# Analyzing Heterogeneous Causal Mediation Effects in Multi-Site Trials With Application to the National Job Corps Study

Xu Qin* Guanglei Hong†

**Abstract**

When a multi-site randomized trial reveals between-site variation in program impact, methods are needed for further investigating heterogeneous mediation mechanisms across the sites. We conceptualize and identify a joint distribution of site-specific direct and indirect effects under the potential outcomes framework. A method-of-moments procedure incorporating ratio-of-mediator-probability weighting (RMPW) consistently estimates the causal parameters. This strategy conveniently relaxes the assumption of no treatment-by-mediator interaction while greatly simplifying the outcome model specification without invoking strong distributional assumptions. We derive asymptotic standard errors that reflect the sampling variability of the estimated weight.

**Key Words:** Direct effect; indirect effect; multilevel modeling; multisite randomized trials; method of moments; RMPW; two-step estimation

## 1. Introduction

Intervention programs in economics, education, political science, public health, and social welfare are usually delivered in organizations or communities. Each local setting can be viewed as an experimental site within which individuals are assigned to different treatment conditions. Multi-site randomized trials and multi-site natural experiments have been pervasive in these fields and often feature longitudinal data collection (Bloom et al., 2005; Raudenbush and Bloom, 2015; Spybrook and Raudenbush, 2009). Different from clustered trials, in which individuals in the same cluster/site are assigned to the same treatment condition, multi-site trials provide unique opportunities for the investigation of how the treatment impact varies across sites. Past research has often reported a considerable amount of cross-site heterogeneity in the total treatment effect possibly due to natural variations in organizational contexts, in participant composition, and in local implementation (Weiss et al., 2014). Assessing between-site variation in the causal mechanisms will generate important information for unpacking and understanding the heterogeneity in the total treatment effects. With the existing statistical methods and analytic tools, however, program evaluators cannot take full advantage of such data.

In the basic mediation framework, the treatment affects a focal mediator, which in turn affects the outcome. To determine the extent to which the focal mediator transmits the treatment effect on the outcome in a single site, one may decompose the total treatment effect into an indirect effect that channels the treatment effect through the hypothesized mediator and a direct effect that works directly or through other unspecified mechanisms. Additional important research questions arise in a multisite study. We illustrate with the National Job Corps Study, a multisite randomized evaluation of the nation's largest job

*University of Chicago, 1126 E. 59th St., Chicago, IL 60637. The authors thank Donald Hedeker, Stephen Raudenbush, Kazuo Yamaguchi, Fan Yang and our colleagues Ed Bein, Jonah Deutsch, Kristin Porter, and Cheng Yang for their contribution of ideas and their comments on earlier versions of this article. Alma Vigil provided research assistance to the analysis of the Job Corps data. This study was supported by a U.S. Department of Education Institute of Education Sciences Statistical and Research Methodology Grant (R305D120020), a subcontract from MDRC funded by the Spencer Foundation, and a Quantitative Methods in Education and Human Development Research Predoctoral Fellowship awarded to the first author.

†University of Chicago, 1126 E. 59th St., Chicago, IL 60637

training program for disadvantaged youth. The Job Corps program theory emphasizes both educational attainment and risk reduction. Previous research has suggested that educational attainment be a potential mediator of the Job Corps impact on earnings (Flores and Flores-Lagunes, 2013). Yet it is unclear whether the treatment mechanism mediated by educational attainment–shown as an indirect effect–operates the same across all the sites; nor is it clear whether the role of other program elements–summarized in a direct effect–is consistent over the sites. Such evidence will be crucial for enriching theoretical understanding and for informing the design and implementation of programs alike.

This study addresses the need for a flexible methodological solution for investigating heterogeneity of causal mediation mechanisms in multi-site trials. We develop concepts and methods for defining, identifying, and estimating (1) population average indirect effect and direct effect that decompose a total treatment effect and (2) between-site variance and covariance of indirect effect and direct effect. Unlike the existing strategies for multi-site mediation analysis, our extension of a weighting method accommodates scenarios in which the mediator-outcome relationship differs across the treatment conditions. The causal parameters are estimated through a two-step procedure. We derive asymptotic variances that reflect the sampling variability of the estimated weight. Applying the proposed analytic strategy to the Job Corps data, we generate new empirical evidence about the program. Our method extends and supplements the existing literature on multi-site causal mediation analysis.

Taking on the challenges of multisite data, researchers (Bauer et al., 2006; Kenny et al., 2003; Krull and MacKinnon, 2001; Preacher et al., 2010; Zhang et al., 2009) have proposed to embed the standard path analysis and SEM in multilevel modeling by including random intercepts and random slopes in the mediator model and the outcome model. Bauer and colleagues have further explored the possibility of quantifying not only the population average but also the between-site variation of the direct effect and the indirect effect through specifying multivariate multilevel models. Path analysis and SEM rely on correct specifications of the mediator model and the outcome model. Covariance adjustment for confounding covariates is crucial for removing selection bias. However, even when the treatment is randomized, results tend to be biased if one misspecifies covariate-outcome relationships or fails to consider possible treatment-by-mediator interaction, mediator-by-covariate interactions, or treatment-by-mediator-by-covariate interactions. In addition, because this approach specifies the average indirect effect as a product of regression coefficients, it becomes particularly challenging to estimate the between-site variance of the indirect effect and the covariance between the site-specific direct and indirect effects. Finally, relying on maximum likelihood estimation, the above strategy typically assumes that the mediator and the outcome are multivariate normal in distribution. As others have pointed out (Imai et al., 010a; MacKinnon and Dwyer, 1993; VanderWeele and Vansteelandt, 2010), applications of path analysis and SEM to discrete mediators and outcomes face many constraints in both single-site and multi-site studies .

Other researchers have specified multilevel path analysis models for analyzing data from group randomized trials (VanderWeele, 010b; Vanderweele et al., 2013) that are useful for evaluating treatments administered at the group level but not for investigating between-site variation in mediation mechanisms in a multisite trial. The multisite instrumental variable (IV) method uses treatment-by-site interactions as instruments for the mediators (Kling et al., 2007; Raudenbush et al., 2012; Reardon and Raudenbush, 2013). With its primary interest in identifying the average effect of each mediator on the outcome, the IV method does not estimate the between-site distributions of the indirect effects. A study by Bind et al. (2016) examined time-varying treatments and mediators nested within individuals. Even though one may view individuals in this longitudinal study as analogous to sites, the

researchers focused only on the population average direct and indirect effects. No solution was provided for estimating and testing the between-individual heterogeneity of these effects. To our knowledge, other methods that allow for a treatment-by-mediator interaction (e.g., Imai et al., 010a; Imai et al., 010b) have not been extended to studies of between-site heterogeneity in mediation mechanisms.

Hong (2010, 2015) and others (Hong et al., 2011, 2015; Hong and Nomi, 2012; Huber, 2014; Lange et al., 2012, 2014; Tchetgen et al., 2012; Tchetgen Tchetgen, 2013) have developed weighting strategies for single-site mediation analysis. Defining direct and indirect effects in terms of potential outcomes (Pearl, 2001; Robins and Greenland, 1992), a ratio-of-mediator-probability weighting (RMPW) analysis identifies and estimates these causal effects each as a mean contrast, along with their standard errors, while adjusting for pretreatment confounding through propensity score-based weighting. The basic rationale is that, among individuals with the same pretreatment characteristics, the distribution of the mediator in the experimental group and that in the control group can be effectively equated through weighting under the assumption of sequential ignorability, which we will explain in Section 2.3. Unlike the regression-based strategies, these weighting methods allow for treatment-by-mediator interaction without having to specify the mediator-outcome relationship and the covariate-outcome relationships. The greatly simplified outcome model minimizes the risk of model misspecification. Simulations (Hong et al., 2015) have shown that, when the outcome model is misspecified, RMPW clearly outperforms path analysis/SEM in bias correction.

By extending the RMPW method to data from a multi-site trial, we aim to reveal between-site differences in the causal mediation mechanism. In doing so, this study provides a new statistical tool that can be applied broadly to multi-site studies in which not only the population average direct and indirect effects but also the between-site variation of the direct and indirect effects are of scientific interest.

In the next section, we define the causal parameters under the counterfactual causal framework, and clarify the identification assumptions, based on which we explain the rationale of RMPW-based multi-site mediation analysis. After delineating the method-of-moments estimation procedure in Section 3, we assess the performance of this estimation approach through simulations in Section 4. Section 5 applies the method to the Job Corps data. In Section 6, we discuss the strengths and limitations of this new approach and raise issues for future research.

## 2. Definition and Identification of the Population Average and Variance of Site-Specific Causal Mediation Effects

### *2.1 The Counterfactual Causal Framework*

Applying the counterfactual framework of causal inference (Neyman and Iwaszkiewicz, 1935; Rubin, 1978), we define the causal parameters of interest in the context of the multi-site Job Corps evaluation. Study participants were assigned at random either to an experimental condition that allowed for immediate enrollment in one of the 103 Job Corps centers or to the control condition that forbade Job Corps enrollment for three years. An individual's weekly earnings 48 months after randomization measures the economic outcome. The focal mediator is whether an individual obtained an education or training credential 30 months after randomization.

*2.1.1. Individual-specific causal effects.* We use $T_{ij} = t$ to indicate the treatment assignment of individual $i$ at site $j$ where $t = 1$ (or $t = 0$) implies the individual was (or was not) assigned to the Job Corps program. Let the mediator value be $m = 1$ if the

individual obatined an education or training credential, and $m = 0$ if not. The potential mediator value for individual $i$ at site $j$ is defined as $M_{ij}(t)$ when the individual's treatment assignment is set to $t$ for $t = 0, 1$. Similarly, we use $Y_{ij}(t, M_{ij}(t))$ to represent the potential outcome value for individual $i$ at site $j$ when $T_{ij} = t$. When $M_{ij}(t) = m$, the individual's potential outcome value can be written as $Y_{ij}(t, m)$.

We have defined an individual's potential educational attainment as a function of the treatment value and have defined his or her potential earnings as a function of the treatment value and the mediator value under the *Stable Unit Treatment Value Assumption* (SUTVA) (Rubin, 1980, 1986, 1990). In the context of a multi-site mediation study, SUTVA implies that (a) there is no interference between sites (Hong and Raudenbush, 2006; Hudgens and Halloran, 2008), i.e. the potential mediators of individual $i$ at site $j$ are independent of the treatment assignments of individuals at site $j'$ for all $j' \neq j$ and, additionally, the potential outcomes of individual $i$ at site $j$ are independent of the treatment assignments and mediator value assignments of individuals at site $j'$; and (b) there is no interference between individuals within a site, i.e. an individual's potential mediators are independent of the treatment assignments of other individuals at the same site and, additionally, the individual's potential outcomes are independent of the treatment assignments and mediator value assignments of other individuals at the same site. In the national Job Corps evaluation, an applicant was usually assigned to a Job Corps center relatively close to his or her original residence. Hence, it seems reasonable to invoke assumption (a). Assumption (b) may be violated if a Job Corps student's performance is affected by the behaviors of other students at a center. Contaminations are also possible between individuals in the treated group and those in the control group who share a social network within a site.

Under SUTVA, for individual $i$ at site $j$, the treatment effect on the outcome (i.e., the ITT effect) is defined as $\beta_{ij}^{(T)} \equiv Y_{ij}(1, M_{ij}(1)) - Y_{ij}(0, M_{ij}(0))$. Decomposing the total treatment effect into a direct effect and an indirect effect, however, involves a third potential outcome $Y_{ij}(1, M_{ij}(0))$. This is the earnings the individual would counterfactually have if assigned to a Job Corps program yet having the same educational attainment as he or she would under the control condition.

The *direct effect* of the treatment on the outcome for individual $i$ at site $j$ is

$$\beta_{ij}^{(D)} \equiv Y_{ij}(1, M_{ij}(0)) - Y_{ij}(0, M_{ij}(0)). \tag{1}$$

The direct effect will be nonzero if the Job Corps program has an impact on earnings even without changing an individual's educational attainment. This is possible because many Job Corps centers provide a range of supplemental services designed to reduce risks and improve participants' overall well-being. This is called "the natural direct effect" by Pearl (2001) and "the pure direct effect" by Robins and Greenland (1992).

