THE UNIVERSITY OF CHICAGO

CAUSAL MEDIATION ANALYSIS IN MULTISITE TRIALS WITH AN EVALUATION OF THE JOB CORPS PROGRAM

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPARATIVE HUMAN DEVELOPMENT

BY

XU QIN

CHICAGO, ILLINOIS

JUNE 2018

To My Family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to extend my deepest gratitude to my advisor, Dr. Guanglei Hong. I feel very lucky that I had the opportunity to work closely with her and learn a lot from her throughout my six years of doctoral study. She led me into the world of causal inference and stimulated my passion for this field. She spent a tremendous amount of time advising me no matter how busy she was. I was able to write this dissertation because of her extremely patient, careful, and insightful guidance and revisions. Thanks to her great support, I attended many conferences. These precious opportunities enabled me to meet with excellent scholars in statistics, education, and other related fields across the country and further improve my dissertation work based on their helpful feedback. I am also deeply grateful to my dissertation committee members, Drs. Stephen Raudenbush, Donald Hedeker, and Margaret Beale Spencer, for their invaluable suggestions and comments. They were always patient and devoted whenever I asked for help. I have been greatly inspired by all the four professors' outstanding work since I entered graduate schools. Their dedication, humility, sincerity, and kindness also generated an influence on me. I feel very fortunate to have them as my lifetime mentors both intellectually and spiritually.

I would also like to thank Drs. Ed Bein, Jonah Deutsch, Ken Frank, Robert Gibbons, Kristin Porter, Peter Schochet, Kazuo Yamaguchi, Cheng Yang, and Fan Yang for their contribution of ideas and their comments on earlier versions of the dissertation. Due to data access restriction, I would not have been able to obtain the analytic results of the Job Corps program without the assistance of Alma Vigil at Mathematica Policy Research. I really appreciate it.

Thanks to the interdisciplinary environment of the University of Chicago, I am lucky to meet many friends from diverse disciplines, such as anthropology, biostatistics, education,

psychology, public policy, social service, sociology, and statistics, etc. I thank them for making my doctoral student life more colorful and fun and for opening my horizon through interdisciplinary communications.

I would also like to thank my family for their great support, encouragement, and love. My parents, Xiaoyin Geng and Zhenliang Qin, have always been my role models. They taught me how to be a good person and stimulated my enthusiasm for research and teaching. I would not have been able to come this far in my academic pursuits without their cultivation, understanding, and support. I feel very fortunate that I have been able to study and grow together with my husband and best friend, Jiebiao Wang, since we met in college. I really appreciate his companionship and encouragement throughout my study in statistics and in my life.

This dissertation was supported by a Quantitative Methods in Education and Human Development Research Predoctoral Fellowship and a Social Science Division Fellowship funded by the University of Chicago and by a National Academy of Education/Spencer Dissertation Fellowship. Additional support came from a U.S. Department of Education Institute of Education Sciences Statistical and Research Methodology Grant "Weighting methods for mediation analysis in experimental and quasi-experimental multilevel data" (R305D120020) and a subcontract from MDRC for the project "Using emerging methods with existing data from multi-site trials to learn about and from variation in education al program effects" funded by the Spencer Foundation.

ABSTRACT

There has been an increasing use of multisite randomized trials in evaluations of educational programs. Multisite designs provide unique opportunities for investigating between-site heterogeneity in the mediation mechanism that characterizes an educational process central to a program theory. Re-analyzing data from the National Job Corps Study, a multisite randomized evaluation, this dissertation develops methods for empirically examining the Job Corps program theory. Job Corps is the nation's largest education and training program for 16-24 year old disadvantaged youths, most of whom had dropped out of high school. Previous research has suggested that Job Corps generated a positive average impact in promoting economic independence. However, the impact was not uniform across all the sites. The multisite data allow us to further investigate whether the central program element, i.e. educational and vocational training, played the same mediating role across sites, and whether the role of other program elements was consistent over the sites. Such evidence will be crucial for enriching theoretical understanding and for informing the design and implementation of education programs alike. However, due to some important constraints of existing analytic tools, analysts have rarely investigated between-site heterogeneity of mediation mechanisms in multisite program evaluations.

To enable researchers to assess the generalizability of an education program theory across a wide range of contexts, this dissertation develops a comprehensive weighting-based analytic procedure for multisite causal mediation analysis. The procedure utilizes a propensity score-based weighting strategy to flexibly decompose the average program impact at each site into a direct effect and one or two indirect effects, the latter being transmitted through one or two hypothesized focal mediators. To enhance the external and internal validity of causal

conclusions, I further incorporate a sample weight to adjust for complex sample and survey designs and employ an estimated nonresponse weight to account for non-random nonresponse in the longitudinal follow-ups. Extending a theoretical model of causal inference under the potential outcomes framework, I conceptualize the population average and the between-site variance of the direct and indirect effects and identify them based on the above weights. For the estimation and inference of the causal parameters, I develop a method-of-moments procedure that takes into account the sampling variability of the estimated weights. Finally, I use a weighting-based balance checking procedure to assess if the weighting adjustment effectively reduces selection bias associated with the observed covariates and adopt a weighting-based sensitivity analysis strategy to assess the consequences of potential violations of key identification assumptions.

I employ the proposed analytic procedure in an in-depth evaluation of Job Corps. The empirical results lend support to the program theory that Job Corps promotes economic well-being among disadvantaged youths through education and training. The results also highlight the crucial role of support services for reducing behavioral and health risks and reveal the need for standardizing the quantity and quality of such services across Job Corps centers.

CHAPTER 1

INTRODUCTION

1.1 An Introduction to Multisite Trials

Intervention programs in education, economics, political science, public health, and social welfare are usually delivered in organizations or communities. Each local setting can be viewed as an experimental site within which individuals are assigned to different treatment conditions. Multisite randomized trials and multisite natural experiments have been pervasive in these fields and are often longitudinal in data collection (Bloom, Hill, & Riccio, 2005; Raudenbush & Bloom, 2015; Spybrook & Raudenbush, 2009). For example, over the past dozen years, the Institute for Education Sciences (IES) of the U.S. Department of Education has funded over 175 large-scale randomized trials. The vast majority of these studies are multisite randomized trials. In some of these studies, districts are taken as sites, and within each site schools are randomized to different treatment conditions which are usually composed of a program condition and a control condition; in some other studies, schools are sites in which classrooms, teachers, or students are randomized to different treatment conditions; sites are sometimes communities or regions in which individuals are randomized to treatment conditions. Most well-known examples include the Tennessee Class Size Study, the Head Start Impact Study, and the National Job Corps Study, among others.

Different from clustered randomized trials (also called “group randomized trials”), which only allow for the estimation of the average treatment effect because individuals in the same cluster are assigned to the same treatment condition, multisite randomized trials offer unique opportunities for investigating how the program impact may vary across a wide range of settings in which a program is implemented. In a multisite randomized trial, a sample of sites represents a

population of sites that may differ in various contextual factors and in program implementation. Due to the randomization within each site, program impacts can be identified without bias at the site level. One may then examine not only the average program impact but also the between-site variation of the impact.

While most evaluation research in the past has focused solely on the average impact, researchers have argued that the average alone is not sufficient for developing policy and practice if the impacts vary from sites to sites. The importance of investigating the heterogeneity of the program impacts across sites has become increasingly appreciated (e.g. Bryk & Raudenbush, 1988; Heckman, Smith, & Clements, 1997; Raudenbush & Bloom, 2015; Bloom et al., 2017; Olsen, 2017). Weiss et al. (2017) provided by far the most comprehensive evidence on the magnitude of between-site impact variation, by evaluating 16 large multisite trials of educational interventions. The results indicate that program impacts do vary for some interventions and some variations can be quite substantial. Such a between-site variation indicates the likely highest and lowest impacts of a program. An investigation of the variation may enable policy makers to gain a better understanding of whether the program impact is generalizable across various contexts and, if not, in what contexts the program is effective and why the impacts vary. This may in turn inform the implementation of the programs under different settings and help to make the programs more equitable.

Causal mechanisms may differ across sites due to natural variations in organizational contexts, in participant composition, and in local implementation (Weiss, Bloom, & Brock, 2014). Assessing between-site variation in the causal mechanisms may generate important information for unpacking and understanding the heterogeneity in the total program impact, may reveal a need to revisit the program theory, and may suggest specific site-level modifications of

the intervention practice. Multisite trials provide rich opportunities for researchers to investigate the mechanisms through which programs produce their intended effects under different local settings. In some cases, evidence may suggest that a program theory is highly generalizable across a wide range of contexts; while in some other cases, such an investigation may reveal important differences between sites in how the program operates, which may explain important between-site variation in program impacts. However, researchers have not taken full advantage of the multisite data to investigate important between-site differences in program mechanisms that may give rise to the unevenness in program impact across the local settings.

Hence, the major goal of this dissertation work is to develop a conceptual framework and a statistical tool for investigating, in a multisite randomized trial, the population average and the between-site variation of the causal mediation mechanisms through which programs produce their intended effects. In the meantime, I apply the proposed methods to an empirical investigation of the causal mediation mechanism of the Job Corps program.

1.2 An Introduction to Causal Mediation Analysis

1.2.1 Causal Mediation Analysis with a Single Mediator

To investigate the mechanisms through which a program operates, that is, to develop and test a program theory explaining the educational processes that shape participants' learning and development, researchers will need to conduct a mediation analysis. A hypothesized mediation mechanism characterizes an educational process central to a program theory. Such a process often involves a change in cognitive or social-emotional behaviors induced by the program participation and subsequently a change in one's developmental outcomes. The variable that transmits the program impact plays a role as a mediator. In the basic mediation framework, a treatment affects a focal mediator, which in turn affects an outcome. The total treatment effect

can be decomposed into an indirect effect that transmits the treatment effect through the hypothesized focal mediator and a direct effect that works directly or through other unspecified mechanisms. In general, an average indirect effect in the desired direction and magnitude lends support to the program theory with regard to the central mechanism.

Let T denote the treatment assignment, M for the focal mediator, and Y for the outcome. Here I use the encouragement design example in Holland (1988) as an illustration. In this example, students were randomly assigned to one of two treatments: those in the experimental group were encouraged to study for a test ($T = 1$) while those in the control group were not ($T = 0$). In the hypothesized mediation mechanism, after being exposed to one of these treatments, the amount of time that a student spent studying for the test, M, would mediate the effect of encouragement on the student's final test score, Y. The causal diagram in Figure 1.1 depicts the causal mediation process. The arrow from T to M and that from M to Y represent how the treatment generates the impact on the outcome through the mediator. The arrow from T to Y captures all the other possible pathways that transmit the treatment effect on the outcome. Hence, the total treatment effect can be decomposed into an indirect effect transmitted through M and a direct effect that operates through all the other possible mechanisms.

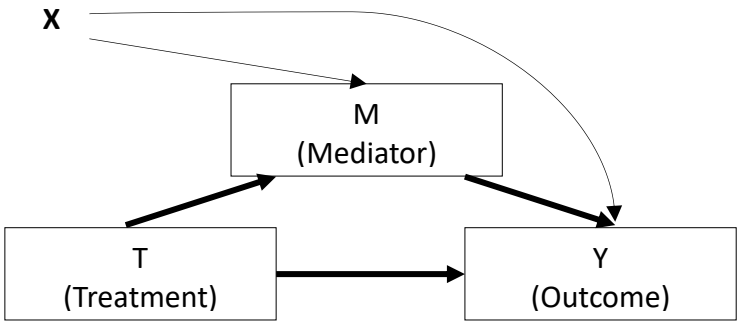


Figure 1.1 Diagram of a Causal Mediation Process

Note. In a randomized experiment, even though the treatment assignment is randomized, the mediator usually is not. Some pretreatment covariates **X** may be correlated with both the mediator and the outcome and thus confound the relationship between the mediator and the outcome.

In social science research, path analysis (Alwin & Hauser, 1975; Baron & Kenny, 1986; Duncan, 1966; Sobel, 1982; Wright, 1934) and structural equation modeling (SEM) (Bollen, 1987; Jo, 2008; Jöreskog, 1970; MacKinnon, 2008; MacKinnon & Dwyer, 1993) have been the primary techniques for mediation analysis in the past several decades. This technique regresses the mediator on the treatment and regresses the outcome on the mediator and the treatment for each individual *i*:

$$M_i = d_M + aT_i + e_{Mi} \tag{1.1}$$

$$Y_i = d_Y + bM_i + cT_i + e_{Yi} \tag{1.2}$$

where $e_{Mi} \sim N\left(0, \sigma^2_{e_M}\right)$ and $e_{Yi} \sim N\left(0, \sigma^2_{e_Y}\right)$. a denotes the association between the treatment and the mediator, b indicates the association between the mediator and the outcome given the treatment condition, and c is the association between the treatment and the outcome given the mediator level. The indirect effect is represented as the product of a and b, and c represents the direct effect, based on the assumptions that there are no confounders of the treatment-mediator, treatment-outcome, or mediator-outcome relationship and that the mediator and outcome models are correctly specified (Holland, 1988).

In presentations and applications of this technique, one major concern that tends to receive little attention is the omission of confounders of the mediator-outcome relationship, shown as $\boldsymbol{X}$ in Figure 1.1. Omitting $\boldsymbol{X}$ from the outcome model may generate biased direct and indirect effect estimates even if the treatment is randomized (Bullock, Green, & Ha, 2010). In the encouragement design example, $\boldsymbol{X}$ is a set of pretreatment or posttreatment covariates that are associated with both the amount of study and the final test score under each encouragement condition. For example, students who had higher pretest scores prior to participating in the experiment might have greater motivation and would study relatively more hours than students

who had lower pretest scores even without the encouragement; students with higher pretest scores are also expected to score higher on the final test, no matter which treatment group they were assigned to. Hence, the observed association between the amount of study and final test score given the treatment condition might be partly attributable to the confounding of the pretest score. Omitting the pretest score from the outcome model would lead to a biased estimate of b and thus a biased estimate of the indirect effect. Similarly, the estimate of c would also be biased, leading to a biased estimate of the direct effect.

Moreover, this method relies heavily on correct specifications of both the mediator model and the outcome model. Even when the treatment is randomized and even when an analyst attempts to make statistical adjustment for all potential confounders of the mediator-outcome relationship, estimation of the indirect and direct effects will nonetheless be biased if the regression models are misspecified. Typically, an analyst may overlook a possible treatment-by-mediator interaction, ignoring the fact that the treatment effect may be generated not only through changing the mediator but also through changing the mediator-outcome relationship (Judd & Kenny, 1981). For example, the impact of study on the final test score might be higher for students who were encouraged to study than for students assigned to the control group even if the two groups of students would study the same number of hours. This is because students who received encouragement to study might display a higher efficiency in study and benefit more from every hour of study. An analyst may also overlook a possible treatment-by-covariate interaction, a mediator-by-covariate interaction, a treatment-by-mediator-by-covariate interaction, a nonlinear covariate-mediator relationship, or a nonlinear covariate-outcome relationship (Hong, 2017). In addition, the strategy typically assumes that the mediator and the outcome are multivariate normal in distribution. As others have pointed out (Imai, Keele, &

Tingley, 2010; MacKinnon & Dwyer, 1993; VanderWeele & Vansteelandt, 2010), their applications to discrete mediators and outcomes face many constraints.

The mainstream literature on path analysis and SEM did not incorporate the causal inference framework until relatively recently (Holland, 1988; Jo, 2008; Sobel, 2008). Since then, serious attempts have been made to reduce selection bias associated with the non-random mediator value assignment. These include two widely-used approaches that have been extended to causal mediation analysis—the instrumental variable (IV) method popular among economists (Heckman & Robb, 1985) and marginal structural models well known to epidemiologists (Coffman & Zhong, 2012; Robins, 2003; Robins & Greenland, 1992; VanderWeele, 2009). However, these methods were built upon the assumption of no treatment-by-mediator interaction. The IV method is employed primarily for estimating the effect of the mediator on the outcome. It relies heavily on the exclusion restriction, which implies that the treatment as an instrument for the mediator does not influence the outcome through any unspecified pathways including a treatment-by-mediator interaction. It is equivalent to assuming that the direct effect is 0, which is unrealistic in many settings. Marginal structural models take the same structural form as path analysis models, while covariates are adjusted for through weighting instead of being directly entered into the structural models. However, as Coffman and Zhong (2012) acknowledged, marginal structural models cannot be used to estimate the indirect effect in the presence of a treatment-by-mediator interaction.

Recently, there have been important extensions that further relax the no treatment-by-mediator assumption. These include the modified regression approaches (Pearl, 2010; Petersen, Sinisi, & van der Laan, 2006; Preacher, Rucker, & Hayes, 2007; Valeri & VanderWeele, 2013; VanderWeele, 2013; VanderWeele & Vansteelandt, 2009, 2010), direct effect models (van der

Laan & Petersen, 2008), conditional structural models (VanderWeele, 2009), and a model-based resampling approach (Imai, Keele, & Yamamoto, 2010; Imai, Keele, & Tingley, 2010). All these strategies accommodate the treatment-by-mediator interaction in the outcome model. This leads to an indirect effect that takes a rather complex form combining more than two parameters and thus adds considerable complications to estimation and statistical inference. Besides, correct model specifications are still crucial for generating consistent causal effect estimates, and challenges involving covariates remain in model specifications.

Unlike the regression-based strategies, the weighting method, proposed for single-site causal mediation analysis by Hong (2010, 2015) and others (Hong, Deutsch, & Hill, 2011, 2015; Hong & Nomi, 2012; Huber, 2014; Lange, Rasmussen, & Thygesen, 2014; Lange, Vansteelandt, & Bekaert, 2012; Tchetgen Tchetgen, 2013; Tchetgen Tchetgen & Shpitser, 2012), has offered an appealing alternative. Defining direct and indirect effects in terms of potential outcomes (Pearl, 2001; Robins & Greenland, 1992) that will be introduced in Section 2.1, a ratio-of-mediator-probability weighting (RMPW) analysis identifies and estimates these causal effects each as a mean contrast, along with their standard errors, while adjusting for pretreatment confounding through propensity score–based weighting. The intuitive rationale is that, among individuals with the same pretreatment characteristics, the distribution of the mediator in the experimental group and that in the control group can be effectively equated through weighting under the assumption of sequential ignorability that will be introduced in Section 2.4. Unlike the regression-based strategies, these weighting methods allow for the treatment-by-mediator interaction without having to specify the mediator–outcome relationship and the covariate–outcome relationship and is suitable for discrete and continuous mediators and outcomes. This is because, unlike other causal mediation methods, the weighting strategy does not require strong

assumptions about the functional form of the outcome model and hence greatly minimizes the risk of model misspecification. Simulations (Hong et al., 2015) have shown that, when the outcome model is misspecified, RMPW clearly outperforms path analysis/SEM in bias correction.

1.2.2 Causal Mediation Analysis with Two Mediators

A theory-based social intervention program tends to have multiple components rather than a single element. This is because a well-developed program theory often recognizes the necessary conditions required for the targeted change and therefore builds condition-changing strategies into the program. A program theory as such may suggest complex mechanisms involving multiple pathways operating jointly to produce a desired outcome.

Causal inference methods for investigating complex mediation mechanisms have only begun to emerge in recent years. Researchers have proposed various methods for rigorously evaluating the causal effects transmitted through two mediators that are either consecutive (i.e., one mediator affecting the other) or concurrent (i.e., two mediators being parallel). Albert and Nelson (2011) used a potential outcomes framework to define effects transmitted through two mediators. They then proposed a system of generalized linear structural models for binary mediators and outcomes. Each model was fit separately using maximum likelihood estimation; and confidence intervals were obtained through bootstrapping. Daniel, De Stavola, Cousens, and Vansteelandt (2015) adopted a similar parametric G-computation approach but implemented it through Monte Carlo simulations that generated multiple random draws of the values of each potential mediator and potential outcome; they estimated standard errors also through the bootstrap. Imai and Yamamoto (2013) presented an idea of setting up bounds for the mediated

impacts. VanderWeele (2015) reviewed two methods, one extending path analysis and the other using a prediction model to impute counterfactual outcome values.

These strategies, however, all require that an analyst correctly specify each mediator model and the outcome model and often further require distributional assumptions about the mediator and outcome measures. As discussed in Section 1.2.1, causal analytic results are sensitive to violations of parametric modeling assumptions. To reduce reliance on model specifications, Lange and colleagues (Lange et al., 2014) extended a propensity score-based weighting method (Hong, 2010; Lange et al., 2012) to the case of multiple concurrent mediators; Hong (2015) further considered the extension of this method to multiple consecutive mediators; so did Huber (2014).

1.3 Causal Mediation Analysis in Multisite Trials

In a single-site study, the population of individuals residing at the site is naturally the target of inference. The 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 & Bloom, 2015; Raudenbush & 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 site-specific subpopulations of individuals are of the same size across all the sites or if the effect does not vary across sites. In this dissertation, with a primary interest in the between-site heterogeneity of the program impacts and of the mediation mechanisms, I focus on the population of sites rather than the overall population of individuals.

In most evaluations of the education programs that are delivered in different local settings, most researchers have focused on the population of individuals and simply ignored the role of multiple sites in their analyses. Hence, little has been done to reveal important between-site heterogeneity of mediation mechanisms. Advancement in this line of research has largely been constrained by existing analytic tools. How can we improve the analysis of between-site heterogeneity in causal mechanisms in multisite evaluations of education programs? This is the fundamental question that motivates my dissertation research.

Taking on the challenges of multisite data, researchers (Bauer, Preacher, & Gil, 2006; Kenny, Korchmaros, & Bolger, 2003; Krull & MacKinnon, 2001; Preacher, Zyphur, & Zhang, 2010; Zhang, Zyphur, & Preacher, 2009) have proposed to embed the standard path analysis and SEM in multilevel modeling by including random intercepts and random slopes in the mediator model and the outcome model as follows:

$$M_{ij} = d_{Mj} + a_j T_{ij} + e_{Mij}, \tag{1.3}$$

$$Y_{ij} = d_{Yj} + b_j M_{ij} + c_j T_{ij} + e_{Yij}, \tag{1.4}$$

for individual i at site j. d_{Mj}, a_j, d_{Yj}, b_j and c_j, indicating the coefficients at site j, are the site-level counterparts for d_M, a, d_Y, b and c in (1.1) and (1.2), and they are assumed to follow a multivariate normal distribution.

Bauer and colleagues have further explored the possibility of quantifying not only the population average but also the between-site variation of the direct effect and the indirect effect through specifying multivariate multilevel models. However, this line of research shares the same limitations as the single-level path analysis and SEM, as explicated in Section 1.2. In addition, although estimating the population average and the between-site variance of the direct effect remains straightforward, estimating the population average and the between-site variance of the indirect effect is nontrivial because it involves estimating the covariance between a_j and b_j; statistical inference is even more challenging. The task would become increasingly daunting if a treatment-by-mediator interaction was under consideration, not to mention possible treatment-by-covariate, mediator-by-covariate, and treatment-by mediator-by-covariate interactions.

Some researchers have incorporated multilevel path analysis models in the causal inference framework, but no solution was provided for estimating and testing the between-site heterogeneity of these effects. The methods developed by VanderWeele (2010b) and Vanderweele, Hong, Jones, and Brown (2013) are useful for evaluating causal mediation mechanisms when treatments are administered at the group level but not for investigating between-site variation in mediation mechanisms in a multisite trial. Bind, Vanderweele, Coull, and Schwartz (2016) examined time-varying treatments and mediators nested within individuals. Even though one may view individuals in this longitudinal study as analogous to sites, the researchers focused only on the population average direct and indirect effects. Other researchers

have extended the instrumental variable (IV) method to multisite trials by using treatment-by-site interactions as instruments for the mediators (Kling, Liebman, &Katz, 2007; Raudenbush, Reardon, & Nomi, 2012; Reardon & Raudenbush, 2013; Reardon, Unlu, Zhu, & Bloom, 2014). With its primary interest in identifying the average effect of each mediator on the outcome, the IV method, when applied to multisite mediation analysis, does not estimate the between-site distributions of the indirect effects. Besides assuming that the exclusion restriction holds at each site, the IV method also assumes that the treatment effect on the mediator is nonzero on average or varies across sites (i.e., $E(a_j) \neq 0$ or $\text{var}(a_j) \neq 0$) and that the site-level treatment effect on the mediator is independent of the site-level mediator effect on the outcome (i.e., $\text{cov}(a_j, b_j) = 0$). This last assumption is plausible only in a limited number of settings. Take the encouragement design as an example and suppose that the experiment was conducted at different schools that serve as sites. In schools that provided high-quality teaching, it is possible that all their students had already been motivated to study extra hours such that the hours of study would be similar between the experimental group and the control group. In these same schools, it is also possible that due to effective teaching, students would have a great amount to gain from their study. In contrast, in schools with low-quality teaching, the control group might study considerably less than the experimental group, and students might have little to gain from their study. This is an example in which the treatment effect on the mediator would be negatively associated with the mediator effect on the outcome, which would violate the last assumption. As far as I know, other methods that allow for a treatment-by-mediator interaction (e.g., Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010) have not been extended to studies of between-site heterogeneity in mediation mechanisms.

To my knowledge, there have been no formal scholarly discussions about the unique research opportunities and methodological challenges that arise in investigations of complex mediation mechanisms that may vary across local settings.

To overcome the limitations of the existing methods and fill the gap in the literature, I extend the novel weighting method, initially developed by Hong (2010, 2015) and others for single-site analysis, to multisite causal mediation analysis with one single mediator or two concurrent mediators. In doing so, I aim to provide a new statistical tool that can be applied broadly to multisite education studies in which not only the population average direct and indirect effects but also the between-site variation of the direct and indirect effects are of scientific interest.

1.4 Empirical Research Questions

The methodological development in this dissertation is motivated by a reanalysis of the multisite experimental data from the National Job Corps Study (NJCS). Job Corps is the largest federal program designed to promote economic well-being among disadvantaged youths in the U.S. who are affiliated with neither school nor work. Intensive education and vocational training are the central elements of the program. Yet unlike most other training programs that were 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. The comprehensive support services provide important protective factors for the vulnerable youths in the process of pursuing education and training and seeking employment. Despite adverse social structural constraints, such supports for risk reduction may enable highly

vulnerable youths to display resilience (Spencer, 2006, 2008; Spencer & Swanson, 2013; Spencer, Swanson, & Harpalani, 2015) and may further reinforce their human capital improvement.

The NJCS sample universe consists of all the 80,883 youths nationwide who applied for Job Corps and were found to be eligible between November 1994 and February 1996 (Schochet, Burghardt, & Glazerman, 2001). Each eligible applicant was associated with one of the more than 100 Job Corps centers that existed at the time of the study. Through a stratified sampling procedure, 15,386 eligible applicants were randomly selected into a nationally representative research sample, among whom 9,409 youths were assigned at random to the program group and 5,977 youths were assigned to 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. Hence, this is a multisite randomized trial in which each Job Corps center served as an experimental site. Participants in the study were interviewed at the time of the random assignment and at 12, 30, and 48 months after randomization. By design, the probability of selection for each follow-up survey differed across individuals.

Population average causal mediation mechanism. By analyzing the NJCS data, researchers (Flores & Flores-Lagunes, 2013; Frumento, Mealli, Pacini, & Rubin, 2012; Lee, 2009; Schochet, Burghardt, & McConnell, 2006, 2008; Zhang et al., 2009) have found that 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. Relying on a bounded local average treatment effect (LATE) approach, Flores and Flores-Lagunes (2013)

has found that for individuals whose educational attainment was improved by the Job Corps program, obtaining an education credential increased employment and earnings. However, no attempt has been made to formally test the Job Corps program theory that emphasizes not only conventional human capital formation through education and vocational training (Becker, 1964; Card, 1999) but also the need to reduce risk exposures and risk behaviors for all the Job Corps participants. To fill the gap, I will address the following research questions in this dissertation:

(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 that are primarily composed of support services for risk reduction?

(3) Given the comprehensiveness of the program and given that support services tend to be lacking under the control condition, did education and training obtained through Job Corps generate a greater impact on earnings than education and training obtained under the control condition? In other words, did Job Corps enhance the economic returns to education and training for disadvantaged youth?

