Journal of Educational and Behavioral Statistics
2017, Vol. 42, No. 3, pp. 308–340
DOI: 10.3102/1076998617694879

A Weighting Method for Assessing Between-Site Heterogeneity in Causal Mediation Mechanism

Xu Qin
Guanglei Hong
University of Chicago

When a multisite 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 × 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. We also offer an easy-to-use R package, `MultisiteMediation`*, that implements the proposed method. It is freely available at the Comprehensive R Archive Network (http://cran.r-project.org/web/packages/MultisiteMediation).*

Keywords: *asymptotic variance; direct effect; indirect effect; multilevel modeling; multisite randomized trials; method of moments; RMPW*

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. Multisite randomized trials and multisite natural experiments have been pervasive in these fields and often feature longitudinal data collection (Bloom, Hill, & Riccio, 2005; Raudenbush & Bloom, 2015; Spybrook & Raudenbush, 2009). Different from clustered randomized trials (also called "group randomized trials"), which only allow for the estimation of the average treatment effect because individuals in the same cluster are assigned to the same treatment condition, multisite randomized trials provide unique opportunities for investigating how the treatment impact may vary 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, Bloom, & Brock, 2014). Assessing between-site variation in the causal mechanisms may 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 (NJCS), a multisite randomized evaluation of the nation's largest job 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 & 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 multisite 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 multisite mediation analysis, our extension of a weighting method accommodates scenarios in which the treatment changes not only the mediator value but also the mediator–outcome relationship (Judd & Kenny, 1981). Because 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. Below, we explain why the new method extends and supplements the existing literature on multisite causal mediation analysis.

Taking on the challenges of multisite data, researchers (Bauer, Preacher, & Gil, 2006; Kenny, Korchmaros, & Bolger, 2003; Krull & MacKinnon, 2001; Preacher, Zyphur, & Zhang, 2010; Zhang, Zyphur, & Preacher, 2009) have proposed to embed the standard path analysis and structural equation modeling (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 in the outcome model or fails to consider possible Treatment × Mediator interaction, Mediator × Covariate interactions, or Treatment × Mediator × 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 (MLE), the above strategy typically assumes that the mediator and the outcome are multivariate normal in distribution. As others have pointed out (Imai, Keele, & Tingley, 2010; MacKinnon & Dwyer, 1993; VanderWeele & Vansteelandt, 2010), applications of path analysis and SEM to discrete mediators and outcomes face many constraints in both single-site and multisite studies.

Other researchers have specified multilevel path analysis models for analyzing data from group randomized trials (VanderWeele, 2010b; Vanderweele, Hong, Jones, & Brown, 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 × Site interactions as instruments for the mediators (Kling, Liebman, & Katz, 2007; Raudenbush, Reardon, & Nomi, 2012; Reardon & Raudenbush, 2013). With its primary interest in identifying the average effect of each mediator on the outcome, the IV method, when applied to multisite mediation analysis, does not estimate the between-site distributions of the indirect effects. A study by Bind, Vanderweele, Coull, and Schwartz (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 × Mediator interaction (e.g., Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010) have not been extended to studies of between-site heterogeneity in mediation mechanisms.

Hong (2010, 2015) and others (Hong, Deutsch, & Hill, 2011, 2015; Hong & Nomi, 2012; Huber, 2014; Lange, Rasmussen, & Thygesen, 2014; Lange, Vansteelandt, & Bekaert, 2012; Tchetgen Tchetgen, 2013; Tchetgen Tchetgen & Shpitser, 2012) have developed weighting strategies for single-site mediation analysis. Defining direct and indirect effects in terms of potential outcomes (Pearl, 2001; Robins & 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 intuitive 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. Unlike the regression-based strategies, these weighting methods allow for Treatment × Mediator interaction without having to specify the mediator–outcome relationship and the covariate–outcome relationship. 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 multisite 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 multisite 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. We have developed an easy-to-use R package, `MultisiteMediation`, that allows users to implement the proposed method.

