Journal of the Statistical Society, Series A: Statistics in Society. (2019) *182*, Part 4, 1343-1370.

Multisite causal mediation analysis in the presence of complex sample and survey designs and non-random nonresponse

Xu Qin
University of Pittsburgh, Pittsburgh, USA.
E-mail: xuqin@pitt.edu

Guanglei Hong
University of Chicago, Chicago, USA.
E-mail: ghong@uchicago.edu

Jonah Deutsch
Mathematica Policy Research, Chicago, USA.
E-mail: JDeutsch@mathematica-mpr.com

Edward Bein
Food and Drug Administration, Maryland, USA.
E-mail: edbein18@gmail.com

Summary. This study provides a template for multisite causal mediation analysis using a comprehensive weighting-based analytic procedure that enhances external and internal validity. The template incorporates a sample weight to adjust for complex sample and survey designs, adopts an IPTW weight to adjust for differential treatment assignment probabilities, employs an estimated nonresponse weight to account for non-random nonresponse, and utilizes a propensity score-based weighting strategy to flexibly decompose not only the population average but also the between-site heterogeneity of the total program impact. Because the identification assumptions are not always warranted, a weighting-based balance checking procedure assesses the remaining overt bias, while a weighting-based sensitivity analysis further evaluates the potential bias related to omitted confounding or to propensity score model misspecification. We derive the asymptotic variance of the estimators for the causal effects that account for the sampling uncertainty in the estimated weights. The method is applied to a re-analysis of the data from the National Job Corps Study.

Keywords: Direct effect; indirect effect; multisite randomized trial; propensity score; ratio-of-mediator-probability weighting; sensitivity analysis; treatment-by-mediator interaction.

1. Introduction

Many program evaluations simply report the estimated average treatment effects without explicitly testing the theories explaining how a program produces its intended effect. One way to test specific theories about program mechanisms is mediation analysis that, in its simplest form, decomposes the total program impact into an indirect effect--transmitted through a hypothesized focal mediator--and a direct effect--attributable to all other possible pathways. Multisite

randomized trials, in which individuals are randomly assigned to treatment and control groups within each site, offer unique opportunities for further testing program theories across a wide range of settings in which a program is implemented. Just as treatment effects may vary across sites, causal mechanisms may differ across sites due to differences in local contexts, in participant composition, and in treatment implementation (Weiss *et al.*, 2014). Hence, assessing between-site variation in the causal mechanisms may generate important information for understanding heterogeneity in the total program impact, may reveal a need to revisit the program theory, and may suggest specific site-level modifications of the intervention practice. However, due to some important constraints of existing analytic tools, analysts have rarely investigated between-site heterogeneity of mediation mechanisms in multisite program evaluations.

In a single-site study, the population of individuals residing at the site is naturally the target of inference. The causal parameter of interest is generally the treatment effect averaged over all the individuals in this site-specific population. In a multisite study, however, there are two potential targets of inference: the population of sites and the overall population of individuals which is the union of all the site-specific subpopulations (Raudenbush and Bloom, 2015; Raudenbush and Schwartz, working paper). When researchers are primarily interested in how a program is implemented at the site level and whether the program impact depends on the local settings, the population of sites clearly becomes the target of inference. In such a case, the population average treatment effect is defined as the average of the site-specific average effect over all the sites. Henceforth we call this "the average effect for the population of sites." Moreover, the between-site variance of the site-specific average effect indicates the extent to which the program impact is generalizable across the sites. In contrast, when researchers are primarily interested in the overall population of individuals served by a particular program, the population average treatment effect is simply an average over the individuals in the overall population regardless of their site membership. We call this "the average effect for the population of individuals". The average effect for the population of sites and that for the population of individuals become equivalent only when the size of the site-specific subpopulation of individuals is the same across all the sites or if the effect does not vary across sites. In this study, with a primary interest in the between-site heterogeneity of the program impacts and of the mediation mechanisms, we focus on the population of sites rather than the overall population of individuals.

The methodological development in this study is motivated by a reanalysis of the multisite experimental data evaluating Job Corps, the largest federal program designed to promote economic well-being among disadvantaged youths in the U.S. who are unemployed and not in school. Intensive education and vocational training are the central elements of the program. Besides, unlike most other training programs that have been generally found ineffective because participants tend to "have more trouble in their lives than the programs could correct" (Pouncy, 2000, p.269), Job Corps is unique in its provision of a comprehensive array of support services including residential living, supervision, behavioral counseling, social skills training, physical and mental health care, and drug and alcohol treatment. According to a nationwide evaluation of all the Job Corps centers in the mid-1990s, known as the National Job Corps Study (NJCS), Job Corps was the only federal program shown to increase earnings of disadvantaged youth; the program also improved educational attainment and employment and reduced criminal involvement (Flores and Flores-Lagunes, 2013; Frumento *et al.*, 2012; Lee, 2009; Schochet *et al.*, 2006, 2008; Zhang *et al.*, 2009).

However, no attempt has been made to formally test the Job Corps program theory. The program is intended to improve disadvantaged youths' economic well-being not only through education and training that form conventional human capital (Becker, 1964; Card, 1999) but also through comprehensive support services for reducing risk exposures and risk behaviors. Given the comprehensiveness of the program and given that support services tend to be lacking under the control condition, another interesting theoretical question is whether education and training obtained through Job Corps generated a greater impact on earnings on average than education and training obtained under the control condition. Therefore, we ask how much of the Job Corps impact on earnings is mediated by education and training and whether Job Corps enhanced the economic returns to education and training for disadvantaged youth.

Moreover, with their primary interest in the population of individuals served by Job Corps, most researchers have simply ignored the role of individual Job Corps centers in their analyses. Yet a recent study (Weiss *et al.*, 2017) reported considerable variation in the program impact on earnings across the sites, with one Job Corps center at each site. This result coincides with findings from a qualitative process analysis (Johnson *et al.*, 1999) revealing important discrepancies between the intended program and the implemented program in service provision at some centers.

In our reanalysis of the NJCS data, we intend to test the Job Corps program theory that focuses on education and training without overlooking the role of support services. Moreover, we will examine how the theory plays out differently at different sites that may explain between-site heterogeneity in the program impact. Given our interest in generating empirical evidence to inform Job Corps operation at the site level, the target of inference in this study is the population of sites rather than the overall population of individuals.

We highlight a number of challenges in such research endeavors:

Potential sampling bias due to differential sampling probabilities. NJCS drew a probability sample of individuals representative of the overall population of eligible applicants to be assigned to each of the Job Corps centers. An individual's probability of being sampled was a function of baseline characteristics. If the analyst overlooks the differential probabilities of sample selection, sample estimates of the average program impacts and of their between-site variance would contain sampling bias.

Potential treatment selection bias due to differential probabilities of treatment assignment. Rather than assigning all sampled individuals with an equal probability to either the program group or the control group, NJCS researchers let the probabilities of treatment assignment differ by personal and site-level characteristics. Ignoring the differential probabilities of treatment assignment would pose a threat to internal validity and lead to treatment selection bias.

Potential nonresponse bias due to differential probabilities of response. In NJCS, some sampled youths were lost to attrition or failed to provide information on education and training or on earnings, while some were not assigned to a specific center prior to random assignment. We define all of these individuals as nonrespondents. The sample estimates would contain nonresponse bias if non-random nonresponse changes the representativeness of the sample of individuals in longitudinal follow-ups or if the remaining sample shows systematic differences between the program group and the control group.

Potential mediator selection bias due to differential probabilities of mediator value assignment. Even if a randomized experiment does not suffer from non-random nonresponse, mediator values are typically generated through a natural process rather than being experimentally manipulated. As a result, individuals displaying different mediator values tend to

differ systematically in many other aspects that would confound the causal mediation analysis and result in mediator selection bias.

Potential bias due to model misspecification. Path analysis and structural equation modeling (SEM) (Alwin and Hauser, 1975; Baron and Kenny, 1986; Duncan, 1966; Sobel, 1982; Wright, 1934) have been the primary technique for mediation analysis in the past several decades with recent extensions to multisite data analysis (Bauer *et al.*, 2006; Kenny *et al.*, 2003; Krull and MacKinnon, 2001). These regression-based methods, however, rely heavily on correct specifications of both the mediator model and the outcome model (Hong, 2017). Recent advances in single-site causal mediation analysis (e.g., Imai *et al.*, 2010; Imai *et al.*, 2010; Pearl, 2010; Petersen *et al.*, 2006; Valeri and VanderWeele, 2013; VanderWeele and Vansteelandt, 2009, 2010; van der Laan and Petersen, 2008) have focused on accommodating treatment-by-mediator interactions within the linear SEM framework; while challenges involving the functional forms of covariates remain in model specifications.

The first three challenges are common in evaluation studies, and are often addressed via sampling weights, inverse probability of treatment weights (IPTW), and nonresponse weights, respectively. We innovatively adapt these weighting adjustments to the context of mediation analysis by combining them with the ratio-of-mediator-probability weighting (RMPW) strategy. The latter is for unpacking the causal mechanism and reducing mediator selection bias. RMPW was initially proposed by Hong (2010, 2015) and others (Bein *et al.*, 2018; Hong *et al.*, 2011, 2015; Hong and Nomi, 2012; Huber, 2014; Lange *et al.*, 2012; Tchetgen Tchetgen and Shpitser, 2012) and was recently extended to multisite studies by Qin and Hong (2017). This strategy, without invoking functional form assumptions for the outcome model, is particularly flexible for accommodating treatment-by-mediator interactions and is suitable for discrete and continuous mediators and outcomes. We assess the remaining overt bias due to possible misspecifications of propensity score models through a weighting-based balance checking procedure; and we adopt a novel weighting-based sensitivity analysis strategy for assessing hidden bias with minimal simplifying assumptions (Hong *et al.*, 2018, working paper). This series of strategies constitute a systematic and coherent template for multisite causal mediation analysis. We also address challenges to estimation and statistical inference when multiple weights are unknown and must be estimated from sample data.

We organize the paper as follows. Section 2 introduces the NJCS sample and data. Section 3 defines the causal parameters under the counterfactual causal framework. Section 4 clarifies the identification assumptions and presents our identification strategy. Section 5 outlines our approaches to estimation, statistical inference, balance checking, and sensitivity analysis. Section 6 reports the analytic results. Section 7 concludes and discusses extensions. In addition, we provide an R package "`MultisiteMediation`" (http://cran.r-project.org/web/packages/MultisiteMediation) that implements the proposed template for multisite causal mediation analysis.