Between-site variance of the causal mediation mechanism. Moreover, 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 Job Corps centers. 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 at some centers. It may be explained by the between-site variation in the causal mediation mechanisms. For example, some Job Corps centers failed to help most participants obtain education or vocational training credentials due to the premature departure of these

participants (Flores, Flores-Lagunes, Gonzalez, & Neumann, 2012). In some other sites, despite an improvement in educational or vocational attainment, it might fail to further enhance the program impact due to important contextual constraints. To be specific, many Job Corps participants obtaining education or training credentials continued to have difficulties securing employment in a sluggish local job market. Moreover, Job Corps programs in the sites where participants have access to and make use of a wider range of support services in addition to education and training might be more effective than those in the sites where service provision and utilization are limited. Hence, a natural next step is to investigate whether the hypothesized program mechanisms operated differently across sites and whether additional investment in education and training or in other support services holds promise for making the program universally effective. I attempt to examine the following questions:

(1) Were Job Corps centers equally effective in increasing earnings through improving educational and vocational attainment?

(2) Were Job Corps centers equally effective in increasing earnings through other pathways that are primarily composed of support services for risk reduction?

(3) Did Job Corps enhance the economic returns to education and training in some centers but not in others?

(4) Did Job Corps centers that increased earnings through improving educational and vocational attainment also tend to be successful in increasing earnings through other pathways?

Causal mediation mechanism involving two concurrent mediators. Becker (1964) made a distinction between *generic human capital* and *job-specific human capital*. The former includes education credentials as a proxy for literacy skills and work ethics; while the latter refers to technical knowledge or skills applicable in a certain vocational trade that may not transfer easily

to other trades. Most job training programs tend to focus solely on vocational training. In contrast, Job Corps places both vocational training and general education at the center of the program. Its general education curriculum prepares those without a high school diploma to become qualified for a GED certificate. The program theory does not clarify, however, whether the education pathway and the vocational training pathway are complementary or mutually reinforcing. Past research has suggested that general education and vocational training are at least complementary (Zimmermannet al, 2013; Blundell, Dearden, Meghir, & Sianesi, 1999). Yet one may argue that, for vocational training to be effective, a student may need basic academic preparation as a pre-requisite. Following this reasoning, high school dropouts may benefit more from vocational training when they work toward a general education credential at the same time rather than receiving vocational training alone. It is of important theoretical interest, therefore, to distinguish the relative contribution of each pathway and determine whether these two types of human capital investments reinforce each other and generate a joint impact greater than the sum of the two separate pathways. Hence, I ask another set of research questions:

(1) What is the average program impact on earnings mediated by vocational training?

(2) What is the average program impact on earnings mediated by general education?

(3) Is the program impact mediated by vocational training reinforced by general education?

(4) What is the average direct effect of the program transmitted through other pathways?

(5) Does the program impact mediated by vocational training vary across the sites?

(6) Does the program impact mediated by general education vary across the sites?

(7) Does the direct effect of the program vary across the sites?

Such evidence will be crucial for enriching theoretical understanding and for informing the design and implementation of education programs alike.

1.5 Organization of the Dissertation

My dissertation is composed of three related studies.

Chapter 2 develops strategies for investigating research questions involving a single mediator. In the Job Corps application, this focal mediator indicates whether an individual had obtained an educational or vocational credential 30 months after the randomization. I incorporate a theoretical model of multisite causal mediation process in the potential outcomes causal framework, define population average indirect effect and direct effect, and conceptualize between-site heterogeneity in the mediation mechanism as novel causal parameters that have not been previously discussed in the causal inference literature. I then develop new statistical methods for the identification, estimation, and inference of not only the population average but also the between-site variation of causal mediation mechanisms in multisite trials. Incorporating RMPW weights, I propose a method-of-moments (MOM) procedure that consistently estimates the causal parameters. I derive asymptotic standard errors for the weighted estimators of the population average effects, reflecting the sampling variability of the RMPW weights estimated based on propensity score models for the mediator. I also conduct a permutation test for the hypothesis testing of the between-site variance of the causal effects. The proposed analytic approach conveniently relaxes the assumption of no treatment-by-mediator interaction while greatly simplifying the outcome model specification without invoking strong distributional assumptions. Hence, it is more broadly applicable than most regression-based strategies. After evaluating the proposed method through simulations, I apply the developed analytic procedures to an empirical investigation of the mediation mechanism of the Job Corps program. The final,

Chapter 3 develops several extensions of the methods proposed in Chapter 2. Large-scale multisite trials are usually conducted based on complex sample and survey designs and are complicated by non-random nonresponse in longitudinal follow-ups. To enhance the external validity and internal validity of multisite causal mediation analysis in multisite trials, I further incorporate into the analytic procedure developed in Chapter 2 a sample weight to adjust for sample and survey designs and an estimated nonresponse weight to account for non-random nonresponse. In addition to decomposing the average program impact into a direct effect and an indirect effect transmitted through a hypothesized focal mediator, as implemented in Chapter 2, I further define, identify, and estimate the treatment-by-mediator interaction effect. The latter addresses the research question regarding whether the economic returns to education and training were greater under Job Corps than under the control condition. I clarify the identification assumptions under which the mediation analysis results are externally and internally valid. I then specify the propensity score models for nonresponse status and those for the mediator and construct the corresponding nonresponse weights and mediator weights. After weighting, I assess balance in the observed covariates between respondents and nonrespondents and between individuals in different mediator categories under each treatment condition. To further evaluate the potential bias related to the omission of confounders or to propensity score model misspecification, I adopt a novel weighting-based sensitivity analysis strategy. Besides, I extend the estimation procedure developed in Chapter 2 to further account for the sampling uncertainty in the estimated nonresponse weights. The updated analytic procedures are expected to

strengthen the external validity and internal validity of the analytic results for the Job Corps application.

In Chapter 4, motivated by a need to explicitly test the theory underlying Job Corps in greater depth, I further extend the refined analytic procedure in Chapter 3 to an investigation of the complex mediation mechanisms that involve two concurrent mediators in multisite trials. To distinguish the relative contributions of vocational training and general education, I consider them as two concurrent mediators. Under the potential outcomes causal framework, I decompose the total Job Corps impact on earnings into an indirect effect transmitted through vocational training, an indirect effect transmitted through general education, and a direct effect attributable to other pathways that are primarily composed of support services.

In the last chapter, I summarize the new methods for the causal mediation analysis in multisite trials and emphasize the methodological contributions of this dissertation. I then highlight the key results obtained from the Job Corps application and discuss their theoretical and practical implications. At the end, I propose an agenda for future research.

The methodological advances will not improve empirical research if they cannot be easily implemented by education researchers. For this reason, I have developed an open-source R package, `MultisiteMediation`, accompanied by a users' manual (http://cran.r-project.org/web/packages/MultisiteMediation). The computer program offers a convenient tool to applied researchers, enabling propensity score analysis, balance checking, estimation of causal parameters, hypothesis testing, as well as sensitivity analysis.

CHAPTER 2

A WEIGHTING METHOD FOR ASSESSING BETWEEN-SITE HETEROGENEITY IN CAUSAL MEDIATION MECHANISM

In this chapter, I present a theoretical model that summarizes key information characterizing the multisite causal mediation process that involves a single mediator under the potential outcomes framework. I then identify a joint distribution of site-specific direct and indirect effects through ratio-of-mediator-probability weighting (RMPW). I also develop new statistical methods for estimation and inference of the causal parameters. In particular, I address challenges when RMPW is unknown and must be estimated from sample data. The weighting-based causal mediation analysis is particularly flexible for accommodating treatment-by-mediator interactions and is suitable for discrete and continuous mediators and outcomes. This is because, unlike other causal mediation methods, the weighting strategy does not require strong assumptions about the functional form of the outcome model. After conducting simulations to evaluate the performance of the proposed approach, I apply the method to a re-analysis of the NJCS data.

2.1 Potential Outcomes Framework for Causal Inference

Rather than defining the causal effects on the basis of arbitrary regression models that often do not hold in reality, I adopt the potential outcomes framework (Holland, 1986, 1988; Neyman & Iwaszkiewicz, 1935; Rubin, 1978) that has previously been extended to causal mediation research (Pearl, 2001; Robins & 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.

As defined in Section 1.2, I use T_{ij} to denote the treatment assignment, M_{ij} for the focal mediator, and Y_{ij} for the outcome, for individual *i* at site *j*, and present the potential outcomes framework in the context of the multisite Job Corps evaluation. Each eligible applicant was assigned at random either to the program group denoted by $t = 1$ or to the control group denoted by $t = 0$. Under either treatment condition, the individual might obtain an education or training credential by the 30-month follow-up, which is denoted by $m = 1$, or might fail to obtain a credential, which is denoted by $m = 0$. For individual i at site j, educational and vocational attainment is a function of the treatment assignment. Hence, $M_{ij}(1)$ and $M_{ij}(0)$ are the individual's respective potential attainment status associated with an assignment to the program group and that to the control group. For each individual, only one potential mediator was observed, depending on which group he or she was actually assigned to.

The individual's earnings in the fourth year after randomization is a final outcome of the treatment. The convention is to use $Y_{ij}(1)$ to represent the potential earnings if one was assigned to the program group and use $Y_{ij}(0)$ for the potential earnings if the same person was assigned to the control group. 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 $t = 0, 1$. Again, only one of the two potential outcomes was observed for each individual while the other remained counterfactual. When $M_{ij}(t) = m$, the individual's potential outcome value associated with treatment t can be written as $Y_{ij}(t, m)$.

In causal mediation analysis, two additional counterfactual outcomes play indispensable roles: $Y_{ij}(1, M_{ij}(0))$ is the individual's potential earnings if assigned to the program group 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 the individual was assigned to the

control group yet counterfactually having the same attainment status as he or she would have under Job Corps. Clearly, neither $Y_{ij}(1, M_{ij}(0))$ nor $Y_{ij}(0, M_{ij}(1))$ was directly observed for any individual.

Table 2.1 Potential Mediators and Outcomes

	Treatment	Potential Mediators		Potential Outcomes			
Individual	T_{ij}	$M_{ij}(1)$	$M_{ij}(0)$	$Y_{ij}(1, M_{ij}(1))$	$Y_{ij}(1, M_{ij}(0))$	$Y_{ij}(0, M_{ij}(1))$	$Y_{ij}(0, M_{ij}(0))$
1	1	0	0	$Y_{ij}(1,0)$	$Y_{ij}(1,0)$	$Y_{ij}(0,0)$	$Y_{ij}(0,0)$
2	1	1	1	$Y_{ij}(1,1)$	$Y_{ij}(1,1)$	$Y_{ij}(0,1)$	$Y_{ij}(0,1)$
3	1	0	1	$Y_{ij}(1,0)$	$Y_{ij}(1,1)$	$Y_{ij}(0,0)$	$Y_{ij}(0,1)$
4	1	0	1	$Y_{ij}(1,0)$	$Y_{ij}(1,1)$	$Y_{ij}(0,0)$	$Y_{ij}(0,1)$
5	0	1	0	$Y_{ij}(1,1)$	$Y_{ij}(1,0)$	$Y_{ij}(0,1)$	$Y_{ij}(0,0)$
6	0	1	1	$Y_{ij}(1,1)$	$Y_{ij}(1,1)$	$Y_{ij}(0,1)$	$Y_{ij}(0,1)$
7	0	0	0	$Y_{ij}(1,0)$	$Y_{ij}(1,0)$	$Y_{ij}(0,0)$	$Y_{ij}(0,0)$
8	0	1	0	$Y_{ij}(1,1)$	$Y_{ij}(1,0)$	$Y_{ij}(0,1)$	$Y_{ij}(0,0)$

Table 2.1 illustrates potential mediators and outcomes with eight individuals when the mediator is binary. $M_{ij}(1)$ and $M_{ij}(0)$ are random variables, taking value 1 or 0. For each individual, I list the potential mediator value under each treatment condition and correspondingly four potential outcomes. For the first four individuals assigned to the program group, only $M_{ij}(1)$ and $Y_{ij}(1, M_{ij}(1))$ are observable, while for the next four individuals assigned to the control group, only $M_{ij}(0)$ and $Y_{ij}(0, M_{ij}(0))$ are observable.

2.2 Stable Unit Treatment Value Assumption

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. 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 & Raudenbush, 2006; Hudgens & Halloran, 2008). That is, the potential intermediate outcomes of individual i at site j are independent of the treatment assignments of individuals at site j' for all $j' \neq j$, and this individual's potential outcomes are independent of the treatment assignments and potential mediator value assignments of individuals at site j'. Because applicants are usually assigned to Job Corps centers relatively close to their original residences and because Job Corps centers are sparsely located on the map, between-site interference seems unlikely.

2.3 Definition of the Causal Parameters

2.3.1 Individual-Specific Causal Effects

Under SUTVA, for individual *i* at site *j*, the intent-to-treatment (ITT) effect of the treatment on the mediator, also known as the total effect of the treatment assignment on the mediator, is defined as $\alpha_{ij} = M_{ij}(1) - M_{ij}(0)$, and the ITT effect of the treatment on the outcome is defined as $\beta_{ij}^{(T)} = Y_{ij}\ (1, M_{ij}(1))\ -\ Y_{ij}(0, M_{ij}(0))$. The superscript in $\beta_{ij}^{(T)}$ serves as a shorthand for the total effect of the treatment assignment on the outcome.

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

$$\beta_{ij}^{(I)}(1) = Y_{ij}\left(1, M_{ij}(1)\right) - Y_{ij}\left(1, M_{ij}(0)\right).$$

The individual-specific NIE represents the Job Corps impact on earnings under Job Corps attributable to the program-induced change in the individual's educational and vocational

attainment from $M_{ij}(0)$ to $M_{ij}\ (1)$. $\beta_{ij}^{(I)}\ (1)$ is called "the total indirect effect" by Robins and Greenland (1992), who distinguished it from the individual-specific "pure indirect effect" (PIE)

$$\beta_{ij}^{(I)}(0) = Y_{ij}\left(0, M_{ij}(1)\right) - Y_{ij}\left(0, M_{ij}(0)\right).$$

The individual-specific PIE represents the impact on earnings under the control condition when educational and vocational attainment is changed from $M_{ij}(0)$ to $M_{ij}(1)$. The superscript in $\beta_{ij}^{(I)}(1)$ and $\beta_{ij}^{(I)}(0)$ serves as a shorthand for the indirect effects. Clearly, mediation does not exist if Job Corps has no impact on one's educational and vocational attainment. In such case, both NIE and PIE are zero.

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

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

The individual-specific NDE represents the Job Corps impact on earnings while holding the individual's educational and vocational attainment at the level that would be realized under the control condition. The direct effect is nonzero if the Job Corps program exerted an impact on earnings without changing an individual's educational and vocational attainment. Robins and Greenland (1992) called $\beta_{ij}^{(D)}(0)$ "the pure direct effect" in contrast with "the total direct effect", $\beta_{ij}^{(D)}(1) = Y_{ij}\ (1, M_{ij}\ (1)) - Y_{ij}\ (0, M_{ij}\ (1))$. The latter is the Job Corps impact on earnings while holding educational and vocational attainment at the level that would be realized under the Job Corps condition. The superscript in $\beta_{ij}^{(D)}(0)$ and $\beta_{ij}^{(D)}(1)$ serves as a shorthand for the individual-specific direct effects.

The individual-specific total treatment effect is the sum of the individual-specific NIE and NDE: $\beta_{ij}^{(T)} = \beta_{ij}^{(I)}(1) + \beta_{ij}^{(D)}(0)$. Alternatively, one may decompose the individual-specific total treatment effect into PIE and the total direct effect: $\beta_{ij}^{(T)} = \beta_{ij}^{(I)}(0) + \beta_{ij}^{(D)}(1)$.

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. Because the comprehensive support services provided by the Job Corps program tend to be lacking under the control condition, obtaining an education or training credential under Job Corps might bring greater economic returns than obtaining a similar credential under the control condition. Therefore, NIE and PIE may not be equal. The difference between the two is defined as the natural treatment-by-mediator interaction effect (Hong, 2015; Hong et al., 2015), which quantifies the treatment effect on the outcome transmitted through a change in the mediator-outcome relationship:

$$\beta_{ij}^{(T\times M)} = \beta_{ij}^{(I)}(1) - \beta_{ij}^{(I)}(0).$$

A nonzero interaction effect will indicate that the program-induced change in educational and vocational attainment influences earnings differently between the Job Corps condition and the control condition.

Table 2.2 summarizes the individual-specific causal effects defined above.

Table 2.2 Definitions of Individual-Specific Causal Effects

	Individual-Specific Effect	Definition
ITT effect on the mediator	$\alpha_{ij} = M_{ij}(1) - M_{ij}(0)$	Effect of treatment assignment on the mediator
ITT effect on the outcome	$\beta_{ij}^{(T)} = Y_{ij}\ (1, M_{ij}(1))\ -\ Y_{ij}(0, M_{ij}(0))$	Effect of treatment assignment on the outcome
NDE	$\beta_{ij}^{(D)}(0)\ =\ Y_{ij}(1, M_{ij}(0))\ -\ Y_{ij}(0, M_{ij}(0))$	Treatment effect on the outcome if the treatment fails to change the mediator
NIE	$\beta_{ij}^{(I)}(1) =\ Y_{ij}(1, M_{ij}(1))\ -\ Y_{ij}(1, M_{ij}(0))$	Treatment effect on the outcome under the experimental condition attributable to the treatment-induced change in the mediator
PIE	$\beta_{ij}^{(I)}(0) =\ Y_{ij}(0, M_{ij}(1))\ -\ Y_{ij}(0, M_{ij}(0))$	Treatment effect on the outcome under the control condition attributable to the treatment-induced change in the mediator
Interaction effect	$\beta_{ij}^{(T\times M)} = \beta_{ij}^{(I)}(1) - \beta_{ij}^{(I)}(0)$	Treatment effect on the outcome transmitted through a change in the mediator-outcome relationship

2.3.2 Site-Specific Causal Effects and Population Parameters

I define the site-specific causal effects including the ITT effects of the treatment on the mediator and the outcome, NDE, NIE, PIE, and the natural treatment-by-mediator interaction effect, by taking an average of the corresponding individual-specific causal effects over the population of eligible Job Corps applicants at a given site. The site-specific effects are listed in the second column in Table 2.3 in which $S_{ij} = j$ indicates the site membership of individual i.

As emphasized earlier, of particular theoretical interest in the Job Corps evaluation is not only the overall average of each of these causal effects but also their possible variations across the sites. Because the composition of applicants, the composition of Job Corps staff, the center operator, and various elements of the control condition tend to be fluid rather than static, I consider a theoretical population of sites that are potentially infinite in number. 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. The population parameters that characterize the distributions of the site-specific causal effects include the ITT effects of the treatment on the

mediator and the outcome, NDE, NIE, PIE, and the natural treatment-by-mediator interaction effect, each averaged over the population of sites, as well as the between-site variance of each site-specific effect.

I have listed in Table 2.3 the research questions with regard to the population average causal effects over all the sites in column 3 and the corresponding notations 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 notations. Besides, I am also interested in the covariance between the site-specific NDE and NIE, $\sigma_{D(0),I(1)} = \text{cov}(\beta_j^{(D)}(0), \beta_j^{(I)}(1))$, 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.

In the rest of this chapter, I focus on the identification and estimation of the population average and between-site variance of NDE and NIE. I will further discuss PIE and the interaction effect in the next chapter.

2.4 Identification Assumptions

For each causal effect, its average over the population of sites and its between-site variance can be easily identified if all the potential mediators and potential outcomes are observed for all the individuals in the population of eligible applicants at every site. However, we are able to observe $M_{ij}(t)$ and $Y_{ij}(t, M_{ij}(t))$ for $t = 0, 1$ only if individual i at site j was assigned to treatment t. In addition, we never directly observe one's potential outcome of assignment to treatment t while the mediator would counterfactually take the value that one would have under 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 I clarify the assumptions under which each of the population

Table 2.3 Definitions of Population Average and Between-Site Variances

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

parameters can be identified from the observed data. These assumptions should not be taken lightly. Rather, they require close scrutiny on scientific grounds.

Assumption 2.1 (Strongly ignorable treatment assignment). Within levels of the observed pretreatment covariates $\mathbf{x}_T$, the treatment assignment is independent of all the potential mediators and potential outcomes at each site.

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

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. 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 $0 < Pr(T_{ij} = t | \mathbf{X}_{Tij} = \mathbf{x}_T, S_{ij} = j) < 1$. That is, each individual had a nonzero probability of being assigned to either treatment group at a given site.

Assumption 2.1 enables the identification of the ITT effects, while identifying NDE and NIE is considerably more challenging. This is because these two mediation-related causal effects involve the counterfactual outcome $Y_{ij}(1, M_{ij}(0))$ that cannot be directly observed; this is additionally because the mediator value assignment under each treatment was not experimentally manipulated. Using the RMPW strategy, Hong (2010) derived an important identification result that requires the following assumption about the strong ignorability of mediator values in a single site.

Assumption 2.2 (Strongly ignorable mediator value assignment). Within levels of the observed pretreatment covariates denoted by $\mathbf{x}_M$, the mediator value assignment for individuals under either treatment condition is independent of the potential outcomes at each site.

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

for all possible values of t and m where $t \neq t'$. Note that $M_{ij}(0)$ is counterfactual for program group members and that $M_{ij}(1)$ is counterfactual for control group members. Under Assumption 2.2, $M_{ij}(1)$ and $M_{ij}(0)$ are both independent of $Y_{ij}(1, m)$ for individuals 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 individuals in the control group at the site who share the same covariate values.

Assumption 2.2 implies that among individuals who share the same observed pretreatment characteristics $\mathbf{x}_M$, the assignment of mediator values is as if randomized within each treatment condition or across treatment conditions at any site. 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, the theoretically important 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, I additionally assume that $0 < Pr(M_{ij}(t) = m | T_{ij} = t, \mathbf{X}_{Mij} = \mathbf{x}_M, S_{ij} = j) < 1$. That is, each individual 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.

Assumptions 2.1 and 2.2 constitute the "sequential ignorability" (Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010) at eat site. In a hypothetical experiment for causal mediation analysis, individuals within each site would be randomized to the experimental or the control condition; subsequently, individuals would be assigned at random to obtain an education credential under each treatment condition. Alternatively, the treatment assignment would be randomized within subgroups of individuals who share the same observed pretreatment characteristics; and subsequently, the randomization to obtain an education credential under each treatment condition would be conducted within subgroups of individuals who share the same observed pretreatment characteristics. These hypothetical sequential randomized designs satisfy the sequential ignorability assumption.

However, in multisite studies such as NJCS, because individuals were not randomized to receive a mediator value after the treatment randomization, Assumption 2.2 becomes particularly strong. The plausibility of this assumption relies heavily on the richness of the observed pretreatment covariates. This assumption also requires that there is no posttreatment covariate that confounds the mediator–outcome relationship (Avin, Shpitser, & Pearl, 2005; VanderWeele, 2010b; Vanderweele et al., 2013). An example of a possible violation is that, if among individuals with the same baseline characteristics, those who are more likely to obtain an education credential are also the ones who tend to receive more counseling services, then the indirect effect mediated by educational attainment would be confounded by the program benefit transmitted through counseling services. The sequential ignorability assumption must hold at every site. If the assumption is violated in one or more sites, the causal parameters will likely be identified with bias. For this reason, the sequential ignorability assumption in the multisite setting is seemingly stronger than that in the single-site setting. Assessing the sensitivity of

analytic results to possible violations of these identification assumptions is a necessary step in applications.

2.5 Identification through Propensity Score-Based Weighting

Under the sequential ignorability, the site-specific average of each potential outcome is identifiable, which then enables the identification of the site-specific causal effects. Here, I discuss the general case in which the treatment assignment and the mediator value assignment under each treatment condition are strongly "ignorable" within each subgroup of individuals who share the same observed pretreatment characteristics.

Weighting adjustment for treatment assignment. When the probability of treatment assignment is determined as a function of individual characteristics, certain subpopulations will become over-represented while others under-represented in a given treatment group. Extending the logic of sample weighting to causal inference, an analyst may apply IPTW (Horvitz & Thompson, 1952; Robins, 1999; Rosenbaum, 1987) to individual i at site j with pretreatment characteristics $\mathbf{x}_T$ who has been assigned to treatment t:

$$W_{Tij} = \frac{Pr(T_{ij} = t|\, S_{ij} = j)}{Pr(T_{ij} = t|\mathbf{X}_{Tij} = \mathbf{x}_T, S_{ij} = j)} \text{ for } t = 0, 1. \quad (2.1)$$

The numerator is the average probability of assigning an 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.

Theorem 2.1. Under Assumption 2.1, the site-specific average potential mediator and potential outcome under treatment t for $t = 0,1$ can be respectively identified by the average of the observed mediator and outcome under treatment condition t at site j, weighted by the IPTW weight:

$$E\left[M_{ij}\,(t)\,|S_{ij}\;=\;j\right]=E\left[W_{Tij}M_{ij}\,|T_{ij}\;=t,S_{ij}=j\right],$$

$$E\left[Y_{ij}(t,M_{ij}\,(t))\,|S_{ij}\;=\;j\right]=E\left[W_{Tij}Y_{ij}\,|T_{ij}\;=t,S_{ij}=j\right].$$

The proof of Theorem 2.1 is presented in Appendix 2.A. This weighting transforms the experimental group composition and the control group composition such that the probability of treatment assignment in the weighted sample would resemble that in a hypothetical randomized design with equal probability of treatment assignment for all individuals. In other words, applying W_{Tij} to individuals with pretreatment characteristics $\mathbf{x}_T$ who have been assigned to treatment *t* at site *j* removes bias due to treatment selection associated with $\mathbf{X}_{Tij}$.

Weighting adjustment for mediator value selection in treatment effect decomposition. In a hypothetical sequential randomized experiment, the randomization of treatment assignment would be followed by a randomization of mediator value assignment within every treatment group at a site. The probability of mediator value assignment under each treatment would reflect the treatment effect on the mediator. The population average of $Y(1,M(0))$ at the site would be identified by the weighted mean observed outcome of the program group, in which the weight $\frac{Pr\left(M_{ij}=m|\,T_{ij}=0,S_{ij}=j\right)}{Pr\left(M_{ij}=m|T_{ij}=1,S_{ij}=j\right)}$, namely ratio-of-mediator-probability weighting (RMPW), would transform the mediator distribution in the program group to resemble that in the control group. In NJCS, only the treatment was experimentally randomized. Yet when Assumption 2.2 holds, the mediator value assignment could be viewed as if it were randomized for individuals sharing the same covariate values $\mathbf{x}_M$. One may apply RMPW within the subpopulations defined by $\mathbf{x}_M$.

Theorem 2.2. Under Assumptions 2.1 and 2.2, the site-specific average potential outcome under the experimental condition yet with the potential mediator counterfactually taking its distribution associated with the alternative control condition can be identified by the average of

the observed outcome in the experimental group at site *j*, weighted by the product of the IPTW and RMPW weights:

$$E\left[Y_{ij}(1, M_{ij}(0)) \,|S_{ij} = j\right] = E\left[W_{Tij}W_{Mij}\, Y_{ij} \mid T_{ij} \,= 1, S_{ij} = j\right]$$

for $t \neq t'$, where

$$W_{Mij} = \frac{Pr\left(M_{ij} = m|T_{ij} = 0, \mathbf{X}_{Mij} = \mathbf{x}_M, S_{ij} = j\right)}{Pr\left(M_{ij} = m|T_{ij} = 1, \mathbf{X}_{Mij} = \mathbf{x}_M, S_{ij} = j\right)} \;\forall\, m \in \mathcal{M}. \tag{2.2}$$

For individual i at site j who were assigned to the experimental group and displayed mediator value m, W_{Mij}, known as RMPW, is a ratio of an experimental individual's conditional probability of displaying mediator value *m* under the counterfactual control condition to that under the assigned experimental condition. For individuals within levels of the pretreatment characteristics $\mathbf{x}_M$ at each site, RMPW transforms the mediator distribution in the experimental group to resemble that in the control group. Applying the product of the IPTW weight and the RMPW weight to the experimental group at each site, we are able to identify $E\left[Y_{ij}\left(1, M_{ij}(0)\right) |S_{ij} = j\right]$, which is the site-specific average counterfactual mean outcome associated with the experimental condition when the mediator counterfactually distributes the same as that under the control condition. RMPW is mathematically equivalent to the inverse probability weight (IPW) proposed by Huber (2014). Appendix 2.A presents a proof of Theorem 2.2.