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 multisite mediation analysis. After delineating the method-of-moments (MOM) 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 & Iwaszkiewicz, 1935; Rubin, 1978), we define the causal parameters of interest in the context of the multisite Job Corps evaluation. The NJCS is based on a national random sample of all eligible applicants to Job Corps in late 1994 and 1995. The sampled youths were assigned randomly either to an experimental condition that allowed for immediate enrollment in one of the Job Corps centers or to the control condition that forbade Job Corps enrollment for 3 years. Which Job Corps center an individual would be assigned to had been determined prior to the treatment randomization. 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 obtained 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 multisite mediation study, SUTVA implies that (a) there is no interference between sites (Hong & Raudenbush, 2006; Hudgens & Halloran, 2008), that is, 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, that is, 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 intention-to-treat [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 Equations 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. 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 σ_D^2, σ_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, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010) 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 multisite 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.

In a hypothetical experiment for causal mediation analysis, individuals within each site would be randomized to the experimental or the control condition; subsequently, individuals would be assigned at random to obtain an education credential under each treatment condition. Alternatively, the treatment assignment would be randomized within subgroups of individuals who share the same observed pretreatment characteristics; and subsequently, the randomization to obtain an education credential under each treatment condition would be conducted within subgroups of individuals who share the same observed pretreatment characteristics. These hypothetical sequential randomized designs satisfy the sequential ignorability assumption.

However, in multisite studies such as NJCS, because individuals were not randomized to receive a mediator value after the treatment randomization, Identification Assumption 2 becomes particularly strong. 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, Shpitser, & Pearl, 2005; VanderWeele, 2010b; 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. For this reason, the sequential ignorability assumption in the multisite setting is seemingly stronger than that in the single-site setting. 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, which then enables the identification of the site-specific direct and indirect effects. Here, we discuss the general case in which the treatment assignment and the mediator value assignment under each treatment condition are "ignorable" within each subgroup of individuals who share the same observed pretreatment characteristics $\mathbf{x}$.

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 & Thompson, 1952; Robins, 2000; Rosenbaum, 1987). This weighting 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 associated with $\mathbf{X}$.

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

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

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 & Nomi, 2012; Tchetgen Tchetgen & Shpitser, 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

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

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

Below we use μ_{0j}, μ_{1j}, and μ_{*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 multisite randomized trial, the average direct effect at site j, $\beta_j^{(D)}$, can be identified by a simple mean contrast:

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

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

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

Once the site-specific direct and indirect effects are identified, their joint distribution in the population can be identified as well. The weighting method, similar to the existing methods that rely on the sequential ignorability assumption, cannot remove bias associated with omitted baseline covariates; nor can it adjust for posttreatment covariates. We will show how to assess the consequences of such potential bias through sensitivity analysis.

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, 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. In Step 2, we employ an MOM estimation procedure to estimate the site-specific direct and indirect effects and the first and second moments of their joint distribution. This procedure estimates the between-site variance of the direct and indirect effects by purging the average sampling variance off the total between-site variance of these effects. However, 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 propose asymptotic variance estimators for the population average direct effect and indirect effect estimators that incorporate the sampling variability in the weight estimation. We also conduct a permutation test for variance testing.

We choose MOM rather than MLE in Step 2 for two 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, 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.

This section starts by introducing the weighted MOM 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. At the end, we explain the estimation and hypothesis testing for the between-site variance of the direct and indirect effects.

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 and indirect effect estimates (e.g., Diggle, Heagerty, Liang, & Zeger, 2002; Raudenbush & Bloom, 2015). 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 control mean outcome μ_{0j} and the experimental mean outcome μ_{1j} 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 MOM estimators of the site-specific direct and 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

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

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 2×2 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. We do not opt for precision weighting in this study.