The *indirect effect* of the treatment on the outcome transmitted through the mediator for individual $i$ at site $j$ is

$$\beta_{ij}^{(I)} \equiv Y_{ij}(1, M_{ij}(1)) - Y_{ij}(1, M_{ij}(0)). \tag{2}$$

The indirect effect represents the Job Corps impact on earnings to be attributed to the program-induced change in educational attainment from $M_{ij}(0)$ to $M_{ij}(1)$. This is called "the natural indirect effect" by Pearl (2001) and "the total indirect effect" by Robins and Greenland (1992). The total treatment effect is the sum of the direct effect and the indirect effect: $\beta_{ij}^{(T)} = \beta_{ij}^{(D)} + \beta_{ij}^{(I)}$.

The above decomposition is not unique. Alternatively, one may decompose the total treatment effect into a "total direct effect", $Y_{ij}(1, M_{ij}(1)) - Y_{ij}(0, M_{ij}(1))$, and a "pure indirect effect", $Y_{ij}(0, M_{ij}(1)) - Y_{ij}(0, M_{ij}(0))$, in Robins and Greenland's terms. The

current study is primarily interested in the impact on earnings when an individual's educational attainment changes from $M_{ij}(0)$ to $M_{ij}(1)$ under the Job Corps program. This is the impact of educational attainment on earnings when the individual has simultaneous access to a range of supplementary services provided by Job Corps. We therefore focus on the causal effects defined in (1) and (2).

*2.1.2. Site-specific causal effects.* There was a Job Corps center at each experimental site. At the time of the study, the 103 Job Corps Centers served eligible participants in almost the entire nation. Rather than viewing the 103 sites in this study as a finite population of sites, we consider a theoretical population of sites that could possibly be infinite in number. This is because the composition of applicants, the composition of Job Corps staff, the center operator, and various elements of the control condition tend to be fluid rather than static. In the National Job Corps Study, which Job Corps center an individual would be assigned to was determined prior to the treatment randomization. Let $S_{ij} = j$ indicate the site membership of individual $i$. We define the site-specific ITT effect $\beta_j^{(T)} = E(\beta_{ij}^{(T)}|S_{ij} = j)$, direct effect $\beta_j^{(D)} = E(\beta_{ij}^{(D)}|S_{ij} = j)$, and indirect effect $\beta_j^{(I)} = E(\beta_{ij}^{(I)}|S_{ij} = j)$.

Given our central interest in between-site heterogeneity, here we focus on the population of sites rather than the population of individuals. We therefore define the key parameters that characterize the distribution of the site-specific causal effects. These include the average ITT effect $\gamma^{(T)} = E(\beta_j^{(T)})$, the average direct effect $\gamma^{(D)} = E(\beta_j^{(D)})$, and the average indirect effect $\gamma^{(I)} = E(\beta_j^{(I)})$ in the population of sites. In addition, the variance of the distribution of the site-specific ITT effect is quantified by $\sigma_T^2 = \text{var}(\beta_j^{(T)}) = E[(\beta_j^{(T)} - \gamma^{(T)})^2]$. The between-site heterogeneity in the ITT effect may be explained by differences between the sites in the direct effect, the indirect effect, or both. We therefore investigate the between-site variance of the direct effect $\sigma_D^2 = \text{var}(\beta_j^{(D)}) = E[(\beta_j^{(D)} - \gamma^{(D)})^2]$, the between-site variance of the indirect effect $\sigma_I^2 = \text{var}(\beta_j^{(I)}) = E[(\beta_j^{(I)} - \gamma^{(I)})^2]$, and the covariance between the site-specific direct effect and indirect effect $\sigma_{D,I} = \text{cov}(\beta_j^{(D)}, \beta_j^{(I)}) = E[(\beta_j^{(D)} - \gamma^{(D)})(\beta_j^{(I)} - \gamma^{(I)})]$. Clearly, $\sigma_T^2 = \text{var}(\beta_j^{(T)}) = \sigma_D^2 + \sigma_I^2 + 2\sigma_{D,I}$.

In summary, we will focus on identifying and estimating the joint distribution of site-specific direct and indirect effects characterized by population means $\gamma^{(D)}$ and $\gamma^{(I)}$ as well as by between-site variances $\sigma_D^2$, $\sigma_I^2$, and covariance $\sigma_{D,I}$.

### *2.2 Identification Assumptions*

The joint distribution of site-specific direct and indirect effects can be identified by observable data under the following two assumptions that constitute the "sequential ignorability" (Imai et al., 010a; Imai et al., 010b) at each site.

*Identification Assumption 1. Ignorable treatment assignment.* This assumption states that, within levels of the observed pretreatment covariates, treatment assignment in each site is independent of all the potential mediators and potential outcomes. In other words, there is no unmeasured confounding of the treatment-mediator relationship or the treatment-outcome relationship at site $j$. This is assumed to be true for all the sites.

$$\{M_{ij}(t), Y_{ij}(t, m)\} \perp\!\!\!\perp T_{ij}|\mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j \qquad \forall j \tag{3}$$

for $t = 0, 1$ and $m = 0, 1$. Here $\mathbf{X}_{ij} = \mathbf{x}$ denotes a vector of observed pretreatment covariates. Additionally, it is assumed that $0 < \Pr(T_{ij} = t|\mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j) < 1$ for $t = 0, 1$. That is, each individual has a nonzero probability of being assigned to either treatment condition in a given site. The assumption of ignorable treatment assignment is easy to satisfy in a multi-site randomized trial such as the Job Corps study.

*Identification Assumption 2. Ignorable mediator value assignment.* This assumption states that, within levels of the observed pretreatment covariates, mediator value assignment under either treatment condition in each site is independent of all the potential outcomes. In other words, there is no unmeasured confounding of the mediator-outcome relationship within a treatment or across the treatment conditions in site $j$. This again is assumed to be true for all the sites.

$$Y_{ij}(t,m) \perp\!\!\!\perp \{M_{ij}(t), M_{ij}(t')\} | T_{ij} = t, \mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j \qquad \forall j \tag{4}$$

for $t$ unequal to $t'$ where $t, t' = 0, 1$ and $m = 0, 1$. It is also assumed that $0 < \Pr(M_{ij}(t) = m | T_{ij} = t, \mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j) < 1$ and $0 < \Pr(M_{ij}(t') = m | T_{ij} = t, \mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j) < 1$. That is, each individual has a nonzero probability of having the mediator value that one would display under the actual or the counterfactual treatment condition. Identification Assumption 2 is particularly strong because usually individuals were not randomized to receive a mediator value after the treatment randomization. The plausibility of this assumption relies heavily on the richness of the observed pretreatment covariates. This assumption also requires that there is no posttreatment covariate that confounds the mediator-outcome relationship (Avin et al., 2005; VanderWeele, 010b; Vanderweele et al., 2013). An example of a possible violation is that, if among individuals with the same baseline characteristics, those who are more likely to obtain an education credential are also the ones who tend to receive more counseling services, then the indirect effect mediated by educational attainment would be confounded by the program benefit transmitted through counseling services. The sequential ignorability assumption must hold in every site. If the assumption is violated in one or more sites, the causal parameters will likely be identified with bias. Assessing the sensitivity of analytic results to possible violations of these identification assumptions is a necessary step in applications.

### *2.3 Identification Results*

Under the sequential ignorability, the site-specific average of each potential outcome is identifiable through weighting, which then enables the identification of the site-specific direct and indirect effects.

In general, when Identification Assumption 1 holds within a site, the average potential outcome associated with treatment condition $t$ at site $j$, $E(Y_{ij}(t, M_{ij}(t)) | S_{ij} = j)$, can be identified by the weighted outcome of individuals actually assigned to treatment t at site j:

$$E(W_{ij}^{(t)} Y_{ij} | T_{ij} = t, S_{ij} = j),$$

where

$$W_{ij}^{(t)} = \frac{\Pr(T_{ij} = t | S_{ij} = j)}{\Pr(T_{ij} = t | \mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j)}. \tag{5}$$

Here $W_{ij}^{(t)}$ is the inverse-probability-of-treatment weight (IPTW) known from past research (Horvitz and Thompson, 1952; Robins, 2000; Rosenbaum, 1987). The weight transforms the experimental group composition and the control group composition such that the probability of treatment assignment in the weighted sample would resemble that in a hypothetical randomized design with equal probability of treatment assignment for all individuals. In other words, applying $W_{ij}^{(t)}$ to individuals with pretreatment characteristics $\mathbf{x}$ who have been assigned to treatment $t$ at site $j$ removes bias due to treatment selection.

When Identification Assumptions 1 and 2 hold within a site, $E(Y_{ij}(1, M_{ij}(0)) | S_{ij} = j)$ can be identified by

$$E(W_{ij} Y_{ij} | T_{ij} = 1, S_{ij} = j),$$

in which

$$W_{ij} = \frac{\Pr(T_{ij} = 1|S_{ij} = j)}{\Pr(T_{ij} = 1|\mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j)} \times \frac{\Pr(M_{ij} = m|T_{ij} = 0, \mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j)}{\Pr(M_{ij} = m|T_{ij} = 1, \mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j)} \quad (6)$$

is the weight applied to individuals with pretreatment characteristics $\mathbf{x}$ who were assigned to the experimental condition in site $j$ and displayed mediator value $m$. Within a single site, this weight is a product of IPTW and RMPW derived by Hong (2010, 2015) and others (Hong et al., 2011, 2015; Hong and Nomi, 2012; Tchetgen et al., 2012). The latter is a ratio of an experimental individual's conditional probability of displaying mediator value $m$ under the counterfactual control condition to that under the experimental condition. For individuals within levels of the pretreatment characteristics $\mathbf{x}$, RMPW transforms the mediator distribution in the experimental group to resemble that in the control group. The weighted experimental group mean outcome therefore identifies the average counterfactual mean outcome associated with the experimental condition when the mediator counterfactually distributes the same as that under the control condition. RMPW is mathematically equivalent to the inverse probability weight (IPW) proposed by Huber (2014). This identification result enables us to relate the observable data to the average counterfactual outcome at a site.

When the treatment assignment is randomized within a site, $\Pr(T_{ij} = t|S_{ij} = j) = \Pr(T_{ij} = t|\mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j)$, we simply have that

$$W_{ij}^{(t)} = 1;$$

$$W_{ij} = \frac{\Pr(M_{ij} = m|T_{ij} = 0, \mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j)}{\Pr(M_{ij} = m|T_{ij} = 1, \mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j)}. \quad (7)$$

Below we use $\mu_{0j}$, $\mu_{1j}$, and $\mu_{*j}$ as shorthand for $E(Y_{ij}|T_{ij} = 0, S_{ij} = j)$, $E(Y_{ij}|T_{ij} = 1, S_{ij} = j)$, and $E(W_{ij}Y_{ij}|T_{ij} = 1, S_{ij} = j)$, respectively. In a multi-site randomized trial, the average direct effect at site $j$, $\beta_j^{(D)}$, can be identified by a simple mean contrast:

$$\beta_j^{(D)} = \mu_{*j} - \mu_{0j}. \quad (8)$$

The average indirect effect at site $j$, $\beta_j^{(I)}$, can be identified by:

$$\beta_j^{(I)} = \mu_{1j} - \mu_{*j}. \quad (9)$$

Once the site-specific direct and indirect effects are identified, their joint distribution in the population can be identified as well.

## 3. Estimation and Inference

The estimation involves two major steps. Step 1 estimates the weight for each individual in the experimental group as a ratio of the conditional probability of mediator value under the experimental condition to that under the control condition corresponding to equation (7). Step 2 estimates the unweighted mean outcome of the control group, the unweighted mean outcome of the experimental group, and the weighted mean outcome of the experimental group for each site and subsequently the site-specific direct effect and indirect effect corresponding to equations (8) and (9). Based on these site-specific estimates, we estimate the population average and the between-site variance of the direct effect and those of the indirect effect.