2. The NJCS Sample and Data

NJCS researchers identified about 80,000 eligible applicants nationwide in the mid-1990s (Schochet *et al.*, 2001). Through a stratified sampling procedure, more than 15,000 eligible applicants were randomly selected into a nationally representative research sample and were assigned at random to either the program group or the control group. Program group members could enroll in Job Corps soon after random assignment; while control group members were barred from enrolling in Job Corps for 3 years. Applicants who were initially assigned to the

same Job Corp center, regardless of their subsequent treatment assignments, constitute the sample of individuals at the given site. Participants in the study were interviewed at baseline and at 12, 30, and 48 months after randomization. By design, the probability of selection for each follow-up survey differed across individuals.

We perform our analysis on the random sample of 14,125 youths who were targeted for the 48-month interview. The mediator, collected at the 30-month follow-up, indicates whether a youth had obtained an education credential—typically a General Educational Development (GED) certificate—or a vocational certificate (or both) since the randomization. The outcome is weekly earnings in the fourth year after randomization. Our sample contains 8,818 respondents (3,491 control group members and 5,327 program group members) and 5,307 nonrespondents (2,235 control group members and 3,072 program group members).

3. A Theoretical Model of Multisite Causal Mediation Process

We investigate the following research questions in relation to Job Corps: 1) To what extent did Job Corps increase earnings through improving educational and vocational attainment? 2) To what extent did Job Corps increase earnings through other pathways? 3) Did the improvement in educational and vocational attainment produce a greater increase in earnings under Job Corps than under the control condition? 4) Were Job Corps centers equally effective in increasing earnings through improving educational and vocational attainment? 5) Were Job Corps centers equally effective in increasing earnings through other pathways? 6) Did Job Corps enhance the economic returns to education and training in some centers but not in others? 7) Did Job Corps centers that increased earnings through improving educational and vocational attainment also tend to be successful in increasing earnings through other pathways?

Here we present a theoretical model that summarizes key information characterizing the multisite causal mediation process. We define the causal parameters under the potential outcomes framework (Holland, 1986, 1988; Neyman and Iwaszkiewicz, 1935; Rubin, 1978) that has previously been extended to causal mediation research (Pearl, 2001; Robins and Greenland, 1992). The extension focuses on the intermediate process in which one's mediator value is a potential natural response to the treatment assigned; and hence mediator values may naturally vary among individuals under the same treatment.

3.1. Potential Mediators and Potential Outcomes

Let T_{ij} denote the treatment assignment of individual i at site j. It takes values $t = 1$ for an assignment to Job Corps and $t = 0$ for the control group. Let M_{ij} denote the focal mediator and Y_{ij} denote the outcome. For individual i at site j, educational and vocational attainment is a function of the treatment assignment t. Hence, we use $M_{ij}(1)$ to represent the individual's potential attainment if assigned to Job Corps and use $M_{ij}(0)$ for the potential attainment if the same person was assigned to the control group. For each individual, only one of these two potential mediators is observable after the treatment assignment. Under treatment condition t, the individual might obtain a credential by the 30-month follow-up ($M_{ij}(t) = 1$) or might fail to do so ($M_{ij}(t) = 0$).

The individual's weekly earnings in the fourth year after randomization also depends on the treatment assignment. The convention is to use $Y_{ij}(1)$ and $Y_{ij}(0)$ to represent the potential earnings associated with an assignment to Job Corps and to the control group, respectively. Alternatively, one may view the potential outcome as a function of both the treatment assignment and the corresponding potential mediator and denote it with $Y_{ij}(t, M_{ij}(t))$ for $t =$

0, 1. When $M_{ij}(t) = m$, where $m = 0,1$, the individual's potential outcome value associated with treatment t can be written as $Y_{ij}(t, m)$. Again, only one of the two potential outcomes is observable for each individual given the treatment assignment.

In causal mediation analysis, two additional counterfactual outcomes play indispensable roles: $Y_{ij}(1, M_{ij}(0))$ is one's potential earnings if assigned to Job Corps yet counterfactually having the same attainment status as he or she would have under the control condition; and $Y_{ij}(0, M_{ij}(1))$ is the potential earnings if one was assigned to the control group yet counterfactually having the same attainment status as he or she would have under Job Corps. Because $M_{ij}(0)$ is counterfactual for program group members and $M_{ij}(1)$ is counterfactual for control group members, neither $Y_{ij}(1, M_{ij}(0))$ nor $Y_{ij}(0, M_{ij}(1))$ is directly observable for any individual.

The above potential mediators and potential outcomes are defined under the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980; Rubin, 1986; Rubin, 1990). In a single site, SUTVA implies (a) that an individual's potential mediators are not functions of the treatment assignments of other individuals, (b) that an individual's potential outcomes are not functions of the treatment assignments and the mediator values of other individuals, and (c) that an individual's potential mediators and potential outcomes do not depend on which program agents (e.g., instructors or counselors) one would encounter, which is also known as "treatment version irrelevance". This assumption would be violated, for example, in the presence of peer influence or if program agents were not equally effective (Hong, 2015). In a multisite study, SUTVA further requires "no interference between sites" (Hong and Raudenbush, 2006; Hudgens and Halloran, 2008). Because applicants are usually assigned to Job Corps centers relatively close to their original residences and because Job Corps centers are sparsely located, between-site interference seems unlikely.

3.2. Individual-Specific Causal Effects

Under SUTVA, for individual i at site j, the *intention-to-treat (ITT) effect of the treatment on the mediator*, i.e. the effect of the treatment assignment on the mediator, is defined as

$$M_{ij}(1) - M_{ij}(0);$$

the *ITT effect of the treatment on the outcome*, also known as the total effect, is defined as

$$Y_{ij}\ (1, M_{ij}(1))\ -\ Y_{ij}(0, M_{ij}(0)).$$

The individual-specific *natural indirect effect* (NIE) of the treatment on the outcome transmitted through the mediator (Pearl, 2001) is defined as

$$Y_{ij}(1, M_{ij}(1)) -\ Y_{ij}(1, M_{ij}(0)).$$

It represents the Job Corps impact on earnings attributable to the program-induced change in the individual's attainment from $M_{ij}(0)$ to $M_{ij}\ (1)$ under Job Corps. This is called "the total indirect effect" by Robins and Greenland (1992), who distinguished it from the individual-specific "*pure indirect effect*" (PIE)

$$Y_{ij}(0, M_{ij}(1)) -\ Y_{ij}(0, M_{ij}(0)).$$

It represents the impact on earnings when the individual's attainment is changed from $M_{ij}(0)$ to $M_{ij}(1)$ under the control condition.

The individual-specific *natural direct effect* of the treatment on the outcome (NDE) is defined as

$$Y_{ij}\ (1, M_{ij}(0))\ -\ Y_{ij}(0, M_{ij}(0)).$$

It represents the Job Corps impact on earnings while holding the individual's attainment at the level that would be realized under the control condition. The direct effect is nonzero if Job Corps exerted an impact on earnings without changing an individual's attainment. Robins and Greenland (1992) called this "the pure direct effect" in contrast with "the total direct effect," $Y_{ij}\ (1, M_{ij}\ (1))\ -\ Y_{ij}\ (0, M_{ij}\ (1))$. The latter is the Job Corps impact on earnings while holding attainment at the level that would be realized under Job Corps.

The individual-specific total treatment effect is the sum of the individual-specific NIE and NDE. Alternatively, one may decompose the individual-specific total treatment effect into PIE and the total direct effect.

As Judd and Kenny (1981) pointed out, a treatment may produce its impact not only through changing the mediator value but also in part by altering the mediational process that produces the outcome. In other words, the treatment may alter the relationship between the mediator and the outcome. We have reasoned that obtaining an education or training credential under Job Corps might bring greater economic returns than obtaining such a credential under the control condition. Therefore, the individual-specific NIE and PIE may not be equal. The difference between the two is defined as the *natural treatment-by-mediator interaction effect* for each individual (Hong, 2015; Hong *et al.*, 2015), which quantifies the treatment effect on the outcome transmitted through a change in the mediator-outcome relationship. A nonzero interaction effect will indicate that the program-induced change in attainment influences earnings differently between the Job Corps condition and the control condition.

3.3. Site-Specific Causal Effects and Population Parameters

We define the site-specific causal effects by taking expectations of the individual-specific causal effects over the population of individuals at a given site. The site-specific effects, represented by β_j in general, are listed in the second column in Table 1 in which $S_{ij}\ =\ j$ indicates the site membership of individual i.

As we emphasized earlier, of particular theoretical interest is not only the overall average of each of these site-specific causal effects but also its possible variation across the sites. NJCS was a census of all the Job Corps centers that existed at the time of the study, which enables us to generalize results to the population of sites. Hence, the population parameters include the population average and the between-site variance of each site-specific effect, respectively represented by γ and σ^2 in general. As shown in Table 1, the superscripts in β_j and γ and subscripts in σ^2, (*T.M*), (*T.Y*), (*I*), (*D*), and ($T \times M$), serve as shorthand for the ITT effect on the mediator, the ITT effect on the outcome, the indirect effects, the direct effects, and the interaction effect, respectively. We have listed in Table 1 the research questions with regard to the population average causal effects over all the sites in column 3 and the corresponding notation in column 4. Column 5 lists the research questions about the between-site variances of the site-specific effects; and column 6 lists the corresponding notation. Besides, we are also interested in the covariance between the site-specific NDE and NIE, $\sigma_{D(0),I(1)} = \mathrm{cov}(\beta_j^{(D)}(0), \beta_j^{(I)}(1))$, indicating whether Job Corps centers that increased earnings through improving educational and vocational attainment also tend to be successful in increasing earnings through other pathways.