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

In a typical multisite 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 MOM 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 intercept 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:

$$\log\left[\frac{p_{tij}}{1-p_{tij}}\right] = \mathbf{X}_{tij}'\boldsymbol{\alpha}_t + \mathbf{C}_{tij}'\mathbf{r}_{tj}, \mathbf{r}_{tj} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_t), \tag{15}$$

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, and $\mathbf{C}_{tij}$ is a vector of covariates with random effects $\mathbf{r}_{tj}$. For computational simplicity, following Hedeker and Gibbons (2006), we standardize the random effects $\mathbf{r}_{tj}$ 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 $\boldsymbol{\Sigma}_t$, 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 MLE using iterative generalized last squares. 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} and 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, random effects 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 estimating equations 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 causal inference 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 et al. (2015) applied the strategy to RMPW-based single-site causal mediation analysis. Here we adapt the estimation procedure to multisite 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, available in the online version of the journal, 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):

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

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 $\widetilde{\mathrm{var}}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})$ can be found in Appendix A.

Subsequently, we derive the sampling variance of the estimators for the direct and indirect effects. 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

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

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, $\mathrm{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is a $2J \times 2J$ matrix with $\mathrm{var}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_j)$ as the jth 2×2 submatrix along the diagonal. The off-diagonal elements $\mathrm{cov}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j, \widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\beta}_{j'})$, where $j \neq j'$, are nonzero due to the use of pooled data from all the sites in estimating the weights in Step 1. Relying on the consistent estimator of $\mathrm{var}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})$, we obtain the consistent estimator of $\mathrm{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. The estimator is composed of $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ and $\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j, \widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\beta}_{j'})$.

Correspondingly, for the population average direct effect and indirect effect estimators given in Equation 14, the sampling variance is

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

in which

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

where $\mathrm{var}(\boldsymbol{\beta}) = \mathbf{I}_J \otimes \mathrm{var}(\boldsymbol{\beta}_j)$. The between-site variance of the direct effect and indirect effect $\mathrm{var}(\boldsymbol{\beta}_j)$ is of key scientific interest. We discuss its estimation in the next subsection. $\mathrm{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ has been defined in Equation 18. After obtaining the consistent estimators of $\mathrm{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\mathrm{var}(\boldsymbol{\beta}_j)$, we will be able to consistently estimate the asymptotic standard errors for the estimators of the population average direct and indirect effects.

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. The total between-site variance of the site-specific effect estimator $\mathrm{var}(\widehat{\boldsymbol{\beta}}_j)$ is equal to the sum of the within-site sampling variance $\mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ and the between-site variance of the site-specific effect $\mathrm{var}(\boldsymbol{\beta}_j)$:

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}_j) = \mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j + \boldsymbol{\beta}_j) = \mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + \mathrm{var}(\boldsymbol{\beta}_j). \tag{21}$$

Hence, by subtracting the average within-site sampling variance estimator from the average total variance estimator, we obtain a consistent estimator of the between-site variance of $\boldsymbol{\beta}_j$. As shown in Appendix B, available in the online version of the journal, this estimator is

$$\begin{aligned}\widehat{\mathrm{var}}(\boldsymbol{\beta}_j) = &\frac{1}{J-1}\sum_{j=1}^{J}(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\gamma}})(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\gamma}})' + \frac{1}{J(J-1)}\sum_{j}\sum_{j'\neq j}\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j, \widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\beta}_{j'})\\ &-\frac{1}{J}\sum_{j=1}^{J}\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j).\end{aligned} \tag{22}$$

In the above equation, the sum of the first two components estimates the average total variance of $\widehat{\boldsymbol{\beta}}_j$. Here the second component provides additional adjustment for the covariance among the sampling errors of $\widehat{\boldsymbol{\beta}}_j$'s between sites. The covariances are nonzero due to the pooling of data from all the sites in Step-1 estimation. The third component estimates the average within-site sampling variance of $\widehat{\boldsymbol{\beta}}_j$. The subtraction removes the sampling variance from the total variance. 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 indirect effects. Taking the direct effect as an example, we prove in Appendix C, available in the online version of the journal, that under $H_0 : \sigma_D^2 = 0$,