In step 1, following the convention of propensity score estimation in multilevel data, we fit multilevel mixed-effects logistic regression models to the sample data in each treatment group pooled from all the sites and estimate the coefficients through maximum likelihood. The analysis in step 2 is complicated by the fact that the causal parameters must be estimated on the basis of the estimated weight rather than the true weight. We employ a method-of-moments (MOM) estimation procedure in step 2 to estimate the site-specific direct and indirect effects and their joint distribution. In the meantime, we propose asymptotic variance estimators for the population average direct effect and indirect effect estimators that incorporate the sampling variability in the weight estimation.

We choose MOM rather than MLE in step 2 for three reasons. First, the likelihood in step 2 is a function of the parameters given both the observed outcome and the estimated individual weight. The unknown distribution of the weight adds difficulty to the specification of the likelihood function. Second, conventional MLE with multilevel data assumes that the estimated site-specific effects are independent between sites, an assumption violated in this case due to the pooling of data from all the sites in the step 1 estimation of the weight. Third, our preliminary results suggest that the site-specific effects are not normally distributed. MOM does not invoke assumptions about the distribution of the site-specific effects and thus has a potential for broad applications.

We start by introducing the weighted method-of-moments estimators of the causal effects in a hypothetical scenario in which the weight is known. We then discuss our strategy of obtaining the asymptotic sampling variance of the causal effect estimates when the weight needs to be estimated. Finally, we estimate the between-site variance of the direct and indirect effects by purging the average sampling variance off the total between-site variance of the direct and indirect effect estimates. We also conduct a permutation test for variance testing.

### *3.1 Method-of-Moments Estimators of the Causal Effects When the Weight Is Known*

To estimate the population average effects, we first estimate the direct and indirect effects site by site and then aggregate the site-specific direct effect estimates and the site-specific indirect effect estimates (e.g., Diggle et al., 2002; Raudenbush and Bloom, 2015). In a hypothetical experiment for causal mediation analysis, individuals within each site would be randomized to the experimental or the control condition. Subsequently, individuals with the same observed pretreatment characteristics would be assigned at random to obtain an education credential under each treatment condition. Such a sequential randomized designs satisfies the sequential ignorability assumption. Suppose that, for sampled individual $i$ in site $j$ with pretreatment characteristics $\mathbf{X}_{ij} = \mathbf{x}$, the probability of obtaining an education credential is $p_{1ij} = \Pr(M_{ij} = 1|T_{ij} = 1, \mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j)$ under the experimental condition and is $p_{0ij} = \Pr(M_{ij} = 1|T_{ij} = 0, \mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j)$ under the control condition. To estimate $\mu_{*j} = E(W_{ij}Y_{ij}|T_{ij} = 1, S_{ij} = j)$, we simply obtain a weighted sample mean outcome of those assigned to the experimental condition at site $j$,

$$\widehat{\mu}_{*j} = \frac{\sum_{i=1}^{n_j} Y_{ij}W_{ij}T_{ij}}{\sum_{i=1}^{n_j} W_{ij}T_{ij}}, \tag{10}$$

where $n_j$ is the sample size at site $j$. The weight is $W_{ij} = p_{0ij}/p_{1ij}$ when $M_{ij} = 1$ and $W_{ij} = (1 - p_{0ij})/(1 - p_{1ij})$ when $M_{ij} = 0$.

The experimental mean outcome $\mu_{1j}$ and the control mean outcome $\mu_{0j}$ can be estimated simply by the corresponding sample mean outcomes at each site:

$$\widehat{\mu}_{0j} = \frac{\sum_{i=1}^{n_j} Y_{ij}(1 - T_{ij})}{\sum_{i=1}^{n_j}(1 - T_{ij})};$$

$$\widehat{\mu}_{1j} = \frac{\sum_{i=1}^{n_j} Y_{ij} T_{ij}}{\sum_{i=1}^{n_j} T_{ij}}. \tag{11}$$

The method-of-moments estimators of the site-specific direct effect and the site-specific indirect effect at site $j$ are:

$$\widehat{\beta}_j^{(D)} = \widehat{\mu}_{*j} - \widehat{\mu}_{0j};$$
$$\widehat{\beta}_j^{(I)} = \widehat{\mu}_{1j} - \widehat{\mu}_{*j}. \tag{12}$$

We then estimate the parameters that characterize the distribution of site-specific causal effects for the population of sites. When the sites have been sampled with equal probability from the population of sites, by taking a simple average of the above unbiased estimates of the site-specific direct and indirect effects across all the $J$ sites in the sample, we obtain unbiased estimators of the average direct and indirect effects for the population of sites,

$$\widehat{\boldsymbol{\gamma}} = \frac{1}{J} \sum_{j=1}^{J} \widehat{\boldsymbol{\beta}}_j, \tag{13}$$

in which $\widehat{\boldsymbol{\beta}}_j = (\widehat{\beta}_j^{(D)}, \widehat{\beta}_j^{(I)})'$ and $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}^{(D)}, \widehat{\gamma}^{(I)})'$. Equivalently, it can be written as

$$\widehat{\boldsymbol{\gamma}} = \left(\boldsymbol{\Psi}'\boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Psi}' \widehat{\boldsymbol{\beta}}, \tag{14}$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1', \ldots, \widehat{\boldsymbol{\beta}}_J')'$, and $\boldsymbol{\Psi} = \mathbf{1}_J \otimes \mathbf{I}_2$, in which $\mathbf{1}_J$ is a $J \times 1$ vector of $1's$, and $\mathbf{I}_2$ is a $2 \times 2$ identity matrix.

An alternative precision-weighted estimator would use the inverse of the covariance matrix of the site-specific effect estimates as the weight. Even though precision-weighting is expected to improve efficiency, it may introduce bias and inconsistency if the precision weight is correlated with the effect size of the site-specific direct or indirect effect.

### *3.2 Asymptotic Sampling Variance of Causal Effect Estimates When Weight Is Unknown*

In a typical multi-site randomized experiment, even though the treatment assignment is randomized, the mediator value assignment is not. Hence the weight is unknown and needs to be estimated from the sample data in step 1 prior to the estimation of the causal effects in step 2. In the analytic procedure that we delineate below, a multilevel logistic regression analysis is employed in step 1 to estimate the weight while step 2 involves site-by-site method-of-moments analysis.

*3.2.1. Two-step estimation procedures.* In step 1, we fit two logistic regression models, one to the sampled individuals in the experimental group and the other to those in the control group. (This is equivalent to fitting one logistic regression model to a combination of these two groups with a submodel for each group.) To maximize the precision of estimation, we pool data from all the sites and include a site-specific random effect in each model. If a covariate predicts the mediator differently across the sites, a site-specific random slope can be included as well. The models take the following form,

$$log\left[\frac{p_{tij}}{1-p_{tij}}\right] = \mathbf{X}_{tij}'\boldsymbol{\alpha}_t + \mathbf{C}_{tij}'\mathbf{F}_t\boldsymbol{\theta}_{tj}, \tag{15}$$

for $t = 0, 1$. Here $\mathbf{X}_{tij}$ is a vector of covariates including the intercept; $\boldsymbol{\alpha}_t$ is the corresponding vector of coefficients; $\mathbf{C}_{tij}$ is a vector of covariates with random effects. For computational simplicity, following Hedeker and Gibbons (2006), we standardize the random effects by representing them as $\mathbf{F}_t\boldsymbol{\theta}_{tj}$. Here $\mathbf{F}_t\mathbf{F}_t' = \boldsymbol{\Sigma}_t$ is the Cholesky factorization

of $\mathbf{\Sigma}_t$, the variance-covariance matrix of the random effects; $\mathbf{F}_t$ is a lower triangular matrix; and $\boldsymbol{\theta}_{tj}$ follows a standardized multivariate normal distribution. The analysis can be conducted through maximum likelihood estimation using iterative generalized last squares (IGLS). In addition to the sequential ignorability, the multilevel logistic regression model comes with its model-based assumptions with regard to the relationships between $\mathbf{X}_{tij}$ and $p_{tij}$ as well as the distribution of the random effects.

We predict $p_{1ij}$ for each individual in the experimental group directly based on the propensity score model fitted to the experimental group data. To predict $p_{0ij}$ for the same individuals, we apply the propensity score model that has been fitted to the control group data. In these two propensity score models, $\boldsymbol{\theta}_{1j}$ and $\boldsymbol{\theta}_{0j}$ are each estimated through an empirical Bayes procedure. Because the treatment assignment was independent of the potential mediators within each site, the independence also holds within levels of the pretreatment covariates. Hence among those with the same pretreatment characteristics, the observed mediator distribution of those assigned to the control condition, in expectation, provides counterfactual information of the mediator distribution that the Job Corps participants would likely have displayed should they have been assigned to the control condition instead. Based on the predicted propensity scores, we obtain the estimated weight $\widehat{W}_{ij} = \widehat{p}_{0ij}/\widehat{p}_{1ij}$ for a Job Corps participant who successfully attained an education credential and $\widehat{W}_{ij} = (1-\widehat{p}_{0ij})/(1-\widehat{p}_{1ij})$ for one who did not. $\widehat{W}_{ij}$ is a consistent estimator of $W_{ij}$ because, as the number of sites and the sample size at each site increase, $\widehat{p}_{0ij}$ and $\widehat{p}_{1ij}$ converge in probability to the corresponding true propensities $p_{0ij}$ and $p_{1ij}$. The estimated weight converges in probability to the true weight accordingly.

The step-2 estimation is similar to that described in Section 3.1 except that we need to replace $W_{ij}$ with $\widehat{W}_{ij}$. In the existing literature on propensity score-based weighting in multilevel settings (e.g., Leite et al., 2015), propensity score estimation and causal effect estimation are conducted separately. In this way, however, the sampling variability of the estimated weight obtained in step 1 will not be represented in the standard errors of the causal effect estimates obtained in step 2. Moreover, because we analyze the propensity score models by pooling data from all the sites, the predicted propensity scores and correspondingly the estimated weights are inevitably correlated between sites. Separating the two steps in analysis would lead to bias in estimating the standard errors for the estimated population average direct and indirect effects. As shown later in the simulation study, the problem becomes salient especially when the site size is small. To deal with this challenge, we extend the strategy that Newey (1984) proposed under the single-level setting. Specifically, we stack the moment functions from the two steps and solve them simultaneously. By doing so, the second order conditions for the site-specific direct effect and indirect effect estimators are considered with respect to the parameters that must be estimated in step 1. Intuitively, the stacking allows the step 1 estimation to be configured into the step 2 estimation. The two-step estimators can be fit into the generalized method of moments (GMM) framework (Hansen, 1982). This idea has been applied in single-level settings. For example, Hirano and Imbens (2001) utilized it in the estimation of the total treatment impact using propensity score weighting. Bein and colleagues (Bein et al., 2015) applied the strategy to RMPW-based single-site causal mediation analysis. However, it has not been employed in the multilevel setting. We innovatively extend the estimation procedure to multi-site causal mediation analysis.

*3.2.2. Asymptotic sampling variance of the causal effect estimates.* Let $\mathbf{h}_{ij}^{(1)}$ denote the moment functions for the step-1 parameter estimators $\widehat{\boldsymbol{\eta}}$. Here $\widehat{\boldsymbol{\eta}}$ includes the estimators of the coefficients in the multilevel logistic regression models as well as the elements on or below the diagonal of $\widehat{\mathbf{F}}_t$. Let $\mathbf{h}_{ij}^{(2)}$ denote the moment functions for the step-2 parameter estimators $\widehat{\boldsymbol{\mu}}$. Here $\widehat{\boldsymbol{\mu}}$ includes the estimators of all the site-specific potential outcome

means. Appendix A provides details of these moment functions. Stacking the moment functions from both steps, we have that

$$\mathbf{h}_{ij} = \begin{bmatrix} \mathbf{h}_{ij}^{(1)} \\ \mathbf{h}_{ij}^{(2)} \end{bmatrix}. \tag{16}$$

Now the estimators in the two steps can be rewritten as a one-step estimator $\widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\eta}}', \widehat{\boldsymbol{\mu}}')'$, which jointly solves $\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{h}_{ij} = \mathbf{0}$. Under the standard regularity conditions, $\widehat{\boldsymbol{\vartheta}}$ is a consistent estimator of $\boldsymbol{\vartheta} = (\boldsymbol{\eta}', \boldsymbol{\mu}')'$ with the asymptotic sampling distribution (Hansen, 1982):

$$\sqrt{N}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}) \xrightarrow{d} N(\mathbf{0}, \widehat{\text{var}}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})). \tag{17}$$

The asymptotic normal distribution enables computation of sensible confidence intervals and tests when the site-specific effects or the outcome are not normally distributed. Details on the consistent estimator of $\widehat{\text{var}}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})$ can be found in Appendix A.