4. Identification

The causal parameters listed in Table 1 could be easily computed if all the potential mediators and potential outcomes were observed for the population of eligible applicants at

Table 1. Causal Parameters

	Site-Specific Effect	Research Question	Average Effect over Population of Sites	Research Question	Between-Site Variance
ITT effect on the mediator	$\beta_j^{(T.M)} = E\left[M_{ij}(1) - M_{ij}(0) \middle\vert S_{ij} = j\right]$	To what extent did Job Corps (JC) improve educational and vocational attainment?	$\gamma^{(T.M)} = E[\beta_j^{(T.M)}]$	Were JC centers equally effective in improving educational and vocational attainment?	$\sigma^2_{T.M} = var\left(\beta_j^{(T.M)}\right)$
ITT effect on the outcome	$\beta_j^{(T.Y)} = E\left[\, Y_{ij}\,(1, M_{ij}(1)) \,-\, Y_{ij}(0, M_{ij}(0)) \middle\vert S_{ij} = j\right]$	To what extent did JC increase earnings?	$\gamma^{(T.Y)} = E\left[\beta_j^{(T.Y)}\right]$	Were JC centers equally effective in increasing earnings?	$\sigma^2_{T.Y} = var\left(\beta_j^{(T.Y)}\right)$
NIE	$\beta_j^{(I)}(1) = E\left[\, Y_{ij}(1, M_{ij}(1)) - \, Y_{ij}(1, M_{ij}(0)) \middle\vert S_{ij} = j\right]$	To what extent did JC increase earnings through improving educational and vocational attainment under the JC condition?	$\gamma^{(I)}(1) = E\left[\beta_j^{(I)}(1)\right]$	Were JC centers equally effective in increasing earnings through improving educational and vocational attainment under the JC condition?	$\sigma^2_{I(1)} = var\left(\beta_j^{(I)}(1)\right)$
NDE	$\beta_j^{(D)}(0) = E\left[Y_{ij}\,(1, M_{ij}(0)) \,-\, Y_{ij}(0, M_{ij}(0)) \middle\vert S_{ij} = j\right]$	To what extent did JC increase earnings through other pathways?	$\gamma^{(D)}(0) = E\left[\beta_j^{(D)}(0)\right]$	Were JC centers equally effective in increasing earnings through other pathways?	$\sigma^2_{D(0)} = var\left(\beta_j^{(D)}(0)\right)$
PIE	$\beta_j^{(I)}(0) = E\left[\, Y_{ij}(0, M_{ij}(1)) - \, Y_{ij}(0, M_{ij}(0)) \middle\vert S_{ij} = j\right]$	To what extent did JC increase earnings through improving educational and vocational attainment under the control condition?	$\gamma^{(I)}(0) = E\left[\beta_j^{(I)}(0)\right]$	Were JC centers equally effective in increasing earnings through improving educational and vocational attainment under the control condition?	$\sigma^2_{I(0)} = var\left(\beta_j^{(I)}(0)\right)$
Interaction effect	$\beta_j^{(T\times M)} = \beta_j^{(I)}(1) - \beta_j^{(I)}(0)$	Did the improvement in educational and vocational attainment produce a greater increase in earnings under JC than under the control condition?	$\gamma^{(T\times M)} = E\left[\beta_j^{(T\times M)}\right]$	Did JC enhance the economic returns to education and training in some centers but not in others?	$\sigma^2_{T\times M} = var\left(\beta_j^{(T\times M)}\right)$

every site. However, $M_{ij}(t)$ and $Y_{ij}(t, M_{ij}(t))$ are observed for $t = 0, 1$ only if individual i at site j was selected into the sample, was assigned to treatment t, and responded to the interviews. In addition, we never directly observe one's potential outcome of assignment to treatment t while the mediator would counterfactually take the value associated with the alternative treatment t' where $t \neq t'$. Causal inference relies exclusively on inferring counterfactual information from the observed information. The inference inevitably invokes one or more assumptions. Here we clarify the assumptions under which each of the causal parameters can be identified from the observed information in the NJCS data. These assumptions should not be taken lightly. Rather, they require close scrutiny on scientific grounds.

4.1. Identification of the ITT Effects

For the ITT effects of the treatment on the mediator and the outcome, identifying their averages over the population of sites along with their between-site variances is complicated by the differential sampling probabilities, treatment assignment probabilities, and nonresponse probabilities, as discussed in the introduction section. We adjust these differential probabilities by applying a series of standard weighting strategies under strong ignorability assumptions about the sampling, treatment assignment, and response mechanisms.

Sampling mechanism. NJCS researchers employed a stratified sampling procedure for individuals. Sampling probabilities varied across strata defined by date of random assignment, gender, residential status, and whether one came from an area with a concentration of nonresidential female students. The probabilities of being included in the follow-up surveys were further determined by a number of factors including population density in one's living area and whether one provided immediate response to the baseline survey. Given this complex sample/survey design, individuals who were included in the 48-month interview sample and those who were not are expected to be comparable in composition only if they share the above mentioned pretreatment characteristics, which we denote with vector $\mathbf{X}_D$. This conclusion also holds within each site. Because the sampling mechanism is known in this study, it is "ignorable" in the sense that we can reasonably make the following assumption:

Assumption 1 (Strongly ignorable sampling mechanism). Within levels of the observed pretreatment covariates $\mathbf{x}_D$, sample selection is independent of all the potential mediators and potential outcomes at each site.

$$\{Y_{ij}(t, m), M_{ij}(t)\} \perp\!\!\!\perp D_{ij} | \mathbf{X}_{Dij} = \mathbf{x}_D, S_{ij} = j,$$

for $t = 0, 1, m \in \mathcal{M}$ where $\mathcal{M}$ is the support for all possible mediator values, and $j = 1, \ldots, J$, where J denotes the total number of sites. Here D_{ij} takes value 1 if individual i at site j was selected into the 48-month interview sample and 0 otherwise. We additionally assume that $0 < Pr\left(D_{ij} = 1 \middle| \mathbf{X}_{Dij} = \mathbf{x}_D, S_{ij} = j\right) < 1$. That is, each eligible applicant at a site had a nonzero probability of being selected (or not being selected) into the sample, an assumption that was guaranteed to hold by the NJCS design. This is also known as the positivity assumption.

Treatment assignment mechanism. NJCS researchers specified an individual's treatment assignment probability as a function of applicants' date of random assignment and residential status among other factors, though not by site. Hence sampled individuals assigned to the program group and those assigned to the control group are expected to be comparable in composition only within each of these predetermined strata, which we denote by $\mathbf{X}_T$. We find that $\mathbf{X}_T$ and $\mathbf{X}_D$ partially overlap.

Assumption 2 (Strongly ignorable treatment assignment). Within levels of the observed pretreatment covariates $\mathbf{x}_T$, the treatment assignment for the sampled individuals is independent of all the potential mediators and potential outcomes at each site.

$$\{Y_{ij}(t,m), M_{ij}(t)\} \perp\!\!\!\perp T_{ij} | D_{ij} = 1, \mathbf{X}_{Tij} = \mathbf{x}_T, S_{ij} = j.$$

Under this assumption, there should be no unmeasured confounding of the treatment-mediator relationship or the treatment-outcome relationship at any site. It is also assumed that $0 < Pr(T_{ij} = t | D_{ij} = 1, \mathbf{X}_{Tij} = \mathbf{x}_T, S_{ij} = j) < 1$. That is, each sampled individual had a nonzero probability of being assigned to either treatment group at a given site. This assumption is similarly guaranteed by the NJCS design.

Response mechanism. NJCS researchers did not have control over an individual's probability of response. Hence, the respondents in the program group and those in the control group are no longer comparable in composition even if they share the same pretreatment characteristics $\{\mathbf{X}_D \cup \mathbf{X}_T\}$. Because response status is possibly a result of the treatment assignment, we find evidence that the response mechanism differs between the program group and the control group. In theory, conditioning on all the pretreatment and posttreatment covariates predicting one's response status under a given treatment at a given site, the respondents and the nonrespondents are expected to be comparable in composition. However, controlling for posttreatment covariates would inevitably introduce bias in identifying the ITT effects of the treatment (Rosenbaum, 1984). Hence in practice, adjustment is made only for the observed pretreatment covariates. We invoke a strong assumption that, among individuals who share the same observed pretreatment characteristics denoted by $\mathbf{X}_R$, one's response status is as if randomized in each treatment group.

Assumption 3 (Strongly ignorable nonresponse). Within levels of the observed pretreatment covariates $\mathbf{x}_R$, the response status of a sampled individual in a given treatment group is independent of the potential mediators and potential outcomes associated with the same treatment at a site.

$$\{Y_{ij}(t,m), M_{ij}(t)\} \perp\!\!\!\perp R_{ij} | T_{ij} = t, D_{ij} = 1, \mathbf{X}_{Rij} = \mathbf{x}_R, S_{ij} = j.$$

Here R_{ij} is equal to 1 if individual i at site j responded and 0 otherwise. This assumption cannot be empirically verified because the potential attainment and the potential earnings were unobserved for the nonrespondents. However, as introduced in Section 5, we could use balance checking and sensitivity analysis to assess the influence of possible violations of the assumption. We also assume that $0 < Pr(R_{ij} = 1 | T_{ij} = t, D_{ij} = 1, \mathbf{X}_{Rij} = \mathbf{x}_R, S_{ij} = j) < 1$. That is, each sampled individual had a nonzero probability of response (or nonresponse) under a given treatment at a given site. This assumption would be violated if certain individuals would always respond or would never do so.

Under these three assumptions, we may equalize the sampling probability and the treatment assignment probability for all the sampled individuals through weighting; by the same logic, the response probability for all the sampled individuals in each treatment group can be equated through weighting as well.

Weighting adjustment for sample selection. Because the sampling probability is predetermined as a function of individual characteristics, certain subpopulations are over-represented while others are under-represented in the sample. The sample representativeness can be restored by applying the stabilized sample weight defined as follows for sampled individual i at site j with pretreatment characteristics $\mathbf{x}_D$,

$$W_{Dij} = \frac{Pr(D_{ij} = 1 | S_{ij} = j)}{Pr(D_{ij} = 1 | \mathbf{X}_{Dij} = \mathbf{x}_D, S_{ij} = j)}. \quad (1)$$

The numerator of the sample weight represents the average sampling probability at site *j*, and the denominator is the individual's sampling probability as a function of the individual's pretreatment characteristics and his or her site membership.

Weighting adjustment for treatment assignment. Similarly, in the presence of treatment selection, certain subpopulations will become over-represented while others are under-represented in a given treatment group. Extending the logic of sample weighting to causal inference, the analyst may apply a stabilized IPTW (Robins et al., 2000) to sampled individual i at site j in treatment group t with pretreatment characteristics $\mathbf{x}_T$,

$$W_{Tij} = \frac{Pr\left(T_{ij} = t|\ D_{ij} = 1, S_{ij} = j\right)}{Pr\left(T_{ij} = t|\mathbf{X}_{Tij} = \mathbf{x}_T, D_{ij} = 1, S_{ij} = j\right)} \text{ for } t = 0, 1. \quad (2)$$

The numerator is the average probability of assigning a sampled individual at site *j* to treatment t; the denominator is the individual's conditional probability of being assigned to treatment t given his or her pretreatment characteristics and site membership, and this probability is pre-determined by design in NJCS.

Weighting adjustment for nonresponse. To remove the observed pretreatment differences between the respondents and the nonrespondents in each treatment group, the analyst may apply a nonresponse weight (see Little and Vartivarian, 2005), which is also stabilized, to sampled individual i at site j in treatment group t with pretreatment characteristics $\mathbf{x}_R$,

$$W_{Rij} = \frac{Pr\left(R_{ij} = r|T_{ij} = t, D_{ij} = 1, S_{ij} = j\right)}{Pr\left(R_{ij} = r|\mathbf{X}_{Rij} = \mathbf{x}_R, T_{ij} = t, D_{ij} = 1, S_{ij} = j\right)} \text{ for } t = 0, 1 \text{ and } r = 0, 1. \quad (3)$$

The numerator is the average probability of response status r among sampled individuals at site *j* who have been assigned to treatment group t; the denominator is the individual's probability of response status r given his or her pretreatment characteristics, treatment assignment, and site membership. This conditional probability is unknown and must be estimated from the observed data, an issue that we will discuss in section 5.