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

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

$$Q^{(D)} = \sum_{j=1}^{J}\frac{(\widehat{\beta}_j^{(D)} - \widehat{\gamma}^{(D)})^2}{\widehat{\mathrm{var}}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})}. \tag{23}$$

As discussed in Section 3.2, as N increases, $\widehat{\mathrm{var}}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})$ converges to $\mathrm{var}(\widehat{\beta}_j^{(D)} - \beta_j^{(D)})$. However, when N is small, the distribution of the sample test statistic may deviate from $\chi^2(J-1)$. The same is true with the between-site variance of the indirect effect. We thus employ a permutation test proposed by Fitzmaurice, Lipsitz, and Ibrahim (2007). The test randomly permutes the site

TABLE 1.
Population Causal Parameter Specification

	Population Average		Between-Site Variation		
Parameters	$\gamma^{(D)}$	$\gamma^{(I)}$	σ_D^2	σ_I^2	$\sigma_{D,I}$
Parameter Set 1	0	0	0	0	0
Parameter Set 2	.08	.08	.04	.04	.02
Parameter Set 3	.19	.19	.06	.06	.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; the between-site variances and covariances have been standardized by the average within-site variance of the outcome in the control group.

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 multicategory mediators and binary outcomes. We implement the estimation in R, using the lme4 package (Bates, Maechler, Bolker, & Walker, 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 the previous simulation studies of multilevel mediational models (Bauer et al., 2006; Krull & MacKinnon, 2001) 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, available in the online version of the journal, 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 multisite studies. For example, the Job Corps study had over 100 sites with an average of about 150 individuals per site in the full sample. The multisite 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 causal 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). 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 positive bias when N is small. The latter apparently has to do with the increase of Heywood cases in small samples. The Type I error rate for variance testing is always close to the nominal rate.

In addition, we compare the estimated standard errors 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 propensity scores and population average direct and indirect effects based on this sample, and repeat this procedure 1,000 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.

Tables 3 and 4 present, respectively, for the population average direct effect estimator and the population average indirect effect estimator, the simulation results for the standard error estimates and confidence interval coverage rates. 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 procedure ignoring the estimation uncertainty in weight when the site size is relatively small and when the between-site variances are nonzero. In those scenarios, we observe a relatively high correlation among the site-specific indirect effect estimates. As shown in Equation 19, the asymptotic variance of the population average effect estimators is a linear combination of the elements in $\mathrm{var}(\widehat{\boldsymbol{\beta}})$ including covariances among the site-specific effect

TABLE 2.
Simulation Results for the Estimation of the Population Average Effects and Between-Site Variances

	$J = 100$			$J = 20$	
Parameter Set	$n_j = 20$	$n_j = 150$	Job Corps Site Size	$n_j = 20$	$n_j = 150$
Parameter Set 1					
Direct effect					
Bias of $\hat{\gamma}^{(D)}$[a]	−0.002	0.000	0.000	−0.007	0.002
Bias of $\hat{\sigma}_D^2$[b]	0.030	0.002	0.004	0.041	0.003
Type I error (%) for $H_0 : \sigma_D^2 = 0$	5.90	5.70	4.90	5.30	4.60
Indirect 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 (%)[c] for $H_0 : \sigma_I^2 = 0$	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
Indirect 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
Indirect 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

[a] To enable comparisons between the different scenarios, bias in the population average effect estimate is computed as the difference between the average of the estimates across the 1,000 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 in the variance estimate is computed as the difference between the average of the variance estimates across the 1,000 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 .05.