Then we are able to derive the sampling variance of the direct and indirect effect estimators. Based on Equations (8), (9) and (12), it is easy to show that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{\Phi}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})$, and thus

$$\text{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \boldsymbol{\Phi}\text{var}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\boldsymbol{\Phi}', \tag{18}$$

where $\boldsymbol{\Phi} = \mathbf{I}_J \otimes \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$, in which $\mathbf{I}_J$ is a $J \times J$ identity matrix. $\text{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is a $2J \times 2J$ matrix, with $\text{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ as the $j$th $2 \times 2$ submatrix along the diagonal. The off-diagonal elements $E[(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)(\widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\beta}_{j'})']$, where $j \neq j'$, are non-zero due to the correlations among the weights estimated in the first step.

Correspondingly, the sampling variance of the population average direct effect and indirect effect estimators, as given in Equation (14), is

$$\text{var}(\widehat{\boldsymbol{\gamma}}) = \left(\boldsymbol{\Psi}'\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}'\text{var}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Psi}\left(\boldsymbol{\Psi}'\boldsymbol{\Psi}\right)^{-1}, \tag{19}$$

in which

$$\text{var}(\widehat{\boldsymbol{\beta}}) = \text{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} + \boldsymbol{\beta}) = \text{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \text{var}(\boldsymbol{\beta}), \tag{20}$$

where $\text{var}(\boldsymbol{\beta}) = \mathbf{I}_J \otimes \text{var}(\boldsymbol{\beta}_j)$. Based on the consistent estimators of $\text{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\text{var}(\boldsymbol{\beta}_j)$, we could consistently estimate the asymptotic standard errors for the population average direct and indirect effect estimators. We explain the estimation of $\text{var}(\boldsymbol{\beta}_j)$ in the next subsection.

### *3.3 Estimation and Inference of Between-Site Variance and Covariance of Causal Effects*

We estimate the between-site variance and covariance of the direct and indirect effects again through the method of moments. We prove in Appendix B that the consistent estimator is:

$$\widehat{\text{var}}(\boldsymbol{\beta}_j) = \frac{1}{J-1}[\sum_{j=1}^{J}(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\gamma}})(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\gamma}})' - \sum_{j=1}^{J}\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + \frac{1}{J}\boldsymbol{\Psi}'\widehat{\text{var}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{\Psi}]. \tag{21}$$

In practice, if a negative variance estimate is obtained, which is known as a Heywood case, both the variance estimate itself and the related covariance estimate will be set to 0.

Previous researchers of multilevel mediation analysis (e.g., Bauer et al., 2006) have not discussed how to conduct hypothesis testing for the between-site variance of the direct and

**Table 1**: Population Causal Parameter Specification

| Parameters | Population Average | | Between-Site Variation | | |
|---|---|---|---|---|---|
| | $\gamma^{(D)}$ | $\gamma^{(I)}$ | $\sigma^2_D$ | $\sigma^2_I$ | $\sigma_{D,I}$ |
| Parameter Set 1 | 0 | 0 | 0 | 0 | 0 |
| Parameter Set 2 | 0.08 | 0.08 | 0.04 | 0.04 | 0.02 |
| Parameter Set 3 | 0.19 | 0.19 | 0.06 | 0.06 | 0.01 |

*Note.* To enable comparisons between the different scenarios, the population average effects have been standardized by the average within-site standard deviation of the outcome in the control group, and the between-site variances and covariances have been standardized by the average within-site variance of the outcome in the control group.

indirect effects. Taking the direct effect as an example, we prove in Appendix C that under $H_0 : \sigma^2_D = 0$,

$$\sum_{j=1}^{J} \frac{(\widehat{\beta}_j^{(D)} - \widehat{\gamma}^{(D)})^2}{\text{var}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})} \xrightarrow{d} \chi^2(J-1).$$

Replacing $\text{var}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})$ with $\widehat{\text{var}}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})$, the test statistic is

$$Q^{(D)} = \sum_{j=1}^{J} \frac{(\widehat{\beta}_j^{(D)} - \widehat{\gamma}^{(D)})^2}{\widehat{\text{var}}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})}. \tag{22}$$

As discussed in Section 3.2.4, as $N$ increases, $\widehat{\text{var}}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})$ converges to $\text{var}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})$. However, when $N$ is small, the distribution of the sample test statistic may deviate from $\chi^2(J-1)$. We thus employ a permutation test proposed by Fitzmaurice et al. (2007). The test randomly permutes the site indices, based on the idea that all permutations of the site indices are equally likely under the null. The details about the algorithm of the permutation test can be found in Appendix C.

## 4. Simulation Study

We conduct a series of Monte Carlo simulations to assess the finite-sample performance of the multilevel RMPW procedure in estimating the population average and between-site variance and covariance of the direct effect and indirect effect. We focus on the case of a binary randomized treatment, a binary mediator, and a continuous outcome, although the estimation procedure can be easily extended to multi-category mediators and binary outcomes. We implement the estimation in `R`, using the `lme4` package (Bates et al., 2014) to fit the multilevel logistic regression models.

We specify three sets of population causal parameters listed in Table 1. The standardized parameter values are similar in magnitude to those used in previous simulation studies of multilevel mediational models (Krull and MacKinnon, 2001; Bauer et al., 2006) and reflect a range of plausible values in real applications. Both the population average and the variance and covariance of the site-specific direct and indirect effects are specified to be 0 in the first scenario, which is designed for examining the Type I error rates in hypothesis testing. All the parameter values increase from set 2 to set 3. Appendix D explains how we generate the simulation data.

The number of sampled sites, $J$, the number of sampled individuals per site, $n_j$, and the probability of treatment assignment at a site, $\Pr(T_{ij} = 1 | S_{ij} = j)$, are manipulated to represent the range observed in past multi-site studies. For example, the Job Corps study

had over 100 sites with an average of about 130 individuals per site. The multi-site sample analyzed by Seltzer (1994) had 20 sites with an average of about 29 individuals per site. Therefore, we generate balanced data sets comprised of 100 or 20 sites of either a small site size ($n_j = 20$) or a moderate site size ($n_j = 150$), while $\Pr(T_{ij} = 1|S_{ij} = j)$ is specified to be 0.5 across all the sites. In addition, we generate an imbalanced data set similar to the Job Corps data with varying site size and varying site-specific probability of treatment assignment.

We make 1,000 replications for each of these scenarios, and then fit analytic models to each data set. We focus on assessing the amount of bias in the parameter estimates when implementing the proposed procedure. Table 2 reports the simulation results for the estimation of the population average effects and the between-site variances with the proposed method under 15 different scenarios (three sets of population causal parameters by five sets of sample sizes) when the propensity score models are correctly specified. As shown in Table 2, the sample estimates of the population average direct effect and indirect effect contain minimal bias. The variance and covariance estimates appear to be unbiased when $N$ is relatively large and show a slight increase in bias when $N$ is small. The type I error rate for variance testing is always close to the nominal rate.

In addition, we compare the estimated standard error for the population average direct effect and indirect effect estimates between the proposed estimation procedure, the procedure that ignores the sampling variability of the weight estimates, and the fully nonparametric bootstrap procedure (Goldstein, 2011). For the latter, we generate a bootstrap sample through a simple random resampling with replacement of the sites, estimate population average direct and indirect effects based on this sample, and repeat this procedure 1000 times. The standard deviation of the bootstrapped estimates provides an estimate of the standard error of each population average causal effect estimate. We construct 95% confidence intervals bounded by the 2.5th and 97.5th percentiles of the bootstrapped estimates.

Table 3 and Table 4 present the simulation results for the standard error estimates and confidence interval coverage rates of the population average direct and indirect effects. For the population average direct effect estimator, all the three approaches to standard error estimation seem to provide acceptable results. For the population average indirect effect estimator, the standard error estimated through the proposed estimation procedure always closely approximates the standard deviation of the sampling distribution. In contrast, the standard error tends to be underestimated by the estimation procedure ignoring the uncertainty in weight when the site size is relatively small. In those scenarios, the weights, estimated through a multilevel logistic regression in the first step, are more correlated across sites, leading to a higher correlation among the site-specific effect estimates. However, the correlation is overlooked in the procedure ignoring the uncertainty in weight. Moreover, the standard error tends to be overestimated by bootstrap when the site size is relatively small. We also note that, the confidence interval coverage rates obtained from all three approaches tend to show some deviations from the nominal rate when the number of sites and the site size are relatively small. The procedure ignoring the uncertainty in weight shows the greatest amount of deviation. In general, the confidence interval coverage rates converge to the nominal rate with the increase of the number of sites and of the site size. Finally, we need to highlight that, with its closed-form expression for the standard error estimator, the proposed method requires much less computation than the bootstrap. For example, it takes less than one minute to run one replication for the scenario of $J = 100$ and $n_j = 150$ with the proposed procedure, while it takes 5.5 hours with the bootstrap.

We also run simulations when the site-specific direct effect and indirect effect are not normal or when the outcome follows other distributions. In all these cases, we obtain simi-

lar findings as above. These additional results suggest that our estimation procedure is not restricted to normally distributed outcomes or normally distributed site-specific effects.

## 5. Empirical Application

In this section, we apply the above estimation procedure to the Job Corps data. Our substantive research questions are, for the population of sites represented in this study: (a) What is the average indirect effect of the treatment assignment on earnings transmitted through educational attainment? (b) What is the direct effect of the treatment assignment on earnings? (c) To what extent did the indirect effect vary across the experimental sites? (d) To what extent did the direct effect vary across the sites? (e) Was there an association between the site-specific indirect effect and direct effect?

The study included 103 experimental sites with one Job Corps center at each site. The sample size at each site ranges from 34 to 656, with a mean of 131. In total, 9,409 applicants were randomly assigned to the experimental group and 5,977 to the control group. The weekly earnings at 48 months after randomization ranged from 0 to $1,289, with a mean of $145.30 and a standard deviation of $115.13. On average, about 38% of the sampled individuals obtained an education or training credential during the 30 months after randomization. We select 26 pretreatment covariates that are theoretically associated with the mediator and the outcome, including age, gender, race, education, criminal involvement, drug use, employment, and earnings at the baseline. The analytic sample includes 8,659 individuals with non-missing outcome and non-missing mediator in the 48-month follow-up interview. We use sample weights to account for the sample and survey designs. Table 5 lists the sample means of the outcome and some pretreatment covariates across the combinations of treatment and mediator levels.

Analyzing the data from each treatment group through a multilevel logistic regression as described in Section 3.2, we predict a Job Corps participant's propensity score for obtaining an education or training credential 30 months after being assigned to Job Corps as a function of the individual's observed pretreatment characteristics and site membership. Applying the coefficient estimates obtained from analyzing the control group data, we predict a Job Corps participant's propensity score for having educational attainment under the counterfactual control condition. We then construct the weight as defined in Equation (7). Subsequently, we estimate the population average direct and indirect effects by aggregating the estimated site-specific effects over all the sites. Finally, we estimate the between-site variance and covariance of these causal effects and conduct hypothesis testing as described in Section 3.3.