Applying the product of W_D, W_T, and W_R to the respondents, we expect that the distributions of the observed pretreatment covariates $\{\mathbf{X}_D \cup \boldsymbol{X}_T \cup \boldsymbol{X}_R\}$ will be balanced between the sampled and the non-sampled, between the program group and the control group, and between the respondents and the nonrespondents in each treatment group. Hence, we obtain the following identification results.

Theorem 1. Under Assumptions 1, 2, and 3, the site-specific average potential mediator and potential outcome under treatment t for $t = 0,1$ can be respectively identified by the sample average of the observed mediator and the sample average of the observed outcome among the respondents assigned to treatment group *t* at site *j*, weighted by the product of the sample weight, IPTW weight, and nonresponse weight.

$$E\left[M_{ij}\ (t)\ |S_{ij}\ =\ j\right] = E\left[W_{ITTij}M_{ij}\ |R_{ij} = 1, T_{ij}\ = t, D_{ij} = 1, S_{ij} = j\right],$$
$$E\left[Y_{ij}(t, M_{ij}\ (t))\ |S_{ij}\ =\ j\right] = E\left[W_{ITTij}Y_{ij}\ |R_{ij} = 1, T_{ij}\ = t, D_{ij} = 1, S_{ij} = j\right].$$

Here $W_{ITTij} = W_{Dij}W_{Tij}W_{Rij}$ removes selection bias in identifying the ITT effects. The proof of Theorem 1 is presented in Appendix A in the supporting web materials.

The weighted mean difference in attainment between the program group and the control group at each site identifies the site-specific ITT effect of the treatment on the mediator; similarly, their weighted mean difference in earnings identifies the site-specific ITT effect of the treatment on the outcome. The population average and the between-site variance of each of these ITT effects can be identified by following standard results without invoking further assumptions.

4.2. Identification of the Mediation-Related Effects

Identifying the population average and the between-site variance of NDE, NIE, PIE, and the natural treatment-by-mediator interaction effect is considerably more challenging. This is because the mediation-related causal effects involve the counterfactual outcomes $Y_{ij}(1, M_{ij}(0))$ and $Y_{ij}(0, M_{ij}(1))$ that cannot be directly observed; this is additionally because the mediator value assignment under each treatment was not experimentally manipulated. We invoke the following assumption about the strong ignorability of mediator values.

Assumption 4 (Strongly ignorable mediator value assignment). Within levels of the observed pretreatment covariates denoted by $\mathbf{x}_M$, the mediator value assignment under either treatment condition for respondents is independent of the potential outcomes at each site.

$$Y_{ij}(t, m) \perp\!\!\!\perp \{M_{ij}(t), M_{ij}(t')\} | R_{ij} = 1, T_{ij} = t, D_{ij} = 1, \mathbf{X}_{Mij} = \mathbf{x}_M, S_{ij} = j,$$

for all possible values of t and m where $t \neq t'$. Under Assumption 4, $M_{ij}(1)$ and $M_{ij}(0)$ are both independent of $Y_{ij}(1, m)$ for respondents in the program group at site j who share the same covariate values; in parallel, they are also independent of $Y_{ij}(0, m)$ for respondents in the control group at the site who share the same covariate values.

Assumption 4 implies that among individuals who share the same observed pretreatment characteristics denoted by $\mathbf{X}_M$, the assignment of mediator values is as if randomized within each treatment condition or across treatment conditions at any site. This is a particularly strong assumption because it requires not only that there are no remaining pretreatment confounding of the mediator-outcome relationship but also that no post-treatment confounding of the mediator-outcome relationship exists. However, this is not entirely implausible. For any Job Corps applicant at a given site, the probability of educational and vocational attainment may be influenced not only by the treatment assignment but also by theoretically important individual characteristics. However, these predictors do not need to determine with certainty whether an individual would obtain a credential under Job Corps or under the control condition. For example, a Job Corps student might successfully complete the program if he or she happened to encounter a highly effective counselor; a student assigned to the control condition might succeed if an alternative training program was launched at about the same time. These possible random events would make the random assignment of mediator values conceivable under each treatment condition. Hence, we additionally assume that $0 < Pr(M_{ij}(t) = m | R_{ij} = 1, T_{ij} = t, D_{ij} = 1, \mathbf{X}_{Mij} = \mathbf{x}_M, S_{ij} = j) < 1$. That is, each respondent has a nonzero probability of displaying a given mediator value under the actual treatment condition at a given site. Given the Job Corps screening procedure, arguably all eligible applicants are expected to have a chance of attainment in the program; their chance of attainment under the control condition would depend on the availability of alternative education and training opportunities in the local community.

Weighting adjustment for mediator value selection in treatment effect decomposition. In NJCS, only the treatment was experimentally randomized. Yet under Assumption 4, the mediator value assignment could be viewed as if it were randomized for individuals sharing the same covariate values $\mathbf{x}_M$. Putting aside the issues of sampling/survey design and non-random nonresponse, the average of $Y(t, M(t'))$ at a site would be identified by a weighted mean of the observed outcome in treatment group t. For individuals who share the same pretreatment characteristics, the weight would transform the mediator distribution in treatment group t to resemble that in treatment group t'. Hong (2010, 2015) and others proved the identification result

for causal mediation analysis in a single site; Qin and Hong (2017) extended this result to multisite causal mediation analysis. Here we extend the result to multisite studies involving complex sample/survey designs and non-random nonresponse by combining the assumptions and the weighting strategies associated with sampling selection, treatment selection, nonresponse selection, and mediator value selection.

Theorem 2. Under Assumptions 1~ 4, the site-specific average counterfactual outcome $E\left[Y_{ij}(t, M_{ij}(t'))\ |S_{ij} = j\right]$ can be identified by the weighted average of the observed outcome among the sample respondents assigned to treatment group *t* at site *j*, the weight being the product of the ITT weight and the RMPW weight,

$$E\left[Y_{ij}(t, M_{ij}(t'))\ |S_{ij} = j\right] = E\left[W_{ITTij}W_{Mij}\ Y_{ij}\mid R_{ij} = 1, T_{ij}\ = t, D_{ij} = 1, S_{ij} = j\right]$$

for $t \neq t'$, where the RMPW weight is

$$W_{Mij} = \frac{Pr\left(M_{ij} = m|\mathbf{X}_{Mij} = \mathbf{x}_M, R_{ij} = 1, T_{ij} = t', D_{ij} = 1, S_{ij} = j\right)}{Pr\left(M_{ij} = m|\mathbf{X}_{Mij} = \mathbf{x}_M, R_{ij} = 1, T_{ij} = t, D_{ij} = 1, S_{ij} = j\right)}\ \forall\, m \in \mathcal{M}. \quad (4)$$

For respondent i at site j who was assigned to treatment group t and displayed mediator value m, W_{Mij} is a ratio of two propensity scores each as a function of the individual's pretreatment characteristics $\mathbf{x}_M$. The numerator is the individual's propensity of displaying mediator value m under the counterfactual treatment t', while the denominator is the individual's propensity of displaying the same mediator value under the assigned treatment t. Applying the product of W_{ITTij} and W_{Mij} to the sample respondents in each treatment group at each site, we identify the site-specific average potential outcomes $E\left[Y_{ij}\left(1, M_{ij}(0)\right)|S_{ij} = j\right]$ and $E\left[Y_{ij}\left(0, M_{ij}(1)\right)|S_{ij} = j\right]$. Appendix A presents a proof of Theorem 2.

To simplify the notation, let

$$\mu_{tj}^{M} = E\left[W_{ITTij}M_{ij}|R_{ij} = 1, T_{ij}\ = t, D_{ij} = 1, S_{ij} = j\right],$$

$$\mu_{tj}^{Y} = E\left[W_{ITTij}Y_{ij}|R_{ij} = 1, T_{ij}\ = t, D_{ij} = 1, S_{ij} = j\right],$$

$$\mu_{tj}^{Y*} = E\left[W_{ITTij}W_{Mij}Y_{ij}|R_{ij} = 1, T_{ij}\ = t, D_{ij} = 1, S_{ij} = j\right]. \quad (5)$$

Here $\boldsymbol{\mu_{tj}^{M}}$ is the weighted average of the observed mediator in treatment group *t* at site *j* that identifies $\boldsymbol{E\left[M_{ij}(t)|S_{ij} = j\right]}$; $\boldsymbol{\mu_{tj}^{Y}}$ is the weighted average of the observed outcome in treatment group *t* at site *j* that identifies $\boldsymbol{E\left[Y_{ij}\left(t, M_{ij}(t)\right)|S_{ij} = j\right]}$; and $\boldsymbol{\mu_{tj}^{Y*}}$ is the weighted average of the observed outcome in treatment group $\boldsymbol{t}$ at site *j*, with additional RMPW weighting, that identifies $\boldsymbol{E\left[Y_{ij}\left(t, M_{ij}(t')\right)|S_{ij} = j\right]}$. With the site-specific mean of each potential mediator and potential outcome identified, we are able to identify the site-specific causal effects through the weighted mean outcome differences at each site. Table 2 summarizes these identification results. The first column lists the site-specific causal effects defined in terms of the counterfactual quantities as explicated in Section 3; the second column lists the corresponding observable quantities. These identification results enable us to equate the average counterfactual quantities with the observable quantities at each site under the assumptions listed in the third column. We then

identify correspondingly the population average and the between-site variance of each causal effect as defined in Section 3.

5. General Analytic Procedure

Based on the above identification results, we develop an analytic procedure and apply it to the NJCS data. As the identification results indicate, the estimation relies on four weights—sample weight W_{Dij}, IPTW weight W_{Tij}, nonresponse weight W_{Rij}, and RMPW weight W_{Mij}. In NJCS, the product of the first two weights was given by design (Schochet *et al.*, 2001), and the nonresponse weight and the RMPW weight need to be estimated. Conceptually, the estimation

Table 2. Identification of the site-specific effects

Site-Specific Effect	Identification Result	Assumptions
ITT effect on the mediator $\beta_j^{(T.M)}$	$\mu_{1j}^{M} - \mu_{0j}^{M}$	Assumptions 1-3
ITT effect on the outcome $\beta_j^{(T.Y)}$	$\mu_{1j}^{Y} - \mu_{0j}^{Y}$	
NIE $\beta_j^{(I)}(1)$	$\mu_{1j}^{Y} - \mu_{1j}^{Y*}$	Assumptions 1-4
NDE $\beta_j^{(D)}(0)$	$\mu_{1j}^{Y*} - \mu_{0j}^{Y}$	
PIE $\beta_j^{(I)}(0)$	$\mu_{0j}^{Y*} - \mu_{0j}^{Y}$	
Interaction effect $\beta_j^{(T\times M)}$	$\left(\mu_{1j}^{Y} - \mu_{1j}^{Y*}\right) - \left(\mu_{0j}^{Y*} - \mu_{0j}^{Y}\right)$	

involves two major steps: (1) estimation of the nonresponse weight and the RMPW weight by fitting mixed-effects logistic regressions, and (2) estimation of the site-specific causal effects and subsequently average and the between-site variance of the causal effects over the population of sites. To produce valid statistical inferences that incorporate the sampling uncertainty of the weights in the estimation of the causal parameters, we adopt a solution that extends an m-estimation procedure for single-site and multisite RMPW analysis (Bein *et al.*, 2018; Qin and Hong, 2017). This approach estimates the weights and the site-specific causal effects jointly under a generalized method of moments (GMM) framework.