This identification result enables us to relate the observable data to the average counterfactual quantities at a site. When the treatment assignment is randomized within a site, $Pr\left(T_{ij} = t|\, S_{ij} = j\right) = Pr\left(T_{ij} = t|\mathbf{X}_{Tij} = \mathbf{x}_T, S_{ij} = j\right)$, we simply have that $W_{Tij} = 1$. To

simplify the notation, let ν_{tj}, μ_{tj}, and μ_{*1j} represent each of the observable quantities at site *j*, which I use to identify the site-specific average counterfactual outcomes.

$$\nu_{tj} = E\left[M_{ij}|T_{ij} = t, S_{ij} = j\right],$$

$$\mu_{tj} = E\left[Y_{ij}|T_{ij} = t, S_{ij} = j\right],$$

$$\mu_{*1j} = E\left[W_{Mij}Y_{ij}|T_{ij} = 1, S_{ij} = j\right].$$

Here ν_{tj} is the average of the observed mediator in treatment group *t* at site *j*; μ_{tj} is the average of the observed outcome in treatment group *t* at site *j*; and μ_{*1j} is the average of the observed outcome weighted by the RMPW weight in the experimental group at site *j*. When the identification assumptions hold, ν_{tj} identifies $E\left[M_{ij}(t)|S_{ij} = j\right]$, μ_{tj} identifies $E\left[Y_{ij}\left(t, M_{ij}(t)\right)|S_{ij} = j\right]$, and μ_{*1j} identifies $E\left[Y_{ij}\left(1, M_{ij}(0)\right)|S_{ij} = j\right]$. Table 2.4 summarizes these identification results. The first column lists the site-specific causal effects defined in terms of the counterfactual quantities as explicated in Section 2.3.1; 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 are then able to identify correspondingly the population average and between-site variance of each causal effect as defined in Section 2.3.

Table 2.4 Identification of the Site-Specific Effects

Site-Specific Effect	Identified by	Assumptions
ITT effect on the mediator α_j	$\nu_{1j} - \nu_{0j}$	Assumption 2.1
ITT effect on the outcome $\beta_j^{(T)}$	$\mu_{1j} - \mu_{0j}$	
NDE $\beta_j^{(D)}(0)$	$\mu_{*1j} - \mu_{0j}$	Assumptions 2.1 ~ 2.2
NIE $\beta_j^{(I)}(1)$	$\mu_{1j} - \mu_{*1j}$	

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

2.6 Estimation and Inference

The estimation involves two major steps. Step 1 estimates the weight for each individual in the experimental group as a ratio of the conditional probability of mediator value under the experimental condition to that under the control condition corresponding to Equation (2.2). Step 2 estimates the unweighted mean outcome of the control group, the unweighted mean outcome of the experimental group, the weighted mean outcome of the experimental group for each site, and subsequently the site-specific ITT effects, NDE and NIE as identified in Table 2.4. Based on these site-specific estimates, I estimate the population average and the between-site variance of the causal effects.

In Step 1, following the convention of propensity score estimation in multilevel data, I fit multilevel mixed-effects logistic regression models to the sample data in each treatment group pooled from all the sites and estimate the coefficients through maximum likelihood. In Step 2, I employ an MOM estimation procedure to estimate the site-specific effects and the first and second moments of their joint distribution. This procedure estimates the between-site variance of the effects by purging the average sampling variance off the total between-site variance of these effects. However, the analysis in Step 2 is complicated by the fact that the causal parameters must be estimated on the basis of the estimated weight rather than the true weight. I propose asymptotic variance estimators for the population average effect estimators that incorporate the sampling variability in the weight estimation. I also conduct a permutation test for variance testing.

I choose MOM rather than maximum likelihood estimator (MLE) in Step 2 for three reasons. First, the likelihood in Step 2 is a function of the parameters, given both the observed outcome and the estimated individual weight. The unknown distribution of the weight adds difficulty to the specification of the likelihood function. Second, the preliminary results suggest that the site-specific effects are not normally distributed. MOM does not invoke assumptions about the distribution of the site-specific effects and thus has a potential for broad applications. Third, 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 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.

This section starts by introducing the weighted MOM estimators of the causal effects in a hypothetical scenario in which the weight is known. I then discuss a strategy of obtaining the asymptotic sampling variance of the causal effect estimates when the weight needs to be estimated. At the end, I explain the estimation and hypothesis testing for the between-site variance of the effects.

2.6.1 Method-of-Moments Estimators of the Population Average Causal Effects When the Weight Is Known

To estimate the population average effects, I first estimate the effects site-by-site and then aggregate the site-specific effect estimates (e.g., Diggle, Heagerty, Liang, & Zeger, 2002; Raudenbush & Bloom, 2015).

To estimate $\mu_{*1j} = E\left[W_{Mij}Y_{ij}|T_{ij} = 1, S_{ij} = j\right]$, I simply obtain a weighted sample mean outcome of those assigned to the experimental condition at site *j*,

$$\hat{\mu}_{*1j} = \frac{\sum_{i=1}^{N} W_{Mij}\, I\left(S_{ij} = j\right) T_{ij} Y_{ij}}{\sum_{i=1}^{N} W_{Mij} I\left(S_{ij} = j\right) T_{ij}}, \tag{2.3}$$

where n_j is the sample size at site *j* and $I\left(S_{ij} = j\right)$ is an indicator for whether individual i was a member of site j. The weight is $W_{Mij} = p_{0ij}/p_{1ij}$ when $M_{ij} = 1$ and $W_{Mij} = (1 - p_{0ij})/(1 - p_{1ij})$ when $M_{ij} = 0$. Here $p_{1ij} = Pr\left(M_{ij} = 1|T_{ij} = 1, \mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j\right)$ is the probability of obtaining an education credential under the experimental condition and $p_{0ij} = Pr\left(M_{ij} = 1|T_{ij} = 0, \mathbf{X}_{ij} = \mathbf{x}, S_{ij} = j\right)$ is the corresponding probability under the control condition, for individual *i* in site *j* with pretreatment characteristics $\mathbf{X}_{ij} = \mathbf{x}$, where $\mathbf{X}_{ij} = \mathbf{X}_{Tij} \cup \mathbf{X}_{Mij}$. Because we can remove mediator selection by controlling for $\mathbf{X}_{Mij}$ under the strongly ignorable mediator value assignment assumption and because $\mathbf{X}_{Mij} \subset \mathbf{X}_{ij}$, p_{1ij} is essentially equal to $Pr\left(M_{ij} = 1|T_{ij} = 1, \mathbf{X}_{Mij} = \mathbf{x}_M, S_{ij} = j\right)$ and p_{0ij} is essentially equal to $Pr\left(M_{ij} = 1|T_{ij} = 0, \mathbf{X}_{Mij} = \mathbf{x}_M, S_{ij} = j\right)$.

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

$$\hat{\mu}_{0j} = \frac{\sum_{i=1}^{N} I\left(S_{ij} = j\right)(1 - T_{ij}) Y_{ij}}{\sum_{i=1}^{N} I\left(S_{ij} = j\right)(1 - T_{ij})}, \tag{2.4}$$

$$\hat{\mu}_{1j} = \frac{\sum_{i=1}^{N} I\left(S_{ij} = j\right) T_{ij} Y_{ij}}{\sum_{i=1}^{N} I\left(S_{ij} = j\right) T_{ij}}, \tag{2.5}$$

Similarly, the estimators of the control mean mediator ν_{0j} and the experimental mean mediator ν_{1j} are:

$$\hat{v}_{0j} = \frac{\sum_{i=1}^{N} I\left(S_{ij} = j\right)(1 - T_{ij})M_{ij}}{\sum_{i=1}^{N} I\left(S_{ij} = j\right)(1 - T_{ij})}, \tag{2.6}$$

$$\hat{v}_{1j} = \frac{\sum_{i=1}^{n_j} I\left(S_{ij} = j\right) T_{ij} M_{ij}}{\sum_{i=1}^{n_j} I\left(S_{ij} = j\right) T_{ij}}, \tag{2.7}$$

Correspondingly, we could obtain the MOM estimator of each site-specific causal effect as identified in Table 2.4.

$$\hat{\alpha}_j = \hat{v}_{1j} - \hat{v}_{0j}, \tag{2.8}$$

$$\hat{\beta}_j^{(T)} = \hat{\mu}_{1j} - \hat{\mu}_{0j}, \tag{2.9}$$

$$\hat{\beta}_j^{(D)}(0) = \hat{\mu}_{*1j} - \hat{\mu}_{0j}, \tag{2.10}$$

$$\hat{\beta}_j^{(I)}(1) = \hat{\mu}_{1j} - \hat{\mu}_{*1j}. \tag{2.11}$$

I then estimate the parameters that characterize the distribution of site-specific causal effects for the population of sites. Because each site-specific effect is estimated through a mean contrast at each site, the estimation procedure for the between-site variance and population average of the ITT effects is essentially the same as that for the between-site variance and population average of NDE and NIE. Hence, for simplicity, I focus on the estimation and inference algorithms for the latter below.

Let the estimates of the site-specific mean potential outcomes across all the sites be $\widehat{\boldsymbol{\mu}} = (\widehat{\boldsymbol{\mu}}_1', \cdots, \widehat{\boldsymbol{\mu}}_J')'$, in which $\widehat{\boldsymbol{\mu}}_j = (\hat{\mu}_{0j}, \hat{\mu}_{*1j}, \hat{\mu}_{1j})'$ for $j = 1, \ldots, J$, and let $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1', \cdots \widehat{\boldsymbol{\beta}}_J')'$, in which $\widehat{\boldsymbol{\beta}}_j = (\hat{\beta}_j^{(D)}(0), \hat{\beta}_j^{(I)}(1))'$. The estimators of site-specific NDE and NIE across all the sites can be written as

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\Phi}\widehat{\boldsymbol{\mu}}, \tag{2.12}$$

where $\mathbf{\Phi} = \mathbf{I}_J \otimes \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$, in which $\mathbf{I}_J$ is a $J \times J$ identity matrix.

When the sites have been sampled with equal probability from the population of sites, by taking a simple average of the above consistent estimates of the site-specific NDE and NIE across all the *J* sites in the sample, we could obtain consistent estimators of the average NDE and NIE for the population of sites,

$$\widehat{\boldsymbol{\gamma}} = \frac{1}{J}\sum_{j=1}^{J} \widehat{\boldsymbol{\beta}}_j, \tag{2.13}$$

in which $\widehat{\boldsymbol{\gamma}} = (\hat{\gamma}^{(D)}(0), \hat{\gamma}^{(I)}(1))'$. Equivalently, it can be written as

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{\Psi}'\mathbf{\Psi})^{-1}\mathbf{\Psi}'\widehat{\boldsymbol{\beta}}, \tag{2.14}$$

where $\mathbf{\Psi} = \mathbf{1}_J \otimes \mathbf{I}_2$, in which $\mathbf{1}_J$ is a $J \times 1$ vector of 1's and $\mathbf{I}_2$ is a 2×2 identity matrix.

An alternative precision-weighted estimator would use the inverse of the covariance matrix of the site-specific effect estimates as the weight. Even though precision weighting is expected to improve efficiency, it may introduce bias and inconsistency if the precision weight is correlated with the effect size of the site-specific effect (Raudenbush and Schwartz, working paper). Hence, I do not opt for precision weighting in this dissertation.

2.6.2 Asymptotic Sampling Variance of the Population Average Causal Effect Estimates When the Weight Is Unknown

In a typical multisite randomized experiment, even though the treatment assignment is randomized, the mediator value assignment is not. Hence, the weight is unknown and needs to be estimated from the sample data in Step 1 prior to the estimation of the causal effects in Step 2. In

the analytic procedure that I delineate below, a multilevel logistic regression analysis is employed in Step 1 to estimate the weight while Step 2 involves site-by-site MOM analysis.

2.6.2.1 Two-step estimation procedures

In Step 1, I estimate the mediator probabilities and correspondingly the RMPW weights. To reflect the differences in mediator selection mechanisms between the program group and the control group, I estimate the mediator probability under each treatment condition by fitting a logistic regression to the corresponding group. To capture between-site differences in the conditional mediator probability in each treatment group, I opt for specifying mixed-effects models each with a site-specific random intercept. Hence, I fit the following mixed-effects logistic regression to the sample of each treatment group to estimate p_{tij}, for t = 0, 1,

$$\log\left[\frac{p_{tij}}{1-p_{tij}}\right] = \boldsymbol{X}'_{ij}\boldsymbol{\pi}_t + r_{tj}, r_{tj} \sim N(0, \sigma_t^2). \tag{2.15}$$

Here, $\mathbf{X}_{ij}$ is a vector of covariates including the intercept, $\boldsymbol{\pi}_t$ is the corresponding vector of coefficients, and r_{tj} is the random intercept which follows a normal distribution with a variance of σ_t^2. If a covariate predicts the mediator differently across the sites, a site-specific random slope can be included as well. In addition to the sequential ignorability, the multilevel logistic regression model comes with its model-based assumptions with regard to the relationships between $\mathbf{X}_{ij}$ and p_{tij} and the distribution of the random intercept. The analysis can be conducted through MLE using iterative generalized last squares. The random intercept is estimated through an empirical Bayes procedure.

I predict p_{1ij} for each individual in the experimental group directly based on the propensity score model fitted to the experimental group data. To predict p_{0ij} for the same individuals, I

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

The Step-2 estimation is similar to that described in Section 2.6.1 except that we need to replace W_{Mij} with $\widehat{W}_{Mij}$. In the existing literature on propensity score–based weighting in multilevel settings (e.g., Leite et al., 2015), propensity score estimation and causal effect estimation are conducted separately. In this way, however, the sampling variability of the estimated weight obtained in Step 1 will not be represented in the standard errors of the causal effect estimates obtained in Step 2. Moreover, because I analyze the propensity score models by pooling data from all the sites, the predicted propensity scores and correspondingly the estimated weights are inevitably correlated between sites. Separating the two steps in analysis would lead to bias in estimating the standard errors for the estimated population average effects. As shown later in the simulation study, the problem becomes salient especially when the site size is small. To deal with this challenge, I extend the strategy that Newey (1984) proposed under the single-

level setting. Specifically, I stack the estimating equations from the two steps and solve them simultaneously. By doing so, the second-order conditions for the site-specific effect estimators are considered with respect to the parameters that must be estimated in Step 1. Intuitively, the stacking allows the Step 1 estimation to be configured into the Step 2 estimation. The two-step estimators can be fit into the generalized method of moments (GMM) framework (Hansen, 1982). This idea has been applied in causal inference in single-level settings. For example, Hirano and Imbens (2001) utilized it in the estimation of the total treatment impact using propensity score weighting. Bein et al. (2018) applied the strategy to RMPW-based single-site causal mediation analysis. Here I adapt the estimation procedure to multisite causal mediation analysis.

2.6.2.2 Asymptotic sampling variance of the causal effect estimates

Let $\boldsymbol{h}_{ij}^{(1)}$ denote the moment functions for the Step-1 parameter estimators $\widehat{\boldsymbol{\eta}}$. Here $\widehat{\boldsymbol{\eta}}$ includes the estimators of the coefficients in the multilevel logistic regression models as well as the standard deviation of the random intercept. Let $\boldsymbol{h}_{ij}^{(2)}$ denote the moment functions for the Step-2 parameter estimators $\widehat{\boldsymbol{\mu}}$. Here $\widehat{\boldsymbol{\mu}}$ includes the estimators of all the site-specific potential outcome means. Appendix 2.B provides details of these moment functions. Stacking the moment functions from both steps, we have that

$$\boldsymbol{h}_{ij} = \begin{bmatrix} \boldsymbol{h}_{ij}^{(1)} \\ \boldsymbol{h}_{ij}^{(2)} \end{bmatrix}. \tag{2.16}$$

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

$$\sqrt{N}\left(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}\right) \xrightarrow{d} N\left(\mathbf{0}, \widehat{\text{var}}\left(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}\right)\right). \tag{2.17}$$

The asymptotic normal distribution enables computation of sensible confidence intervals and tests when the site-specific effects or the outcome are not normally distributed. Details on the consistent estimator of $\widehat{\text{var}}\left(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}\right)$ can be found in Appendix 2.B.

Subsequently, I derive the sampling variance of the site-specific NDE and NIE estimators. Based on Equation (2.12), it is easy to derive that

$$\text{var}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) = \boldsymbol{\Phi}\text{var}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})\boldsymbol{\Phi}'. \tag{2.18}$$

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

Correspondingly, for the population average NDE and NIE estimators given in Equation (2.14), the sampling variance is

$$\text{var}(\widehat{\boldsymbol{\gamma}}) = (\boldsymbol{\Psi}'\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'\text{var}\left(\widehat{\boldsymbol{\beta}}\right)\boldsymbol{\Psi}(\boldsymbol{\Psi}'\boldsymbol{\Psi})^{-1}, \tag{2.19}$$

in which

$$\text{var}\left(\widehat{\boldsymbol{\beta}}\right) = \text{var}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}+\boldsymbol{\beta}\right) = \text{var}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) + \text{var}(\boldsymbol{\beta}), \tag{2.20}$$

where $\text{var}(\boldsymbol{\beta}) = \mathbf{I}_J \otimes \text{var}(\boldsymbol{\beta}_j)$. The between-site variance of NDE and NIE $\text{var}(\boldsymbol{\beta}_j)$ is of key scientific interest. I discuss its estimation in the next subsection. The consistent estimator of $\text{var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ has been obtained as shown above. After further obtaining the consistent estimator of $\text{var}(\boldsymbol{\beta}_j)$, we will be able to consistently estimate the asymptotic standard errors for the estimators of the population average NDE and NIE.

2.6.3 Estimation and Inference of Between-Site Variance and Covariance of Causal Effects

I estimate the between-site variance and covariance of NDE and NIE again through the method of moments. Although a simple average of $\widehat{\boldsymbol{\beta}}_j$ is consistent for $\boldsymbol{\gamma}$, a simple variance of $\widehat{\boldsymbol{\beta}}_j$ is biased for $\text{var}(\boldsymbol{\beta}_j)$ because this variance estimator contains the sampling variance of $\widehat{\boldsymbol{\beta}}_j$. To be specific, the total between-site variance of the site-specific effect estimator $\text{var}(\widehat{\boldsymbol{\beta}}_j)$ is equal to the sum of the within-site sampling variance $\text{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ and the between-site variance of the site-specific effect $\text{var}(\boldsymbol{\beta}_j)$:

$$\text{var}(\widehat{\boldsymbol{\beta}}_j) = \text{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j + \boldsymbol{\beta}_j) = \text{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + \text{var}(\boldsymbol{\beta}_j). \tag{2.21}$$

Hence, by subtracting the average within-site sampling variance estimator from the average total variance estimator, I obtain a consistent estimator of the between-site variance of $\boldsymbol{\beta}_j$. As shown in Appendix 2.C, this estimator is

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

$$-\frac{1}{J}\sum_{j=1}^{J}\widehat{\text{var}}\,(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j). \tag{2.22}$$

In the above equation, the sum of the first two components estimates the average total variance of $\widehat{\boldsymbol{\beta}}_j$. Here the second component provides additional adjustment for the covariance among the sampling errors of $\widehat{\boldsymbol{\beta}}_j$'s between sites. The covariances are nonzero due to the pooling of data from all the sites in Step-1 estimation. The third component estimates the average within-site sampling variance of $\widehat{\boldsymbol{\beta}}_j$. The subtraction removes the sampling variance from the total variance. In practice, if a negative variance estimate is obtained, which is known as a Heywood case, both the variance estimate itself and the related covariance estimate will be set to 0.

Previous researchers of multilevel mediation analysis (e.g., Bauer et al., 2006) have not discussed how to conduct hypothesis testing for the between-site variance of NDE and NIE. Taking NDE as an example, I prove in Appendix 2.D, available in the online version of the journal, that under $H_0: \sigma^2_{D(0)} = 0$,

$$\sum_{j=1}^{J} \frac{\left(\hat{\beta}_j^{(D)}(0) - \hat{\gamma}^{(D)}(0)\right)^2}{\text{var}\left(\hat{\beta}_j^{(D)}(0) - \beta_j^{(D)}(0)\right)} \xrightarrow{d} \chi^2(J-1). \tag{2.23}$$

Replacing $\text{var}\left(\hat{\beta}_j^{(D)}(0) - \beta_j^{(D)}(0)\right)$ with $\widehat{\text{var}}\left(\hat{\beta}_j^{(D)}(0) - \beta_j^{(D)}(0)\right)$, the test statistic is

$$Q^{(D)}(0) = \sum_{j=1}^{J} \frac{\left(\hat{\beta}_j^{(D)}(0) - \hat{\gamma}^{(D)}(0)\right)^2}{\widehat{\text{var}}\left(\hat{\beta}_j^{(D)}(0) - \beta_j^{(D)}(0)\right)}. \tag{2.24}$$

As discussed in Section 2.6.2, as *N* increases, $\widehat{\text{var}}\left(\hat{\beta}_j^{(D)}(0) - \beta_j^{(D)}(0)\right)$ converges to $\text{var}\left(\hat{\beta}_j^{(D)}(0) - \beta_j^{(D)}(0)\right)$. However, when *N* is small, the distribution of the sample test statistic may deviate from $\chi^2(J-1)$. The same is true with the between-site variance of the NIE. I thus employ a permutation test proposed by Fitzmaurice, Lipsitz, and Ibrahim (2007). The test randomly permutes the site indices based on the idea that all permutations of the site indices are

equally likely under the null. The details about the algorithm of the permutation test can be found in Appendix 2.D.

2.7 Simulations

I conduct a series of Monte Carlo simulations to assess the finite-sample performance of the multilevel RMPW procedure in estimating the population average and between-site variance and covariance of NDE and NIE. I focus on the case of a binary randomized treatment, a binary mediator, and a continuous outcome, although the estimation procedure can be easily extended to multicategory mediators and binary outcomes. I implement the estimation in `R`, using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014) to fit the multilevel logistic regression models.

I specify three sets of population causal parameters listed in Table 2.5. The standardized parameter values are similar in magnitude to those used in the previous simulation studies of multilevel mediational models (Bauer et al., 2006; Krull & MacKinnon, 2001) and reflect a range of plausible values in real applications. Both the population average and the variance and covariance of the site-specific NDE and NIE are specified to be 0 in the first scenario, which is designed for examining the Type I error rates in hypothesis testing. All the parameter values increase from Set 2 to Set 3. Appendix 2.E explains how I generate the simulation data.

Table 2.5 Population Causal Parameter Value Specification for Simulations

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

SOURCE: Reprinted from Qin and Hong (2017). © 2017 by Sage Publications.

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

The number of sampled sites, *J*, the number of sampled individuals per site, n_j, and the probability of treatment assignment at a site, $Pr(T_{ij} = 1 | S_{ij} = j)$, are manipulated to represent the range observed in past multisite studies. For example, the Job Corps study had over 100 sites with an average of about 150 individuals per site in the full sample. The multisite sample analyzed by Seltzer (1994) had 20 sites with an average of about 29 individuals per site. Therefore, I generate balanced data sets comprised of 100 or 20 sites of either a small site size ($n_j = 20$) or a moderate site size ($n_j = 150$), while $Pr(T_{ij} = 1 | S_{ij} = j)$ is specified to be 0.5 across all the sites. In addition, I generate an imbalanced data set similar to the Job Corps data with varying site size and varying site-specific probability of treatment assignment.

I make 1,000 replications for each of these scenarios and then fit analytic models to each data set. I focus on assessing the amount of bias in the causal parameter estimates when implementing the proposed procedure. Table 2.6 reports the simulation results for the estimation of the population average effects and the between-site variances with the proposed method under 15 different scenarios (three sets of population causal parameters by five sets of sample sizes). As shown in Table 2.6, the sample estimates of the population average NDE and NIE contain minimal bias. The variance and covariance estimates appear to be unbiased when *N* is relatively large and show a slight increase in positive bias when *N* is small. The latter apparently has to do with the increase of Heywood cases in small samples. The Type I error rate for variance testing is always close to the nominal rate.

In addition, I compare the estimated standard errors for the population average NDE and NIE estimates between the proposed estimation procedure, the procedure that ignores the sampling variability of the weight estimates, and the fully nonparametric bootstrap procedure (Goldstein, 2011). For the latter, I generate a bootstrap sample through a simple random

resampling with replacement of the sites, estimate propensity scores and population average NDE and NIE based on this sample, and repeat this procedure 1,000 times. The standard deviation of the bootstrapped estimates provides an estimate of the standard error of each population average causal effect estimate. I construct 95% confidence intervals bounded by the 2.5th and 97.5th percentiles of the bootstrapped estimates.

Tables 2.7 and 2.8 present, respectively, for the population average NDE estimator and the population average NIE estimator, the simulation results for the standard error estimates and confidence interval coverage rates. For the population average NDE estimator, all the three approaches to standard error estimation seem to provide acceptable results. For the population average NIE estimator, the standard error estimated through the proposed estimation procedure always closely approximates the standard deviation of the sampling distribution. In contrast, the standard error tends to be underestimated by the procedure ignoring the estimation uncertainty in weight when the site size is relatively small and when the between-site variances are nonzero. In those scenarios, I observe a relatively high correlation among the site-specific NIE estimates. As shown in Equation (2.19), the asymptotic variance of the population average effect estimators is a linear combination of the elements in $\text{var}(\widehat{\boldsymbol{\beta}})$ including covariances among the site-specific effect estimates. However, these covariances are overlooked in the procedure ignoring the uncertainty in weight. I also observe that, when the between-site variance of the NIE increases, the magnitude of the covariance between the site-specific NIE estimates tends to increase accordingly, which then aggravates the bias in the standard error estimates. In the simulated scenarios, the standard error tends to be overestimated by bootstrap when the site size is relatively small.

I also note that the proposed estimation procedure generates acceptable confidence interval coverage rates. This is generally true for the bootstrapping procedure as well except for one case in which the bootstrapped standard error is a severe overestimate. The procedure ignoring the uncertainty in weight, however, generates coverage rates for the population average NIE that deviate notably from the nominal rate when the number of sites and the site size are relatively small. In general, for all three estimation approaches, the confidence interval coverage rates tend to converge to the nominal rate with the increase of the number of sites and of the site size.

Finally, I need to highlight that, with its closed-form expression for the standard error estimator, the proposed method requires much less computation than the bootstrap. For example, it takes less than 1 min to run one replication for the scenario of $J = 100$ and $n_j = 150$ with the proposed procedure, while it takes 5.5 hr with the bootstrap.

I also run simulations when the site-specific NDE and NIE are not normal or when the outcome follows other distributions. In all these cases, I obtain similar findings as above. Applying the proposed procedure, I have found that the estimates of the causal parameters contain minimal bias and the estimated standard errors always closely approximate the empirical standard errors. These additional results suggest that the proposed estimation procedure is not restricted to normally distributed outcomes or normally distributed site-specific effects.

Table 2.6 Simulation Results for the Estimation of the Population Average Effects and Between-Site Variances

Parameter Set	$J = 100$			$J = 20$	
	$n_j = 20$	$n_j = 150$	Job Corps Site Size	$n_j = 20$	$n_j = 150$
Parameter Set 1					
Direct effect					
Bias of $\hat{\gamma}^{(D)}$[a]	-0.002	0.000	0.000	-0.007	0.002
Bias of $\hat{\sigma}^2_{D(0)}$[b]	0.030	0.002	0.004	0.041	0.003
Type I error (%)[c] for $H_0: \sigma^2_{D(0)} = 0$	5.90	5.70	4.90	5.30	4.60
Indirect effect					
Bias of $\hat{\gamma}^{(I)}$	0.002	0.000	0.000	0.001	0.000
Bias of $\hat{\sigma}^2_{I(1)}$	0.002	0.000	0.000	0.003	0.000
Type I error (%) for $H_0: \sigma^2_{I(1)} = 0$	5.10	6.00	4.90	5.30	5.40
Bias of $\hat{\sigma}_{D(0),I(1)}$	-0.004	0.000	0.000	-0.007	0.000
Parameter Set 2					
Direct effect					
Bias of $\hat{\gamma}^{(D)}$	0.004	0.000	0.001	0.001	-0.001
Bias of $\hat{\sigma}^2_{D(0)}$	0.022	0.000	0.001	0.027	-0.002
Indirect effect					
Bias of $\hat{\gamma}^{(I)}$	-0.004	0.000	-0.001	-0.004	-0.003
Bias of $\hat{\sigma}^2_{I(1)}$	-0.002	0.001	0.001	-0.004	0.000
Bias of $\hat{\sigma}_{D(0),I(1)}$	0.001	0.000	0.000	0.000	-0.001
Parameter Set 3					
Direct effect					
Bias of $\hat{\gamma}^{(D)}$	0.011	-0.001	0.001	0.003	-0.004
Bias of $\hat{\sigma}^2_{D(0)}$	0.017	-0.003	-0.002	0.013	-0.003
Indirect effect					
Bias of $\hat{\gamma}^{(I)}$	-0.010	0.000	-0.001	-0.004	0.001
Bias of $\hat{\sigma}^2_{I(1)}$	-0.005	0.002	0.001	-0.005	0.001
Bias of $\hat{\sigma}_{D(0),I(1)}$	0.007	0.000	0.000	0.007	0.001

SOURCE: Reprinted from Qin and Hong (2017). © 2017 by Sage Publications.