*Total Program Impact.* The results indicate that the probability of obtaining an education or training credential during the 30 months after randomization among the individuals assigned to the Job Corps program was 18.27% higher than those assigned to the control group (SE = 0.01, t = 18.27, P¡0.001), but the difference did not vary significantly across sites. Job Corps programs had a significant positive impact on earnings on average; the impact varied considerably across the sites. The estimated population average ITT effect is $22.12 ($SE = 5.01, t = 4.42, p < 0.001$), which amounts to about 11.80% of a standard deviation of the outcome. The between-site standard deviation of the ITT effect is estimated to be $25.21 ($p = 0.05$). Therefore, if we assume that the site-specific ITT effect is approximately normally distributed, in 95% of the sites, the ITT effect may range from -$27.30 to $71.54. Apparently, the Job Corps centers were not equally effective in improving earnings.

*Population Average Direct and Indirect Effects.* We decompose the total ITT effect on earnings into an indirect effect mediated through educational attainment and a direct

**Table 2**: Simulation Results for the Estimation of the Population Average Effects and Between-Site Variances

| | $J = 100$ | | | $J = 20$ | |
|---|---|---|---|---|---|
| | $n_j = 20$ | $n_j = 150$ | Job Corps site size | $n_j = 20$ | $n_j = 150$ |
| | | | Parameter Set 1 | | |
| *Direct Effect* | | | | | |
| Bias of $\hat{\gamma}^{(D)}$[1] | -0.002 | 0.000 | 0.000 | -0.007 | 0.002 |
| Bias of $\hat{\sigma}_D^2$[2] | 0.030 | 0.002 | 0.004 | 0.041 | 0.003 |
| Type I error (%)[3] of $H_0 : \sigma_D^2 = 0$ | 5.90 | 5.70 | 4.90 | 5.30 | 4.60 |
| *Inirect Effect* | | | | | |
| Bias of $\hat{\gamma}^{(I)}$ | 0.002 | 0.000 | 0.000 | 0.001 | 0.000 |
| Bias of $\hat{\sigma}_I^2$ | 0.002 | 0.000 | 0.000 | 0.003 | 0.000 |
| Type I error (%) of $H_0 : \sigma_I^2 = 0$ | 5.10 | 6.00 | 4.90 | 5.30 | 5.40 |
| Bias of $\hat{\sigma}_{D,I}$ | -0.004 | 0.000 | 0.000 | -0.007 | 0.000 |
| | | | Parameter Set 2 | | |
| *Direct Effect* | | | | | |
| Bias of $\hat{\gamma}^{(D)}$ | 0.004 | 0.000 | 0.001 | 0.001 | -0.001 |
| Bias of $\hat{\sigma}_D^2$ | 0.022 | 0.000 | 0.001 | 0.027 | -0.002 |
| *Inirect Effect* | | | | | |
| Bias of $\hat{\gamma}^{(I)}$ | -0.004 | 0.000 | -0.001 | -0.004 | -0.003 |
| Bias of $\hat{\sigma}_I^2$ | -0.002 | 0.001 | 0.001 | -0.004 | 0.000 |
| Bias of $\hat{\sigma}_{D,I}$ | 0.001 | 0.000 | 0.000 | 0.000 | -0.001 |
| | | | Parameter Set 3 | | |
| *Direct Effect* | | | | | |
| Bias of $\hat{\gamma}^{(D)}$ | 0.011 | -0.001 | 0.001 | 0.003 | -0.004 |
| Bias of $\hat{\sigma}_D^2$ | 0.017 | -0.003 | -0.002 | 0.013 | -0.003 |
| *Inirect Effect* | | | | | |
| Bias of $\hat{\gamma}^{(I)}$ | -0.010 | 0.000 | -0.001 | -0.004 | 0.001 |
| Bias of $\hat{\sigma}_I^2$ | -0.005 | 0.002 | 0.001 | -0.005 | 0.001 |
| Bias of $\hat{\sigma}_{D,I}$ | 0.007 | 0.000 | 0.000 | 0.007 | 0.001 |

*Note.* a. To enable comparisons between the different scenarios, bias of the population average effect estimate is computed as the difference between the average of the estimates across the 1000 replications and the true value, standardized by the average within-site standard deviation of the outcome in the control group. b. To make different scenarios comparable, bias of the variance estimate is computed as the difference between the average of the variance estimates across the 1000 replications and the true value, standardized by the average within-site variance of the outcome in the control group. c. The Type I error rate is computed for the null hypothesis test of the between-site variance of the direct effect and that of the indirect effect when the nominal level is set to 0.05.

**Table 3**: Simulation Results for the Standard Error Estimate and Confidence Interval Coverage Rate of the Population Average Direct Effect Estimate ($\hat{\gamma}^{(D)}$)

| | $J = 100$ | | | $J = 20$ | |
|---|---|---|---|---|---|
| | $n_j = 20$ | $n_j = 150$ | Job Corps site size | $n_j = 20$ | $n_j = 150$ |
| | | | Parameter Set 1 | | |
| Empirical SE[1] | 0.045 | 0.016 | 0.020 | 0.101 | 0.037 |
| Relative bias of SE (%)[2] | | | | | |
| Proposed Method | -1.90 | 1.10 | 1.60 | -3.50 | -3.00 |
| Ignore Uncertainty in $\hat{W}_{ij}$ | -1.80 | 1.30 | 1.70 | -3.40 | -2.90 |
| Bootstrap | 3.40 | 3.30 | 3.40 | -1.60 | -5.00 |
| 95% CI coverage (%)[3] | | | | | |
| Proposed Method | 94.30 | 94.50 | 94.70 | 92.50 | 94.10 |
| Ignore Uncertainty in $\hat{W}_{ij}$ | 94.20 | 94.70 | 94.70 | 92.60 | 94.10 |
| Bootstrap | 94.00 | 95.10 | 95.00 | 93.50 | 93.30 |
| | | | Parameter Set 2 | | |
| Empirical SE | 0.047 | 0.025 | 0.026 | 0.104 | 0.056 |
| Relative bias of SE (%) | | | | | |
| Proposed Method | -1.10 | -1.30 | 6.40 | -0.10 | -2.80 |
| Ignore Uncertainty in $\hat{W}_{ij}$ | -0.30 | -0.80 | 6.90 | 0.90 | -2.20 |
| Bootstrap | -1.90 | -0.40 | -4.50 | -1.80 | -5.70 |
| 95% CI coverage (%) | | | | | |
| Proposed Method | 94.50 | 94.80 | 96.10 | 93.80 | 92.80 |
| Ignore Uncertainty in $\hat{W}_{ij}$ | 94.70 | 94.90 | 96.10 | 94.20 | 92.90 |
| Bootstrap | 94.20 | 94.80 | 93.10 | 94.60 | 92.20 |
| | | | Parameter Set 3 | | |
| Empirical SE | 0.047 | 0.029 | 0.033 | 0.104 | 0.063 |
| Relative bias of SE (%) | | | | | |
| Proposed Method | 1.40 | -0.70 | -4.50 | -0.10 | 0.20 |
| Ignore Uncertainty in $\hat{W}_{ij}$ | 6.50 | 1.70 | -2.30 | 5.60 | 2.90 |
| Bootstrap | -6.70 | 0.10 | -2.90 | -1.80 | -2.80 |
| 95% CI coverage (%) | | | | | |
| Proposed Method | 94.40 | 95.00 | 93.70 | 93.50 | 92.80 |
| Ignore Uncertainty in $\hat{W}_{ij}$ | 95.90 | 95.60 | 94.10 | 95.20 | 93.80 |
| Bootstrap | 94.40 | 96.10 | 95.20 | 93.70 | 92.30 |

*Note.* a. Empirical SE (standard error), $\mathrm{SE}(\hat{\gamma}^{(D)})$, is the standard deviation of the sampling distribution of the average direct effect estimates, approximated by the standard deviation of the sample estimates of direct effects over the 1,000 replications. It is also standardized. b. Relative bias of SE is the relative bias of the estimated standard error, computed as $E[\widehat{\mathrm{SE}}(\hat{\gamma}^{(D)})]/\mathrm{SE}(\hat{\gamma}^{(D)}) - 1$. c. 95% CI coverage rate is the coverage probability of the 95% confidence interval estimate of the direct effect. We construct the bootstrap confidence intervals nonparametrically from the 2.5th and 97.5th percentiles of the set of empirical bootstrap values.

**Table 4**: Simulation Results for the Standard Error Estimate and Confidence Interval Coverage Rate of the Population Average Indirect Effect Estimate ($\hat{\gamma}^{(I)}$)

| | $J = 100$ | | | $J = 20$ | |
|---|---|---|---|---|---|
| | $n_j = 20$ | $n_j = 150$ | Job Corps site size | $n_j = 20$ | $n_j = 150$ |
| | | | Parameter Set 1 | | |
| Empirical SE[1] | 0.011 | 0.004 | 0.005 | 0.029 | 0.009 |
| Relative bias of SE (%)[2] | | | | | |
| Proposed Method | -2.30 | -2.20 | -1.10 | -3.80 | -0.50 |
| Ignore Uncertainty in $\hat{W}_{ij}$ | -1.00 | 0.60 | 0.90 | -2.10 | 2.50 |
| Bootstrap | 43.5 | 6.60 | 5.40 | 38.10 | 6.40 |
| 95% CI coverage (%)[3] | | | | | |
| Proposed Method | 94.40 | 94.80 | 94.70 | 94.50 | 93.70 |
| Ignore Uncertainty in $\hat{W}_{ij}$ | 94.40 | 95.10 | 94.80 | 93.90 | 94.70 |
| Bootstrap | 97.60 | 95.00 | 95.00 | 99.40 | 95.80 |
| | | | Parameter Set 2 | | |
| Empirical SE | 0.022 | 0.020 | 0.021 | 0.056 | 0.045 |
| Relative bias of SE (%) | | | | | |
| Proposed Method | 2.40 | 2.70 | -0.90 | -3.90 | -0.20 |
| Ignore Uncertainty in $\hat{W}_{ij}$ | -5.00 | 1.30 | -2.40 | -9.80 | -1.10 |
| Bootstrap | 29.50 | 5.80 | 3.80 | 21.30 | 0.90 |
| 95% CI coverage (%) | | | | | |
| Proposed Method | 92.90 | 96.60 | 94.40 | 92.10 | 93.10 |
| Ignore Uncertainty in $\hat{W}_{ij}$ | 91.90 | 96.10 | 93.80 | 90.40 | 92.80 |
| Bootstrap | 95.30 | 95.90 | 94.00 | 97.20 | 93.50 |
| | | | Parameter Set 3 | | |
| Empirical SE | 0.033 | 0.027 | 0.028 | 0.078 | 0.063 |
| Relative bias of SE (%) | | | | | |
| Proposed Method | 1.40 | -0.40 | -1.70 | 1.10 | -3.50 |
| Ignore Uncertainty in $\hat{W}_{ij}$ | -21.60 | -5.30 | -7.20 | -18.90 | -8.10 |
| Bootstrap | 18.30 | 4.40 | 1.60 | 17.00 | -3.10 |
| 95% CI coverage (%) | | | | | |
| Proposed Method | 93.10 | 95.40 | 94.50 | 92.10 | 93.70 |
| Ignore Uncertainty in $\hat{W}_{ij}$ | 82.70 | 93.80 | 92.10 | 84.40 | 91.90 |
| Bootstrap | 96.20 | 95.30 | 94.10 | 96.00 | 93.50 |

*Note.* a. Empirical SE (standard error), $\mathrm{SE}(\widehat{\gamma}^{(I)})$, is the standard deviation of the sampling distribution of the average indirect effect estimates, approximated by the standard deviation of the sample estimates of indirect effects over the 1,000 replications. It is also standardized. b. Relative bias of SE is the relative bias of the estimated standard error, computed as $E[\widehat{\mathrm{SE}}(\widehat{\gamma}^{(I)})]/\mathrm{SE}(\widehat{\gamma}^{(I)}) - 1$. c. 95% CI coverage rate is the coverage probability of the 95% confidence interval estimate of the indirect effect. We construct the bootstrap confidence intervals nonparametrically from the 2.5th and 97.5th percentiles of the set of empirical bootstrap values.