However, the analytic results cannot be given causal interpretations if the identification assumptions are violated. We therefore use balance checking to assess if the estimated weights effectively reduce selection bias associated with the observed covariates. To examine if possible violations of the identification assumptions due to omitting confounders or due to overlooking between-site heterogeneity in the selection mechanisms would easily alter the analytic conclusions, we further conduct a sensitivity analysis.

5.1. Weight estimation

As clarified above, the estimation of the causal parameters depends on the estimates of the nonresponse weight and the RMPW weight. We selected the pretreatment covariates on theoretical grounds (see Appendix B in the supporting web materials for a list of the 51 covariates). We categorized all the continuous covariates to reduce the potential risk of misspecifying the functional form of a model. To preserve the probability sampling and the randomized experimental design, we create a missing indicator for each covariate with missing values. Incorporating the missing indicators, as suggested by Rosenbaum and Rubin (1984), tends to balance not only the observed pretreatment covariates but also the missing patterns. One alternative approach to dealing with missing data is complete case analysis that deletes all the

observations with missing values. This approach is suboptimal because, besides reducing statistical power, it would generally introduce bias except when the missing is completely at random, a particularly strong assumption that rarely holds in reality. Another approach is multiple imputation, which requires the assumption of missing at random -- that is, **the probability that a variable is observed can depend only on the values of those other variables which have been observed** (Little & Rubin, 1989). The missing indicator approach that we have chosen requires a different assumption, namely, that given other observed covariates, the missing values in a covariate are independent of the key variable of interest; or in other words, within levels of other observed covariates, the unobserved values in a covariate do not differ in distribution between those in different categories of the key variable (Groenwold et al, 2012; Jones, 1996). In estimating the nonresponse weight, response status R is the key variable of interest; in estimating the RMPW weight, the key variable is an individual's mediator value assignment. In these two cases, the missing indicator approach assumes strongly ignorable nonresponse or strongly ignorable mediator value assignment among those whose covariate values are missing, conditional on all the observed information.

Nonresponse weight estimation. Following Equation (3), let $p_{Rtj} = Pr\big(R_{ij} = 1|T_{ij} = t, D_{ij} = 1, S_{ij} = j\big)$ denote the average response rate among sampled individuals in treatment group t at site j. To reflect the differences in response mechanisms between the program group and the control group, we fit a logistic regression to each treatment group. The between-site difference in the conditional response rate in each treatment group is captured by a site-specific random intercept in a mixed-effects model. The model specified below estimates the numerator of the weight:

$$\log\left[\frac{p_{Rtj}}{1-p_{Rtj}}\right] = \pi^*_{Rt} + r^*_{Rtj}, \qquad r^*_{Rtj} \sim N(0, \sigma^{*2}_{Rt}),$$

in which π^*_{Rt} indicates the average log-odds of response among the sampled individuals assigned to treatment group t across all the sites; the random intercept, r^*_{Rtj}, assumed to be normally distributed, indicates the deviance of the log-odds of response in each treatment group t at site j from its overall mean; the variance of r^*_{Rtj} is σ^{*2}_{Rt}. To estimate the denominator of the nonresponse weight, we further control for the observed pretreatment covariates $\boldsymbol{X}_{Rij}$ in the mixed-effects logistic regressions.

$$\log\left[\frac{p_{Rtij}}{1-p_{Rtij}}\right] = \boldsymbol{X}'_{Rij}\boldsymbol{\pi}_{Rt} + r_{Rtj}, \quad r_{Rtj} \sim N(0, \sigma^2_{Rt}),$$

in which $p_{Rtij} = Pr\big(R_{ij} = 1|\mathbf{X}_{Rij} = \mathbf{x}_R, T_{ij} = t, D_{ij} = 1, S_{ij} = j\big)$; $\boldsymbol{X}_{Rij}$ includes the intercept; $\boldsymbol{\pi}_{Rt}$ is the corresponding vector of coefficients; and r_{Rtj} is the random intercept with variance σ^2_{Rt}. By fitting each response model through maximum likelihood estimation (MLE) (e.g. Goldstein, 2011), as shown in Appendix C in the supporting web materials, we estimate the coefficients in the response models and obtain the Empirical Bayes estimates of the random intercepts. Based on these estimates, we obtain $\hat{p}_{Rtj}$ and $\hat{p}_{Rtij}$ and the nonresponse weights $\widehat{W}_{Rij} = \hat{p}_{Rtj}/\hat{p}_{Rtij}$ for the respondents and $\widehat{W}_{Rij} = \big(1-\hat{p}_{Rtj}\big)/\big(1-\hat{p}_{Rtij}\big)$ for the nonrespondents.

RMPW weight estimation. To obtain the RMPW weight as defined in equation (4), we need to estimate each respondent's probability of attaining an education or training credential under Job Corps and the probability of obtaining such a credential under the control condition. Let $p_{Mtij} = Pr\big(M_{ij} = 1|\mathbf{X}_{Mij} = \mathbf{x}_M, R_{ij} = 1, T_{ij} = t, D_{ij} = 1, S_{ij} = j\big)$ and $p_{Mt'ij} = Pr\big(M_{ij} = 1|\mathbf{X}_{Mij} = \mathbf{x}_M, R_{ij} = 1, T_{ij} = t', D_{ij} = 1, S_{ij} = j\big)$ denote respondent i's probabilities

of attaining a credential at site j if assigned to treatment t and treatment t', respectively, for $t \neq t'$. We fit the following mediator model to each treatment group, allowing the mediator value selection mechanisms to differ between Job Corps and the control condition:

$$\log\left[\frac{p_{Mtij}}{1-p_{Mtij}}\right] = \boldsymbol{X}'_{Mij}\boldsymbol{\pi}_{Mt} + r_{Mtj}, r_{Mtj} \sim N(0, \sigma^2_{Mt}),$$

in which $\boldsymbol{X}_{Mij}$ includes the intercept; $\boldsymbol{\pi}_{Mt}$ is the corresponding vector of coefficients; and r_{Mtj} is the random intercept with variance σ^2_{Mt}. Importantly, the denominator of the RMPW weight is one's mediator probability under the treatment that he or she was actually assigned to and can be obtained directly by fitting the mediator model to the corresponding treatment group. The numerator of the weight, however, is one's counterfactual probability of having the same mediator value under the alternative treatment. This is obtained by fitting the second mediator model to the alternative treatment group and then applying the coefficient estimates and the empirical Bayes estimate of the random intercept to the focal individual. The estimated RMPW weight is $\widehat{W}_{Mij} = \hat{p}_{Mt'ij}/\hat{p}_{Mtij}$ for respondents in treatment group t at site j who attained a credential and is $\widehat{W}_{Mij} = \left(1-\hat{p}_{Mt'ij}\right)/\left(1-\hat{p}_{Mtij}\right)$ for respondents in the same group at the same site who did not.

5.2. Causal Parameter Estimation and Inference

In accordance with the identification results as shown in Equation (5), the sample estimators for the site-specific average potential mediators and potential outcomes are

$$\hat{\mu}^M_{tj} = \frac{\sum_{i=1}^N \widehat{W}_{ITTij} D_{ij} R_{ij} I\left(S_{ij}=j\right) I\left(T_{ij}=t\right) M_{ij}}{\sum_{i=1}^N \widehat{W}_{ITTij} D_{ij} R_{ij} I\left(S_{ij}=j\right) I\left(T_{ij}=t\right)},$$

$$\hat{\mu}^Y_{tj} = \frac{\sum_{i=1}^N \widehat{W}_{ITTij} D_{ij} R_{ij} I\left(S_{ij}=j\right) I\left(T_{ij}=t\right) Y_{ij}}{\sum_{i=1}^N \widehat{W}_{ITTij} D_{ij} R_{ij} I\left(S_{ij}=j\right) I\left(T_{ij}=t\right)},$$

$$\hat{\mu}^{Y*}_{tj} = \frac{\sum_{i=1}^N \widehat{W}_{ITTij} \widehat{W}_{Mij} D_{ij} R_{ij} I\left(S_{ij}=j\right) I\left(T_{ij}=t\right) Y_{ij}}{\sum_{i=1}^N \widehat{W}_{ITTij} \widehat{W}_{Mij} D_{ij} R_{ij} I\left(S_{ij}=j\right) I\left(T_{ij}=t\right)}.$$

Here $\widehat{W}_{ITTij} = W_{Dij} W_{Tij} \widehat{W}_{Rij}$, where W_{Dij} and W_{Tij} are given by design, and $\widehat{W}_R$ is estimated from the sample data; $\widehat{W}_M$ also needs to be estimated; $I\left(S_{ij}=j\right)$ is an indicator for whether individual i was a member of site j; $I\left(T_{ij}=t\right)$ is an indicator for whether the individual was assigned to treatment t for $t = 0, 1$. Under the identification assumptions 1 ~ 4, mean contrasts between the estimated average potential mediators and potential outcomes at each site consistently estimate the site-specific causal effects listed in Table 2.

We then obtain method-of-moments (MOM) estimates of the causal parameters that characterize the distribution of the site-specific effects in a theoretical population of sites (e.g. Cameron and Trivedi, 2005). For simplicity, we use γ as a general form of each population average causal effect standing for $\gamma^{(T.M)}, \gamma^{(T.Y)}, \gamma^{(D)}(0), \gamma^{(I)}(1), \gamma^{(I)}(0)$, and $\gamma^{(T\times M)}$ and use β_j as a general form of each site-specific causal effect standing for $\beta_j^{(T.M)}, \beta_j^{(T.Y)}, \beta_j^{(D)}(0), \beta_j^{(I)}(1), \beta_j^{(I)}(0)$, and $\beta_j^{(T\times M)}$. By definition, the average of each causal effect over the population of sites, γ, is a simple average of the corresponding site-specific effect, β_j. Hence, the estimate of γ is

$$\hat{\gamma} = \frac{1}{J}\sum_{j=1}^{J}\hat{\beta}_j,$$

where $\hat{\beta}_j$, a mean contrast as described above, is a consistent estimate of β_j.