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

	$J = 100$			$J = 20$	
Parameter Set	$n_j = 20$	$n_j = 150$	Job Corps Site Size	$n_j = 20$	$n_j = 150$
Parameter Set 1					
Empirical *SE*[a]	0.045	0.016	0.020	0.101	0.037
Relative bias of *SE* (%)[b]					
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 (%)[c]					
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

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

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

	$J = 100$			$J = 20$	
Parameter Set	$n_j = 20$	$n_j = 150$	Job Corps Site Size	$n_j = 20$	$n_j = 150$
Parameter Set 1					
Empirical *SE*	0.011	0.004	0.005	0.029	0.009
Relative bias of *SE* (%)					
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 (%)					
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	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

estimates. However, these covariances are overlooked in the procedure ignoring the uncertainty in weight. We also observe that, when the between-site variance of the indirect effect increases, the magnitude of the covariance between the site-specific indirect effect estimates tends to increase accordingly, which then aggravates the bias in the standard error estimates. In the simulated scenarios, the standard error tends to be overestimated by bootstrap when the site size is relatively small.

We also note that the proposed estimation procedure generates acceptable confidence interval coverage rates. This is generally true for the bootstrapping procedure as well except for one case in which the bootstrapped standard error is a severe overestimate. The procedure ignoring the uncertainty in weight, however, generates coverage rates for the population average indirect effect that deviate notably from the nominal rate when the number of sites and the site size are relatively small. In general, for all three estimation approaches, the confidence interval coverage rates tend to 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 1 min to run one replication for the scenario of $J = 100$ and $n_j = 150$ with the proposed procedure, while it takes 5.5 hr 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 similar findings as above. Applying the proposed procedure, we have found that the estimates of the causal parameters contain minimal bias and the estimated standard errors always closely approximate the empirical standard errors. 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 for the population of sites represented in this study are (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? and (e) Was there an association between the site-specific indirect effect and direct effect?

The analytic sample includes 8,659 individuals with nonmissing outcome and nonmissing mediator in the 48-month follow-up interview. There are 100 total experimental sites with one Job Corps center at each site. The sample size at each site ranges from 24 to 417. Of all, 5,202 applicants were randomly assigned to the experimental group and 3,457 to the control group. The application that we present here has not incorporated the NJCS sample weight. Therefore, the analytic results are only illustrative. 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. Table 5 lists the sample means and standard deviations 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.

5.1 Total Program Impact

The results indicate that, 30 months after randomization, about 40% of the individuals assigned to Job Corps obtained an education or training credential; only about 22% of those assigned to the control condition obtained a credential. This stark contrast (coefficient = .18, standard error [*SE*] = 0.01, t = 18.27, p < .001) did not vary significantly across sites. Job Corps programs had a significant positive impact on earnings on average; this impact, however, varied considerably across the sites. The estimated population average ITT effect is US$16.41 ($SE = 5.30,\ t = 3.10,\ p = .002$), which amounts to about 8.75% of a standard deviation of the outcome. The between-site standard deviation of the ITT effect is estimated to be US$24.81 ($p = 0.03$). 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 −US$32.22 to US$65.04. Apparently, the Job Corps centers were not equally effective in improving earnings.

5.2 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 effect that channels the Job Corps impact through other services. The estimated population average indirect effect is US$8.68 ($SE = 1.61,\ t = 5.39,\ p < .001$), about 4.63% of a standard deviation of the outcome. The estimated population average direct effect is US$7.74 ($SE = 5.38,\ t = 1.44,\ p = .15$), about 4.13% 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,

