

Little, R. J., & Yau, L. H. Y. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3*(2), 147-159.

Rubin, D. B. (1986). Comment: Which Ifs Have Causal Answers. *Journal of the American Statistical Association, 81*(396), 961-962.

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*(3).