Although a simple average of $\hat{\beta}_j$ is consistent for γ, a simple average of the squared deviation of $\hat{\beta}_j$ from $\hat{\gamma}$ is biased for $var(\beta_j)$ because this variance estimator contains the sampling variance of $\hat{\beta}_j$ as well as the sampling variance of $\hat{\gamma}$. The estimation of $var(\beta_j)$ is further complicated due to the fact that $\hat{\beta}_j$ is obtained on the basis of the estimated nonresponse weight and the estimated RMPW weight. This is known as the two-step estimation problem in which nuisance parameters must be estimated in the first step and are then used to obtain estimates of the parameters of interest in the second step. The nuisance parameters in this case are the coefficients in the propensity score models for the response and for the mediator. Moreover, although the site-specific effects are to be estimated with only the observed data at a given site, the nuisance estimators are estimated with the data pooled from all the sites, which leads to a nonzero correlation of the sampling errors in the site-specific effect estimates.

Earlier research has extended a two-step estimation procedure (Newey, 1984) to single-site (Bein *et al.*, 2018) and multisite (Qin and Hong, 2017) RMPW analysis in which the RMPW weights are estimated. The rationale is to stack the estimating equations from both steps and solve them simultaneously in the spirit of one-step GMM estimation (Hansen, 1982). The current study makes a further extension to incorporate the estimated nonresponse weights. Under the standard regularity conditions, we derive the asymptotic sampling variance matrix for the site-specific causal effect estimates $var(\hat{\beta}_j - \beta_j)$, and then obtain a consistent estimate of the standard error for each estimated population average causal effect $\hat{\gamma}$. The details can be found in Appendix C. Even though the standard errors can be alternatively estimated through a bootstrap procedure, the closed-form method is favored because it requires much less computation.

The between-site variance of each site-specific effect $var(\beta_j)$ is a population parameter of key interest because it quantifies between-site heterogeneity in the causal mechanism. Its estimation involves subtracting the estimated average within-site sampling variance of the site-specific effect estimates (i.e. the second component of the following equation) from the estimated between-site variance of the site-specific effect estimates (i.e. the first component of the following equation), with adjustment for the between-site sampling covariance of the site-specific effect estimates (i.e. the third component of the following equation):

$$\widehat{var}(\beta_j) = \frac{1}{J-1}\sum_{j=1}^{J}(\hat{\beta}_j - \hat{\gamma})^2 - \frac{1}{J}\sum_{j=1}^{J}\widehat{var}(\hat{\beta}_j - \beta_j) + \frac{1}{J(J-1)}\sum_{j}\sum_{j'\neq j}\widehat{cov}(\hat{\beta}_j - \beta_j, \hat{\beta}_{j'} - \beta_{j'})$$

The estimate of the variance covariance matrix of all the site-specific effects can be found in Appendix C. When a variance estimate is negative, known as a Heywood case, we set the variance estimate as well as the related covariance estimate to 0.

We adopt a permutation procedure (Fitzmaurice *et al.*, 2007) for hypothesis testing. All possible permutations of the site memberships are equally likely under the null hypothesis that the between-site variance is 0. By randomly permuting the site indices while holding the site sizes fixed, we are able to approximate the null distribution of the test statistic and empirically

determine the probability of obtaining values greater than the sample test statistic. Appendix C provides technical details about the estimation and inference.

5.3. Balance Checking

The nonresponse weight and the RMPW weight are estimated and are subjected to potential model misspecification errors. Major errors in model misspecification can be detected if, within a treatment group, the estimated nonresponse weight fails to balance the distribution of the observed covariates between the respondents and the nonrespondents, or if the estimated RMPW weight fails to balance the distribution of the observed covariates between those who succeeded in attaining a credential and those who did not. A substantial reduction in the imbalance in each case would indicate that the estimated weight is effective in reducing selection bias associated with the observed covariates. If some observed covariates are still imbalanced after weighting, balance checking results would indicate how much bias might be remaining and in which direction it might affect the analytic results.

Balance after nonresponse weighting adjustment. Having estimated the nonresponse weight as defined in Equation (3), we use the standardized bias to quantify the balance in an observed covariate between the respondents and the nonrespondents in each treatment group after weighting. The standardized bias is calculated by dividing the weighted mean difference in each covariate by the standard deviation of the covariate (Harder *et al.* 2010). By convention, a covariate is considered to be balanced if the standardized bias is less than 0.25 and preferably less than 0.10 in magnitude. To evaluate whether the balance is achieved across most or all of the sites, it is essential to further estimate the between-site standard deviation of the standardized bias. Under the assumption that the site-specific standardized bias is normally distributed, we compute the 95% plausible value range of the site-specific standardized bias, which is expected to be within the range of $[-0.25, 0.25]$ if the covariate has acceptable balance at each site. Because all the observed pretreatment covariates in this study are categorical, we obtain the results for each treatment group by fitting a weighted mixed-effects logistic model regressing a binary indicator for each covariate category on the response indicator R. The model includes a site-specific random intercept and a random slope that are assumed to be bivariate normal.

Balance after RMPW adjustment. We further assess the extent to which the estimated RMPW weights balance the distribution of the observed covariates between mediator categories in each treatment group at each site. Regressing a binary indicator for each covariate category on the mediator M in a weighted mixed-effects logistic model for each treatment group, we obtain estimates that allow us to calculate the population average and the between-site standard deviation of the standardized bias.

5.4. Sensitivity Analysis

The analytic procedure described above would generate causally valid results only when the identification assumptions hold. In the current study, although the sampling mechanism and the treatment assignment mechanism are ignorable, the assumptions of strongly ignorable nonresponse and strongly ignorable mediator value assignment are likely untenable. A sensitivity analysis is necessary for determining whether potential violations of these assumptions due to omitted confounders would easily alter the causal conclusions. A conclusion is considered to be sensitive if the inference can be easily reversed by additional adjustment for an omitted confounder.

We apply a weighting-based approach to sensitivity analysis that has been extended from single-site to multisite causal mediation studies (Hong *et al.*, 2018, working paper). This approach reduces the reliance on functional form assumptions characteristic of most other existing sensitivity analysis methods. The hidden bias associated with one or more omitted confounders is summarized by a function of a small number of weighting-based sensitivity parameters. In a single-site mediation study in which the treatment is randomized, there are two sensitivity parameters: one is the standard deviation of the discrepancy between a new weight that adjusts for a confounder and an initial weight that omits the confounder; and the other is the correlation between the weight discrepancy and the outcome within a treatment group. Intuitively, the former is associated with the degree to which the omitted confounder predicts the mediator and the latter is associated with the degree to which it predicts the outcome.

In the current study, we consider potential violations of Assumption 3 (strongly ignorable nonresponse) and Assumption 4 (strongly ignorable mediator value assignment). The former are posed by omitted pretreatment or posttreatment confounders of the response-mediator or response-outcome relationships. Such omissions may bias all the causal parameters of interest. The latter are posed by possible omissions of pretreatment and posttreatment confounders of the mediator-outcome relationships. These omissions threaten to bias the population average NDE, NIE, PIE, and interaction effect, and their between-site variances. Moreover, both assumptions need to hold within each site; yet the response models and the mediator models have assumed the same response mechanism and mediation mechanism across all the sites for keeping the models parsimonious. If the response mechanism or the mediation mechanism associated with an observed pretreatment confounder in fact varied across the sites for a given treatment group, omitting the site-specific increment to the coefficient for the confounder in the response model or the mediator model would introduce bias as well. In addition, the original analysis only adjusted for pretreatment covariates, because in the presence of treatment-by-mediator interactions, posttreatment confounders of the mediator-outcome relationship cannot be directly adjusted for in the mediator model (Avin *et al.*, 2005). Similarly, in the presence of treatment-by-response interactions, posttreatment confounders of the response-mediator or response-outcome relationship cannot be directly adjusted for in the response model. We adopt a weighting-based strategy that offers a solution to sensitivity analysis concerning posttreatment confounders (Hong *et al.*, 2018, working paper). Appendix D in the supporting web materials provides a list of weighting-based sensitivity parameters relevant to multisite causal mediation research. For each type of omission, we assess its potential impact on the causal conclusion with regard to each of the population parameters of interest.

6. Analytic Results

6.1. Estimated Nonresponse and RMPW Weights

Appendix B compares the distribution of the outcome and of the 51 covariates between the program group and the control group, between the respondents and the nonrespondents in each treatment group, and between the two mediator categories among the respondents in each treatment group. Average pretreatment differences are notable between the columns. All these covariates are included in the propensity score models for response status and those for the mediator. The estimated nonresponse weight ranges from 0.57 to 3.95 among the respondents and from 0.33 to 4.48 among the nonrespondents, both with a mean equal to 1 in each treatment group. The estimated RMPW weight ranges from 0.13 to 2.53 among the respondents in the program group and from 0.40 and 6.62 among those in the control group, again with a mean

equal to 1 in each case. The stabilized ITT weight ranges from 0.40 to 6.30 among the respondents in the program group and from 0.54 to 5.13 among those in the control group. The stabilized product of the ITT weight and RMPW weight ranges from 0.11 to 6.74 among the respondents in the program group and from 0.27 to 8.17 among those in the control group. An overly large weight may indicate possible violations of the positivity assumption or suggest computational error and may pose a threat to the stability of the estimation results. Our results do not flag such a concern.

6.2. Results of Causal Parameter Estimation and Inference

Table 3 presents the results of estimation and inference for the population average and the between-site standard deviation of the causal effects. These results are generalizable to a theoretical population of Job Corps centers serving disadvantaged youth, most of whom had not acquired a labor market-worthy qualification in education and training at the time of application.

Population average ITT effects of Job Corps. The population average ITT effects of Job Corps on educational and vocational attainment and on earnings are both positive and statistically significant. Job Corps increased the rate of educational and vocational attainment from 22% to 40% within 30 months after randomization, and increased weekly earnings by about $21 (in 1994 dollars) in the fourth year after randomization. The original study estimated an ITT effect of close to $16 dollars (Schochet *et al.*, 2006). This was estimated in the population of individuals, while we estimate the average ITT effect in the population of sites. We do not expect these two parameters to have the same values as we discussed in the second paragraph of the introduction section. Besides, there were other differences between the analyses. We conducted an analysis to decompose which differences between our analysis and the original one led to this difference in estimated impacts (results available upon request). We found that most of the difference was due to how we classified nonrespondents (those missing a site ID, the mediator, or the outcome) compared to the original study classification (those missing the outcome).