TABLE 5.
Sample Statistics by Treatment and Mediator

	Treatment Group		Control Group	
Education Attainment	Yes	No	Yes	No
Outcome (in 1995 dollars), mean (standard deviation)				
Weekly earnings	245.41	192.84	227.83	187.09
	(214.33)	(189.31)	(189.44)	(182.63)
Pretreatment covariates[a] (proportions)				
Gender				
Female	.46 (.50)	.47 (.50)	.42 (.49)	.38 (.49)
Male	.54 (.50)	.53 (.50)	.58 (.49)	.62 (.49)
Age				
16–17	.43 (.49)	.38 (.48)	.50 (.50)	.41 (.49)
18–19	.33 (.47)	.31 (.46)	.32 (.47)	.31 (.46)
20–24	.24 (.43)	.31 (.46)	.19 (.39)	.28 (.45)
Race				
Hispanic	.18 (.38)	.16 (.37)	.19 (.39)	.17 (.37)
Black	.46 (.50)	.52 (.50)	.46 (.50)	.50 (.50)
White	.30 (.46)	.25 (.43)	.28 (.45)	.26 (.44)
Other	.07 (.26)	.07 (.26)	.07 (.26)	.07 (.26)
Arrest				
Serious	.04 (.20)	.04 (.21)	.05 (.22)	.04 (.20)
Nonserious	.18 (.38)	.17 (.38)	.20 (.40)	.18 (.38)
Never arrested	.74 (.44)	.75 (.43)	.71 (.45)	.75 (.44)
Baseline earnings				
No earnings	.33 (.47)	.36 (.48)	.33 (.47)	.36 (.48)
0–1,000	.10 (.30)	.11 (.31)	.13 (.33)	.10 (.31)
1,000–5,000	.29 (.46)	.27 (.44)	.29 (.45)	.27 (.44)
5,000–10,000	.16 (.37)	.13 (.33)	.13 (.33)	.13 (.34)
≥10,000	.06 (.23)	.07 (.25)	.07 (.25)	.06 (.25)
Baseline education				
Had HS[b] diploma	.14 (.35)	.24 (.43)	.10 (.30)	.21 (.41)
Had GED[c]	.03 (.17)	.05 (.23)	.03 (.17)	.06 (.23)
Vocational degree	.01 (.12)	.02 (.15)	.02 (.12)	.02 (.14)
Other degree	.01 (.07)	.01 (.09)	.00 (.05)	.01 (.09)
None	.81 (.39)	.69 (.46)	.85 (.35)	.72 (.45)
Sample size	2,081	3,121	779	2,678

[a] Due to the page limit, here we display an incomplete list of the pretreatment covariates. Additional information about other covariates is available from the authors. [b] HS stands for High School. [c] GED stands for General Educational Development.

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.

5.3 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 US\$7.12 ($p = .06$), while the estimated between-site standard deviation of the direct effect is as large as US\$23.76 ($p = .055$). We have additionally found that the estimated covariance between the site-specific direct and indirect effects is -48.38, which corresponds to a correlation of -0.29. Based on these estimates, we can infer that the mediating role of educational attainment was similar 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. Our results corroborate the previous qualitative findings and suggest a need to improve the quantity and quality of supplementary services especially in the Job Corps centers in which the estimated direct effect is relatively small or even negative.

5.4 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. If a pretreatment covariate that affects both the mediator and the outcome is unobserved, sensitivity analysis could be employed (Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010; VanderWeele, 2010a) to assess the extent to which the omission might invalidate inference about the direct and indirect effects. We extend the bias formulas proposed by VanderWeele (2010a) to multisite mediation analysis. In addition to assessing the potential bias in the estimated population average direct effect and indirect effect, we assess the potential bias in the between-site variance of the direct effect and indirect effect.