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

Table 2.7 Simulation Results for the Standard Error Estimate and Confidence Interval Coverage Rate of the Population Average Natural Direct Effect Estimate ($\hat{\gamma}^{(D)}(0)$)

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

SOURCE: Reprinted from Qin and Hong (2017). © 2017 by Sage Publications.

Note. [a]"Empirical SE", SE($\hat{\gamma}^{(D)}(0)$), is the standard deviation of the sample estimates of direct effects over the 1,000 replications and is standardized. It approximates the standard deviation of the sampling distribution of the average direct effect estimates. [b]"Relative bias of SE" is the relative bias of the estimated standard error, computed as E[$\widehat{\text{SE}}$ ($\hat{\gamma}^{(D)}(0)$),)]/SE($\hat{\gamma}^{(D)}(0)$),) − 1. [c]"95% CI coverage rate" is the coverage probability of the 95% confidence interval estimate of the direct effect. I construct the bootstrap confidence intervals nonparametrically from the 2.5th and 97.5th percentiles of the set of empirical bootstrap values.

Table 2.8 Simulation Results for the Standard Error Estimate and Confidence Interval Coverage Rate of the Population Average Natural Indirect Effect Estimate ($\hat{\gamma}^{(I)}(1)$)

Parameter Set	$J = 100$			$J = 20$	
	$n_j = 20$	$n_j = 150$	Job Corps Site Size	$n_j = 20$	$n_j = 150$
Parameter Set 1					
Empirical SE[a]	0.011	0.004	0.005	0.029	0.009
Relative bias of SE (%)[b]					
Proposed method	-2.30	-2.20	-1.10	-3.80	-0.50
Ignore uncertainty in $\widehat{W}_{Mij}$	-1.00	0.60	0.90	-2.10	2.50
Bootstrap	43.50	6.60	5.40	38.10	6.40
95% CI coverage (%)[c]					
Proposed method	94.40	94.80	94.70	94.50	93.70
Ignore uncertainty in $\widehat{W}_{Mij}$	94.40	95.10	94.80	93.90	94.70
Bootstrap	97.60	95.00	95.00	99.40	95.80
Parameter Set 2					
Empirical SE	0.022	0.020	0.021	0.056	0.045
Relative bias of SE (%)					
Proposed method	2.40	2.70	-0.90	-3.90	-0.20
Ignore uncertainty in $\widehat{W}_{Mij}$	-5.00	1.30	-2.40	-9.80	-1.10
Bootstrap	29.50	5.80	3.80	21.30	0.90
95% CI coverage (%)					
Proposed method	92.90	96.60	94.40	92.10	93.10
Ignore uncertainty in $\widehat{W}_{Mij}$	91.90	96.10	93.80	90.40	92.80
Bootstrap	95.30	95.90	94.00	97.20	93.50
Parameter Set 3					
Empirical SE	0.033	0.027	0.028	0.078	0.063
Relative bias of SE (%)					
Proposed method	1.40	-0.40	-1.70	1.10	-3.50
Ignore uncertainty in $\widehat{W}_{Mij}$	-21.60	-5.30	-7.20	-18.90	-8.10
Bootstrap	18.30	4.40	1.60	17.00	-3.10
95% CI coverage (%)					
Proposed method	93.10	95.40	94.50	92.10	93.70
Ignore uncertainty in $\widehat{W}_{Mij}$	82.70	93.80	92.10	84.40	91.90
Bootstrap	96.20	95.30	94.10	96.00	93.50

SOURCE: Reprinted from Qin and Hong (2017). © 2017 by Sage Publications.

2.8 Empirical Application

In this section, I apply the above estimation procedure to the Job Corps data. The substantive research questions for the population of sites represented in this section are 1) What is the average indirect effect of the treatment assignment on earnings transmitted through educational attainment? 2) What is the average direct effect of the treatment assignment on earnings? 3) To what extent did the indirect effect vary across the experimental sites? 4) To what extent did the direct effect vary across the sites? and 5) Was there an association between the site-specific indirect effect and direct effect?

The analytic sample includes 8,659 individuals with non-missing outcome and non-missing mediator in the 48-month follow-up interview. There are 100 total experimental sites with one Job Corps center at each site. The sample size at each site ranges from 24 to 417. Of all, 5,202 applicants were randomly assigned to the experimental group and 3,457 to the control group. The application that I present here has not incorporated the NJCS sample weight and non-random nonresponse. Therefore, the analytic results in this section are only illustrative. I select 26 pretreatment covariates that are theoretically associated with the mediator and the outcome, including age, gender, race, education, criminal involvement, drug use, employment, and earnings at the baseline.

Analyzing the data from each treatment group through a multilevel logistic regression as described in Section 2.6.2, I predict a Job Corps participant's propensity score for obtaining an education or training credential 30 months after being assigned to Job Corps as a function of the individual's observed pretreatment characteristics and site membership. Applying the coefficient estimates obtained from analyzing the control group data, I predict a Job Corps participant's propensity score for having educational attainment under the counterfactual control condition. I

then construct the weight as defined in Equation (2.2). Subsequently, I estimate the population average direct and indirect effects by aggregating the estimated site-specific effects over all the sites. Finally, I estimate the between-site variance and covariance of these causal effects and conduct hypothesis testing as described in Section 2.6.3.

2.8.1 Total Program Impact

The results indicate that, 30 months after randomization, about 40% of the individuals assigned to Job Corps obtained an education or training credential; only about 22% of those assigned to the control condition obtained a credential. This stark contrast (coefficient = .18, standard error [SE] = 0.01, $t = 18.27$, $p < .001$) did not vary significantly across sites. Job Corps programs had a significant positive impact on earnings on average; this impact, however, varied considerably across the sites. The estimated population average ITT effect is \$16.41 (SE = 5.30, $t = 3.10$, $p = .002$), which amounts to about 8.75% of a standard deviation of the outcome. The between-site standard deviation of the ITT effect is estimated to be \$24.81 ($p = .03$). Therefore, if we assume that the site-specific ITT effect is approximately normally distributed, in 95% of the sites, the ITT effect may range from -\$32.22 to \$65.04. Apparently, the Job Corps centers were not equally effective in improving earnings.

2.8.2 Population Average Direct and Indirect Effects

I decompose the total ITT effect on earnings into an indirect effect mediated through educational attainment and a direct effect that channels the Job Corps impact through other services. The estimated population average indirect effect is \$8.68 (SE = 1.61, $t = 5.39$, $p < .001$), about 4.63% of a standard deviation of the outcome. The estimated population average direct effect is \$7.74 (SE = 5.38, $t = 1.44$, $p = .15$), about 4.13% of a standard deviation of the outcome. According to these results, on average, the change in educational attainment induced

by the program significantly increased earnings, while other supplemental services available to the Job Corps participants in contrast with services available to those under the control condition also seemed to play a crucial role in explaining the program mechanisms.

2.8.3 Between-Site Variance of Direct and Indirect Effects

To explain why some sites seemed to be more effective than others, I further investigate between-site heterogeneity in the causal mediation mechanism. The between-site standard deviation of the indirect effect is estimated to be only \$7.12 (p = .06), while the estimated between-site standard deviation of the direct effect is as large as \$23.76 (p = .055). I have additionally found that the estimated covariance between the site-specific direct and indirect effects is - 48.38, which corresponds to a correlation of - 0.29. Based on these estimates, we can infer that the mediating role of educational attainment was similar over all the sites. Yet the site-specific direct effect may range widely from negative to positive, suggesting that some sites were much more effective than others in promoting economic independence through services above and beyond increasing educational attainment. Hence, the variation in the Job Corps impact across the sites is mainly explained by the heterogeneity in the direct effect. Indeed, the National Job Corps office and regional offices centrally standardized the provision of education and strictly regulated vocational training programs for all the Job Corps centers, which might greatly limit between-site variation in education and training. In contrast, the management of other services was left largely to the discretion of each local center. As revealed in a qualitative process analysis (Johnson et al., 1999), the quantity and quality of supplementary services varied by a great amount across the Job Corps centers. The above analysis results corroborate the previous qualitative findings and suggest a need to improve the quantity and quality of

supplementary services especially in the Job Corps centers in which the estimated direct effect is relatively small or even negative.

2.9 Remaining Issues

As discussed in Section 2.4, the proposed procedure identifies the causal parameters only when the sequential ignorability assumption holds. In a multisite randomized trial, the assumption of ignorable treatment assignment within each site may be easy to satisfy. However, the assumption of ignorable mediator value assignment under each treatment condition within levels of the observed pretreatment covariates at each site is particularly strong. In the next chapter, I will use balance checking to assess if the propensity score-based weighting adjustment effectively reduces selection bias associated with the observed covariates.

Even though the distributions of the observed covariates achieve balance between mediator levels, the assumption of strongly ignorable mediator value assignment will still become implausible if posttreatment or omitted pretreatment covariates imply hidden bias that could alter the conclusion. If a pretreatment covariate that affects both the mediator and the outcome is omitted, sensitivity analysis could be employed (Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010; VanderWeele, 2010a; Hong, Qin, & Yang, 2018) to assess the extent to which the omission might invalidate inference about the direct and indirect effects. In the next Chapter, I will adopt a novel weighting-based sensitivity analysis strategy that Hong et al. (2018; working paper) developed for causal mediation analysis. In addition to assessing the potential bias in the estimated population average direct effect and indirect effect, I will assess the potential bias in the between-site variance of the direct effect and indirect effect.

In addition, I ignore the complex sample and survey designs and non-random nonresponse in longitudinal follow-ups in this chapter. Without considering these common issues in large-

scale multisite trials, we are not able to generalize the analytic results to the population of interest; the internal validity of the causal conclusions may also be contaminated. Hence, in the next chapter, I will clarify the identification assumptions under which the mediation analysis results are externally and internally valid and incorporate into the proposed analytic procedure a sample weight to adjust for sample and survey designs and an estimated nonresponse weight to account for non-random nonresponse.

CHAPTER 3

MULTISITE CAUSAL MEDIATION ANALYSIS IN THE PRESENCE OF COMPLEX SAMPLE AND SURVEY DESIGNS AND NON-RANDOM NONRESPONSE

To enhance the external and internal validity of the causal conclusions in Chapter 2, I refine the proposed methods in this chapter by overcoming the following obstacles:

***Sampling bias*.** NJCS drew a probability sample of individuals representative of the overall population of eligible applicants and then associated each sampled individual with one of the more than 100 Job Corps centers. The treatment was randomized within each site. In theory, random sampling of individuals and within-site treatment randomization allow an analyst to estimate without bias the average program impact at every site. Because all the Job Corps centers were included in the study, inference could be made for the population of sites as well. Yet in practice, external validity would be compromised if an analyst pays little attention to the complex sample and survey designs or if non-random nonresponse changes the representativeness of the sample of individuals in longitudinal follow-ups. In such cases, sample estimates of the average program impacts and of their between-site variance would contain sampling bias.

***Treatment selection bias*.** Nonresponse also poses a familiar threat to the internal validity of the causal conclusions if the remaining sample shows systematic differences between the program group and the control group. Additional threats to internal validity arise in a natural experiment in which the treatment is not strictly randomized. In either case, estimates of the average program impact and its between-site variance would contain selection bias.

***Mediator selection bias*.** Finally, even if a randomized experiment does not suffer from nonrandom 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.

***Bias due to model misspecification*.** Even when the treatment is randomized and even when an analyst attempts to make statistical adjustment for all potential confounders of the mediator-outcome relationship, estimation of the indirect and direct effects will nonetheless be biased if the mediator model as shown in Equation (2.15) is misspecified.

To address these major challenges, this chapter develops a systematic and coherent template for multisite causal mediation analysis by innovatively integrating a series of weighting-based strategies in a multisite causal mediation analysis. (1) I incorporate a sample weight in combination with a nonresponse weight to maintain sample representativeness. (2) The nonresponse weight, applied to each treatment group, is also used to restore the comparability of composition between the program group and the control group that has been jeopardized by differential nonresponse in a multisite randomized trial. Combining the sample weight and the nonresponse weight reduces sampling bias and treatment selection bias. (3) To unpack the causal mechanism and reduce mediator selection bias, I employ the RMPW strategy, as introduced in Chapter 2, for identification and estimation. (4) I emphasize the importance of empirically assessing the remaining overt bias associated with the observed pretreatment covariates through a weighting-based balance checking procedure. (5) I implement a novel weighting-based sensitivity analysis strategy (Hong et al., 2018, working paper) for evaluating the potential consequences of hidden bias due to omitting certain pretreatment or posttreatment confounders

or overlooking between-site differences in the selection mechanisms. This method distinguishes itself from other sensitivity analysis methods by minimizing its reliance on additional model-based assumptions.

3.1 Identification of the Causal Parameters

In a multisite study such as NJCS, we are able to observe $M_{ij}(t)$ and $Y_{ij}(t, M_{ij}(t))$ 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. Hence, to identify the causal parameters as defined in Chapter 2, we need additional sets of identification assumptions and correspondingly additional propensity score-based weights.

3.1.1 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 would be relatively straightforward if, at each site, all individuals in the site's eligible population had the same sampling probability, if all sampled individuals had the same treatment assignment probability, and if all sampled individuals in a given treatment group had the same probability of response. This is because, under the above hypothetical conditions, the respondents in each treatment group at each site would be representative of the population of eligible applicants at the site. However, an examination of the sampling mechanism, the treatment assignment mechanism, and the response mechanism in NJCS suggests otherwise.

***Sampling mechanism*.** NJCS researchers employed a stratified sampling procedure for individuals. Sampling probabilities varied across strata defined by individuals' 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 the population density in one's living area and whether one responded to the baseline survey (if selected to complete it) within a relatively short time frame. 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 I denote with vector $\mathbf{X}_D$. This conclusion also holds within each site. Because the sampling mechanism is known in NJCS, it is "ignorable" in the sense that we can reasonably make the following assumption.

Assumption 3.1 (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 $j = 1, \dots, J$, 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. I additionally assume that $0 < Pr\big(D_{ij} = 1 \big| \mathbf{X}_{Dij} = \mathbf{x}_D, S_{ij} = j\big) < 1$. That is, each individual in the population of eligible applicants 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*.** Rather than assigning all sampled individuals to the program group with an equal probability, NJCS researchers let the probabilities of treatment assignment differ by 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 I denote by $\mathbf{x}_T$. I find that $\mathbf{X}_T$ and $\mathbf{X}_D$ in NJCS partly overlap.

Assumption 3.2 (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.

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

In Chapter 2, Assumption 2.2 assumes the treatment assignment to be strongly ignorable in the whole population at each site. However, the individuals who were not sampled did not participate in the multisite trial. Instead, Assumption 3.2 is made only for sampled individuals at each site. Under this assumption, there should be no unmeasured confounding of the treatment-mediator relationship or the treatment-outcome relationship among sampled individuals 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*.** Among the individuals who were selected into the study sample, some failed to respond to the questions about educational and vocational attainment at the 30-month interview or to the questions about earnings at the 48-month interview. Due to nonrandom nonresponse, 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 $\{\mathbf{X}_D \cup \mathbf{X}_T\}$. NJCS researchers did not have control over an individual's probability of response. Nonetheless, some important distinctions can be identified between the respondents and the nonrespondents, for example, in gender composition and in education participation in the pre-randomization year.

Because response status is possibly a result of the treatment assignment, I 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 a subset of the pretreatment covariates that have been observed. I 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 3.3 (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 3.2, we could use balance checking and sensitivity analysis to assess the influence of possible violations of the assumption. I 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 the above assumptions, the site-specific averages of the potential mediator and potential outcome under each treatment condition are identifiable, which then enables the identification of the site-specific ITT effects. The key idea is to 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. For simplicity, suppose that the average sampling probability is 0.1. However, females in the population are assigned a sampling probability equal to 0.15 and thus are over-represented in the sample; while males in the population are assigned a sampling probability equal to 0.05 and are under-represented. To correct for sampling bias, an analyst may construct a sample weight that is equal to 0.1/0.15 = 2/3 for each sampled female and is equal to 0.1/0.05 = 2 for each sampled male. The weighted sample is then expected to have the same gender composition as the population. To generalize, for individual i at site j with pretreatment characteristics $\mathbf{X}_{Dij} = \mathbf{x}_D$, the sample weight is

$$W_{Dij} = \frac{Pr\big(D_{ij} = 1 |\, S_{ij} = j\big)}{Pr\big(D_{ij} = 1 | \mathbf{X}_{Dij} = \mathbf{x}_D, S_{ij} = j\big)}. \tag{3.1}$$

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, we may apply IPTW to sampled individual i at site j in treatment group t with pretreatment characteristics $\mathbf{X}_{Tij} = \mathbf{x}_T$:

$$W_{Tij} = \frac{Pr\left(T_{ij} = t|\ D_{ij} = 1, S_{ij} = j\right)}{Pr\left(T_{ij} = t|\mathbf{X}_{Tij} = \mathbf{x}_T, D_{ij} = 1, S_{ij} = j\right)} \text{ for } t = 0, 1. \quad (3.2)$$

Different than the IPTW as defined in (2.2), the IPTW in (3.2) is constructed based on the information of sampled individuals rather than the whole population. Here, 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, an analyst may apply the following nonresponse weight to sampled individual i at site *j* in treatment group t with pretreatment characteristics $\mathbf{X}_{Rij} = \mathbf{x}_R$:

$$W_{Rij} = \frac{Pr\left(R_{ij} = 1|T_{ij} = t, D_{ij} = 1, S_{ij} = j\right)}{Pr\left(R_{ij} = 1|\mathbf{X}_{Rij} = \mathbf{x}_R, T_{ij} = t, D_{ij} = 1, S_{ij} = j\right)} \text{ for } t = 0, 1. \quad (3.3)$$

The numerator is the average probability of response among sampled individuals at site *j* who have been assigned to treatment group t; the denominator is the individual's probability of response 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 I will discuss in Section 3.2.

Applying the product of the sample weight W_D, the IPTW W_T, and the nonresponse weight W_R to the respondents, I 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, I obtain the following identification results.

Theorem 3.1. Under Assumptions 3.1, 3.2, and 3.3, 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,$E\left[M_{ij}\ (t)\ |S_{ij}\ =\ j\right] =$

$E\left[W_{Dij}W_{Tij}W_{Rij}M_{ij}\ |R_{ij} = 1, T_{ij}\ = t, D_{ij} = 1, S_{ij} = j\right]$,

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

The proof of Theorem 3.1 is presented in Appendix 3.A.

Correspondingly, the weighted mean difference in educational and vocational 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.

3.1.2 Identification of the Mediation-Related Effects

In Chapter 2, I mainly discussed the identification and estimation of the population average and between-site variance of NDE and NIE. In this chapter, I further investigate whether the improvement in educational and vocational attainment produced a greater increase in earnings under Job Corps than under the control condition and whether the finding applies to all sites. Therefore, I will also discuss the identification of the population average and between-site variance of PIE and the interaction effect that have already been defined in Chapter 2.

As explicated in Chapter 2, the identification of the mediation-related effects requires an additional assumption about the strong ignorability of mediator values. In a multisite trial study such as NJCS, mediator and outcome values are missing among nonrespondents and those who were not sampled. Hence, it is inappropriate to impose such an assumption on the whole population at each site. Therefore, I replace Assumption 2.2 with the following Assumption 3.4, which is applied only to the sample respondents at each site.

Assumption 3.4 (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 $t \neq t'$. I additionally assume that $0 < Pr(M_{ij}(t) = m | R_{ij}=1, T_{ij}=t, D_{ij}=1, \mathbf{X}_{Mij}=\mathbf{x_M}, S_{ij}=j) < 1$.

We can now combine the assumptions and the weighting strategies associated with the sampling selection, treatment selection, nonresponse selection, and mediator value selection.

Theorem 3.2. Under Assumptions 3.1 ~ 3.4, 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 sample weight, IPTW weight, nonresponse weight, and RMPW weight,$E\left[Y_{ij}(t, M_{ij}(t')) \,|S_{ij}=j\right] = E\left[W_{Dij}W_{Tij}W_{Rij}W_{Mij}\, Y_{ij} \mid R_{ij}=1, T_{ij}=t, D_{ij}=1, S_{ij}=j\right]$ for $t \neq t'$, where

$$W_{Mij} = \frac{Pr\left(M_{ij}=m|\mathbf{X}_{Mij}=\mathbf{x}_M, R_{ij}=1, T_{ij}=t', D_{ij}=1, S_{ij}=j\right)}{Pr\left(M_{ij}=m|\mathbf{X}_{Mij}=\mathbf{x}_M, R_{ij}=1, T_{ij}=t, D_{ij}=1, S_{ij}=j\right)} \;\forall\, m \in \mathcal{M}. \qquad (3.4)$$

Different than the RMPW defined in (2.2), which ignores the fact that the mediator values of nonrespondents and non-sampled individuals are unobservable, the RMPW in (3.4) is constructed based on sample respondents. Besides, the RMPW in (2.2), constructed for those assigned to the experimental group, is only for identifying $E\left[Y_{ij}\left(1, M_{ij}(0)\right)|S_{ij}=j\right]$. Applying the same logic as explicated in Chapter 2, we could also identify $E\left[Y_{ij}\left(0, M_{ij}(1)\right)|S_{ij}=j\right]$ by constructing an additional RMPW for sample respondents who were assigned to the control group, so that we are able to identify the site-specific average PIE and interaction effect. In the modified RMPW, as shown in (3.4), the numerator is a respondent's propensity of displaying mediator value m under the counterfactual treatment t', while the denominator is his propensity of displaying the same mediator value under the assigned treatment t. At each site, for respondents in treatment group t who share the same covariate values $\mathbf{x}_M$, RMPW transforms their mediator distribution to resemble the mediator distribution of their counterparts in treatment group t'. Applying the product of W_{Dij}, W_{Tij}, W_{Rij}, 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 3.A presents a proof of Theorem 3.2.

To simplify the notation, let v'_{tj}, μ'_{tj}, and μ'_{*tj} represent each of the observable quantities at site *j* under treatment *t*, which I use to identify the site-specific average counterfactual outcomes.

$$v'_{tj} = E\left[W_{Dij}W_{Tij}W_{Rij}M_{ij}|R_{ij}=1, T_{ij}=t, D_{ij}=1, S_{ij}=j\right],$$

$$\mu'_{tj} = E\left[W_{Dij}W_{Tij}W_{Rij}Y_{ij}|R_{ij}=1, T_{ij}=t, D_{ij}=1, S_{ij}=j\right],$$

$$\mu'_{*tj} = E\left[W_{Dij}W_{Tij}W_{Rij}W_{Mij}Y_{ij}|R_{ij}=1, T_{ij}=t, D_{ij}=1, S_{ij}=j\right].$$

Table 3.1 summarizes the identification of the site-specific causal effects. Different than the identification results in Table 2.4, the results here carefully take into account the complex sample and survey designs and non-random nonresponse in longitudinal follow-ups. Correspondingly, we are able to identify the population average and between-site variance of each causal effect as defined in Chapter 2.

Table 3.1 Identification of the Site-Specific Effects

Site-Specific Effect	Identification Result	Assumptions
ITT effect on the mediator α_j	$v'_{1j} - v'_{0j}$	Assumptions 3.1 ~ 3.3
ITT effect on the outcome $\beta_j^{(T)}$	$\mu'_{1j} - \mu'_{0j}$	
NDE $\beta_j^{(D)}(0)$	$\mu'_{*1j} - \mu'_{0j}$	Assumptions 3.1 ~ 3.4
NIE $\beta_j^{(I)}(1)$	$\mu'_{1j} - \mu'_{*1j}$	
PIE $\beta_j^{(I)}(0)$	$\mu'_{*0j} - \mu'_{0j}$	
Interaction effect $\beta_j^{(T\times M)}$	$(\mu'_{1j} - \mu'_{*1j}) - (\mu'_{*0j} - \mu'_{0j})$	

3.2 General Analytic Procedure

Based on the above identification results, I 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.

Same as in Chapter 2, the estimation 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, I adopt the solution that estimates the weights and the site-specific causal effects jointly under a GMM framework, as proposed in Chapter 2.

However, the analytic results cannot be given causal interpretations if the identification assumptions are violated. I 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, I further conduct a sensitivity analysis.

3.2.1 Weight estimation

As clarified above, the estimation of the causal parameters depends on the estimates of the nonresponse weight and RMPW weight, W_{Rij} and W_{Mij}. I have selected, on theoretical grounds, pretreatment predictors of earnings $\mathbf{X} = \mathbf{X}_D \cup \mathbf{X}_T \cup \boldsymbol{X}_R \cup \mathbf{X}_M$. These include demographic characteristics such as gender, age, and race, fertility and living arrangements, education and training experiences, employment and earnings, public assistance receipt, and motivation and support for joining Job Corps prior to random assignment. For example, the baseline data suggest that, compared to male applicants, female applicants were more likely to have children, and females with children were highly dependent on public assistance; compared to older applications, younger applicants were more likely to have used drugs, have been arrested, and come from single-parent families; compared to white applications, minority applicants tended to have less work experience, and they were more likely to be high school dropouts, have received public assistance, and have children. These facts indicate that female, younger, or minority Job Corps applicants were relatively more disadvantaged. The more disadvantaged an individual was, the less he or she might earn in the long run.

I have categorized all the continuous covariates, and have imputed the missing information of each covariate with a missing category (see Appendix 3.B for a list of 51 covariates, more than those selected in Chapter 2). Categorizing the continuous covariates reduces the potential risk of misspecifying the functional form of a model; 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.

Sample weight and IPTW weight. In NJCS, sampling and treatment assignment were conducted simultaneously within subpopulations of eligible applicants. Each individual's joint probability of being selected into the sample and being assigned to one of the two treatment groups is pre-determined by design. I make use of the 48-month sample/survey weight, which is the product of W_D and W_T as defined in Equations (3.1) and (3.2).

Nonresponse weight estimation. Following Equation (3.3), 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; and let $p_{Rtij} = Pr\big(R_{ij} = 1|\mathbf{X}_{ij} = \mathbf{x}, T_{ij} = t, D_{ij} = 1, S_{ij} = j\big)$ denote individual i's conditional probability of response in treatment group t at site j, where $t = 0, 1$ denotes the treatment group that the individual was assigned to. Because we can remove nonresponse selection by controlling for $\mathbf{X}_R$ under the strongly ignorable nonresponse assumption and $\mathbf{X}_{Rij} \subset \mathbf{X}_{ij}$, p_{Rtij} is essentially equal to $Pr\big(R_{ij} = 1|\mathbf{X}_{Rij} = \mathbf{x}_R, T_{ij} = t, D_{ij} = 1, S_{ij} = j\big)$. To reflect the differences in response mechanisms between the program group and the control group, I 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:

$$\log\left[\frac{p_{Rtj}}{1-p_{Rtj}}\right] = \pi^*_{Rt} + r^*_{Rtj}, \quad r^*_{Rtj} \sim N(0, \sigma^{*2}_{Rt}), \tag{3.5}$$

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, I further control for the observed pretreatment covariates in the mixed-effects logistic regressions.

$$\log\left[\frac{p_{Rtij}}{1-p_{Rtij}}\right] = \boldsymbol{X}'_{ij}\boldsymbol{\pi}_{Rt} + r_{Rtj}, \quad r_{Rtj} \sim N(0, \sigma^2_{Rt}), \tag{3.6}$$

in which $\boldsymbol{X}_{ij}$ 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), I estimate the coefficients in the response models and obtain the Empirical Bayes estimates of the random effects. Based on these estimates, I 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 $\widehat{W}_{Rij} = \left(1-\hat{p}_{Rtj}\right)/\left(1-\hat{p}_{Rtij}\right)$ for the nonrespondents.