Population average mediation mechanism. The ITT effect of Job Corps on earnings is partly transmitted through educational and vocational attainment. The estimated average NIE is $8.47 (standard error [*SE*] = 1.61; $t = 5.26$; $p < 0.001$). This result suggests that human capital formation is not the only pathway through which Job Corps generated its impact on earnings. The estimated average NDE is $12.56 (*SE* = 5.73; $t = 2.19$; $p = 0.028$), accounting for nearly 60% of the ITT effect. According to our earlier reasoning, NDE transmits the Job Corps impact primarily through a wide array of support services. The estimated difference between NDE and NIE is not statistically significant, indicating that the support services played a role at least as important as general education and vocational training in promoting economic well-being among disadvantaged youths. The estimated natural treatment-by-mediator interaction effect $2.27 (*SE* = 2.50; $t = 0.91$; $p = 0.36$) is simply the difference between the estimated NIE and the estimated PIE, the latter being $6.20 (*SE* = 1.78; $t = 3.48$; $p = 0.001$). The interaction effect is not statistically significant. Therefore, the economic returns to the program-induced increase in educational and vocational attainment are indistinguishable between Job Corps and the control condition.

Between-site variance of the ITT effects. The ITT effect of Job Corps on educational and vocational attainment did not vary significantly across sites. However, there is considerable between-site variation in the ITT effect on earnings. Its between-site standard deviation is estimated to be $29.60 ($p < 0.05$). Under the assumption that the site-specific ITT effects are normally distributed, these effects range from -$37 to $79 in 95% of the sites. This result indicates that even though Job Corps significantly improved earnings on average, not all the

centers generated positive impacts. An estimated negative correlation (-0.18) between the site-specific control group mean and the ITT effect (result not tabulated) suggests that Job Corps tended to have a greater positive impact on earnings in the sites where economic prospects were particularly dire under the control condition.

Between-site variance of the mediation mechanism. To explain the between-site heterogeneity in the ITT effect on earnings, we further investigate how the causal mediation mechanism varied across sites. The estimated between-site standard deviation of NDE is as large as \$29 ($p = 0.07$), nearly equal to the estimated between-site standard deviation of the ITT effect on earnings; the estimated site-specific NDE ranges from -\$44 to \$69 in 95% of the sites. In contrast, the estimated between-site standard deviation of NIE is only about \$5 ($p = 0.14$). The estimated between-site standard deviation of PIE and that of the interaction effect are similarly negligible. According to these results, not only did Job Corps universally improved the rate of attaining a certificate, the economic benefit of such an improvement was also comparable across the sites. However, the program impact transmitted through support services appeared to be uneven across the sites. The site-specific NDE seems to largely coincide with the site-specific ITT effect on earnings, their correlation being greater than 0.9 (result not tabulated). Therefore, the between-site variation in the ITT effect on earnings is primarily explained by the heterogeneity in support services. This result is consistent with a qualitative process analysis (Johnson *et al.*, 1999) showing that, unlike the provision of education and vocational training that was strictly regulated by the national and regional Job Corps offices, the quantity and quality of support services were left largely to the discretion of agents at each local center.

Table 3. Estimated causal parameters using the Job Corps data

	Population Average Effect			Between-Site Standard Deviation		95% Plausible Value Range of Site-Specific Effects
	Estimate	Effect Size	*p*-Value	Estimate	*p*-Value	
ITT effect on the mediator (difference in probability)	0.186 (0.014)	0.445	<0.001	0.087	0.035	[0.015, 0.357]
ITT effect on the outcome (dollars)	21.030 (5.684)	0.114	<0.001	29.603	0.035	[-36.992, 79.052]
NDE (dollars)	12.561 (5.730)	0.068	0.028	28.985	0.070	[-44.250, 69.372]
NIE (dollars)	8.469 (1.612)	0.046	<0.001	5.407	0.135	[-2.129, 19.067]
PIE (dollars)	6.198 (1.781)	0.034	0.001	4.351	0.215	[-2.330, 14.726]
Interaction effect (dollars)	2.270 (2.503)	0.012	0.364	11.083	0.220	[-19.453, 23.993]

Note. 1. For the point estimate of each population average effect, the corresponding standard error is provided in parentheses. 2. The effect size of each population average effect estimate is calculated by dividing the point estimate by the standard deviation of the outcome in the control group. 3. The bounds for the 95% plausible value range of the site-specific effects are 1.96 times the between-site standard deviation estimate away from the population average effect estimate, under the assumption that the site-specific effects are normally distributed.

Summary. Our empirical evidence supports the Job Corps program theory and suggests necessary modifications in program practice. Job Corps distinguishes itself from other training programs by emphasizing both human capital formation and risk reduction as complementary

pathways for improving the economic well-being of disadvantaged youths. Our results have indicated that the latter mechanism is no less if not more important than the former. Although all Job Corps centers succeeded in increasing educational and vocational attainment which subsequently led to an increase in average earnings, they were not equally successful in promoting economic well-being through countering a wide range of risk factors. One implication seems clear: regularizing the quantity and ensuring the quality of support services is likely the key to achieving universal effectiveness of Job Corps.

6.3. Results of Balance Checking

As expected, the nonresponse weighting adjustment substantially improved the balance between respondents and nonrespondents on average in both treatment groups. Before weighting, the magnitude of the standardized bias averaged over all the sites was greater than 0.25 for one variable and greater than 0.1 for six other variables in the program group and was greater than 0.25 for two variables and greater than 0.1 for seven other variables in the control group. After weighting, the average standardized bias becomes less than 0.1 in magnitude for all the variables in both groups. The 95% plausible value range of the site-specific standardized bias, initially exceeding the -0.25 and 0.25 thresholds for five variables in the program group and for four variables in the control group, is kept between these thresholds for all but two variables in the program group and for all but one variable in the control group. We notice that the nonresponse weighting increased the plausible value range for some variables due to the increase in estimation uncertainty. These balance checking results are illustrated in Figures E.1 ~ E.4 in Appendix E in the supporting web materials.

Figures E.5 ~ E.8 in the same appendix summarize the balance between mediator categories among the respondents in each treatment group after RMPW weighting. The weighting reduced the number of variables with an average standardized bias exceeding 0.1 in magnitude from nine to zero in the program group and from ten to three in the control group. The number of variables with the plausible value range falling beyond the thresholds of -0.25 and 0.25 is reduced from six to three in the program group and is, however, increased from eight to ten in the control group. This is because, in some of the sites, relatively few respondents in the control group successfully attained a credential. Such noise may reduce the precision in estimating the between-site variance of the standardized bias.

6.4. Results of Sensitivity Analysis

It is straightforward to assess the sensitivity of the original conclusions to the omission of an observed pretreatment covariate, because we can directly calculate its sensitivity parameters based on the observed data. However, to determine if the initial results are sensitive to the existence of an unmeasured pretreatment confounder, it is important to further reason whether the confounding impact of the unmeasured covariate would be comparable to that of an observed pretreatment confounder. For example, characteristics of peer network might influence a Job Corps applicant's response status, educational attainment, and job prospect. Even though peer network was unmeasured in NJCS, we may reason that its confounding impact is comparable to one of the most important observed pretreatment confounders such as baseline earnings and thereby obtain a plausible reference value of the bias caused by the omission of peer network.

Take the population average NIE as an example. An omission of the indicator for upper-middle level baseline earnings would result in a negative bias -$3.39. Our original estimate of the NIE effect is $8.47, with a 95% confidence interval (CI) [$5.31, $11.63]. With an additional

adjustment for an unmeasured pretreatment confounder that is assumed to be comparable to upper-middle level baseline earnings, the new estimate of NIE would become $11.86; the 95% CI of the adjusted NIE estimate is [$8.70, $15.02]. Here we consider the plausible reference value of bias associated with the omission to be given rather than estimated, and thus the additional adjustment does not change the width of the 95% CI. This hypothetical adjustment would lead to an increase in the magnitude of the NIE estimate without changing the initial conclusion about the significant positive NIE. For the population average NDE, the omission would contribute a positive bias of $1.85. With an additional adjustment for this hypothetical bias, the estimate of NDE would change from $12.56 (95% CI = [$1.33, $23.79]) to $10.71 (95% CI = [-$0.52, $21.94]). The adjusted CI now contains zero. Hence the original conclusion about the significant positive NDE is potentially sensitive to an unmeasured confounder comparable to baseline earnings. Among the 51 observed pretreatment covariates, ten of them each provides a plausible reference value of bias that would lead to a statistically insignificant NDE once the hypothetical bias is additionally removed. This is also true with five observed pretreatment covariates when we assess the sensitivity of the population average PIE. In addition, nine covariates would overturn the statistical significance of the population average natural treatment-by-mediator interaction effect. Nevertheless, none of the between-site variance estimates is sensitive to the omission of pretreatment confounders.

In many cases, the analyst might not have enough scientific knowledge to equate the potential bias of an omitted confounder with that of an observed covariate. Yet other data sources might supply values of its sensitivity parameters. Applying the bias formula as represented in Appendix D, the analyst can compute the approximate amount of bias associated with the omission and then assess the sensitivity of the original conclusion to the omission. In addition to assessing the amount of bias that each single omitted covariate might contribute, we could also assess how much bias a set of omitted covariates might introduce jointly.

We further assess the sensitivity of the original conclusions to the omission of the site-specific increment to the coefficient for each pretreatment confounder that has been adjusted for in the response model or the mediator model. The population average ITT effect estimate is insensitive to such an omission. In contrast, with an additional adjustment for the site-specific increment to the coefficient for some of the covariates, the estimated population average NIE, NDE, or PIE would become insignificant, while the population average natural treatment-by-mediator interaction effect, originally tested to be insignificant, would become either significantly negative or significantly positive. Nevertheless, none of the between-site variance estimates is sensitive to the omission of the site-specific increment.

The above discussions are focused on the omission of pretreatment confounders. As explicated in section 5.4, the omission of a posttreatment confounder would also pose threats to the identification assumptions. Because the NJCS data do not have any measurement of a potential posttreatment confounder of the response-mediator, response-outcome, or mediator-outcome relationship, we are unable to assess the potential influence of omitted posttreatment confounders in this study.

7. Conclusion

This article presents a comprehensive template for multisite causal mediation analysis that integrates a series of weighting-based strategies. These include using a sample weight to adjust for complex sample and survey designs, using an IPTW weight to adjust for differential treatment assignment probabilities, using a nonresponse weight to adjust for non-random nonresponse, and using RMPW weights to adjust for mediator value selection while unpacking

the causal mechanisms. Under the identification assumptions clarified in the article, these weighting strategies promise to enhance the external validity and internal validity of the conclusions with regard to the population average and the between-site variance of the causal effects. In addition to decomposing the average ITT effect of the treatment on the outcome into direct and indirect effects, the template further investigates the heterogeneity in causal mechanisms across sites that explains the between-site variation in the ITT effect. Weighting-based balance checking assesses the amount of overt bias associated with the observed covariates. And finally, a weighting-based sensitivity analysis allows for a flexible assessment of the causal conclusions in light of possible violations of the identification assumptions due to hidden bias. The article is accompanied by the "`MultisiteMediation`" R package for implementing the entire analytic procedure.