To assess the consequence of omitting an unmeasured pretreatment covariate denoted by U in each experimental site, we quantify the potential bias in the estimated site-specific direct and indirect effects attributable to U. In the Job Corps study, conceivably, the potential bias associated with U is perhaps comparable to the confounding impact of any of the observed pretreatment covariates. We use Bias_j to denote the potential confounding impact of U on the site-specific direct effect estimate at site j. Under a series of simplifying assumptions specified in Appendix E, available in the online version of the journal, Bias_j can be represented as a product of two terms: The first is the association between U and Y conditioning on the treatment condition t, the mediator value m, and the values of other pretreatment covariates $\mathbf{x}$ at a given site; the second is the association between U and the treatment indicator T conditioning on m and $\mathbf{x}$ at a given site. Due to the randomization of the treatment assignment at each site, U does not bias the site-specific ITT effect estimate. Hence, the potential confounding impact of U on the site-specific indirect effect estimate at site j is simply $-\text{Bias}_j$. Extending this result to the multisite mediation analysis, we assess the potential impact of U on the population average direct effect estimate $\text{Bias} = E[\text{Bias}_j]$ and that on the population average indirect effect estimate $-\text{Bias} = E[-\text{Bias}_j]$. Appendix E additionally derives the respective bias in the estimated between-site variance of the direct effect, in the estimated between-site variance of the indirect effect, and in the estimated covariance of the site-specific direct effect and indirect effect.

We have found that gender, race, age, and baseline earnings are among the observed pretreatment covariates that display the greatest confounding impact. We speculate that an omitted pretreatment confounder such as academic achievement or self-regulation skills might have a comparable confounding impact. After an additional removal of the potential bias of such an omitted confounder, the statistical inference conclusions about the population average direct effect and indirect effect would remain unchanged. The same would be true with the estimated between-site standard deviation of the direct effect and that of the indirect effect. Hence, we tentatively conclude that our results are insensitive to the existence of unmeasured pretreatment confounders. We do not rule out the possibility of alternative conclusions should the simplifying assumptions invoked in the bias formula be unwarranted.

Posttreatment covariates may be viewed as additional mediators that precede or are concurrent with the current focal mediator. 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. When the research interest is focused on a single mediator such as educational attainment, the omission of other mediators seems inevitable and can be potentially consequential. 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 another topic of emerging interest (e.g., Albert & Nelson, 2011; Imai & Yamamoto, 2013; Tchetgen Tchetgen & Shpitser, 2012; VanderWeele & Chiba, 2014).

6. Discussion

This article has shown that, aided by methodological development in multisite 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 multisite 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 multisite 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 covary between the sites. The complexity increases exponentially in the presence of Treatment × Mediator interaction as well as Treatment × Covariate or Mediator × 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 multisite 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 site-specific direct and indirect 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. An MOM procedure applied to the weighted data generates estimates of all the causal parameters that define the first two moments of 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 comparisons suggest that, for the population average indirect effect estimator in particular, the estimated asymptotic standard errors often outperform not only the standard error estimators that ignore 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 χ^2 distribution, we have implemented a permutation test that produces valid statistical inference.

We acknowledge several potential limitations of the proposed procedure. First, past research has shown that misspecifying the functional form of a propensity score model will bias the RMPW results in single-site mediation analysis (Hong, 2015; Hong et al., 2015). This is because an omission of nonlinear or nonadditive terms in the propensity score model can be viewed as an omission of potential confounders. Multisite RMPW analysis is subjected to a similar requirement that the multilevel logistic regression model must be correctly specified. This includes correct specifications of the functional form and the distribution form of random effects. We will investigate in future research the extent to which results are sensitive to possible deviations from these model-based assumptions and whether the sensitivity depends on the number of sites and the number of individuals per site in a sample. We will also explore alternative nonparametric strategies for propensity score estimation and weight estimation.

Second, although our simulations have shown satisfactory results under a number of common scenarios represented by past multisite trials, we anticipate that the current procedure may not be optimal when site sizes are extremely small. As indicated by the simulation results, a bias seems to arise, albeit small in magnitude, in the between-site variance of direct effect when the number of individuals per site is as small as 20. 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 multisite trial.