**Table 5**: Sample Statistics by Treatment and Mediator

| Variable | Treatment Group | | Control Group | |
|---|---|---|---|---|
| Educational Attainment | Yes | No | Yes | No |
| Outcome Measure (in 1995 dollars) | | | | |
| Weekly Earnings | 244.89 | 193.14 | 224.28 | 181.94 |
| Part of the Pretreatment Covariates (percentage) | | | | |
| Gender | | | | |
| Female | 0.44 | 0.45 | 0.47 | 0.44 |
| Male | 0.56 | 0.55 | 0.53 | 0.56 |
| Age | | | | |
| 16-17 | 0.44 | 0.39 | 0.50 | 0.41 |
| 18-19 | 0.33 | 0.31 | 0.32 | 0.32 |
| 20-24 | 0.23 | 0.3 | 0.19 | 0.27 |
| Race | | | | |
| Hispanic | 0.17 | 0.16 | 0.20 | 0.16 |
| Black | 0.45 | 0.51 | 0.45 | 0.49 |
| Arrested Before Application | | | | |
| Serious | 0.04 | 0.05 | 0.05 | 0.04 |
| Non-Serious | 0.18 | 0.17 | 0.19 | 0.18 |
| Baseline Earnings | | | | |
| No Earnings | 0.32 | 0.36 | 0.33 | 0.36 |
| 0 to 1,000 | 0.10 | 0.11 | 0.13 | 0.11 |
| 1,000 to 5,000 | 0.30 | 0.27 | 0.28 | 0.27 |
| 5,000 to 10,000 | 0.16 | 0.12 | 0.13 | 0.13 |
| ¿=10,000 | 0.06 | 0.07 | 0.07 | 0.06 |
| Baseline Education | | | | |
| Had a HS diploma | 0.13 | 0.23 | 0.10 | 0.21 |
| Had a GED | 0.03 | 0.06 | 0.03 | 0.06 |
| Had a vocational degree | 0.01 | 0.02 | 0.01 | 0.02 |
| Sample Size | 2,081 | 3,121 | 779 | 2,678 |

effect that channels the Job Corps impact through other services. The estimated population average indirect effect is \$8.48 ($SE = 1.40, t = 6.06, p < 0.001$), about 4.52% of a standard deviation of the outcome. The estimated population average direct effect is \$13.36 ($SE = 5.20, t = 2.57, p = 0.01$), about 7.12% of a standard deviation of the outcome. According to these results, on average, the change in educational attainment induced by the program significantly increased earnings, while other supplemental services available to the Job Corps participants in contrast with services available to those under the control condition also seemed to play a crucial role in explaining the program mechanisms.

*Between-Site Variance of Direct and Indirect Effects.* To explain why some sites seemed to be more effective than others, we further investigate between-site heterogeneity in the causal mediation mechanism. The between-site standard deviation of the indirect effect is estimated to be only \$5.60 ($p = 0.05$), while the estimated between-site standard deviation of the direct effect is as large as \$24.65 ($p = 0.045$). Based on these estimates, we can infer that the mediating role of educational attainment was nearly universal over all the sites. Yet the site-specific direct effect may range widely from negative to positive, suggesting that some sites were much more effective than others in promoting economic independence through services above and beyond increasing educational attainment. Hence, the variation in the Job Corps impact across the sites is mainly explained by the heterogeneity in the direct effect. Indeed, the national Job Corps office and regional offices centrally standardized the provision of education and strictly regulated vocational training programs for all the Job Corps centers, which might greatly limit between-site variation in education and training. In contrast, the management of other services was left largely to the discretion of each local center. As revealed in a qualitative process analysis (Johnson et al., 1999), the quantity and quality of supplemental services varied by a great amount across the Job Corps centers. We have additionally found that the estimated covariance between the site-specific direct and indirect effects is only 3.61, which corresponds to a correlation of 0.03.

*Sensitivity Analysis.* As discussed in Section 2.2, the proposed procedure identifies the causal parameters only when the sequential ignorability assumption holds. In a multisite randomized trial, the assumption of ignorable treatment assignment within each site may be easy to satisfy. However, the assumption of ignorable mediator value assignment under each treatment condition within levels of the observed pretreatment covariates is particularly strong. This assumption becomes implausible if posttreatment or unmeasured pretreatment covariates imply hidden bias that could alter the conclusion. Hence, comprehensive measurement of pretreatment and posttreatment covariates is an essential premise for valid causal inference in mediation analysis. If a pretreatment covariate that affects both the mediator and the outcome is unobserved, sensitivity analysis could be employed (Imai et al., 010a; Imai et al., 010b;VanderWeele, 010a) to assess the extent to which the possible omission might invalidate inference about the joint distribution of the site-specific direct and indirect effects. We extend the bias formulas proposed by VanderWeele (010a) to multisite mediation analysis. In addition to assessing the potential bias in the estimated population average direct effect and indirect effect, we also assess the potential bias in the between-site variance of the direct effect and indirect effect. The details can be found in Appendix E. We speculate that an omitted pretreatment confounder, such as academic achievement or self-regulation skills at the baseline, might have an impact comparable to the confounding impact of race, gender or baseline earnings that, if omitted, would contribute the greatest amount of bias among the observed pretreatment covariates. After additionally removing the potential bias of such an omitted confounder, the population average direct effect and indirect effect estimates remain statistically significant. The same is true with the between-site variance of the direct effect. Hence, we tentatively conclude that our results are insensitive to the existence of unmeasured pretreatment confounders.

If a posttreatment covariate exists, one may extend the proposed approach to a causal

mediation analysis involving two consecutive mediators by viewing the posttreatment covariate as a mediator that precedes the focal mediator (Hong, 2015; Huber, 2014). For example, we have found that Job Corps programs reduced victimization and criminal involvement and in the meantime increased access to drug and alcohol treatment during the 12 months after randomization. These intermediate experiences, in theory, might remove barriers to educational attainment and to future earnings. Extending the RMPW strategy to an analysis of multiple mediators (Hong, 2015; Huber, 2014; Lange et al., 2014) in multisite trials is an immediate topic on the research agenda. Sensitivity analysis for unobserved posttreatment confounders is a topic of emerging interest (e.g., Albert and Nelson, 2011; Tchetgen et al., 2012; Imai and Yamamoto, 2013).

## 6. Discussion

This paper has shown that, aided by methodological development in multi-site causal mediation analysis, researchers can generate new empirical evidence important for advancing social scientific knowledge. Interventions such as Job Corps must be delivered by local agents who differ in their professional capacity for engaging participants in critical elements of the program. The composition of the client population and their needs may not be identical across the sites. Moreover, the job market and alternative programs available to the client population may differ across the localities as well. A multi-site randomized trial offers unique opportunities to empirically examine the program theory across these different contexts.

Estimating and testing the between-site variance of the indirect effect in addition to that of the direct effect and quantifying the correlation between the two have been a major challenge in multi-site causal mediation analysis. This is because, in the standard regression-based approach, the indirect effect is represented as a product of multiple regression coefficients that may vary and co-vary between the sites. The complexity increases exponentially in the presence of treatment-by-mediator interaction as well as treatment-by-covariate or mediator-by-covariate interactions. The standard regression approach tends to be constrained, with few exceptions, to mediators and outcomes that are multivariate normal. A computationally intensive bootstrap procedure has been typically recommended for assessing the standard error of each causal effect estimate.

In this study, we have extended the RMPW strategy to multi-site causal mediation analysis. The simplicity of this weighting strategy brings multiple benefits. It does not require any assumption about the functional form of the outcome model; nor does it invoke any distributional assumption about the site-specific effects. Therefore, the method can be applied to outcomes measured on various scales as long as each causal effect can be defined as a mean contrast between two potential outcomes. A method-of-moments procedure applied to the weighted data generates estimates of all the causal parameters that define the joint distribution of the site-specific direct effect and indirect effect. In addition, there is virtually no constraint on the mediator distribution because RMPW is suitable for any discrete mediators (Hong, 2015; Hong et al., 2011, 2015) and because a mathematical equivalent of RMPW (Huber, 2014) easily handles continuous mediators. Hence, we conclude that the proposed strategy has considerably greater applicability than the existing methods.

We have additionally made several improvements to the estimation and hypothesis testing. The propensity score-based weights must be estimated from the sample data pooled over all the sites in the first step before the causal parameters can be estimated in the second step. To fully account for the sampling variability in the two-step estimation, we have derived a consistent estimator of the asymptotic standard error for each causal effect estimator. This solution may be applied generally to other propensity score-based two-step estimation problems in analyses of multilevel data. The results of our simulation compar-

isons suggest that the estimated asymptotic standard errors often outperform not only the standard error estimators ignoring the step 1 estimation but also the bootstrapped standard errors. Finally, given that the test statistic for the between-site variance of the direct effect and that for the indirect effect do not follow a theoretical Chi-squared distribution, we have implemented a permutation test that produces valid statistical inference. We acknowledge other potential limitations of the proposed procedure. Although our simulations have shown satisfactory results under a number of common scenarios represented by past multi-site trials, we anticipate that the current procedure may not be optimal when site sizes and the number of sites are extremely small. Moreover, when selection mechanisms vary across sites each of a relatively small sample size, propensity score models may become overfitted. In such scenarios, the lack of precision of the site-specific causal effect estimates would likely destabilize the estimation of the between-site variance-covariance matrix. In general, a reduction in site size reduces the amount of information and hence minimizes the statistical power for detecting meaningful between-site differences regardless of what analytic strategy one employs. This has direct implications for the design of a multi-site trial. The proposed MOM procedure is robust to the violation of distributional assumptions, at the cost of losing efficiency. In contrast, MLE improves efficiency by relying on stronger assumptions, as discussed at the beginning of Section3. In future research, we will investigate the feasibility of employing MLE in step 2 and derive the asymptotic standard error estimator accordingly. We will also explore an alternative estimation procedure based on Bayesian methods. The Bayesian perspective views parameters as random and naturally accounts for uncertainty in the propensity score weighting through the specification of prior distributions of propensity score model parameters. Compared to the proposed MOM approach, the Bayesian method is unconstrained by a small sample size per site and is expected to be more flexible for investigating complex mediation mechanisms and their between-site heterogeneity.

## 7. Appendix A. Asymptotic sampling variance of the estimators in the two steps

As a supplement to Section 3.2.4, this Appendix shows the details of the asymptotic sampling variance of the estimators in the two steps.

Specifically, the step-1 estimators $\widehat{\boldsymbol{\eta}}_t = (\widehat{\boldsymbol{\alpha}}_t', \mathrm{v}(\widehat{\mathbf{F}}_t)')'$ for $t = 0, 1$, where $\mathrm{v}(\widehat{\mathbf{F}}_t)$ is a vector of all the elements on or below the diagonal of $\widehat{\mathbf{F}}_t$, solve the following estimating equations,

$$\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{h}_{tij}^{(1)}(M_{ij}, T_{ij}, \mathbf{X}_{tij}, \mathbf{C}_{tij}, \boldsymbol{\theta}_{tj}, \boldsymbol{\eta}_t) = \mathbf{0},$$

where $N = \sum_{j=1}^{J} n_j$ is the total sample size of individuals and $\mathbf{h}_{tij}^{(1)}$ are score functions with the same dimension as $\boldsymbol{\eta}_t = (\boldsymbol{\alpha}_t', \mathrm{v}(\mathbf{F}_t)')'$. The above equation is essentially the first-order condition for the maximum-likelihood estimators in multilevel logistic regression. We use $\mathbf{h}_{ij}^{(1)} = (\mathbf{h}_{0ij}^{(1)'}, \mathbf{h}_{1ij}^{(1)'})'$ to denote the moment functions for the step-1 estimators $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\eta}}_0', \widehat{\boldsymbol{\eta}}_1')'$. Details on the derivation of $\mathbf{h}_{ij}^{(1)}$ can be found in the supplementary material.