RMPW weight estimation. To obtain the RMPW weight as defined in Equation (3.4), I follow a similar procedure as described in Section 2.6.2.1. I first estimate each respondent's probability of attaining a credential under Job Corps and that under the control condition. Let $p_{Mtij} = Pr\left(M_{ij} = 1 | \mathbf{X}_{ij} = \mathbf{x}, R_{ij} = 1, T_{ij} = t, D_{ij} = 1, S_{ij} = j\right)$ and $p_{Mt'ij} = Pr\left(M_{ij} = 1 | \mathbf{X}_{ij} = \mathbf{x}, R_{ij} = 1, T_{ij} = t', D_{ij} = 1, S_{ij} = j\right)$ denote respondent i's probabilities of attaining a credential at site j if assigned to treatment t and treatment t', respectively, for $t \neq t'$. I fit the

following mediator model to each treatment group, which allows the mediator value selection mechanisms to differ between Job Corps and the control condition:

$$\log\left[\frac{p_{Mtij}}{1-p_{Mtij}}\right] = \boldsymbol{X}'_{ij}\boldsymbol{\pi}_{Mt} + r_{Mtj}, r_{Mtj} \sim N(0, \sigma^2_{Mt}), \tag{3.7}$$

in which $\boldsymbol{X}_{ij}$ 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 effect 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.

3.2.2 Causal Parameter Estimation and Inference

In accordance with the identification results summarized in Table 3.1, the sample estimators for the site-specific average potential mediators and potential outcomes are

$$\hat{v}'_{tj} = \frac{\sum_{i=1}^{N} W_{Dij} W_{Tij} \widehat{W}_{Rij} D_{ij} R_{ij} I\left(S_{ij}=j\right) I\left(T_{ij}=t\right) M_{ij}}{\sum_{i=1}^{N} W_{Dij} W_{Tij} \widehat{W}_{Rij} D_{ij} R_{ij} I\left(S_{ij}=j\right) I\left(T_{ij}=t\right)}, \tag{3.8}$$

$$\hat{\mu}'_{tj} = \frac{\sum_{i=1}^{N} W_{Dij} W_{Tij} \widehat{W}_{Rij} D_{ij} R_{ij} I\left(S_{ij}=j\right) I\left(T_{ij}=t\right) Y_{ij}}{\sum_{i=1}^{N} W_{Dij} W_{Tij} \widehat{W}_{Rij} D_{ij} R_{ij} I\left(S_{ij}=j\right) I\left(T_{ij}=t\right)}, \tag{3.9}$$

$$\hat{\mu}'_{*tj} = \frac{\sum_{i=1}^{N} W_{Dij} W_{Tij} \widehat{W}_{Rij} \widehat{W}_{Mij} D_{ij} R_{ij} I(S_{ij} = j) I(T_{ij} = t) Y_{ij}}{\sum_{i=1}^{N} W_{Dij} W_{Tij} \widehat{W}_{Rij} \widehat{W}_{Mij} D_{ij} R_{ij} I(S_{ij} = j) I(T_{ij} = t)} \quad (3.10)$$

Here $I(S_{ij} = j)$ is an indicator for whether individual i was a member of site j; and $I(T_{ij} = t)$ is an indicator for whether the individual was assigned to treatment t for $t = 0, 1$. $\widehat{W}_R$ and $\widehat{W}_M$ are estimates from the sample data. Under the identification assumptions 3.1 ~ 3.4, mean contrasts between the estimated average potential mediators and potential outcomes consistently estimate the site-specific causal effects listed in Table 3.1.

Following the same procedure as described in Sections 2.6.1 ~ 2.6.3, we could then obtain MOM estimates of the population parameters that characterize the distribution of the site-specific effects in a theoretical population of sites (e.g. Cameron & Trivedi, 2005). To be specific, the average of each causal effect over the population of sites is estimated through a simple average of the corresponding site-specific effect estimates; the between-site variance of each site-specific effect is estimated by subtracting the estimated average within-site sampling variance of the site-specific effect estimates from the estimated between-site variance of the site-specific effect estimates, with adjustment for the between-site sampling covariance of the site-specific effect estimates. Here, each site is given an equal weight. Unlike NJCS, which included all the sites at the time of study, some multisite studies sample sites first and then sample individual within the sampled sites. In such a case, we will need to further incorporate a site-level sample weight to adjust for the sample selection of sites.

The major difference between the estimation procedure in this chapter and that in Chapter 2 is that we have an additional nonresponse weight to estimate in the first step. Nevertheless, the rationale of the derivation of the asymptotic sampling variance matrix for the site-specific causal effect estimates $var(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ stays the same. To take into account the uncertainty in the

estimated weights, I still stack the estimating equations from both steps and solve them simultaneously in the spirit of one-step generalized method of moments (GMM) estimation. The only change is to further incorporate the estimating equations of the response models in the first step. Correspondingly, we could obtain a consistent estimate of the standard error for each estimated population average causal effect, as explicated in Chapter 2.

To test the significance of the between-site variances, I adopt the same permutation procedure as described in Chapter 2.

3.2.3 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 could indicate how much bias might be remaining and in which direction it could affect the analytic results.

Balance after nonresponse weighting adjustment. If the weighting adjustment was successful, we would expect the weighted data to approximate data from a design in which sampled individuals are randomized to respond in each treatment group at each site, so that balance between response levels is achieved across all the sites for each pretreatment covariate after adjustment. I first estimate the nonresponse weight as defined in Equation (3.3),

$\frac{Pr(R_{ij}=1|T_{ij}=t,D_{ij}=1,S_{ij}=j)}{Pr(R_{ij}=1|\mathbf{X}_{Rij}=\mathbf{x}_R,T_{ij}=t,D_{ij}=1,S_{ij}=j)}$ and apply it to the respondents and then estimate $\frac{Pr(R_{ij}=0|T_{ij}=t,D_{ij}=1,S_{ij}=j)}{Pr(R_{ij}=0|\mathbf{X}_{Rij}=\mathbf{x}_R,T_{ij}=t,D_{ij}=1,S_{ij}=j)}$ and apply it to the nonrespondents in treatment group t. I then quantify the balance after adjustment with standardized bias, which is calculated by dividing the weighted mean difference in each observed pretreatment covariate between the respondents and nonrespondents in each treatment group by the standard deviation of the covariate (Harder, Stuart, & Anthony, 2010). By convention, a covariate is considered to be balanced on average 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. Based on the estimated population average and standard deviation of the site-specific standardized bias and by assuming that the site-specific standardized bias is normally distributed, I 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 are categorical, I 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. I 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. Here I estimate the weight $\frac{Pr(M_{ij}=1|R_{ij}=1,T_{ij}=t,D_{ij}=1,S_{ij}=j)}{Pr(M_{ij}=1|\mathbf{X}_{Mij}=\mathbf{x}_M,R_{ij}=1,T_{ij}=t,D_{ij}=1,S_{ij}=j)}$ and apply it to the respondents who attained a credential in treatment group t at site *j* and then estimate the weight $\frac{Pr(M_{ij}=0|R_{ij}=1,T_{ij}=t,D_{ij}=1,S_{ij}=j)}{Pr(M_{ij}=0|\mathbf{X}_{Mij}=\mathbf{x}_M,R_{ij}=1,T_{ij}=t,D_{ij}=1,S_{ij}=j)}$ and apply it to those who did not.

Regressing a binary indicator for each covariate category on the mediator *M* in a weighted mixed-effects logistic model for each treatment group, I obtain estimates that allow us to calculate the population average and the between-site standard deviation of the standardized bias.

3.2.4 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 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.

I 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) because 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, I consider potential violations of Assumption 3.3 (strongly ignorable nonresponse) and Assumption 3.4 (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. I adopt a weighting-based strategy that offers a solution to sensitivity analysis concerning posttreatment confounders (Hong et al., 2018, working paper).

Appendix 3.C provides a list of weighting-based sensitivity parameters relevant to multisite causal mediation research. For each type of omission, I assess its potential impact on the causal conclusion with regard to each of the population parameters of interest.

3.3 Analytic Results

In this section, by applying the refined estimation procedure to the Job Corps data, I investigate the following research questions: 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?

I perform the analysis on the random sample of 14,125 youths who were targeted for the 48-month interview. However, 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. I define all of these individuals as nonrespondents because site membership and the mediator – in addition to the outcome – play crucial roles in the analysis. The 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).

3.3.1 Estimated Nonresponse and RMPW Weights

Appendix 3.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 normalized product of the sample weight, IPTW weight, and nonresponse weight, which is for identifying and estimating $E\left[M_{ij}\,(t)\,|S_{ij}\,=\,j\right]$ or $E\left[Y_{ij}(t, M_{ij}\,(t))\,|S_{ij}\,=\,j\right]$, 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 normalized product of all the fours weights, which is for identifying and estimating $E\left[Y_{ij}(t, M_{ij}\,(t'))\,|S_{ij}\,=\,j\right]$ where $t \neq t'$, 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.

3.3.2 Results of Causal Parameter Estimation and Inference

Table 3.2 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.

Table 3.2. 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	0.186 (0.014)	0.445	<0.001	0.087	0.035	[0.015, 0.357]
ITT effect on the outcome	21.030 (5.684)	0.114	<0.001	29.603	0.035	[-36.992, 79.052]
NDE	12.561 (5.730)	0.068	0.028	28.985	0.070	[-44.250, 69.372]
NIE	8.469 (1.612)	0.046	<0.001	5.407	0.135	[-2.129, 19.067]
PIE	6.198 (1.781)	0.034	0.001	4.351	0.215	[-2.330, 14.726]
Interaction effect	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.

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. The rate of educational and vocational attainment 30 months after randomization is estimated to be 22% in the control group and 40% in the program group. In the fourth year after randomization, the average weekly earnings are estimated to be $193.68 (in 1994 dollars) in the control group, while the amount is estimated to be $21.03 higher in the program group, an at least 10% increase. In contrast, the ITT effect of Job Corps on earnings was estimated to be $16.41 in Chapter 2. The difference in the estimation results between the two analyses reveals that ignoring the complex sample and survey design and non-random nonresponse does have a substantial influence on the estimation results.

Population average mediation mechanism. The ITT effect of Job Corps on earnings is partly explained by the program impact on educational and vocational attainment. The estimated

average NIE is $8.47 (standard error [SE] = 1.61, t = 5.26, p < .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 = .028), accounting for nearly 60% of the ITT effect. According to my 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 = .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 = .001). This difference 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 < .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, the impact was negligible or even negative in some of the sites. It is possible that, at some of these sites, the Job Corps centers might not have adequate resources to serve the needs of a concentration of highly vulnerable youths, who might otherwise receive better services from alternative programs under the control condition. It is also possible that the vocational training provided by the Job Corps

centers might not match the latest changes in the occupation structure at some of the sites. 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, I further investigate how the causal mediation mechanism varied across sites. The estimated between-site standard deviation of NDE is as large as \$29 (p = .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 = .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 educational and vocational attainment, 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.

3.3.3 Results of Balance Checking

According to the balance checking results, the nonresponse weighting adjustment has 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. 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. I notice that the nonresponse weighting has increased the plausible value range for some variables due to the increase in estimation uncertainty. These balance checking results are illustrated in Figures 3.D.1 to 3.D.4 in Appendix 3.D.

Figures 3.D.5 to 3.D.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.

3.3.4 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. The 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 I 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 I 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 3.C, 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.

I 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 3.2.4, 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, I am not able to assess the potential influence of omitted posttreatment confounders.

3.4 Remaining Issues

The mediator that I have been using so far is a combination of two central elements of the Job Corps program. It takes value 1 if an individual obtained either education or training credential within 30 months after randomization. However, the selection mechanism of getting 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. Besides, it is of important theoretical interest to distinguish the relative contribution of each of these two types of human capital investments and determine whether they are complementary or mutually reinforcing. Hence, in the next Chapter, I treat vocational training attainment and general education attainment as two concurrent mediators and decompose the total Job Corps impact on earnings into an indirect effect transmitted through vocational training, an indirect effect transmitted through general education, and a direct effect attributable to other pathways.

CHAPTER 4

UNPACKING COMPLEX MEDIATION MECHANISMS AND ITS HETEROGENEITY BETWEEN SITES

In this chapter, I extend the refined analytic procedure in Chapter 3 to an investigation of the complex mediation mechanisms that involve two parallel pathways in multisite trials. Conceptually, I characterize three basic types of mechanisms involving two parallel pathways. By design, the pathways transmitting the impacts of different program components are intended to be at least complementary to one another. For example, program developers may expect that the mediated impact of program component B will add onto the mediated impact of program component A. An absence of the mediated impact of component B, in this case, does not necessarily undermine the mediated impact of component A. If however, the mediated impact of component A is strengthened in the presence of the mediated impact of component B and is weakened in the absence of the latter, and if the same is true vice versa, then the two pathways are mutually reinforcing. Sometimes, unforeseen by the program developers, an unintended pathway may produce a side effect that offsets the program impact transmitted through the intended pathway. In such cases, a null ITT effect may disguise a mechanism featuring two counteracting mediators.

Building directly on the previous chapters, this chapter makes several methodological contributions. Most importantly, concepts and methods are developed for defining, identifying, and estimating not only the population average but also the between-site variance of the program impact transmitted through each of the two concurrent mediators of focal interest as well as the between-site variance of the direct effect. I further examine whether the two concurrent

mediators are complementary or mutually reinforcing. The identification strategy does not require parametric assumptions about the outcome model specification.

4.1 Definition of the Causal Parameters

4.1.1 Individual-Specific Causal Effects

Let T_{ij} denote the treatment assignment of individual *i* at site *j*. It takes values 1 or 0 indicating the individual was assigned to the Job Corps program or the control group, respectively. One of the concurrent mediators is a binary indicator of whether or not the youths gained a vocational training certificate (as a measure of job-specific human capital); the other indicates whether the youth obtained a high school diploma or GED (as a measure of generic human capital), within the 30 months after random assignment. Let M_{Vij} represent vocational training attainment and M_{Eij} for general education attainment. Each mediator takes value 1 if the individual obtained a credential in the corresponding domain, and 0 otherwise. Because vocational and educational attainment can be affected by the treatment assignment, I define the individual's potential mediators under the Job Corps condition as $M_{Vij}(1)$ and $M_{Eij}(1)$ and those under the control condition as $M_{Vij}(0)$ and $M_{Eij}(0)$. For each individual, we can only observe the potential mediators associated with the treatment condition that the individual was actually assigned to.

I use the same measure of the outcome as defined in the previous chapters. Similarly, I use $Y_{ij}(t)$ to represent the potential outcome associated with treatment condition *t*, where $t = 0,1$, for individual *i* at site *j*. Because the potential outcome depends on both the treatment assignment and the corresponding potential mediators, it can be equivalently written as $Y_{ij}(t, M_{Vij}(t), M_{Eij}(t))$. When $M_{Vij}(t) = m_V$ and $M_{Eij}(t) = m_E$, the individual's potential

outcome associated with treatment t can be written as $Y_{ij}(t, m_V, m_E)$. Again, only one potential outcome is observable for each individual, depending on which treatment condition the individual was actually assigned to.

By taking the difference in each potential outcome between the two treatment conditions, I define under SUTVA the total effect of the treatment assignment on the outcome (i.e. the ITT effect) for individual *i* at site *j* as,

$$\beta_{ij}^{(T)} = Y_{ij}\left(1, M_{Vij}(1), M_{Eij}(1)\right) - Y_{ij}\left(0, M_{Vij}(0), M_{Eij}(0)\right).$$

The is equivalent to the ITT effect of the treatment on the outcome defined in Chapter 2.

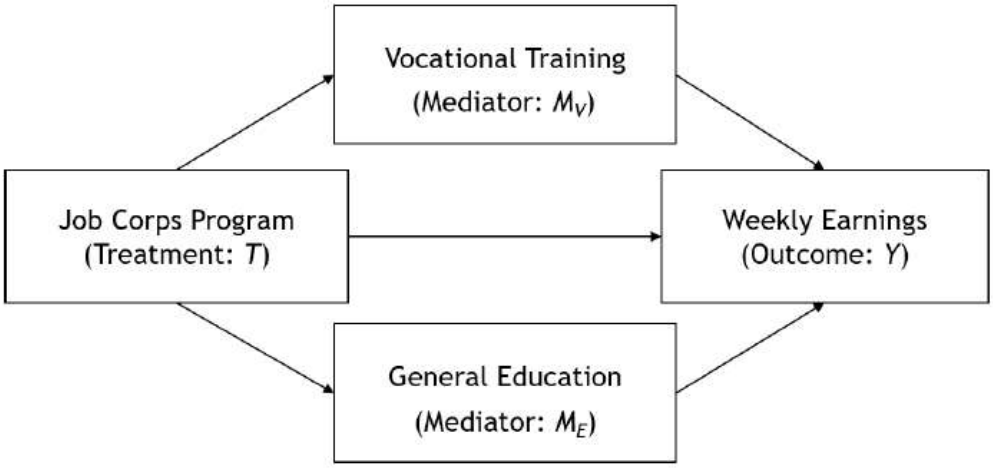


Figure 4.1 Causal Mediation in the Presence of Two Concurrent Mediators

As illustrated in Figure 4.1, the ITT effect of Job Corps on the outcome can be decomposed into two indirect effects, one transmitted through vocational training and the other through general education, and a direct effect. The decomposition involves another two potential outcomes: $Y_{ij}\left(1, M_{Vij}(1), M_{Eij}(0)\right)$ denotes the individual's potential outcome under the Job Corps condition when the treatment counterfactually fails to improve general education attainment; $Y_{ij}\left(1, M_{Vij}(0), M_{Eij}(0)\right)$ denotes the individual's potential outcome under the Job

Corps condition when the treatment counterfactually brought improvement in the attainment of neither general education nor vocational training.

I define the indirect effect that operates through improving vocational training without improving general education for individual *i* at site *j* as

$$\beta_{ij}^{(I.V)}(0) = Y_{ij}\left(1, M_{Vij}(1), M_{Eij}(0)\right) - Y_{ij}\left(1, M_{Vij}(0), M_{Eij}(0)\right).$$

It represents the Job Corps impact on earnings to be attributed to the program-induced change in the person's vocational training attainment from $M_{Vij}(0)$ to $M_{Vij}(1)$, while general education attainment is counterfactually not improved. The indirect effect transmitted through general education beyond improving vocational training can be defined for individual *i* at site *j* as

$$\beta_{ij}^{(I.E)}(1) = Y_{ij}\left(1, M_{Vij}(1), M_{Eij}(1)\right) - Y_{ij}\left(1, M_{Vij}(1), M_{Eij}(0)\right).$$

This is the Job Corps impact attributable to the program-induced change in the person's general education attainment from $M_{Eij}(0)$ to $M_{Eij}(1)$ above and beyond the counterfactual improvement in vocational training attainment. I further define the direct effect for individual *i* at site *j* as

$$\beta_{ij}^{(D)}(0) = Y_{ij}\left(1, M_{Vij}(0), M_{Eij}(0)\right) - Y_{ij}\left(0, M_{Vij}(0), M_{Eij}(0)\right),$$

which is the effect that operates through all the other unspecified possible mechanisms.

The ITT effect on the outcome is equal to the sum of the direct effect $\beta_{ij}^{(D)}(0)$ and the total indirect effect, the latter being the sum of the two indirect effects, $\beta_{ij}^{(I.V)}(0)$ and $\beta_{ij}^{(I.E)}(1)$. The total indirect effect can be alternatively decomposed into

$$\beta_{ij}^{(I.V)}(1) = Y_{ij}\left(1, M_{Vij}(1), M_{Eij}(1)\right) - Y_{ij}\left(1, M_{Vij}(0), M_{Eij}(1)\right),$$

$$\beta_{ij}^{(I.E)}(0) = Y_{ij}\left(1, M_{Vij}(0), M_{Eij}(1)\right) - Y_{ij}\left(1, M_{Vij}(0), M_{Eij}(0)\right).$$

The two decompositions are not equivalent in the presence of an interaction between the two mediators. This is because, if the program impact mediated by vocational training depended on the improvement in general education, $\beta_{ij}^{(I.V)}(1)$ would be unequal to $\beta_{ij}^{(I.V)}(0)$; and similarly, $\beta_{ij}^{(I.E)}(1)$ and $\beta_{ij}^{(I.E)}(0)$ would be unequal. For example, although both $\beta_{ij}^{(I.V)}(1)$ and $\beta_{ij}^{(I.V)}(0)$ represent the indirect effects transmitted through improving vocational training, general education attainment is kept at the level under the Job Corps condition in $\beta_{ij}^{(I.V)}(1)$ but kept at the level under the control condition in $\beta_{ij}^{(I.V)}(0)$. If the ITT effects on educational attainment and vocational attainment are both positive, and if $\beta_{ij}^{(I.V)}(1)$ is positive and greater than $\beta_{ij}^{(I.V)}(0)$, it will suggest that the indirect effect transmitted through improving vocational training may be reinforced by general education attainment. The degree of this potential reinforcement, known as the interaction effect between the two mediators, is defined as the difference between $\beta_{ij}^{(I.V)}(1)$ and $\beta_{ij}^{(I.V)}(0)$, which is numerically equivalent to the difference between $\beta_{ij}^{(I.E)}(1)$ and $\beta_{ij}^{(I.E)}(0)$,

$$\beta_{ij}^{(I.V\times E)}(1) = \left[Y_{ij}\left(1, M_{Vij}(1), M_{Eij}(1)\right) - Y_{ij}\left(1, M_{Vij}(1), M_{Eij}(0)\right)\right] - \left[Y_{ij}\left(1, M_{Vij}(0), M_{Eij}(1)\right) - Y_{ij}\left(1, M_{Vij}(0), M_{Eij}(0)\right)\right].$$

The above two decompositions are not unique. Daniel et al. (2015) showed as many as six different ways of decomposing the total treatment effect in the case of two concurrent mediators. I have chosen the above decomposition to answer the research questions proposed at the beginning of the empirical analysis section, Section 4.4.

4.1.2 Site-Specific Causal Effects and Population Parameters

Given my central interest in not only the prevalent causal mechanisms but also how the mechanisms may vary across local settings, I consider a population of sites, as in the previous

chapters. Taking the expectation of each individual-specific causal effect over the population of individuals at each site, I define the corresponding site-specific causal effect. Again using S_{ij} to indicate the site membership of individual *i* at site *j*, I list the site-specific causal effects in the first column of Table 4.1. Taking the expectation and the variance of each site-specific causal effect over the population of sites, I define the corresponding population average and between-site variance of the causal effect, as listed in the second and third columns of Table 4.1. These causal parameters correspond to the research questions proposed at the beginning of Section 4.4. To evaluate the average program impact on earnings mediated by vocational training and investigate whether this effect varies across the sites, I identify and estimate $\gamma^{(I.V)}(0)$ and $\sigma^2_{I.V(0)}$ if general education remains at the level under the control condition, or $\gamma^{(I.V)}(1)$ and $\sigma^2_{I.V(1)}$ if general education remains at the level under the Job Corps condition. To evaluate the average indirect effect that operates through improving general education and investigate whether this effect varies across the sites, I identify and estimate $\gamma^{(I.E)}(0)$ and $\sigma^2_{I.E(0)}$ if the interest is in the indirect effect mediated by general education without improving vocational training, or $\gamma^{(I.E)}(1)$ and $\sigma^2_{I.E(1)}$ if the indirect effect mediated by general education beyond improving vocational training is of interest. $\gamma^{(D)}(0)$ and $\sigma^2_{D(0)}$ are respectively defined for answering questions about the average and between-site variance of the direct effect transmitted through other unspecified mechanisms. By identifying and estimating $\gamma^{(I.V \times E)}$, I investigate if the program impact mediated by vocational training is reinforced by general education.

Table 4.1 Definition of Population Average Effects and Variance of Site-Specific Effects

Site-Specific Effect	Population Average	Between-Site Variance
$\beta_j^{(T)} = E\left[\beta_{ij}^{(T)} \middle\| S_{ij} = j\right]$	$\gamma^{(T)} = E\left[\beta_j^{(T)}\right]$	$\sigma_T^2 = var\left(\beta_j^{(T)}\right)$
$\beta_j^{(I.V)}(0) = E\left[\beta_{ij}^{(I.V)}(0) \middle\| S_{ij} = j\right]$	$\gamma^{(I.V)}(0) = E\left[\beta_j^{(I.V)}(0)\right]$	$\sigma_{I.V(0)}^2 = var\left(\beta_j^{(I.V)}(0)\right)$
$\beta_j^{(I.E)}(1) = E\left[\beta_{ij}^{(I.E)}(1) \middle\| S_{ij} = j\right]$	$\gamma^{(I.E)}(1) = E\left[\beta_j^{(I.E)}(1)\right]$	$\sigma_{I.E(1)}^2 = var\left(\beta_j^{(I.E)}(1)\right)$
$\beta_j^{(D)}(0) = E\left[\beta_{ij}^{(D)}(0) \middle\| S_{ij} = j\right]$	$\gamma^{(D)}(0) = E\left[\beta_j^{(D)}(0)\right]$	$\sigma_{D(0)}^2 = var\left(\beta_j^{(D)}(0)\right)$
$\beta_j^{(I.V)}(1) = E\left[\beta_{ij}^{(I.V)}(1) \middle\| S_{ij} = j\right]$	$\gamma^{(I.V)}(1) = E\left[\beta_j^{(I.V)}(1)\right]$	$\sigma_{I.V(1)}^2 = var\left(\beta_j^{(I.V)}(1)\right)$
$\beta_j^{(I.E)}(0) = E\left[\beta_{ij}^{(I.E)}(0) \middle\| S_{ij} = j\right]$	$\gamma^{(I.E)}(0) = E\left[\beta_j^{(I.E)}(0)\right]$	$\sigma_{I.E(0)}^2 = var\left(\beta_j^{(I.E)}(0)\right)$
$\beta_j^{(I.V \times E)} = E\left[\beta_{ij}^{(I.V \times E)} \middle\| S_{ij} = j\right]$	$\gamma^{(I.V \times E)} = E\left[\beta_j^{(I.V \times E)}\right]$	

4.2 Identification of the Causal Parameters

As explicated in Chapter 3, the NJCS sample was selected through a stratified sampling procedure to represent the entire national population of eligible Job Corps applicants at the time of the study. Groups of youths in the study population had different probabilities of being selected into the research sample. The external validity of causal conclusions would be compromised if little attention is paid to the complex sample design. In the longitudinal follow-ups, selective nonresponses to measures of the mediators and the outcome may cause some groups to be over- or under-represented among the respondents in the sample and thus pose an additional threat to the external validity. If the nonresponse also leads to a systematic difference between the treatment group and the control group in the sample of respondents, this would further introduce treatment selection bias and thus undermine the internal validity of the causal conclusions. Furthermore, in a multisite randomized trial, even if the treatment is randomized, mediators are usually generated in a natural process. Hence, mediator selection bias would arise if one pays little attention to the confounding factors of the mediator-outcome relationships.

The potential outcome and potential mediators associated with each treatment condition are observable only among sample respondents in the corresponding treatment group. To relate

the counterfactual quantities to the observable data, I invoke a series of assumptions about the sampling mechanism, the treatment assignment mechanism, the response mechanism, and the mediator selection mechanism, by extending those proposed in Chapter 3. All these assumptions share the notion of "selection on the observables." Let $\mathbf{X} = \mathbf{X}_D \cup \mathbf{X}_T \cup \boldsymbol{X}_R \cup \mathbf{X}_V \cup \mathbf{X}_E$ be a set of pretreatment covariates, where $\mathbf{X}_D$ predicts sample selection, $\mathbf{X}_T$ predicts treatment selection, $\boldsymbol{X}_R$ predicts nonresponse selection, $\mathbf{X}_V$ predicts selection into vocational attainment, and $\mathbf{X}_E$ predicts selection into educational attainment. Under these assumptions, when $\mathbf{X}$ is observed, we are able to remove sampling bias, treatment selection bias, nonresponse bias, and mediator selection bias by applying a series of propensity score-based weights to the observed data, thereby identifying the causal parameters defined in Table 4.1.