Developed under the framework of potential outcomes, the template presented in this article provides an important alternative to the existing methods. Multilevel path analysis and SEM (Bauer *et al.*, 2006) with MLE have difficulties in estimating and testing the between-site variance of NIE and the natural treatment-by-mediator interaction effect. Consistent estimation further requires that both the mediator model and the outcome model be correctly specified and that the mediator, the outcome, and the site-specific effects be normally distributed. In contrast, with each causal effect identified through a mean contrast between the weighted outcomes, the proposed MOM strategy invokes no assumptions about the functional form of the outcome model or about the distributions of the mediator, the outcome, and the site-specific effects. The issue of possible misspecifications of the functional forms of the response models or of the mediator models can be evaluated through balance checking and weighting-based sensitivity analysis. We opt for the MOM estimators also because the MLE of a population average effect is essentially a precision weighted average of the site-specific effect estimates. As discussed in Raudenbush and Schwartz (working paper), the MLE will be biased if the precision is correlated with the site-specific effect, which is likely in a multisite trial in which the sites that are more effective in implementing the program may attract more applicants. In contrast, the MOM estimator ensures consistency at some cost of efficiency. In general, efficiency becomes less of a concern in studies with a larger number of sites. We have also derived an asymptotic standard error for each population average effect estimate that fully accounts for the sampling variability in the two-step estimation. The permutation test for variance testing also fills a gap in the literature on multilevel mediation analysis.

There are important topics remaining. First, we have conceptualized the site-specific causal effects under SUTVA. This assumption will need to be relaxed if an individual's potential mediators and potential outcomes could be affected by other individuals' treatment assignments, or if an individual's potential outcomes could additionally be affected by other individuals' mediator values (Hong, 2015; Vanderweele *et al.*, 2013). For example, about half-way into the NJCS sample intake period, the Job Corps centers nationwide implemented a "zero tolerance" policy eliminating students involved in drug abuse or violence. The removal of such "problem" students would presumably improve the institutional environment and would increase allocation of resources to other students who were not directly targeted by the policy. Hence expelling "problem" peers from the program might contribute positively to one's potential earnings. To test this hypothesis will require a major revision of the conceptual framework and creative extensions of the current template, which we will explore in future work.

Second, the proposed template is directly applicable to multisite trial data similar to the Job Corps data, in which all the sites at the time of study were included and at least a moderate number of individuals were selected into the sample at each site. Unlike NJCS, some multisite

studies may sample sites first and then sample individuals within the sampled sites. One may further incorporate a site-level sample weight to adjust for the sample selection of sites. The sample design has important implications for causal inference. For example, researchers would not be able to obtain results generalizable to the population of sites if individuals were sampled from the overall population with a relatively small probability while the number of sites was relatively large and the site sizes were uneven. This is because sampled observations might become too sparse or even non-existent in some of the relatively small sites, in which case the sample of sites would not be representative of the population of sites.

Third, we have adopted the missing-indicator strategy to handle missingness in the pretreatment covariates while using the inverse probability weighting strategy to account for nonresponse in the mediator and the outcome. If the true values of the missing cases were highly variant, the missing-indicator strategy would underestimate the variance and covariance of the covariates. When the missing-at-random assumption holds, an alternative is to use multiple imputation to impute the missing values in the covariates as well as in the mediator and the outcome. A product of the sample weight, IPTW weight, and RMPW weight, i.e. $W_D W_T W_M$, will then be applied to each imputed data set. The final estimation results can be obtained by combining the estimates from multiple imputed data sets.

Fourth, as acknowledged by Qin and Hong (2017), the MOM estimation procedure may not be optimal if there are fewer than 20 individuals at each site. Moreover, when site sizes are relatively small, propensity score models may be overfitted if selection mechanisms vary across sites. In such cases, there might be a lack of statistical power for detecting between-site heterogeneity in the causal mediation mechanism.

Fifth, our mediator is a combination of two central elements of the Job Corps program. It takes value 1 if an individual obtained either an education or a training credential within 30 months after randomization. However, the selection mechanism that led to an education credential might be different from the mechanism that led to a training credential. Combining these two distinct types of credentials into one mediator may result in misspecified propensity score models for the mediator and correspondingly biased estimates of the causal parameters. This problem can be addressed by viewing vocational training attainment and general education attainment as two concurrent mediators, a topic that we investigate in a separate study.

Acknowledgment

The authors thank Donald Hedeker, Luke Miratrix, Stephen Raudenbush, James Rosenbaum, Peter Schochet, Yongyun Shin, Margaret Beale Spencer, Kazuo Yamaguchi, and Fan Yang for their contribution of ideas and their comments on earlier versions of this article. Alma Vigil provided invaluable research assistance to the analysis of the NJCS data. This study was supported by a grant from the National Science Foundation (SES 1659935), a U.S. Department of Education Institute of Education Sciences Statistical and Research Methodology Grant (R305D120020), and a subcontract from MDRC funded by the Spencer Foundation. In addition, the first author received a Quantitative Methods in Education and Human Development Research Predoctoral Fellowship from the University of Chicago and a National Academy of Education/Spencer Foundation Dissertation Fellowship. This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

References

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

Baron, R. M. and Kenny, D. A. (1986) The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6), 1173.

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

Becker, G. S. (1964) Human capital theory. *Columbia, New York*.

Bein, E., Deutsch, J., Hong, G., Porter, K., Qin, X., & Yang, C. (2018). Two-step estimation in ratio-of-mediator-probability weighted causal mediation analysis. *Statistics in Medicine,* 37(8), 1304–1324.

Bloom, H. S., Hill, H. and Riccio, J. A. (2005) Modeling cross-site experimental differences to find out why program effectiveness vary. In *Learning more from social experiments: Evolving analytic approaches*. (eds Howard S. Bloom). NY: Russell Sage Foundation.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Card, D. (1999) The causal effect of education on earnings. In *Handbook of Labor Economics*. (eds Orley Ashenfelter and David Card), vol.3A, pp.1801-1863. Amsterdam: Elsevier Science, North-Holland.

Duncan, O. D. (1966) Path analysis: Sociological examples. *American journal of Sociology*, 1-16.

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

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

Frumento, P., Mealli, F., Pacini, B., & Rubin, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*, *107*(498), 450-466.

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

Groenwold, R. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., & Moons, K. G. (2012). Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Canadian Medical Association Journal*, 184(11), 1265-1269.

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

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, *15*(3), 234.

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

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

Hong, G. (2017) A review of "Explanation in causal inference: Methods of mediation and interaction." *Journal of Educational and Behavioral Statistics*, *42*(4), 491-495.

Hong, G., Deutsch, J. and 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. American Statistical Association.

Hong, G., Deutsch, J. and 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(3), 307-340.

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

Hong, G., Qin, X., & Yang, F. (2018). Weighting-based sensitivity analysis in causal mediation studies. *Journal of Educational and Behavioral Statistics*.

Hong, G., Qin, X., & Yang, F. (working paper). Sensitivity Analysis for Multisite Causal Mediation Studies. Technical Report.

Hong, G. and 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(475), 901-910.

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

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

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

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

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

Johnson, T., Gritz, M., Jackson, R., Burghardt, J., Boussy, C., Leonard, J. and Orians, C. (1999) National Job Corps Study: Report on the Process Analysis. Research and Evaluation Report Series.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433), 222-230.

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

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

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

Lee, D. S. (2009) Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3), 1071-1102.

Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3), 292-326.

Little, R. J., & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means?. *Survey Methodology*, *31*(2), 161.Newey, W. K. (1984) A method of moments interpretation of sequential estimators. *Economics Letters*, 14(2), 201-206.

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

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

Petersen, M. L., Sinisi, S. E. and van der Laan, M. J. (2006) Estimation of direct causal effects. *Epidemiology*, 17(3), 276-284.

Qin, X., & Hong, G. (2017). A weighting method for assessing between-site heterogeneity in causal mediation mechanism. *Journal of Educational and Behavioral Statistics, 42(3), 308-340*.

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

Raudenbush, S.W. and Schwartz, D. (working paper) Estimation in Multisite Randomized Trials with Heterogeneous Treatment Effects.

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

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, *11*(5), 551.

Rosenbaum, P. R. (1984) The consequence of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society*, *Series A (General)*, *147*(5), 656-666.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, *79*(387), 516-524.
Rubin, D. B. (1978) Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1), 34-58.
Rubin, D. B. (1980) Randomization Analysis of Experimental Data: The Fisher Randomization Test: Comment. *Journal of the American Statistical Association*, 75(371), 591-593.
Rubin, D. B. (1986) Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396), 961-962.
Rubin, D. B. (1990) Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3), 279-292.
Schochet, P., Burghardt, J., & Glazerman, S. (2001). National Job Corps Study: The impacts of Job Corps on participants' employment and related outcomes.
Schochet, P. Z., Burghardt, J. and McConnell, S. (2006) National job corps study and longer-term follow-up study: impact and benefit-cost findings using survey and summary earnings records data. *Mathematica Policy Research, Inc*.
Schochet, P. Z., Burghardt, J. and McConnell, S. (2008) Does Job Corps Work? Impact Findings from the National Job Corps Study. *The American Economic Review*, 98(5), 1864-1886.
Sobel, M. E. (1982) Asymptotic confidence intervals for indirect effects in structural models, in *Sociological Methodology* (ed S. Leinhardt), Jossey-Bass, San Francisco, CA, pp.290-312.
Spybrook, J. and 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(3), 298-318.
Tchetgen Tchetgen, E. J. and Shpitser, I. (2012) Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness and sensitivity analysis. *The Annals of Statistics*, 40(3), 1816-1845.
Valeri, L. and VanderWeele T. J. (2013). Mediation analysis allowing for exposure– mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological methods*, 18 (2), 137.
Vanderweele, T. J., Hong, G., Jones, S. M. and 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(502), 469-482.
VanderWeele and Vansteelandt (2009) Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2, 457-468.
VanderWeele and Vansteelandt (2010) Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, 171 (12), 1339-1348.
Weiss, M. J., Bloom, H. S. and Brock, T. (2014) A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, 33(3), 778-808.
Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence from Past Multisite Randomized Trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843-876.
Wright, S. (1934) The method of path coefficients. The Annals of Mathematical Statistics, 5(3), 161-215.
Zhang, J. L., Rubin, D. B. and Mealli, F. (2009) Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, 104(485), 166-176.