In step 2, in order to estimate the site-specific direct and indirect effects, we estimate the site-specific means of the three potential outcomes identified by $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_J')'$, in which $\boldsymbol{\mu}_j = (\mu_{0j}, \mu_{*j}, \mu_{1j})'$, for $j = 1, \ldots, J$. We obtain the estimators specifically for site $s$, $\widehat{\boldsymbol{\mu}}_s = (\widehat{\mu}_{0s}, \widehat{\mu}_{*s}, \widehat{\mu}_{1s})'$, by solving the following moment conditions:

$$\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}h_{ij,0s}^{(2)}(Y_{ij}, T_{ij}, \mu_{0s}) = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \mu_{0s})(1 - T_{ij})I(S_{ij} = s) = 0,$$

$$\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}h_{ij,*s}^{(2)}(Y_{ij}, T_{ij}, W_{ij}, \mu_{*s}) = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \mu_{*s})W_{ij}T_{ij}I(S_{ij} = s) = 0,$$

$$\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}h_{ij,1s}^{(2)}(Y_{ij}, T_{ij}, \mu_{1s}) = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \mu_{1s})T_{ij}I(S_{ij} = s) = 0,$$

in which $W_{ij}$ is estimated based on the first-step estimators, $\widehat{\boldsymbol{\eta}}$, while $I(S_{ij} = s)$ is an indicator taking value 1 if individual $i$ is from site $s$ and 0 otherwise. In this second-step estimation, the moment functions are $\mathbf{h}_{ij}^{(2)} = (h_{ij,01}^{(2)}, h_{ij,*1}^{(2)}, h_{ij,11}^{(2)}, \ldots, h_{ij,0J}^{(2)}, h_{ij,*J}^{(2)}, h_{ij,1J}^{(2)})'$.

The estimators in the two steps can be rewritten as a one-step estimator $\widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\eta}}', \widehat{\boldsymbol{\mu}}')'$. Stacking the moment functions from both steps, we have that $\mathbf{h}_{ij} = (\mathbf{h}_{ij}^{(1)'}, \mathbf{h}_{ij}^{(2)'})'$. The asymptotic covariance matrix of $\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}$ is $\widetilde{\mathrm{var}}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})/N$, as shown in Equation (17).

$$\widetilde{\mathrm{var}}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}) = \begin{pmatrix} \widetilde{\mathrm{var}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) & \widetilde{\mathrm{cov}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}, \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \\ \widetilde{\mathrm{cov}}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) & \widetilde{\mathrm{var}}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \end{pmatrix} = \mathbf{R}^{-1}\mathbf{H}(\mathbf{R}^{-1})',$$

where

$$\mathbf{H} = E\left[\mathbf{h}_{ij}\mathbf{h}_{ij}'\right] = E\begin{bmatrix} \mathbf{h}_{ij}^{(1)}\mathbf{h}_{ij}^{(1)\prime} & \mathbf{h}_{ij}^{(1)}\mathbf{h}_{ij}^{(2)\prime} \\ \mathbf{h}_{ij}^{(2)}\mathbf{h}_{ij}^{(1)\prime} & \mathbf{h}_{ij}^{(2)}\mathbf{h}_{ij}^{(2)\prime} \end{bmatrix};$$

$$\mathbf{R} = E\left[\frac{\partial \mathbf{h}_{ij}}{\partial \boldsymbol{\vartheta}}\right] = E\begin{bmatrix} \frac{\partial \mathbf{h}_{ij}^{(1)}}{\partial \boldsymbol{\eta}} & \mathbf{0} \\ \frac{\partial \mathbf{h}_{ij}^{(2)}}{\partial \boldsymbol{\eta}} & \frac{\partial \mathbf{h}_{ij}^{(2)}}{\partial \boldsymbol{\mu}} \end{bmatrix}.$$

Details on the derivation of $\widetilde{\mathrm{var}}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})$ are included in the supplementary material. We estimate $\mathbf{H}$ with $\widehat{\mathbf{H}} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\widehat{\mathbf{h}}_{ij}\widehat{\mathbf{h}}_{ij}'$, and estimate $\mathbf{R}$ with $\widehat{\mathbf{R}} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{\partial \mathbf{h}_{ij}}{\partial \boldsymbol{\vartheta}}|_{\widehat{\boldsymbol{\vartheta}}}$. According to Lemma 3.3 of Hansen (1982), plim $\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{H}}(\widehat{\mathbf{R}}^{-1})' = \mathbf{R}^{-1}\mathbf{H}(\mathbf{R}^{-1})'$. We thus obtain the consistent estimator of the asymptotic sampling variance of the estimators in the two steps.

## 8. Appendix B. Method-of-moments estimator for the between-site variance

Applying the method-of-moments approach, we estimate the between-site variance as follows.

Let $\mathbf{G} = \sum_{j=1}^{J}(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\gamma}})(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\gamma}})'$, then

$$E(\mathbf{G}) = \sum_{j=1}^{J} E[(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma}) - (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})][(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma}) - (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})]'$$

$$= \sum_{j=1}^{J} E[(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})' - (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})' - (\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})' + (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})']$$

in which

$$E(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})' = \mathrm{var}(\widehat{\boldsymbol{\beta}}_j) = \mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j + \boldsymbol{\beta}_j) = \mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + \mathrm{var}(\boldsymbol{\beta}_j);$$

$$E(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})' = E(\frac{1}{J}\sum_{j'}\widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\gamma})(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})' = \frac{1}{J}\sum_{j'} E(\widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\gamma})(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})';$$

$$E(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})' = E(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})(\frac{1}{J}\sum_{j'}\widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\gamma})' = \frac{1}{J}\sum_{j'} E(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})(\widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\gamma})';$$

$$E(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})' = E(\frac{1}{J}\sum_{j}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})(\frac{1}{J}\sum_{j'}\widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\gamma})' = \frac{1}{J^2}\sum_{j}\sum_{j'} E(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})(\widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\gamma})'.$$

Therefore,

$$\begin{aligned}
E(\mathbf{G}) &= \sum_{j=1}^{J}(\mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + \mathrm{var}(\boldsymbol{\beta}_j)) - \frac{1}{J}\sum_{j}\sum_{j'} E(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})(\widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\gamma})' \\
&= \sum_{j=1}^{J}\mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + J\mathrm{var}(\boldsymbol{\beta}_j) - \frac{1}{J}\boldsymbol{\Psi}'\mathrm{var}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Psi} \\
&= \sum_{j=1}^{J}\mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + J\mathrm{var}(\boldsymbol{\beta}_j) - \frac{1}{J}\boldsymbol{\Psi}'(\mathrm{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathrm{var}(\boldsymbol{\beta}))\boldsymbol{\Psi} \\
&= \sum_{j=1}^{J}\mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + J\mathrm{var}(\boldsymbol{\beta}_j) - \frac{1}{J}\boldsymbol{\Psi}'(\mathrm{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}))\boldsymbol{\Psi} - \mathrm{var}(\boldsymbol{\beta}_j) \\
&= (J-1)\mathrm{var}(\boldsymbol{\beta}_j) + \sum_{j=1}^{J}\mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) - \frac{1}{J}\boldsymbol{\Psi}'\mathrm{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{\Psi}
\end{aligned}$$

Replacing $\mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ and $\mathrm{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ with the corresponding consistent estimators, as shown in Section 3.2.4, we obtain the consistent estimator for the between-site variance:

$$\widehat{\mathrm{var}}(\boldsymbol{\beta}_j) = \frac{1}{J-1}[\sum_{j=1}^{J}(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\gamma}})(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\gamma}})' - \sum_{j=1}^{J}\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + \frac{1}{J}\boldsymbol{\Psi}'\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{\Psi}]$$

## 9. Appendix C. Test for the between-site variance

As a supplement to Section 3.3, this Appendix explicates a hypothesis testing procedure for the between-site variance of the direct and indirect effects. Under the null hypothesis that the between-site variance $\sigma_D^2$ is zero, that is, $\beta_j^{(D)} = \gamma^{(D)}$ for all $j$, according to the Central Limit Theorem, $\widehat{\beta}_j^{(D)}$ converges in distribution to a normal distribution as the sample size at the site goes to infinity,

$$\frac{\widehat{\beta}_j^{(D)} - \gamma^{(D)}}{\sqrt{\text{var}(\widehat{\beta}_j^{(D)})}} \xrightarrow{d} N(0,1),$$

in which

$$\text{var}(\widehat{\beta}_j^{(D)}) = \text{var}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)}).$$

As the sample size at each site goes to infinity, the weights estimated in the first step are independent across sites, so that $\widehat{\beta}_1^{(D)}, \ldots, \widehat{\beta}_J^{(D)}$ can be viewed as independent. Therefore, the sum of squares of the standardized site-specific effect estimates converges to a $\chi^2$ distribution,

$$\sum_{j=1}^{J} \frac{(\widehat{\beta}_j^{(D)} - \gamma^{(D)})^2}{\text{var}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})} \xrightarrow{d} \chi^2(J).$$

We lose one degree of freedom by replacing $\gamma^{(D)}$ with $\widehat{\gamma}^{(D)}$,

$$\sum_{j=1}^{J} \frac{(\widehat{\beta}_j^{(D)} - \widehat{\gamma}^{(D)})^2}{\text{var}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})} \xrightarrow{d} \chi^2(J-1).$$

In the test statistic, we replace $\text{var}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})$ with $\widehat{\text{var}}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})$ :

$$Q^{(D)} = \sum_{j=1}^{J} \frac{(\widehat{\beta}_j^{(D)} - \widehat{\gamma}^{(D)})^2}{\widehat{\text{var}}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})}.$$

Due to this approximation, the distribution of the sample test statistic is not exactly $\chi^2(J-1)$. We therefore employ a permutation test proposed by (Fitzmaurice et al., 2007). The test randomly permutes the site indices, based on the idea that all permutations of the site indices are equally likely under the null. The algorithm is as follows:

*Step 1*. Calculate the test statistic, $Q_{obs}^{(D)}$, for the original sample.

*Step 2*. Randomly permute the site indices while holding fixed the site size, $n_j$. Calculate the test statistic for the permutation sample. By repeating this step 200 times, we can obtain 200 test statistics, $Q_p^{(D)}$, $p = 1, \ldots, 200$.

*Step 3*. Calculate the $p$-value of this test as the proportion of the permutation samples with $Q_p^{(D)} \geq Q_{obs}^{(D)}$.

Although many have suggested generating 1,000 permutation samples (Manly, 1997; Drikvandi et al., 2013), our simulation results have replicated the finding in Fitzmaurice et al. (2007) that 200 permutation samples are enough to give a nominal type I error rate.

## 10. Appendix D. Generation of simulation data

This Appendix explains how we generate the simulation data in Section 4. The goal is to assess the finite-sample performance of the multilevel RMPW procedure in estimating the population average and between-site variance of the direct effect and indirect effect. In the

basic mediation framework, the treatment affects the mediator, which in turn affects the outcome. Therefore, we generate the data using the following models:

$$\begin{aligned} T_{ij}|j &\sim \mathrm{B}(1, \Pr(T_{ij}=1|j)), \\ logit\{\Pr(M_{ij}=1|T_{ij}, \mathbf{X}_{ij})\} &= \alpha_{0j} + \alpha_{1j}T_{ij} + \alpha_j^{(1)}X_{1ij} + \alpha_j^{(2)}X_{2ij} + \alpha_j^{(3)}X_{3ij} \\ &\quad + \alpha_j^{(4)}X_{1ij}T_{ij} + \alpha_j^{(5)}X_{2ij}T_{ij} + \alpha_j^{(6)}X_{3ij}T_{ij}, \end{aligned}$$

$$Y_{ij} = \theta_{0j} + \theta_{1j}T_{ij} + \theta_{2j}M_{ij} + \theta_{3j}T_{ij}M_{ij} + \theta_j^{(1)}X_{1ij} + \theta_j^{(2)}X_{2ij} + \theta_j^{(3)}X_{3ij} + \varepsilon_{ij},$$

in which the confounding factors $X_1$, $X_2$, and $X_3$ are generated from identical distributions: $X_{kij} = \bar{X}_{kj} + e_{X_{kij}}$ for individual $i$ in site $j$, in which $\bar{X}_{kj} \sim N(0, 0.1)$ and $e_{X_{kij}} \sim N(0,1)$ for $k = 1, 2, 3$, so that the ICC of each confounding factor is 0.09, similar to that in the Job Corps data.