4.2.1 Identification of the ITT Effects

The identification of the ITT effects requires the same identification assumptions as listed in Chapter 3. The only change is that, in the mediation mechanism involving two concurrent mediators, potential outcomes are replaced with $Y_{ij}(t, m_V, m_E)$, and potential mediators are replaced with $M_{Vij}(t)$ and $M_{Eij}(t)$.

Assumption 4.1 (Strongly ignorable sampling mechanism). Given the observed pretreatment covariates $\mathbf{x}_D$, sample selection is independent of the potential mediators and potential outcomes at each site.

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

for $t = 0, 1, m_V, m_E \in \mathcal{M}$ where $\mathcal{M}$ is the support for all possible mediator values, and $j = 1, \ldots, J$, where J denotes the total number of sites. Same as in Chapter 3, D_{ij} takes value 1 if

individual i at site j was selected into the analysis sample and 0 otherwise, and it is additionally assumed that $0 < Pr(D_{ij} = 1|\mathbf{X}_{Dij} = \mathbf{x}_D, S_{ij} = j) < 1$.

Assumption 4.2 (Strongly ignorable treatment assignment). Given the observed pretreatment covariates $\mathbf{x}_T$, the treatment assignment is independent of the potential mediators and outcomes among the sampled individuals at each site.

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

It is also assumed that $0 < Pr(T_{ij} = t|D_{ij} = 1, \mathbf{X}_{Tij} = \mathbf{x}_T, S_{ij} = j) < 1$.

Assumption 4.3 (Strongly ignorable nonresponse). Given the observed baseline covariates $\mathbf{x}_R$, the response status of a sampled individual in a treatment group is independent of the potential mediators and potential outcomes associated with the same treatment at each site.

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

Same as defined in Chapter 3, R_{ij} is equal to 1 if individual i at site j responded and 0 otherwise, and $0 < \Pr(R_{ij} = 1|T_{ij} = t, D_{ij} = 1, \mathbf{X}_{Rij} = \mathbf{x}_R, S_{ij} = j) < 1$.

Under the above assumptions, applicants sharing the same pretreatment characteristics are as if randomized to be sampled, to be assigned to a given treatment group, and to respond. Hence, we are able to relate the counterfactual quantities to the observable data of the sample respondents and obtain the following identification results.

Theorem 4.1. Under Assumptions 4.1, 4.2, and 4.3, the site-specific means of each potential mediator and the potential outcome under treatment *t* can be respectively identified by the weighted averages of each observed mediator and the observed outcome among the respondents assigned to treatment group *t* at site *j*, as follows:

$$E\left[M_{Vij}\,(t)\,|S_{ij}\;=\;j\right]=E\left[W_{Dij}W_{Tij}W_{Rij}M_{Vij}\,|R_{ij}=1,T_{ij}\;=t,D_{ij}=1,S_{ij}=j\right],$$

$$E\left[M_{Eij}\,(t)\,|S_{ij}\;=\;j\right]=E\left[W_{Dij}W_{Tij}W_{Rij}M_{Eij}\,|R_{ij}=1,T_{ij}\;=t,D_{ij}=1,S_{ij}=j\right],$$

$$E\left[Y_{ij}(t,M_{Vij}(t),M_{Eij}(t))\,|S_{ij}\;=\;j\right]=E\left[W_{Dij}W_{Tij}W_{Rij}Y_{ij}\,|R_{ij}=1,T_{ij}\;=t,D_{ij}=1,S_{ij}=j\right],$$

where

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

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

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

As introduced in detail in Chapter 3, the sample weight W_{Dij}, IPTW weight W_{Tij}, and nonresponse weight W_{Rij} are used to adjust for sample selection, treatment selection, and nonresponse selection, respectively. By applying the product of these weights, we expect that, when the identification assumptions hold, the respondents in each treatment group will have the same composition of pretreatment characteristics as that in the entire population. This is because weighting will balance the joint distribution of the observed baseline covariates between the sampled and the non-sampled, between the treated and the untreated, and between the respondents and the non-respondents in each treatment group. Hence, the site-specific means of the potential mediator and potential outcome under each treatment condition can be identified on the basis of the observed information of the sample respondents after weighting. The proof of Theorem 4.1 is presented in Appendix 4.A. Subsequently, we can identify the site-specific ITT effect of the treatment on each mediator (or the outcome) through computing the weighted mean difference in the mediator (or the outcome) between the program group and the control group at

each site. Correspondingly, the population average and the between-site variance of each of these ITT effects can be identified as well.

4.2.2 Identification of the Mediation-Related Effects

Identifying the mediation-related effects is more challenging for several reasons. First of all, these effects involve the counterfactual outcomes that are never directly observable for any individual. Secondly, the mediator-outcome relationships are likely confounded by pretreatment and posttreatment differences between individuals in different mediator categories. And thirdly, the two concurrent mediators may act as confounders for each other. To identify these effects, we need two additional assumptions. Besides the strong ignorability assumption of each mediator value assignment, similar to that in Chapters 2 and 3, I also assume that the two potential mediators are conditionally independent.

Assumption 4.4 (Strongly ignorable mediator selection mechanism). Given the observed pretreatment covariates denoted by $\mathbf{x}_V$, whether one obtains a vocational credential under either treatment condition is independent of the potential outcomes among sample respondents at each site. Similarly, given the observed pretreatment covariates denoted by $\mathbf{x}_E$, whether one obtains an education credential under either treatment condition is independent of the potential outcomes among sample respondents at each site.

$$Y_{ij}(t, m_V, m_E) \perp\!\!\!\perp \{M_{Vij}(t), M_{Vij}(t')\} | R_{ij} = 1, T_{ij} = t, D_{ij} = 1, \mathbf{X}_{Vij} = \mathbf{x}_V, S_{ij} = j,$$

$$Y_{ij}(t, m_V, m_E) \perp\!\!\!\perp \{M_{Eij}(t), M_{Eij}(t')\} | R_{ij} = 1, T_{ij} = t, D_{ij} = 1, \mathbf{X}_{Eij} = \mathbf{x}_E, S_{ij} = j,$$

for all possible values of t, m_V, and m_E, where $t \neq t'$. It is also assumed that $0 < Pr\big(M_{Vij} = m_V | R_{ij} = 1, T_{ij} = t, D_{ij} = 1, \mathbf{X}_{Vij} = \mathbf{x}_V, S_{ij} = j\big) < 1$ and $0 < Pr\big(M_{Eij} = m_E | R_{ij} = 1, T_{ij} = t, D_{ij} = 1, \mathbf{X}_{Eij} = \mathbf{x}_E, S_{ij} = j\big) < 1$. That is, within levels of the observed pretreatment

covariates, every sample respondent at site *j* has a nonzero probability of obtaining (or not obtaining) a vocational (or an education) credential.

Assumption 4.5 (Conditional independence between two potential mediators). Given the observed pretreatment covariates $\mathbf{x}_V$ and $\mathbf{x}_E$, whether one obtains a vocational credential under one treatment condition is independent of whether one obtains an education credential under the same or the alternative treatment condition among sample respondents at each site.

$$M_{Vij}(t) \perp\!\!\!\perp M_{Eij}(t') | R_{ij} = 1, T_{ij} = t, D_{ij} = 1, \mathbf{X}_{Vij} = \mathbf{x}_V, \mathbf{X}_{Eij} = \mathbf{x}_E, S_{ij} = j$$

for $t, t' = 0,1$. This assumption is necessary for distinguishing the indirect effect transmitted through M_V from the indirect effect transmitted through M_E. It might be violated in the Job Corps study due to omitted pretreatment or posttreatment confounders of the two mediators. For example, whether one obtained a vocational certificate or an educational credential may depend on one's motivation to learn at the baseline or after randomization. Failure to control for pretreatment or posttreatment motivation would lead to a correlation between vocational training attainment and general education attainment.

Theorem 4.2. Under Assumptions 4.1 ~ 4.5, the site-specific average potential outcome associated with treatment condition t while one mediator or both mediators take the values associated with the counterfactual condition can be identified by the average of the observed outcome among the sample respondents assigned to treatment group *t* at site *j* weighted by the product of the sample weight, IPTW weight, nonresponse weight, and RMPW weights:

$$E\left[Y_{ij}(t, M_{Vij}(t'), M_{Eij}(t'')) \,|S_{ij} = j\right]$$
$$= E\left[W_{Dij} W_{Tij} W_{Rij} W_{Vt'ij} W_{Et''ij} Y_{ij} \,|R_{ij} = 1, T_{ij} = t, D_{ij} = 1, S_{ij} = j\right],$$

where $t' \neq t$ or $t'' \neq t$, and

$$W_{Vt'ij} = \frac{Pr\left(M_{Vij} = m_V | \mathbf{X}_{Vij} = \mathbf{x}_V, R_{ij} = 1, T_{ij} = t', D_{ij} = 1, S_{ij} = j\right)}{Pr\left(M_{Vij} = m_V | \mathbf{X}_{Vij} = \mathbf{x}_V, R_{ij} = 1, T_{ij} = t, D_{ij} = 1, S_{ij} = j\right)}, \quad (4.1)$$

$$W_{Et''ij} = \frac{Pr\left(M_{Eij} = m_E | \mathbf{X}_{Eij} = \mathbf{x}_E, R_{ij} = 1, T_{ij} = t'', D_{ij} = 1, S_{ij} = j\right)}{Pr\left(M_{Eij} = m_E | \mathbf{X}_{Eij} = \mathbf{x}_E, R_{ij} = 1, T_{ij} = t, D_{ij} = 1, S_{ij} = j\right)}, \quad (4.2)$$

for $m_V, m_E \in \mathcal{M}$. If $t' = t$, then $W_{Vt'ij} = 1$; if $t'' = t$, then $W_{Et''ij} = 1$. Otherwise, $W_{Vt'ij}$ and $W_{Et''ij}$ are the RMPW weights, constructed in the same way as in Chapter 3. Each weight is a ratio of a sample respondent's conditional probability of displaying a given mediator value under the counterfactual treatment condition to that under the actual treatment condition. For respondents who were assigned to treatment group *t* and displayed the pretreatment characteristics **x** at site *j*, W_{Vij} transforms their distribution of vocational attainment to resemble that of their counterparts in the counterfactual group t' if $t' \neq t$; W_{Eij} transforms their distribution of educational attainment to resemble that of their counterparts in the counterfactual group t'' if $t'' \neq t$. Hence, by applying the product of the sample weight, the IPTW weight, the nonresponse weight, and the RMPW weights to the sample respondents in treatment group t at each site, we are able to identify the site-specific population average counterfactual outcomes including $E\left[Y_{ij}(1, M_{Vij}(1), M_{Eij}(0)) \,|S_{ij} = j\right]$ and $E\left[Y_{ij}(1, M_{Vij}(0), M_{Eij}(0)) \,|S_{ij} = j\right]$. Appendix 4.A presents a proof of Theorem 4.2.

I use the same notation for the identified potential outcome under each treatment condition as in Chapter 3,

$$\mu'_{tj} = E\left[W_{Dij} W_{Tij} W_{Rij} Y_{ij} | R_{ij} = 1, T_{ij} = t, D_{ij} = 1, S_{ij} = j\right].$$

In addition, let

$$\mu_{t.t'.t''j} = E\left[W_{Dij} W_{Tij} W_{Rij} W_{Vt'ij} W_{Et''ij} Y_{ij} \,|R_{ij} = 1, T_{ij} = t, D_{ij} = 1, S_{ij} = j\right].$$

The latter identifies $E\left[Y_{ij}(t, M_{Vij}(t'), M_{Eij}(t'')) \,|S_{ij} = j\right]$, which is the expected potential outcome if one was assigned to treatment group *t*, while his vocational training attainment takes the same value as he would take if assigned to treatment group t' and his general education attainment takes the value that he would take under treatment condition t'', where either $t' \neq t$ or $t'' \neq t$. With the site-specific mean of each potential outcome identified, we are able to identify the site-specific causal effects through the weighted mean outcome differences at each site. Table 4.2 summarizes the identification results. It is then straightforward to identify the joint distribution of the site-specific effects in the population as defined in Table 4.1.

Table 4.2 Identification of the Site-Specific Causal Effects

Site-Specific Effect	Identification Result	Assumptions
ITT effect on the outcome *Y*	$\beta_j^{(T)} = \mu'_{1j} - \mu'_{0j}$	Assumptions 4.1 ~ 4.3
Indirect effect through M_V when M_E remains at $M_E(0)$	$\beta_j^{(I.V)}(0) = \mu_{1.1.0j} - \mu_{1.0.0j}$	Assumptions 4.1 ~ 4.5
Indirect effect through M_E when M_V remains at $M_V(1)$	$\beta_j^{(I.E)}(1) = \mu'_{1j} - \mu_{1.1.0j}$	
Direct effect	$\beta_j^{(D)}(0) = \mu_{1.0.0j} - \mu'_{0j}$	
Indirect effect through M_V when M_E remains at $M_E(1)$	$\beta_j^{(I.V)}(1) = \mu'_{1j} - \mu_{1.0.1j}$	
Indirect effect through M_E when M_V remains at $M_V(0)$	$\beta_j^{(I.E)}(0) = \mu_{1.0.1j} - \mu_{1.0.0j}$	
Interaction effect between M_V and M_E	$\beta_j^{(I.V \times E)} = \beta_j^{(I.E)}(1) - \beta_j^{(I.E)}(0)$	

4.3 Estimation and Inference of the Causal Parameters

For estimating and testing the population average and between-site variance of the total program impact and the mediation mechanism, we follow the same analytic procedure as in Chapter 3. The essential difference is that, with two concurrent mediators involved, the estimation now relies on two RMPW weights rather than one. Hence, we need to fit a separate propensity score model for each mediator in the first step, instead of fitting one single propensity score model for a combined mediator. Given that the selection mechanism of getting an

education credential can be quite different from that of getting a training credential, as discussed in Section 3.4, this procedure effectively differentiates the two and thus reduces possible bias.

By applying the product of the sample weight and IPTW weight given by design and the estimated nonresponse weight and RMPW weights, we could estimate the site-specific mean of each counterfactual potential outcome as

$$\hat{\mu}_{t.t'.t''j} = \frac{\sum_{i=1}^{N} W_{Dij} W_{Tij} \widehat{W}_{Rij} \widehat{W}_{Vt'ij} \widehat{W}_{Et''ij} D_{ij} R_{ij} I(S_{ij} = j) I(T_{ij} = t) Y_{ij}}{\sum_{i=1}^{N} W_{Dij} W_{Tij} \widehat{W}_{Rij} \widehat{W}_{Vt'ij} \widehat{W}_{Et''ij} D_{ij} R_{ij} I(S_{ij} = j) I(T_{ij} = t)} \quad (4.3)$$

Based on $\hat{\mu}'_{1j}$ and $\hat{\mu}'_{0j}$ as shown in Equation (3.9) and $\hat{\mu}_{1.1.0j}$, $\hat{\mu}_{1.0.0j}$, and $\hat{\mu}_{1.0.1j}$, we could estimate the site-specific causal effects through mean contrasts as represented in Table 4.2 and obtain MOM estimates of the population average and between-site variance of the effects by following the same procedure as described in Sections 2.6.1 ~ 2.6.3.

When all the identification assumptions hold, the estimates are consistent and generalizable to the population of Job Corps centers at the time of study. The proposed procedure greatly simplifies the outcome model specification without invoking strong functional and distributional assumptions.

4.4 Empirical Analysis

In this section, by applying the extended method, I investigate the following substantive research questions: 1) What is the average program impact on earnings mediated by vocational training? 2) What is the average program impact on earnings mediated by general education? 3) Is the program impact mediated by vocational training reinforced by general education? 4) What is the average direct effect of the program transmitted through other pathways? 5) Does the program impact mediated by vocational training vary across the sites? 6) Does the program

impact mediated by general education vary across the sites? 7) Does the direct effect of the program vary across the sites?

I begin with the sample of 14,125 youths who were targeted for the 48-month survey. However, individuals who had already completed high school at the baseline were unlikely to gain an additional high school diploma or GED. Thus the research questions about the indirect effect transmitted through general education attainment clearly have no relevance to them. In addition, a number of individuals had already obtained a vocational certificate at baseline. The NJCS survey questions at the 30 month follow-up were not specific enough for us to determine whether these individuals obtained an additional vocational certificate during 30 months after randomization. Hence, I exclude 3,272 youths who had a high school diploma, GED, or a vocational certificate at baseline, which leaves a sample of 10,853 individuals. Among them, 6,614 individuals responded to measures of the mediators and the outcome and were assigned to a specific center prior to the random treatment assignment. The rest 4,239 individuals are considered to be non-respondents.

Because the sample and survey weights by design are not a function of baseline educational and vocational attainment, the sample weights can be applied such that, at each site, the subsample of Job Corps applicants who lacked an education or training credential at baseline represents the corresponding subpopulation of youths. I use the same set of baseline covariates as described in Chapter 3 to account for selective nonresponse and for selection into different levels of education and training attainment. Combining the sample and survey design weights with the estimated nonresponse weights, I aim to generalize the analytic results to a theoretical population of Job Corps centers serving disadvantaged youths who lacked an education or training credential at baseline.

ITT effect of Job Corps on each mediator. The results show that, indeed, Job Corps improved educational attainment and vocational attainment among disadvantaged youth. During the 30 months after randomization, 37.1% of the individuals assigned to Job Corps obtained an education credential; in contrast, only 22.5% of those assigned to the control group had the same level of attainment. The proportion difference is 0.146 and is statistically significant (SE = 0.017, $p < 0.001$). The between-site standard deviation of this proportion difference is estimated to be 0.108. In the meantime, the probability of obtaining a vocational credential is 0.186 if one was assigned to the Job Corps group and is only 0.052 if one was assigned to the control group instead. This proportion difference is 0.134 and is also statistically significant (SE = 0.010, $p < 0.001$). The standard deviation of this proportion difference across sites is 0.058.

ITT effect of Job Corps on earnings. The Job Corps program generated a significant average impact on earnings 48 months after randomization. The population average weekly earnings among individuals assigned to Job Corps is estimated to be $21.26 (SE = 5.94, $p < 0.001$) higher than that among individuals assigned to the control group. This impact accounts for about 12.9% of a standard deviation of the outcome. Yet the impact on earnings varied significantly across sites. If we assume that the site-specific ITT effect is normally distributed in the population of sites, then the impact on earnings would range from -$30.54 to $73.06 in 95% of the sites.

Population average mediation mechanism. I further decompose the total Job Corps impact on earnings into an indirect effect transmitted through vocational training, an indirect effect transmitted through general education, and a direct effect primarily attributable to supplementary services. The population average indirect effect that operated through improving vocational training without improving education is estimated to be $2.77 (SE = 1.62, $p = 0.09$), which

amounts to about 1.7% of a standard deviation of the outcome. With vocational training already improved, the population average indirect effect that operated through Job Corps' education is estimated to be \$7.14 (SE = 1.58, $p < 0.001$), about 4.3% of a standard deviation of the outcome. The first indirect effect accounts for 13.0% of the total ITT effect, while the second indirect effect is about a third of the total ITT effect. Vocational training and general education together mediated 46.6% of the total program impact on earnings. The mediating role of educational attainment, mainly in the form of a GED certificate, is seemingly greater than that for vocational attainment. This appears to be in contrast with the past findings that GED recipients earn less than ordinary high school graduates and they even earn less than the high school dropouts who do not have GED certificates and are at the same level of cognitive ability (e.g. Heckman, Hsse, & Rubinstein, 2000; Heckman & Rubinstein, 2001). As Heckman and colleagues reasoned, this is mainly caused by the relatively low non-cognitive skills, such as persistence and self-discipline, of GED recipients. However, for those who obtained GED certificates in the experimental group, the comprehensive training that they received at the Job Corps centers emphasized both cognitive and non-cognitive skills, which was distinguished from typical GED programs available to students in the control group. In addition, unlike many other adult education programs, Job Corps tailored the pace of GED instruction to students' individual abilities. Many Job Corps centers provided individualized tutorial assistance to students who were not performing at the expected pace. Some centers further offered "enrichment" courses beyond the basic curriculum to the relatively advanced students. The comprehensiveness and the flexibility of the educational opportunities offered by Job Corps may effectively facilitate individual academic growth and thus enhance the mediating role of GED attainment.

Under the improvement in education, the estimated indirect effect operating through improving vocational training would increase from $2.77 to $3.84 and the latter will become statistically significant (SE = 1.72, $p < .05$). However, this amount of increase is not statistically significant (estimated interaction effect = 1.06, SE = 0.69, $p = 0.12$). Hence, we can infer that vocational training and general education were complementary to each other when transmitting the Job Corps impact on earnings, while the two did not mutually reinforce each other.

The population average effect that operated through other mechanisms without improving vocational training and education is estimated to be $11.35 (SE = 6.28, $p = 0.07$), which amounts to about 6.9% of a standard deviation of the outcome. This effect accounts for 53.4% of the total ITT effect, indicating that the supplementary services and other local factors played a role at least as important as vocational training and general education in promoting economic well-being.

Between-site variance of the mediation mechanism. To explain the between-site heterogeneity in the total Job Corps impact on earnings, I further investigate how the causal mediation mechanism varied across sites. The estimated between-site standard deviation of the Job Corps impact mediated by vocational training without improving education is $7.85, and that of the Job Corps impact mediated by general education beyond improving vocational training is $3.59. The remaining bulk of the between-site heterogeneity in the Job Corps impact, estimated to be as large as $28.47 for the between-site standard deviation of the direct effect, is explained by the local implementation of supplementary services among other factors. The 95% plausible values of the site-specific direct effect range from -$44.51 to $67.21. According to these results, the variation in the Job Corps impact across the sites is mainly explained by the heterogeneity in supplementary services and other local factors. This result reflects the fact that supplementary

services were provided at the discretion of each local Job Corps center. In contrast, education and vocational training programs were standardized by the national Job Corps office and regional offices.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

5.1 A Summary of the Proposed Multisite Causal Mediation Analysis Methods

Estimating and testing the between-site variance of the indirect effect in addition to that of the direct effect and quantifying the correlation between the two have been a major challenge in multisite causal mediation analysis. This is because, in the standard regression-based approach, the indirect effect is represented as a product of multiple regression coefficients that may vary and covary between the sites. The complexity increases exponentially in the presence of treatment-by-mediator interaction as well as treatment-by-covariate or mediator-by-covariate interactions. A consistent estimation requires that both the mediator model and the outcome model be correctly specified. Moreover, the standard regression approach tends to be constrained, with few exceptions, to mediators and outcomes that are multivariate normal. A computationally intensive bootstrap procedure has been typically recommended for assessing the standard error of each causal effect estimate. Developed under the framework of potential outcomes, the methods presented in this dissertation provides an important alternative to the existing methods.

In Chapter 2, I have extended the RMPW strategy to multisite causal mediation analysis. The simplicity of this weighting strategy brings multiple benefits. It does not require any assumption about the functional form of the outcome model; nor does it invoke any distributional assumption about site-specific direct and indirect effects. Therefore, the method can be applied to outcomes measured on various scales as long as each causal effect can be defined as a mean contrast between two potential outcomes. An MOM procedure applied to the weighted data

generates estimates of all the causal parameters that define the first two moments of the joint distribution of the site-specific direct effect and indirect effect. In addition, there is virtually no constraint on the mediator distribution because RMPW is suitable for any discrete mediators (Hong, 2015; Hong et al., 2011, 2015) and because a mathematical equivalent of RMPW (Huber, 2014) easily handles continuous mediators. Hence, I conclude that the proposed strategy has considerably greater applicability than the existing methods.

I have additionally made several improvements to the estimation and hypothesis testing. The propensity score–based weights must be estimated from the sample data pooled over all the sites in the first step before the causal parameters can be estimated in the second step. The uncertainty of the estimated weights is usually ignored in causal inference studies that rely on propensity score-based weighting strategies in multilevel settings (Leite et al., 2015). To fully account for the sampling variability in the two-step estimation, I have derived a consistent estimator of the asymptotic standard error for each causal effect estimator. This solution may be applied generally to other propensity score–based two-step estimation problems in analyses of multilevel data. The results of my simulation comparisons suggest that, for the population average indirect effect estimator in particular, the estimated asymptotic standard errors often outperform not only the standard error estimators that ignore the Step-1 estimation but also the bootstrapped standard errors. Finally, given that the test statistic for the between-site variance of the direct effect and that for the indirect effect do not follow a theoretical χ^2 distribution, I have implemented a permutation test that produces valid statistical inference.

In Chapter 3, I further refine the proposed analytic procedure by integrating a series of weighting-based strategies. These include using a sample weight to adjust for complex sample and survey designs, using a nonresponse weight to adjust for nonrandom nonresponse, and using

RMPW weights to adjust for mediator value selection while unpacking the causal mechanisms. 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 only when the identification assumptions hold. Hence, it is crucial to assess the causal conclusions in light of possible violations of the identification assumptions. I conduct a weighting-based balance checking procedure to examine the bias associated with the observed pretreatment covariates. To further evaluate if omitting certain confounders or overlooking between-site heterogeneity in the selection mechanism would easily alter the conclusions, I adopt a weighting-based sensitivity analysis.

Chapter 4 extends the refined analytic procedure for decomposing the population average and the between-site variance of the total treatment effect in the presence of two concurrent mediators in multisite trials. By integrating a sample weight and a nonresponse weight with a separate RMPW weight for each mediator, I unpack the total treatment effect into two indirect effects, each transmitted through a mediator, and a direct effect transmitted through other possible mechanisms. Besides, I examine whether the indirect effect transmitted through one mediator is reinforced by the other mediator. In addition to decomposing the average total treatment effect, the analytic procedure also enables analysts to investigate the heterogeneity in the complex causal mechanisms across sites that explains the between-site variation in the total treatment effect.

5.2 A Summary of the Job Corps Program Evaluation

The development of the methods for investigating mediation mechanisms in multisite trials fills an important gap in the literature and enables researchers to ask a new set of empirical questions crucial for testing the generalizability of an intervention theory across a wide range of

settings. Interventions such as Job Corps must be delivered by local agents who differ in their professional capacity for engaging participants in critical elements of the program. The composition of the client population and their needs may not be identical across the sites. Moreover, the job market and alternative programs available to the client population may differ across the localities as well. A multisite randomized trial offers unique opportunities to empirically examine the program theory across these different contexts.

Applying the proposed analytic procedure to the NJCS data, I have found empirical evidence that supports the Job Corps program theory and suggests necessary modifications in program practice. The analytic results reveal that both vocational training and general education play pivotal roles in transmitting the program impact on earnings for disadvantaged youth. These two pathways are complementary rather than mutually reinforcing. 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. The results have indicated that the latter mechanism is no less if not more important than the former. By further examining the between-site variance of each causal effect, I have found that the between-site heterogeneity in the total program impact is mainly explained by the variation in the latter mechanism shown as the direct effect. To be specific, compared to the control condition, most Job Corps centers successfully increased educational and vocational attainment which subsequently increased earnings. However, Job Corps centers were not equally successful in promoting economic welling-being through countering a wide range of risk factors.

It is of theoretical importance to further investigate features of sites that may explain the substantial between-site variation in the direct effect. Potential features may include the quality and quantity of the supplementary services offered by a Job Corps center, the racial and ethnic

composition of eligible Job Corps applicants at a site, the proportion of nonresidential participants, and the local unemployment rate. I elaborate the theoretical reasoning as follows.

I hypothesize that differences between Job Corps centers in the management of supplementary services may explain most of the between-site variation in the direct effect. This is because Job Corps centers differed in counselors' caseload and the availability of counseling services during evening hours and weekends and in the dormitories (Johnson et al, 1999). Moreover, center staff had different levels of training and commitment; and the supplementary services were provided under different contextual constraints. Job Corps centers with more high-quality support resources are expected to more effectively reduce risk exposures and risk behaviors and generate greater positive impacts on participants' developmental trajectories. Regularizing the quantity and ensuring the quality of supplementary services is likely the key to achieving universal effectiveness of Job Corps.