Third, in causal mediation analysis in general, without a randomization of mediator value assignment in addition to the randomization of treatment assignment, the causal validity of analytic conclusions relies entirely on the statistical adjustment for observed pretreatment covariates. Even in the absence of omitted pretreatment covariates, the results will be invalid if the focal mediator is not independent of other mediators that constitute additional pathways. This is likely the case when an intervention program has multiple complementary components and when the focal mediator singles out only one of these components. For example, there are concerns that the indirect effect transmitted through educational attainment might be confounded by an unspecified indirect effect transmitted through individual counseling. This would be the case if, among individuals sharing the same pretreatment characteristics at a site, those who are more likely to obtain an education credential are also more likely to seek counseling. Identifying and estimating indirect effects transmitted by correlated mediators remains a major methodological challenge.

As we discussed at the beginning of Section 3, the proposed MOM procedure produces robust results when the site-specific direct effect and indirect effect are not normally distributed, though at the cost of losing efficiency. In contrast, MLE improves efficiency by relying on stronger assumptions. 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 likely 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.

Authors' Note

Alma Vigil provided research assistance to the analysis of the Job Corps data.

Acknowledgments

The authors thank Donald Hedeker; Stephen Raudenbush; Kazuo Yamaguchi; Fan Yang; and their colleagues Ed Bein, Jonah Deutsch, Kristin Porter, and Cheng Yang for their contribution of ideas and their comments on earlier versions of this article. In addition, comments from the Editor and the three anonymous reviewers have led to major improvements in this article.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by 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.

References

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

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and s4* (R package version 1.1-7). Retrievable from https://cran.r-project.org/web/packages/lme4/index.html.

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

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

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

Bloom, H., Hill, C. J., & Riccio, J. (2005). Modeling cross-site experimental differences to find out why program effectiveness varies. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 37–74). New York, NY: Russell Sage Foundation.

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

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

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

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). Chichester, England: John Wiley.

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

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis* (Vol. 451). Hoboken, NJ: John Wiley.

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

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

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

Hong, G., Deutsch, J., & Hill, H. D. (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). Alexandria, VA: American Statistical Association.

Hong, G., Deutsch, J., & Hill, H. D. (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*, 307–340.

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

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

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

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

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

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

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

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

Johnson, T., Gritz, M., Jackson, R., Burghardt, J., Boussy, C., Leonard, J., & Orians, C. (1999). *National job corps study: Report on the process analysis* (Research and Evaluation Report Series 8140-510). Princeton, NJ: Mathematica Policy Research.

Judd, C. M., & Kenny, D. A. (1981). Process analysis estimating mediation in treatment evaluations. *Evaluation Review*, *5*, 602–619.

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

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

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

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

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

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research*, *50*, 265–284.

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

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

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

Pearl, J. (2001). Direct and indirect effects. In J. Breese & D. Koller (Eds.), *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufmann.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*, 209.

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

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

Reardon, S. F., & Raudenbush, S. W. (2013). Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociological Methods & Research*. doi:10.1177/0049124113494575

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp. 95–133). New York, NY: Springer.

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

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

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

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

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

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

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

Spybrook, J., & Raudenbush, S. W. (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*, 298–318.

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

Tchetgen Tchetgen, E. J., & Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, *40*, 1816.

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

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

VanderWeele, T. J., & Chiba, Y. (2014). Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator–outcome confounders. *Epidemiology, Biostatistics, and Public Health*, *11*, 1–16.

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

VanderWeele, T. J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, *172*, 1339–1348.

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

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

Authors

XU QIN is a PhD candidate at the Department of Comparative Human Development and the Committee on Education at the University of Chicago, 1126 E. 59th St., Chicago, IL 60637; email: xuqin@uchicago.edu. Her research interests are causal inference, causal mediation analysis, multilevel modeling, and program evaluation.

GUANGLEI HONG is an associate professor at the Department of Comparative Human Development and the Committee on Education at the University of Chicago, 1126 E. 59th St., Chicago, IL 60637; email: ghong@uchicago.edu. Her research interests are causal inference theories and methods, causal mediation analysis, multilevel modeling, longitudinal data analysis, policy analysis and program evaluation, and instructional effectiveness.

Manuscript received May 5, 2016
Revision received October 20, 2016
Accepted December 10, 2016