In the mediator model, we specify the values of the parameters as $\alpha_{0j} \sim N(-1, 0.01)$, $\alpha_j^{(1)} \sim N(0.4, 0.01)$, $\alpha_j^{(2)} = 0.15$, $\alpha_j^{(3)} = 0.02$, $\alpha_{1j} \sim N(0.8, 0.01)$, $\alpha_j^{(4)} \sim N(0.01, 0.0001)$, $\alpha_j^{(5)} = 0.05$, $\alpha_j^{(6)} = 0.1$, so that the population average of $\Pr(M_{ij}=1|T_{ij}=0, \mathbf{X}_{ij})$ is 0.28, with a standard deviation of 0.09, and the population average of $\Pr(M_{ij}=1|T_{ij}=1, \mathbf{X}_{ij})$ is 0.46, with a standard deviation of 0.12, which resemble the Job Corps data. We then generate for each individual observation a binary mediator $M_{ij}$ from $\mathrm{B}(1, \Pr(M_{ij}=1|T_{ij}, \mathbf{X}_{ij}))$.

In the outcome model, $\theta_{1j}$, $\theta_{2j}$ and $\theta_{3j}$ are determined by the specified values of the site-specific direct and indirect effects. Based on the expressions derived by Valeri and VanderWeele (2013) for the direct effect and indirect effect under the potential outcomes causal framework, as defined in Section 2.1, these parameters can be computed as follows:

$$\theta_{2j} = \frac{\beta_j^{(I)}}{E(\Pr(M_{ij}=1|T_{ij}=1, \mathbf{X}_{ij}, j)) - E(\Pr(M_{ij}=1|T_{ij}=0, \mathbf{X}_{ij}, j))} - \theta_{3j}$$

$$\theta_{1j} = \beta_j^{(D)} - \theta_{3j}E(\Pr(M_{ij}=1|T_{ij}=0, \mathbf{X}_{ij}, j))$$

To resemble the Job Corps data, we specify the values of the other parameters in the outcome model as $\theta_{0j} \sim N(2, 9)$, $\theta_j^{(1)} \sim N(1,1)$, $\theta_j^{(2)} = 1.6$, $\theta_j^{(3)} = 1.9$, and $\varepsilon_{ij} \sim N(0, 100)$.

## 11. Appendix E. Bias formulas for sensitivity analysis

As a supplement to sensitivity analysis in Section 5, this Appendix quantifies the bias for both the population average and betwen-site variance of the natural direct and indirect effects caused by the omission of a pretreatment confounder of the mediator-outcome relationship. We extend extend the bias fomulas proposed by VanderWeele (010a) to the multi-site trials.

Directly applying the result from VanderWeele (010a) for sensitivity analysis in single-site mediation analysis, we first quantify the bias at site $j$ caused by the omission of a pretreatment confounder of the mediator-outcome relationship denoted by $U$. Under the assumptions that (1) $U \perp\!\!\!\perp \mathbf{X}|S=j$, (2) $Y(t,m) \perp\!\!\!\perp T|S=j$, (3) $Y(t,m) \perp\!\!\!\perp M|T, \mathbf{X}, U, S=j$, (4) $M(t) \perp\!\!\!\perp T|S=j$, (5) $Y(t,m) \perp\!\!\!\perp M(t')|\ T, \mathbf{X}, U,\ S=j$, (6) $U$ is binary, (7) $E(Y|t,m,\mathbf{x}, U=1, S=j) - E(Y|t,m,\mathbf{x}, U=0, S=j) = \lambda_j$ across strata of $t, m, \mathbf{x}$ at site $j$, and (8) $\Pr(U=1|t,m,\mathbf{x}, S=j) - \Pr(U=1|t',m,\mathbf{x}, S=j) = \delta_j$ across strata of $m, \mathbf{x}$ at site $j$, in which $t=1, t'=0$, the bias for the direct effect at site $j$ is $\lambda_j\delta_j$. The bias for the indirect effect at site j is $-\lambda_j\delta_j$. Correspondingly, the

bias in the population average direct effect and indirect effect are $E(\lambda_j\delta_j)$ and $-E(\lambda_j\delta_j)$, respectively. Finally, the respective bias in the between-site variance of the direct effect and that of the indirect effect are

$$\text{var}(\beta_j^{(D)} + \lambda_j\delta_j) - \text{var}(\beta_j^{(D)}) = \text{var}(\lambda_j\delta_j) + 2\text{cov}(\beta_j^{(D)}, \lambda_j\delta_j),$$

$$\text{var}(\beta_j^{(I)} - \lambda_j\delta_j) - \text{var}(\beta_j^{(I)}) = \text{var}(\lambda_j\delta_j) - 2\text{cov}(\beta_j^{(I)}, \lambda_j\delta_j).$$

## References


Albert, J. M. and S. Nelson (2011). Generalized causal mediation analysis. *Biometrics 67*(3), 1028–1038.

Avin, C., I. Shpitser, and J. Pearl (2005). Identifiability of path-specific effects. *Department of Statistics, UCLA*.

Bates, D., M. Maechler, B. Bolker, S. Walker, et al. (2014). lme4: Linear mixed-effects models using eigen and s4. *R package version 1*(7).

Bauer, D. J., K. J. Preacher, and K. M. Gil (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations. *Psychological methods 11*(2), 142.

Bein, E., J. Deutsch, K. Porter, X. Qin, C. Yang, and G. Hong (2015). *Technical report on two-step estimation in RMPW analysis*. Oakland, CA: MDRC.

Bind, M.-A., T. Vanderweele, B. Coull, and J. Schwartz (2016). Causal mediation analysis for longitudinal data with exogenous exposure. *Biostatistics 17*(1), 122–134.

Bloom, H., C. J. Hill, and J. Riccio (2005). *Modeling cross-site experimental differences to find out why program effectiveness varies*. New York, NY: Russell Sage Foundation.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger (2002). *Analysis of longitudinal data*. Oxford University Press.

Drikvandi, R., G. Verbeke, A. Khodadadi, and V. P. Nia (2013). Testing multiple variance components in linear mixed-effects models. *Biostatistics 14*(1), 144–159.

Fitzmaurice, G. M., S. R. Lipsitz, and J. G. Ibrahim (2007). A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics 63*(3), 942–946.

Flores, C. A. and A. Flores-Lagunes (2013). Partial identification of local average treatment effects with an invalid instrument. *Journal of Business & Economic Statistics 31*(4), 534–545.

Goldstein, H. (2011). *Multilevel statistical models*, Volume 922. John Wiley & Sons.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society 50*(4), 1029–1054.

Hedeker, D. and R. D. Gibbons (2006). *Longitudinal data analysis*, Volume 451. John Wiley & Sons.

Hirano, K. and G. W. Imbens (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology 2*(3-4), 259–278.

Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. In *Proceedings of the American Statistical Association, Biometrics Section*, pp. 2401–2415. American Statistical Association.

Hong, G. (2015). *Causality in a social world: Moderation, mediation and spill-over*. John Wiley & Sons.

Hong, G., J. Deutsch, and H. D. Hill (2011). Parametric and non-parametric weighting methods for estimating mediation effects: An application to the national evaluation of welfare-to-work strategies. In *Proceedings of the American Statistical Association, Social Statistics Section*, pp. 3215–3229. American Statistical Association.

Hong, G., J. Deutsch, and H. D. Hill (2015). Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-by-mediator interaction. *Journal of Educational and Behavioral Statistics 40*(3), 307–340.

Hong, G. and T. Nomi (2012). Weighting methods for assessing policy effects mediated by peer change. *Journal of Research on Educational Effectiveness 5*(3), 261–289.

Hong, G. and S. W. Raudenbush (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association 101*(475), 901–910.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics 29*(6), 920–943.

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–842.

Imai, K., L. Keele, and D. Tingley (2010a). A general approach to causal mediation analysis. *Psychological methods 15*(4), 309.

Imai, K., L. Keele, and T. Yamamoto (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*(1), 51–71.

Imai, K. and T. Yamamoto (2013). Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. *Political Analysis 21*(2), 141–171.

Johnson, T., M. Gritz, R. Jackson, J. Burghardt, C. Boussy, J. Leonard, and C. Orians (1999). National job corps study: Report on the process analysis. research and evaluation report series.

Kenny, D. A., J. D. Korchmaros, and N. Bolger (2003). Lower level mediation in multilevel models. *Psychological methods 8*(2), 115.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007). Experimental analysis of neighborhood effects. *Econometrica 75*(1), 83–119.

Krull, J. L. and D. P. MacKinnon (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate behavioral research 36*(2), 249–277.

Lange, T., M. Rasmussen, and L. Thygesen (2014). Assessing natural direct and indirect effects through multiple pathways. *American journal of epidemiology 179*(4), 513.

Lange, T., S. Vansteelandt, and M. Bekaert (2012). A simple unified approach for estimating natural direct and indirect effects. *American journal of epidemiology 176*(3), 190–195.

Leite, W. L., F. Jimenez, Y. Kaya, L. M. Stapleton, J. W. MacInnes, and R. Sandbach (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research*.

MacKinnon, D. P. and J. H. Dwyer (1993). Estimating mediated effects in prevention studies. *Evaluation review 17*(2), 144–158.

Manly, B. F. (1997). *Randomization, bootstrap and Monte Carlo methods in biology, 2nd edition*. London: Chapman & Hall.

Newey, W. K. (1984). A method of moments interpretation of sequential estimators. *Economics Letters 14*(2), 201–206.

Neyman, J. and K. Iwaszkiewicz (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society 2*(2), 107–180.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pp. 411–420. Morgan Kaufmann Publishers Inc.

Preacher, K. J., M. J. Zyphur, and Z. Zhang (2010). A general multilevel sem framework for assessing multilevel mediation. *Psychological methods 15*(3), 209.

Raudenbush, S. W. and H. Bloom (2015). Using multi-site randomized trials to learn about and from a distribution of program impacts. *American Journal of Evaluation 36*(4), 475–499.

Raudenbush, S. W., S. F. Reardon, and T. Nomi (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of research on Educational Effectiveness 5*(3), 303–332.

Reardon, S. F. and S. W. Raudenbush (2013). Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociological Methods & Research 33*(4), 974–987.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pp. 95–133. Springer.

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

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association 82*(398), 387–394.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics 6*(1), 34–58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association 75*(371), 591–593.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association 81*(396), 961–962.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference 25*(3), 279–292.

Seltzer, J. A. (1994). Consequences of marital dissolution for children. *Annual Review of Sociology 20*(1), 235–266.

Spybrook, J. and S. W. Raudenbush (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the institute of education sciences. *Educational Evaluation and Policy Analysis 31*(3), 298–318.

Tchetgen, E. J. T., I. Shpitser, et al. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness and sensitivity analysis. *The Annals of Statistics 40*(3), 1816–1845.

Tchetgen Tchetgen, E. J. (2013). Inverse odds ratio-weighted estimation for causal mediation analysis. *Statistics in medicine 32*(26), 4567–4580.

Valeri, L. and T. J. VanderWeele (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with sas and spss macros. *Psychological methods 18*(2), 137.

VanderWeele, T. J. (2010a). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology (Cambridge, Mass.) 21*(4), 540.

VanderWeele, T. J. (2010b). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological methods & research 38*(4), 515–544.

Vanderweele, T. J., G. Hong, S. M. Jones, and J. L. Brown (2013). Mediation and spillover effects in group-randomized trials: a case study of the 4rs educational intervention. *Journal of the American Statistical Association 108*(502), 469–482.

VanderWeele, T. J. and S. Vansteelandt (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American journal of epidemiology 172*(12), 1339–1348.

Weiss, M. J., H. S. Bloom, and T. Brock (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management 33*(3), 778–808.

Zhang, Z., M. J. Zyphur, and K. J. Preacher (2009). Testing multilevel mediation using hierarchical linear models problems and solutions. *Organizational Research Methods 12*(4), 695–719.