The site-specific direct effect may also depend on the racial and ethnic composition of the Job Corps applicants at a site. Large differences were found in the racial and ethnic composition between regions and between Job Corps centers within a region. Compared to white applicants, minority applicants tended to have higher vulnerability due to their experiences with discrimination and inequality; and thus supplementary services are expected to play a more important role in promoting their economic well-being. Hence, the program impact transmitted through supplementary services might be different across racial and ethnical groups. Job Corps centers with more minority applicants might be more successful in promoting economic well-being through countering risk factors.

The between-site variation in the direct effect may also be partly due to differences in the proportion of residential living. Residential living is a unique feature of Job Corps. Past research

has suggested that it plays an important role in helping disadvantaged youths become more employable (Schochet et al., 2001). Because nonresidential participants were less involved in dormitory life, recreational activities, and student government and had less access to counseling services, they might benefit less from the comprehensive supplementary services. Hence, Job Corps centers with a higher proportion of nonresidential participants might generate a smaller positive direct effect.

In addition, site-specific program impacts may also depend on the local labor market condition. Analyzing the NJCS data, researchers reported that, at a site with a relatively high local unemployment rate, Job Corps tended to shield whites but not blacks and Hispanics (Flores-Lagnunes, Gonzalez, & Neumann, 2010). This result indicates that the influence of local unemployment rate on the between-site variation in the program impact may differ between whites and minorities.

5.3 Directions for Future Methodological Research

I acknowledge several potential limitations of the proposed procedure and will develop new methods in my future research to overcome the limitations.

First, as highlighted throughout this dissertation, the proposed propensity score-based weighting strategy for causal mediation analysis reduces reliance on the outcome model specification and avoids distributional assumptions about the outcome, the mediators, and the site-specific causal effects. Nonetheless, the weighting methods require that the propensity score models for the mediators are correctly specified. Yet in contrast, most other existing methods rely heavily on parametric assumptions about both the mediator models and the outcome model and are particularly vulnerable to bias when any of these models are misspecified. I acknowledge

that a bias-variance tradeoff is likely because weighting has a tendency of reducing efficiency in estimation. I will develop methods to further improve estimation efficiency.

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

remains a major methodological challenge. Moreover, in many studies, researchers may be interested in more than two mediators. It is important to further extend the methods for applications to multiple concurrent or consecutive mediators and develop sensitivity analysis strategies to assess the potential consequences if the independence assumption between concurrent mediators within and across treatment conditions at each site does not hold.

Finally, I 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 I will explore in future work.

APPENDICES

APPENDIX 2.A

Proof of Theorems 2.1 and 2.2

This Appendix provides a supplement to Section 2.5. The derivation below proves that, under Assumptions 2.1 $\sim$ 2.2, the expectation of each potential outcome at site j, $E[Y(t, M(t'))|S = j]$, for $t, t' = 0, 1$, can be identified with a weighted average of the observed outcome at that site. Let $\mathbf{X} = \{\mathbf{X}_T \cup \mathbf{X}_M\}$ be the union of all the observed pretreatment confounders. To simplify notations, I drop the subscript ij of each variable.

$$\begin{aligned} E[Y(t, M(t'))|S = j] &= E\{E[Y(t, M(t'))|\mathbf{X} = \mathbf{x}, S = j]\} \\ &= \int_{\mathbf{x}} \int_m \int_y y \times f(Y(t, m) = y|M(t') = m, \mathbf{X} = \mathbf{x}, S = j) \\ &\quad \times Pr(M(t') = m|\mathbf{X} = \mathbf{x}, S = j) \times g(\mathbf{X} = \mathbf{x}|S = j) dy dm d\mathbf{x}. \end{aligned}$$

Under Assumption 2.1, $\{Y(t, m), M(t)\} \perp\!\!\!\perp T|\mathbf{X}_T = \mathbf{x}_T, S = j$. Because $\mathbf{X}_T \subset \mathbf{X}$, $\{Y(t, m), M(t)\} \perp\!\!\!\perp T|\mathbf{X} = \mathbf{x}, S = j$ also holds. Hence,

$$\begin{aligned} E[Y(t, M(t'))|S = j] = \int_{\mathbf{x}} \int_m \int_y & y \times f(Y(t, m) = y|T = t, M(t') = m, \mathbf{X} = \mathbf{x}, S = j) \\ & \times Pr(M(t') = m|T = t', \mathbf{X} = \mathbf{x}, S = j) \times g(\mathbf{X} = \mathbf{x}|S = j) dy dm d\mathbf{x}. \end{aligned}$$

By Bayes theorem,

$$g(\mathbf{X} = \mathbf{x}|S = j) = g(\mathbf{X} = \mathbf{x}|T = t, S = j) \times \frac{Pr(T = t|S = j)}{Pr(T = t|\mathbf{X} = \mathbf{x}, S = j)}.$$

Because we can remove treatment selection bias by only controlling for $\mathbf{X}_T$ and $\mathbf{X}_T \subset \mathbf{X}$, $Pr(T = t|\mathbf{X} = \mathbf{x}, S = j) = Pr(T = t|\mathbf{X}_T = \mathbf{x}_T, S = j)$. Hence, the above equation can be simplified as

$$g(\mathbf{X} = \mathbf{x}|S = j) = g(\mathbf{X} = \mathbf{x}|T = t, S = j) \times \frac{Pr(T = t|S = j)}{Pr(T = t|\mathbf{X}_T = \mathbf{x}_T, S = j)},$$

where $0 < Pr(T = t|\mathbf{X}_T = \mathbf{x}_T, S = j) < 1$. Let $W_T = \frac{Pr(T=t|S=j)}{Pr(T=t|\mathbf{X}_T=\mathbf{x}_T,S=j)}$, and thus

$$E[Y(t, M(t'))|S = j] = \int_{\mathbf{x}} \int_m \int_y W_T \times y \times f(Y(t, m) = y|T = t, M(t') = m, \mathbf{X} = \mathbf{x}, S = j)$$
$$\times Pr(M(t') = m|T = t', \mathbf{X} = \mathbf{x}, S = j) \times g(\mathbf{X} = \mathbf{x}|T = t, S = j) dy dm d\mathbf{x}.$$

When $t = t'$, it is easy to obtain the following identification result,

$$E[Y(t, M(t))|S = j] = E[W_T Y|T = t, S = j].$$

When $t \neq t'$ and if Assumption 2.2 holds, i.e. $Y(t, m) \perp\!\!\!\perp \{M(t), M(t')\}|T = t, \mathbf{X}_M = \mathbf{x}_M, S = j$, because $\mathbf{X}_M \subset \mathbf{X}$, $Y(t, m) \perp\!\!\!\perp \{M(t), M(t')\}|T = t, \mathbf{X} = \mathbf{x}, S = j$ also holds. Hence,

$$E[Y(t, M(t'))|S = j] = \int_{\mathbf{x}} \int_m \int_y W_T \times y \times f(Y(t, m) = y|T = t, M(t) = m, \mathbf{X} = \mathbf{x}, S = j)$$
$$\times Pr(M(t') = m|T = t', \mathbf{X} = \mathbf{x}, S = j) \times g(\mathbf{X} = \mathbf{x}|T = t, S = j) dy dm d\mathbf{x}.$$

Under the assumption that $0 < Pr(M(t) = m|\mathbf{X} = \mathbf{x}, T = t, S = j) < 1$, let $W_M = \frac{Pr(M(t')=m|\mathbf{X}=\mathbf{x},T=t',S=j)}{Pr(M(t)=m|\mathbf{X}=\mathbf{x},T=t,S=j)} = \frac{Pr(M=m|\mathbf{X}=\mathbf{x},T=t',S=j)}{Pr(M=m|\mathbf{X}=\mathbf{x},T=t,S=j)}$, because $M(t) = M$ when $T = t$ and, similarly, $M(t') = M$ when $T = t'$. Because we can remove mediator selection bias by only controlling for $\mathbf{X}_M$ and $\mathbf{X}_M \subset \mathbf{X}$, $W_M = \frac{Pr(M=m|\mathbf{X}=\mathbf{x},T=t',S=j)}{Pr(M=m|\mathbf{X}=\mathbf{x},T=t,S=j)} = \frac{Pr(M=m|\mathbf{X}_M=\mathbf{x}_M,T=t',S=j)}{Pr(M=m|\mathbf{X}_M=\mathbf{x}_M,T=t,S=j)}$, where $0 < Pr(M(t) = m|\mathbf{X}_M = \mathbf{x}_M, T = t, S = j) < 1$. Then

$$E[Y(t, M(t'))|S = j] = \int_{\mathbf{x}} \int_m \int_y W_T W_M \times y \times f(Y(t, m) = y|T = t, M(t) = m, \mathbf{X} = \mathbf{x}, S = j)$$
$$\times Pr(M(t) = m|T = t, \mathbf{X} = \mathbf{x}, S = j) \times g(\mathbf{X} = \mathbf{x}|T = t, S = j) dy dm d\mathbf{x}$$
$$= E[W_T W_M Y|T = t, S = j].$$

APPENDIX 2.B

Asymptotic Sampling Variance of the Estimators in the Two Steps

As described in Section 2.6.2, the estimation of the causal effects is based on initial estimation of the RMPW weight. A multilevel logistic regression analysis is employed in step 1 to estimate the weight while step 2 involves site-by-site method-of-moments analysis. To represent the sampling variability of the estimated weight obtained in step 1 in the standard errors of the causal effect estimates obtained in step 2, I stack the moment functions from the first step and the second step and solve them simultaneously. This Appendix provides details on the moment functions in both steps and correspondingly the asymptotic sampling variance of the estimators in the two steps, as a supplement to Section 2.6.2.

2.B.1 Moment functions in step 1

In step 1, I fit a logistic regression model to each treatment group, as shown in Equation (2.15), through maximum likelihood estimation. For computational simplicity, following Hedeker & Gibbons (2006), I standardize the random intercept r_{tj} by representing it as $\sigma\theta_{tj}$, where θ_{tj} follows a standardized normal distribution. The step-1 estimators $\widehat{\boldsymbol{\eta}}_t = (\widehat{\boldsymbol{\pi}}_t', \widehat{\sigma}_t)'$, for $t = 0, 1$, solve the following estimating equations,

$$\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{h}_{tij}^{(1)}(M_{ij}, T_{ij}, \mathbf{X}_{ij}, \theta_{tj}, \boldsymbol{\eta}_t) = \mathbf{0}, \tag{2.B1}$$

where $N = \sum_{j=1}^{J} n_j$ is the total sample size of individuals, and $\mathbf{h}_{tij}^{(1)}$ are score functions with the same dimension as $\boldsymbol{\eta}_t = (\boldsymbol{\pi}_t', \sigma_t)'$. Equation (2.B1) is essentially the first-order conditions for the maximum-likelihood estimators in multilevel logistic regression,

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{h}_{tij}^{(1)} = \frac{\partial log\mathcal{L}_t}{\partial\boldsymbol{\eta}_t} = \frac{\partial\sum_{j=1}^{J} log\, l_t(\mathbf{M}_j)}{\partial\boldsymbol{\eta}_t} = \sum_{j=1}^{J}\frac{1}{l_t(\mathbf{M}_j)}\cdot\frac{\partial l_t(\mathbf{M}_j)}{\partial\boldsymbol{\eta}_t}, \tag{2.B2}$$

in which

$$l_t(\mathbf{M}_j) = \int_{\theta_{tj}} f_t(\mathbf{M}_j|\theta_{tj}) g_t(\theta_{tj}) d\theta_{tj}, \tag{2.B3}$$

where

$$f_t(\mathbf{M}_j|\theta_{tj}) = \prod_{i=1}^{n_j} \left[\left(p_{tij}\right)^{M_{ij}} \left(1 - p_{tij}\right)^{1-M_{ij}} \right]^{(1-T_{ij})(1-t)+T_{ij}t}.$$

Correspondingly,

$$\frac{\partial l_t(\mathbf{M}_j)}{\partial \boldsymbol{\eta}_t} = \int_{\theta_{tj}} \left\{ \sum_{i=1}^{n_j} \left[(1-T_{ij})(1-t) + T_{ij}t\right] \left[\frac{M_{ij} - p_{tij}}{p_{tij}\left(1-p_{tij}\right)} \cdot \frac{\partial p_{tij}}{\partial \boldsymbol{\eta}_t} \right] \right\} \cdot f_t(\mathbf{M}_j|\theta_{tj}) g_t(\theta_{tj}) d\theta_{tj}. \tag{2.B4}$$

To approximate the above integral, I use Gauss-Hermite quadrature (Stroud & Secrest, 1966) by summing over Q quadrature points for the integration, given that θ_{tj} is assumed to follow a standardized normal distribution (Hedeker & Gibbons, 2006). Let the optimal points be B_{tq} and the weights be $A_t(B_{tq})$, for $q = 1, \ldots, Q$, under treatment condition $t = 0, 1$. With these points and weights, the approximated marginal likelihood becomes

$$l_t(\mathbf{M}_j) = \int_{\theta_{tj}} f_t(\mathbf{M}_j|\theta_{tj}) g_t(\theta_{tj}) d\theta_{tj} \approx \sum_{q=1}^{Q} f_t(\mathbf{M}_j|B_{tq}) A_t(B_{tq}). \tag{2.B5}$$

Hence,

$$\frac{\partial log\mathcal{L}_t}{\partial \boldsymbol{\eta}_t} \approx \sum_{j=1}^{J} \frac{1}{l_t(\mathbf{M}_j)} \cdot \sum_{q=1}^{Q} \sum_{i=1}^{n_j} \left[(1-T_{ij})(1-t) + T_{ij}t\right] \left[\frac{M_{ij} - p_{tijq}}{p_{tijq}\left(1-p_{tijq}\right)} \cdot \frac{\partial p_{tijq}}{\partial \boldsymbol{\eta}_t} \right] \cdot f_t(\mathbf{M}_j|B_{tq}) A_t(B_{tq}), \tag{2.B6}$$

in which $\frac{\partial p_{tijq}}{\partial \boldsymbol{\eta}_t} = (\frac{\partial p_{tijq}}{\partial \boldsymbol{\pi}_t}, \frac{\partial p_{tijq}}{\partial \sigma_t})$, and

$$p_{tijq} = 1/[1 + \exp(-(\mathbf{X}'_{ij}\boldsymbol{\pi}_t + \sigma_t B_{tq}))];$$

$$\frac{\partial p_{tijq}}{\partial \boldsymbol{\pi}_t} = p_{tijq}\left(1 - p_{tijq}\right)\mathbf{X}'_{ij};$$

$$\frac{\partial p_{tijq}}{\partial \sigma_t} = p_{tijq}\left(1 - p_{tijq}\right)B_{tq}.$$

Therefore,

$$\mathbf{h}^{(1)}_{tij} \approx \frac{1}{l_t(\mathbf{M}_j)} \cdot \sum_{q=1}^{Q}\left[(1 - T_{ij})(1 - t) + T_{ij}t\right]\left[\frac{M_{ij} - p_{tijq}}{p_{tijq}\left(1 - p_{tijq}\right)} \cdot \frac{\partial p_{tijq}}{\partial \boldsymbol{\eta}_t}\right] f_t(\mathbf{M}_j|B_{tq})A_t(B_{tq}). \tag{2.B7}$$

I use $\mathbf{h}^{(1)}_{ij} = (\mathbf{h}^{(1)'}_{0ij}, \mathbf{h}^{(1)'}_{1ij})'$ to denote the moment functions for the step-1 estimators $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\eta}}'_0, \widehat{\boldsymbol{\eta}}'_1)'$.

2.B.2 Moment functions in step 2

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

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

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

$$\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j} h^{(2)}_{ij,1s}(Y_{ij}, T_{ij}, \mu_{1s}) = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \mu_{1s})T_{ij}I(S_{ij} = s) = 0, \tag{2.B8}$$

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

2.B.3 Asymptotic sampling variance of the two-step estimators

Stacking the moment functions from both steps, I have that $\mathbf{h}_{ij} = (\mathbf{h}_{ij}^{(1)'}, \mathbf{h}_{ij}^{(2)'})'$. The estimators in the two steps can be rewritten as a one-step estimator $\widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\eta}}', \widehat{\boldsymbol{\mu}}')'$ which jointly solves $\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{h}_{ij} = \mathbf{0}$. Under the standard regularity conditions, $\widehat{\boldsymbol{\vartheta}}$ is a consistent estimator of $\boldsymbol{\vartheta} = (\boldsymbol{\eta}', \boldsymbol{\mu}')'$ with the asymptotic sampling distribution (Hansen, 1982):

$$\sqrt{N}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \widetilde{\text{var}}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})). \tag{2.B9}$$

The asymptotic covariance matrix of $\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}$ is $\widetilde{\text{var}}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})/N$, in which

$$\widetilde{\text{var}}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}) = \begin{pmatrix} \widetilde{\text{var}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) & \widetilde{\text{cov}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}, \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \\ \widetilde{\text{cov}}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) & \widetilde{\text{var}}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \end{pmatrix} = \mathbf{R}^{-1}\mathbf{H}(\mathbf{R}^{-1})', \tag{2.B10}$$

where

$$\mathbf{H} = E\left[\mathbf{h}_{ij}\mathbf{h}_{ij}'\right] = E\begin{bmatrix} \mathbf{h}_{ij}^{(1)}\mathbf{h}_{ij}^{(1)\prime} & \mathbf{h}_{ij}^{(1)}\mathbf{h}_{ij}^{(2)\prime} \\ \mathbf{h}_{ij}^{(2)}\mathbf{h}_{ij}^{(1)\prime} & \mathbf{h}_{ij}^{(2)}\mathbf{h}_{ij}^{(2)\prime} \end{bmatrix}; \tag{2.B11}$$

$$\mathbf{R} = E\left[\frac{\partial \mathbf{h}_{ij}}{\partial \boldsymbol{\vartheta}}\right] = E\begin{bmatrix} \frac{\partial \mathbf{h}_{ij}^{(1)}}{\partial \boldsymbol{\eta}} & \mathbf{0} \\ \frac{\partial \mathbf{h}_{ij}^{(2)}}{\partial \boldsymbol{\eta}} & \frac{\partial \mathbf{h}_{ij}^{(2)}}{\partial \boldsymbol{\mu}} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{0} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}, \tag{2.B12}$$

in which

$$\mathbf{R}_{11} = E\left[\frac{\partial \mathbf{h}_{ij}^{(1)}}{\partial \boldsymbol{\eta}}\right] = E\begin{bmatrix} \frac{\partial \mathbf{h}_{0ij}^{(1)}}{\partial \boldsymbol{\eta}_0} & \frac{\partial \mathbf{h}_{0ij}^{(1)}}{\partial \boldsymbol{\eta}_1} \\ \frac{\partial \mathbf{h}_{1ij}^{(1)}}{\partial \boldsymbol{\eta}_0} & \frac{\partial \mathbf{h}_{1ij}^{(1)}}{\partial \boldsymbol{\eta}_1} \end{bmatrix} = \begin{bmatrix} -E\left[\mathbf{h}_{0ij}^{(1)}\mathbf{h}_{0ij}^{(1)\prime}\right] & \mathbf{0} \\ \mathbf{0} & -E\left[\mathbf{h}_{1ij}^{(1)}\mathbf{h}_{1ij}^{(1)\prime}\right] \end{bmatrix};$$

$$\mathbf{R}_{22} = E\left[\frac{\partial \mathbf{h}_{ij}^{(2)}}{\partial \boldsymbol{\mu}}\right] = E\begin{bmatrix} \mathbf{R}_{221} & & & & \\ & \ddots & & & \\ & & \mathbf{R}_{22j} & & \\ & & & \ddots & \\ & & & & \mathbf{R}_{22J} \end{bmatrix},$$

in which

$$\mathbf{R}_{22j} = E\begin{bmatrix} \frac{\partial h_{0ij}^{(2)}}{\partial \mu_{0j}} & 0 & 0 \\ 0 & \frac{\partial h_{*ij}^{(2)}}{\partial \mu_{*j}} & 0 \\ 0 & 0 & \frac{\partial h_{1ij}^{(2)}}{\partial \mu_{1j}} \end{bmatrix},$$

where $\frac{\partial h_{0ij}^{(2)}}{\partial \mu_{0j}} = -(1-T_{ij})I(site=j)$, $\frac{\partial h_{*ij}^{(2)}}{\partial \mu_{*j}} = -W_{Mij}T_{ij}I(site=j)$, $\frac{\partial h_{1ij}^{(2)}}{\partial \mu_{1j}} = -T_{ij}I(site=j)$;

$$\mathbf{R}_{21} = E\left[\frac{\partial \mathbf{h}_{ij}^{(2)}}{\partial \boldsymbol{\eta}}\right] = E\begin{bmatrix} \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' \\ \frac{\partial h_{*i1}^{(2)}}{\partial \boldsymbol{\pi}_0} & \frac{\partial h_{*i1}^{(2)}}{\partial \sigma_0} & \frac{\partial h_{*i1}^{(2)}}{\partial \boldsymbol{\pi}_1} & \frac{\partial h_{*i1}^{(2)}}{\partial \sigma_1} \\ \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' \\ \frac{\partial h_{*iJ}^{(2)}}{\partial \boldsymbol{\pi}_0} & \frac{\partial h_{*iJ}^{(2)}}{\partial \sigma_0} & \frac{\partial h_{*iJ}^{(2)}}{\partial \boldsymbol{\pi}_1} & \frac{\partial h_{*iJ}^{(2)}}{\partial \sigma_1} \\ \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' \end{bmatrix},$$

in which

$$\frac{\partial h_{*ij}^{(2)}}{\partial \boldsymbol{\pi}_0} = (Y_{ij} - \mu_{*j})T_{ij}I(site=j)\frac{\partial W_{Mij}}{\partial \boldsymbol{\pi}_0},$$

$$\frac{\partial h_{*ij}^{(2)}}{\partial \sigma_0} = (Y_{ij} - \mu_{*j})T_{ij}I(site=j)\frac{\partial W_{Mij}}{\partial \sigma_0},$$

$$\frac{\partial h_{*ij}^{(2)}}{\partial \boldsymbol{\pi}_1} = (Y_{ij} - \mu_{*j})T_{ij}I(site=j)\frac{\partial W_{Mij}}{\partial \boldsymbol{\pi}_1},$$

$$\frac{\partial h^{(2)}_{*ij}}{\partial \sigma_1} = (Y_{ij} - \mu_{*j}) T_{ij} I(site = j) \frac{\partial W_{Mij}}{\partial \sigma_1},$$

where

$$W_{Mij} = M_{ij} \frac{p_{0ij}}{p_{1ij}} + (1 - M_{ij}) \frac{1 - p_{0ij}}{1 - p_{1ij}},$$

and thus

$$\frac{\partial W_{Mij}}{\partial \boldsymbol{\pi}_0} = \left[\frac{M_{ij}}{p_{1ij}} - \frac{1 - M_{ij}}{1 - p_{1ij}} \right] \frac{\partial p_{0ij}}{\partial \boldsymbol{\pi}_0} = \left[\frac{M_{ij}}{p_{1ij}} - \frac{1 - M_{ij}}{1 - p_{1ij}} \right] p_{0ij}(1 - p_{0ij}) \mathbf{X}'_{ij};$$

$$\frac{\partial W_{Mij}}{\partial \sigma_0} = \left[\frac{M_{ij}}{p_{1ij}} - \frac{1 - M_{ij}}{1 - p_{1ij}} \right] \frac{\partial p_{0ij}}{\partial \sigma_0} = \left[\frac{M_{ij}}{p_{1ij}} - \frac{1 - M_{ij}}{1 - p_{1ij}} \right] p_{0ij}(1 - p_{0ij}) \theta_{0j};$$

$$\begin{aligned}
\frac{\partial W_{Mij}}{\partial \boldsymbol{\pi}_1} &= \left[-M_{ij} \frac{p_{0ij}}{(p_{1ij})^2} + (1 - M_{ij}) \frac{1 - p_{0ij}}{(1 - p_{1ij})^2} \right] \frac{\partial p_{1ij}}{\partial \boldsymbol{\pi}_1} \\
&= \left[-M_{ij} \frac{p_{0ij}}{(p_{1ij})^2} + (1 - M_{ij}) \frac{1 - p_{0ij}}{(1 - p_{1ij})^2} \right] p_{1ij}(1 - p_{1ij}) \mathbf{X}'_{ij} \\
&= \left[-M_{ij} \frac{p_{0ij}}{p_{1ij}} (1 - p_{1ij}) + (1 - M_{ij}) \frac{1 - p_{0ij}}{1 - p_{1ij}} p_{1ij} \right] \mathbf{X}'_{ij};
\end{aligned}$$

$$\begin{aligned}
\frac{\partial W_{Mij}}{\partial \sigma_1} &= \left[-M_{ij} \frac{p_{0ij}}{(p_{1ij})^2} + (1 - M_{ij}) \frac{1 - p_{0ij}}{(1 - p_{1ij})^2} \right] \frac{\partial p_{1ij}}{\partial \sigma_1} \\
&= \left[-M_{ij} \frac{p_{0ij}}{(p_{1ij})^2} + (1 - M_{ij}) \frac{1 - p_{0ij}}{(1 - p_{1ij})^2} \right] p_{1ij}(1 - p_{1ij}) \theta_{1j} \\
&= \left[-M_{ij} \frac{p_{0ij}}{p_{1ij}} (1 - p_{1ij}) + (1 - M_{ij}) \frac{1 - p_{0ij}}{1 - p_{1ij}} p_{1ij} \right] \theta_{1j}.
\end{aligned}$$

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

APPENDIX 2.C

Method-of-Moments Estimator of the Between-Site Variance

In this appendix, as a supplement to Section 2.6.3, I derive the sample estimator of the between-site variance of $\boldsymbol{\beta}_j$. Let $\mathbf{G} = \sum_{j=1}^{J}(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\gamma}})(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\gamma}})'$, we can show that

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

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

in which

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

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

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

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

where

$$\begin{aligned}
& \sum_j \sum_{j'} E(\widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\gamma})(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\gamma})' \\
= & \sum_j \sum_{j'} E[(\widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\beta}_{j'}) + (\boldsymbol{\beta}_{j'} - \boldsymbol{\gamma})][(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + (\boldsymbol{\beta}_j - \boldsymbol{\gamma})]' \\
= & \sum_j \sum_{j'} E[(\widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\beta}_{j'})(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)'] + \sum_j \sum_{j'} E[(\boldsymbol{\beta}_{j'} - \boldsymbol{\gamma})(\boldsymbol{\beta}_j - \boldsymbol{\gamma})'] \\
= & \sum_j \sum_{j' \neq j} \mathrm{cov}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j, \widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\beta}_{j'}) + \sum_j \mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + J\mathrm{var}(\boldsymbol{\beta}_j).
\end{aligned}$$

Therefore,

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

Replacing $\mathrm{var}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ and $\mathrm{cov}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j, \widehat{\boldsymbol{\beta}}_{j'} - \boldsymbol{\beta}_{j'})$ with the corresponding consistent estimators, as derived in Appendix 2.B, I obtain the consistent estimator for the between-site variance:

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

APPENDIX 2.D

Permutation Test for the Between-Site Variance

As a supplement to Section 2.6.3, this Appendix explicates a hypothesis testing procedure for the between-site variance of the direct and indirect effects. Under the null hypothesis that the between-site variance $\sigma^2_{D(0)}$ is zero, that is, $\beta_j^{(D)} = \gamma^{(D)}$ for all j, according to the Central Limit Theorem, $\widehat{\beta}_j^{(D)}$ converges in distribution to a normal distribution as the sample size at the site goes to infinity,

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

in which

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

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

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

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

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

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

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

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

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

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

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

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

APPENDIX 2.E

Generation of Simulation Data

This Appendix explains the generation of simulation data in Section 2.7. The goal is to assess finite-sample performance of the multilevel RMPW procedure in estimating population average and between-site variance of the direct effect and indirect effect. In the basic mediation framework, the treatment affects the mediator, which in turn affects the outcome. Therefore, I generate the data using the following models:

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

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

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

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

In the outcome model, θ_{1j}, θ_{2j} and θ_{3j} are determined by the specified values of the site-specific direct and indirect effects. Based on the expressions derived by Valeri & VanderWeele (2013) for the direct effect and indirect effect under the potential outcomes causal framework,

as defined in Section 2.3, these parameters can be computed as follows:

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

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

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

APPENDIX 3.A

Proof of Theorems 3.1 and 3.2

This Appendix provides a supplement to Section 3.1. The derivation below proves that, under Assumptions 3.1 $\sim$ 3.4, the expectation of each potential outcome at site j, $E[Y(t, M(t'))|S = j]$, for $t, t' = 0, 1$, can be identified with a weighted average of the observed outcome at that site. Let $\mathbf{X} = \{\mathbf{X}_D \cup \mathbf{X}_T \cup \mathbf{X}_R \cup \mathbf{X}_M\}$ be the union of all the observed pretreatment confounders. To simplify notations, I drop the subscript ij of each variable.

$$\begin{aligned} E[Y(t, M(t'))|S = j] &= E\{E[Y(t, M(t'))|\mathbf{X} = \mathbf{x}, S = j]\} \\ &= \int_{\mathbf{x}} \int_m \int_y y \times f(Y(t, m) = y|M(t') = m, \mathbf{X} = \mathbf{x}, S = j) \\ &\quad \times Pr(M(t') = m|\mathbf{X} = \mathbf{x}, S = j) \times g(\mathbf{X} = \mathbf{x}|S = j) dy dm d\mathbf{x}. \end{aligned}$$

Under Assumption 3.1, $\{Y(t, m), M(t)\} \perp\!\!\!\perp D|\mathbf{X}_D = \mathbf{x}_D, S = j$. Because $\mathbf{X}_D \subset \mathbf{X}$, $\{Y(t, m), M(t)\} \perp\!\!\!\perp D|\mathbf{X} = \mathbf{x}, S = j$ also holds. Hence,

$$\begin{aligned} E[Y(t, M(t'))|S = j] = \int_{\mathbf{x}} \int_m \int_y & y \times f(Y(t, m) = y|D = 1, M(t') = m, \mathbf{X} = \mathbf{x}, S = j) \\ & \times Pr(M(t') = m|D = 1, \mathbf{X} = \mathbf{x}, S = j) \times g(\mathbf{X} = \mathbf{x}|S = j) dy dm d\mathbf{x}. \end{aligned}$$

By Bayes theorem,

$$g(\mathbf{X} = \mathbf{x}|S = j) = g(\mathbf{X} = \mathbf{x}|D = 1, S = j) \times \frac{Pr(D = 1|S = j)}{Pr(D = 1|\mathbf{X} = \mathbf{x}, S = j)}.$$

Because we can remove sampling selection bias by only controlling for $\mathbf{X}_D$ and $\mathbf{X}_D \subset \mathbf{X}$, $Pr(D = 1|\mathbf{X} = \mathbf{x}, S = j) = Pr(D = 1|\mathbf{X}_D = \mathbf{x}_D, S = j)$. Hence, the above equation can be

simplified as

$$g(\mathbf{X}=\mathbf{x}|S=j) = g(\mathbf{X}=\mathbf{x}|D=1,S=j) \times \frac{Pr(D=1|S=j)}{Pr(D=1|\mathbf{X}_D=\mathbf{x}_D,S=j)},$$

where $0 < Pr(D=1|\mathbf{X}_D=\mathbf{x}_D,S=j) < 1$. Let $W_D = \frac{Pr(D=1|S=j)}{Pr(D=1|\mathbf{X}_D=\mathbf{x}_D,S=j)}$, and thus

$$E[Y(t,M(t'))|S=j] = \int_{\mathbf{x}}\int_m\int_y W_D \times y \times f(Y(t,m)=y|D=1,M(t')=m,\mathbf{X}=\mathbf{x},S=j)$$
$$\times Pr(M(t')=m|D=1,\mathbf{X}=\mathbf{x},S=j) \times g(\mathbf{X}=\mathbf{x}|D=1,S=j)dydmd\mathbf{x}.$$

Similarly, under the assumption that $0 < Pr(T=1|\mathbf{X}_T=\mathbf{x}_T,D=1,S=j) < 1$, let $W_T = \frac{Pr(T=t|D=1,S=j)}{Pr(T=t|\mathbf{X}_T=\mathbf{x}_T,D=1,S=j)}$. When Assumption 3.2 holds, i.e. $\{Y(t,m),M(t)\} \perp\!\!\!\perp T|D=1,\mathbf{X}_T=\mathbf{x}_T,S=j$, and by Bayes theorem, the above equation is equal to

$$\int_{\mathbf{x}}\int_m\int_y W_D W_T \times y \times f(Y(t,m)=y|T=t,D=1,M(t')=m,\mathbf{X}=\mathbf{x},S=j)$$
$$\times Pr(M(t')=m|T=t',D=1,\mathbf{X}=\mathbf{x},S=j) \times g(\mathbf{X}=\mathbf{x}|T=t,D=1,S=j)dydmd\mathbf{x}.$$

Under the assumption that $0 < Pr(R=1|\mathbf{X}_R=\mathbf{x}_R,T=t,D=1,S=j) < 1$, let $W_R = \frac{Pr(R=1|T=t,D=1,S=j)}{Pr(R=1|\mathbf{X}_R=\mathbf{x}_R,T=t,D=1,S=j)}$. When Assumption 3.3 holds, i.e. $\{Y(t,m),M(t)\} \perp\!\!\!\perp R|T=t,D=1,\mathbf{X}_R=\mathbf{x}_R,S=j$, and by Bayes theorem, the above equation is equal to

$$\int_{\mathbf{x}}\int_m\int_y W_D W_T W_R \times y \times f(Y(t,m)=y|R=1,T=t,D=1,M(t')=m,\mathbf{X}=\mathbf{x},S=j)$$
$$\times Pr(M(t')=m|R=1,T=t',D=1,\mathbf{X}=\mathbf{x},S=j)$$
$$\times g(\mathbf{X}=\mathbf{x}|R=1,T=t,D=1,S=j)dydmd\mathbf{x}.$$

When $t=t'$, it is easy to obtain the following identification result,

$$E[Y(t,M(t))|S=j] = E[W_D W_T W_R Y|R=1,T=t,D=1,S=j].$$

When $t \neq t'$ and if Assumption 3.4 holds, i.e.

$$Y(t,m) \perp\!\!\!\perp \{M(t), M(t')\} | R=1, T=t, D=1, \mathbf{X}_M = \mathbf{x}_M, S=j,$$

$$\begin{aligned} E[Y(t,M(t'))|S=j] = \int_{\mathbf{x}} \int_m \int_y & W_D W_T W_R \times y \\ & \times f(Y(t,m)=y|R=1,T=t,D=1,M(t)=m,\mathbf{X}=\mathbf{x},S=j) \\ & \times Pr(M(t')=m|R=1,T=t',D=1,\mathbf{X}=\mathbf{x},S=j) \\ & \times g(\mathbf{X}=\mathbf{x}|R=1,T=t,D=1,S=j) dy dm d\mathbf{x}. \end{aligned}$$

Under the assumption that $0 < Pr(M(t)=m|\mathbf{X}=\mathbf{x},R=1,T=t,D=1,S=j) < 1$, let $W_M = \frac{Pr(M(t')=m|\mathbf{X}=\mathbf{x},R=1,T=t',D=1,S=j)}{Pr(M(t)=m|\mathbf{X}=\mathbf{x},R=1,T=t,D=1,S=j)} = \frac{Pr(M=m|\mathbf{X}=\mathbf{x},R=1,T=t',D=1,S=j)}{Pr(M=m|\mathbf{X}=\mathbf{x},R=1,T=t,D=1,S=j)}$, because $M(t) = M$ when $T = t$ and, similarly, $M(t') = M$ when $T = t'$. Because we can remove mediator selection bias by only controlling for $\mathbf{X}_M$ and $\mathbf{X}_M \subset \mathbf{X}$, $W_M = \frac{Pr(M=m|\mathbf{X}=\mathbf{x},R=1,T=t',D=1,S=j)}{Pr(M=m|\mathbf{X}=\mathbf{x},R=1,T=t,D=1,S=j)} = \frac{Pr(M=m|\mathbf{X}_M=\mathbf{x}_M,R=1,T=t',D=1,S=j)}{Pr(M=m|\mathbf{X}_M=\mathbf{x}_M,R=1,T=t,D=1,S=j)}$, where $0 < Pr(M(t) = m|\mathbf{X}_M = \mathbf{x}_M, R=1, T=t, D=1, S=j) < 1$. Then

$$\begin{aligned} E[Y(t,M(t'))|S=j] = \int_{\mathbf{x}} \int_m \int_y & W_D W_T W_R W_M \times y \\ & \times f(Y(t,m)=y|R=1,T=t,D=1,M(t)=m,\mathbf{X}=\mathbf{x},S=j) \\ & \times Pr(M(t)=m|R=1,T=t,D=1,\mathbf{X}=\mathbf{x},S=j) \\ & \times g(\mathbf{X}=\mathbf{x}|R=1,T=t,D=1,S=j) dy dm d\mathbf{x} \\ & = E[W_D W_T W_R W_M Y|R=1,T=t,D=1,S=j]. \end{aligned}$$

APPENDIX 3.B. Sample Statistics by Treatment, Response Status, and Mediator

		T = 1							T = 0
		R = 1			R = 0	R = 1			R = 0
Variable Name	**Description**	**M = 1**	**M = 0**	**Combined**		**M = 1**	**M = 0**	**Combined**	
Outcome, mean (standard deviation)									
earny4	Weekly earnings in the fourth year	244.48	191.89	213.04		228.25	186.97	196.15	
		(212.92)	(187.43)	(199.73)		(189.61)	(182.56)	(184.92)	
Pretreatment Covariates, mean (standard deviation)									
Demographic characteristics									
female	Female	0.46	0.48	0.47	0.35	0.42	0.38	0.39	0.26
		(0.5)	(0.5)	(0.5)	(0.48)	(0.49)	(0.49)	(0.49)	(0.44)
age_1617	Age was 16-17 at application	0.42	0.37	0.39	0.41	0.50	0.41	0.43	0.38
		(0.49)	(0.48)	(0.49)	(0.49)	(0.5)	(0.49)	(0.49)	(0.49)
age_1819	Age was 18-19 at application	0.33	0.31	0.32	0.30	0.32	0.31	0.31	0.30
		(0.47)	(0.46)	(0.47)	(0.46)	(0.47)	(0.46)	(0.46)	(0.46)
race_w	Race/ethnicity is White	0.30	0.25	0.27	0.25	0.28	0.26	0.27	0.24
		(0.46)	(0.43)	(0.44)	(0.43)	(0.45)	(0.44)	(0.44)	(0.43)
race_b	Race/ethnicity is Black	0.46	0.52	0.49	0.46	0.46	0.50	0.49	0.45
		(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)
race_h	Race/ethnicity is Hispanic	0.18	0.16	0.17	0.19	0.19	0.17	0.17	0.19
		(0.38)	(0.37)	(0.37)	(0.39)	(0.39)	(0.37)	(0.38)	(0.39)
otherling	If native language is not English	0.14	0.13	0.14	0.14	0.13	0.14	0.13	0.14
		(0.35)	(0.34)	(0.34)	(0.35)	(0.34)	(0.34)	(0.34)	(0.35)
inpers	Lived in in-person areas	0.66	0.68	0.67	0.72	0.73	0.72	0.72	0.78
		(0.47)	(0.47)	(0.47)	(0.45)	(0.45)	(0.45)	(0.45)	(0.41)
in57	In 57 areas from which many nonresidential students came from	0.32	0.36	0.34	0.35	0.38	0.35	0.36	0.37
		(0.46)	(0.48)	(0.47)	(0.48)	(0.49)	(0.48)	(0.48)	(0.48)
nonres	Designated for a nonresidential slot	0.18	0.22	0.21	0.17	0.17	0.17	0.17	0.14
		(0.38)	(0.42)	(0.4)	(0.37)	(0.38)	(0.37)	(0.37)	(0.35)
area_d	From a dense area	0.29	0.30	0.30	0.27	0.31	0.28	0.29	0.27
		(0.45)	(0.46)	(0.46)	(0.45)	(0.46)	(0.45)	(0.45)	(0.44)
area_nd	From a nondense area	0.37	0.35	0.36	0.32	0.34	0.38	0.37	0.30

					T = 1				T = 0
			R = 1		R = 0		R = 1		R = 0
Variable Name	**Description**	**M = 1**	**M = 0**	**Combined**		**M = 1**	**M = 0**	**Combined**	
		(0.48)	(0.48)	(0.48)	(0.47)	(0.47)	(0.49)	(0.48)	(0.46)
jcmsa_1	Residence status is PMSA at baseline	0.32	0.29	0.30	0.35	0.33	0.28	0.29	0.37
		(0.47)	(0.45)	(0.46)	(0.48)	(0.47)	(0.45)	(0.46)	(0.48)
jcmsa_2	Residence status is MSA at baseline	0.46	0.48	0.47	0.40	0.48	0.46	0.47	0.39
		(0.5)	(0.5)	(0.5)	(0.49)	(0.5)	(0.5)	(0.5)	(0.49)
app_1	Applied to Job Corps in quarter 1	0.22	0.25	0.24	0.24	0.23	0.22	0.22	0.21
		(0.41)	(0.43)	(0.42)	(0.42)	(0.42)	(0.41)	(0.41)	(0.41)
app_2	Applied to Job Corps in quarter 2	0.30	0.29	0.29	0.25	0.31	0.30	0.31	0.24
		(0.46)	(0.45)	(0.46)	(0.43)	(0.46)	(0.46)	(0.46)	(0.43)
app_3	Applied to Job Corps in quarter 3	0.30	0.27	0.28	0.25	0.27	0.28	0.28	0.27
		(0.46)	(0.44)	(0.45)	(0.44)	(0.45)	(0.45)	(0.45)	(0.44)
phaseII_demo_nmiss	Phase II demographics not missing	0.01	0.03	0.02	0.05	0.01	0.03	0.02	0.05
		(0.09)	(0.16)	(0.14)	(0.21)	(0.08)	(0.16)	(0.14)	(0.21)
Fertility and living arrangements at the baseline interview									
haschld	Had children at baseline	0.19	0.23	0.21	0.17	0.18	0.18	0.18	0.16
		(0.39)	(0.42)	(0.41)	(0.38)	(0.38)	(0.38)	(0.38)	(0.37)
r_head	Sample member is head of household at baseline	0.10	0.14	0.12	0.11	0.13	0.11	0.12	0.12
		(0.3)	(0.35)	(0.33)	(0.32)	(0.34)	(0.31)	(0.32)	(0.32)
livespou	Lived with spouse or partner at baseline	0.06	0.07	0.06	0.06	0.06	0.06	0.06	0.06
		(0.23)	(0.25)	(0.24)	(0.24)	(0.23)	(0.24)	(0.24)	(0.23)
publich	Lived in public or rent-subsidized housing at baseline	0.19	0.21	0.20	0.19	0.18	0.19	0.19	0.19
		(0.39)	(0.41)	(0.4)	(0.39)	(0.39)	(0.39)	(0.39)	(0.39)
hh1_3	Family size is 1-3 at baseline	0.34	0.34	0.34	0.32	0.35	0.33	0.34	0.35
		(0.47)	(0.47)	(0.47)	(0.47)	(0.48)	(0.47)	(0.47)	(0.48)
hh4_6	Family size is 4-6 at baseline	0.53	0.50	0.51	0.48	0.54	0.49	0.50	0.45
		(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)
Education and training experiences prior to random assignment									
hs_ged	Had high school diploma or GED certificates at baseline	0.17	0.30	0.25	0.19	0.13	0.26	0.23	0.21
		(0.38)	(0.46)	(0.43)	(0.4)	(0.34)	(0.44)	(0.42)	(0.4)

					T = 1				T = 0
			R = 1		R = 0		R = 1		R = 0
Variable Name	**Description**	**M = 1**	**M = 0**	**Combined**		**M = 1**	**M = 0**	**Combined**	
any_ed	Joined any education program in the past year	0.71	0.61	0.65	0.58	0.78	0.64	0.67	0.56
		(0.45)	(0.49)	(0.48)	(0.49)	(0.41)	(0.48)	(0.47)	(0.5)
reassch02	Reason for leaving school	0.11	0.16	0.14	0.14	0.09	0.15	0.13	0.14
		(0.32)	(0.37)	(0.35)	(0.35)	(0.29)	(0.35)	(0.34)	(0.34)
reasschm	Reason for leaving school is missing	0.72	0.63	0.67	0.67	0.78	0.66	0.69	0.68
		(0.45)	(0.48)	(0.47)	(0.47)	(0.41)	(0.47)	(0.46)	(0.47)
arrst	Ever arrested at baseline	0.22	0.21	0.21	0.21	0.25	0.21	0.22	0.21
		(0.41)	(0.41)	(0.41)	(0.41)	(0.43)	(0.41)	(0.42)	(0.41)
sick	Had physical or emotional problems that limited the amount of work that could be done at baseline	0.04	0.05	0.05	0.04	0.05	0.05	0.05	0.05
		(0.2)	(0.22)	(0.21)	(0.2)	(0.22)	(0.22)	(0.22)	(0.22)
badhlth	Bad health at baseline	0.12	0.14	0.13	0.12	0.14	0.13	0.13	0.11
		(0.32)	(0.34)	(0.33)	(0.32)	(0.35)	(0.33)	(0.34)	(0.32)
potreg	Used Marijuana regularly in the past year	0.09	0.08	0.08	0.09	0.08	0.08	0.08	0.09
		(0.28)	(0.27)	(0.28)	(0.28)	(0.28)	(0.27)	(0.27)	(0.28)
potocc	Used Marijuana occasionally in the past year	0.22	0.20	0.21	0.21	0.24	0.21	0.22	0.19
		(0.41)	(0.4)	(0.41)	(0.41)	(0.43)	(0.41)	(0.41)	(0.39)
drugtrt	Ever in drug treatment	0.05	0.04	0.05	0.04	0.05	0.05	0.05	0.06
		(0.21)	(0.21)	(0.21)	(0.21)	(0.21)	(0.22)	(0.22)	(0.24)
Employment and earnings prior to random assignment									
evworkb	Had full or part time job at baseline	0.82	0.78	0.79	0.74	0.82	0.77	0.78	0.71
		(0.39)	(0.42)	(0.4)	(0.44)	(0.38)	(0.42)	(0.41)	(0.45)
currjob	Employed at baseline	0.23	0.20	0.21	0.19	0.22	0.20	0.20	0.17
		(0.42)	(0.4)	(0.41)	(0.39)	(0.42)	(0.4)	(0.4)	(0.38)
numbjob_0	No jobs in the past year	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
		(0.09)	(0.08)	(0.08)	(0.07)	(0.07)	(0.08)	(0.08)	(0.09)
job3_9	Months employed in the past year is 3-9	0.28	0.24	0.26	0.24	0.27	0.24	0.25	0.22
		(0.45)	(0.43)	(0.44)	(0.43)	(0.45)	(0.43)	(0.43)	(0.42)

					T = 1				T = 0
			R = 1		R = 0		R = 1		R = 0
Variable Name	**Description**	**M = 1**	**M = 0**	**Combined**		**M = 1**	**M = 0**	**Combined**	
job9_12	Months employed in the past year is 9-122	0.18	0.16	0.17	0.14	0.17	0.17	0.17	0.14
		(0.38)	(0.37)	(0.37)	(0.35)	(0.37)	(0.37)	(0.37)	(0.35)
earn0	Earnings in the past year is 0	0.33	0.36	0.35	0.34	0.33	0.36	0.35	0.34
		(0.47)	(0.48)	(0.48)	(0.47)	(0.47)	(0.48)	(0.48)	(0.47)
earn1	Earnings in the past year is 0-1000	0.10	0.10	0.10	0.10	0.13	0.10	0.11	0.10
		(0.3)	(0.3)	(0.3)	(0.3)	(0.34)	(0.3)	(0.31)	(0.29)
earn2	Earnings in the past year is 1000-5000	0.30	0.26	0.27	0.24	0.29	0.26	0.27	0.21
		(0.46)	(0.44)	(0.45)	(0.43)	(0.45)	(0.44)	(0.44)	(0.41)
earn3	Earnings in the past year is 5000-10000	0.16	0.12	0.14	0.12	0.13	0.13	0.13	0.13
		(0.36)	(0.33)	(0.34)	(0.33)	(0.33)	(0.34)	(0.33)	(0.34)
Public assistance receipt prior to random assignment									
afdc_fs	Received AFDC or food stamp in the past year	0.45	0.50	0.48	0.40	0.47	0.48	0.48	0.38
		(0.5)	(0.5)	(0.5)	(0.49)	(0.5)	(0.5)	(0.5)	(0.48)
otherwelf	Received other welfare in the past year	0.11	0.09	0.10	0.09	0.10	0.11	0.11	0.08
		(0.31)	(0.28)	(0.3)	(0.28)	(0.29)	(0.31)	(0.31)	(0.27)
anywelf_miss	Whether receiving any welfare in the past year is missing	0.04	0.07	0.06	0.15	0.03	0.07	0.06	0.18
		(0.2)	(0.26)	(0.24)	(0.36)	(0.18)	(0.26)	(0.24)	(0.38)
mostwelf	Family was on welfare most of the time when youth was growing up	0.18	0.21	0.20	0.18	0.18	0.19	0.19	0.17
		(0.38)	(0.41)	(0.4)	(0.38)	(0.38)	(0.39)	(0.39)	(0.38)
Motivation and support for joining Job Corps (JC)									
jobmotiv_8_10	Motivation for joining JC is moderate	0.39	0.35	0.37	0.34	0.41	0.36	0.37	0.34
		(0.49)	(0.48)	(0.48)	(0.47)	(0.49)	(0.48)	(0.48)	(0.48)
jobmotiv_11_14	Motivation for joining JC is strong	0.44	0.46	0.45	0.43	0.42	0.46	0.45	0.40
		(0.5)	(0.5)	(0.5)	(0.49)	(0.49)	(0.5)	(0.5)	(0.49)
nencouragement_3_5	Received moderate support for JC participation	0.51	0.48	0.49	0.46	0.52	0.50	0.51	0.44
		(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)
nencouragement_6_8	Received strong support for JC participation	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.02
		(0.14)	(0.14)	(0.14)	(0.16)	(0.14)	(0.14)	(0.14)	(0.13)

APPENDIX 3.C. Sensitivity Analysis

As a supplement to Section 3.2.4, this Appendix introduces a weighting-based approach to sensitivity analysis that has been extended from single-site to multisite causal mediation studies (Hong, Qin, & Yang, 2018, working paper).

In Chapter 3, the causal conclusions may be easily altered by a plausible violation of the strongly ignorable nonresponse (Assumption 3.3) or the strongly ignorable mediator value assignment (Assumption 3.4). Hidden bias may arise due to possible omissions of pretreatment and posttreatment confounders or omissions of possible between-site variations in the selection mechanism. The latter happens when pretreatment selection into different mediator values vary across the sites but site-specific increments to pretreatment confounders are omitted from the propensity score models. The weighting-based approach to sensitivity analysis assesses the consequences of such omissions. It quantifies the amount of bias due to the omission by comparing an initial weight with a new weight that adjusts for the omissions. Below I focus on discussing the weighting-based sensitivity parameters for the population average and between-site variance of the natural indirect effect (NIE) and extend the same logic to the other causal parameters.

3.C.1 Sensitivity Analysis for the Natural Indirect Effect

I derive the bias in identifying the site-specific NIE due to each type of omission. Although I focus on the omission of a single confounder, the result can be extended to a set of confounders.

Adjusting for an Omitted Pretreatment Confounder

Let *P* be an omitted pretreatment confounder of the response-mediator, response-outcome, or mediator-outcome relationships. The site-specific NIE, denoted by $\beta_j^{(I)}(1)$, is to be identified when both **X** and *P* are adjusted for in the propensity score models:

$$A_{Pj} = E\left[W_{Dij}W_{Tij}W_{P.Rij}Y_{ij}\,|T_{ij}\ = 1, S_{ij} = j\right] - E\left[W_{Dij}W_{Tij}W_{P.Rij}W_{P.Mij}Y_{ij}\,|T_{ij}\ = 1, S_{ij} = j\right],$$

where $W_{P.Rij} = \frac{Pr\left(R_{ij}=1|T_{ij}=t, D_{ij}=1, S_{ij}=j\right)}{Pr\left(R_{ij}=1|\mathbf{X}_{ij}=\mathbf{x}, P_{ij}=p, T_{ij}=t, D_{ij}=1, S_{ij}=j\right)}$ for $t = 0, 1$ and $W_{P.Mij} = \frac{Pr\left(M_{ij}=m|\ T_{ij}=0, \mathbf{X}_{ij}=\mathbf{x}, P_{ij}=p, S_{ij}=j\right)}{Pr\left(M_{ij}=m|T_{ij}=1, \mathbf{X}_{ij}=\mathbf{x}, P_{ij}=p, S_{ij}=j\right)}$. If *P* only confounds the mediator-outcome relationship, the adjusted nonresponse weight $W_{P.Rij}$ is equal to the initial weight W_{Rij}. Similarly, if *P* only confounds the response-mediator or response-outcome relationship, the adjusted RMPW weight $W_{P.Mij}$ is equal to the initial weight W_{Mij}.

Adjusting for the Omitted Site-Specific Increment for a Pretreatment Confounder

I use *X* to denote an element in the vector of observed pretreatment covariates **X** included in the original adjustment. The predictive relationship between *X* and the indicator for response status *R* or that between *X* and the mediator *M* under either or both treatment conditions may vary across the sites. In such cases, a response model or a mediator model omitting the site-specific increment for *X* is misspecified. To allow the predictive relationships to be different across the sites, I specify the nonresponse model as

$$\log\left[\frac{p_{X^\star.Rtij}}{1 - p_{X^\star.Rtij}}\right] = \boldsymbol{X}_{ij}'\boldsymbol{\pi}_{Rt} + r_{Rtj} + r_{1Rtj}X, \left(r_{Rtj}, r_{1Rtj}\right)' \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\upsilon_{Rt}^2 & \upsilon_{01Rt}\\ \upsilon_{01Rt} & \upsilon_{1Rt}^2\end{pmatrix}\right),$$

and the mediator model as

$$\log\left[\frac{p_{X^\star.Mtij}}{1-p_{X^\star.Mtij}}\right] = \boldsymbol{X}_{ij}'\boldsymbol{\pi}_{Mt} + r_{Mtj} + r_{1Mtj}X, \left(r_{Mtj}, r_{1Mtj}\right)' \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\upsilon_{Mt}^2 & \upsilon_{01Mt}\\ \upsilon_{01Mt} & \upsilon_{1Mt}^2\end{pmatrix}\right).$$

Here r_{1Rtj} and r_{1Mtj} are the random site-specific increments to the coefficients for *X* in the nonresponse model and the mediator model, respectively. With these adjustments, $\beta_j^{(I)}(1)$ will be identified by:

$$A_{X^\star j} = E\left[W_{Dij}W_{Tij}W_{X^\star.Rij}Y_{ij} \mid T_{ij} = 1, S_{ij} = j\right] - E\left[W_{Dij}W_{Tij}W_{X^\star.Rij}W_{X^\star.Mij}Y_{ij} \mid T_{ij} = 1, S_{ij} = j\right].$$

where $W_{X^\star.Rij} = p_{Rtj}/p_{X^\star.Rtij}$ for the respondents, $W_{X^\star.Mij} = p_{X^\star.Mt'ij}/p_{X^\star.Mtij}$ for respondents with $M = 1$ in treatment group t at site *j*, and $W_{Mij} = \left(1 - p_{X^\star.Mt'ij}\right)/\left(1 - p_{X^\star.Mtij}\right)$ for respondents with $M = 0$ in the same group at the same site.

This result also applies to an observed or unobserved pretreatment covariate *P* that has been omitted from the original analysis. If we allow the predictive relationship between *P* and *R* or that between *P* and *M* under each treatment to be different across the sites, we may replace $W_{X^\star.Rij}$ with $W_{P^\star.Rij}$ and replace $W_{X^\star.Mij}$ with $W_{P^\star.Mij}$.

Adjusting for an Omitted Posttreatment Confounder

Because posttreatment confounders are intermediate outcomes of the treatment, they cannot be adjusted for through standard regression in causal mediation analysis. A posttreatment confounder of the mediator-outcome relationship *Q* can be viewed as an additional mediator that transmits the treatment effect both directly on the outcome and indirectly through affecting the focal mediator *M* and then the outcome. In the single-level setting, Hong, Qin, and Fan (2018) defined the population average NIE transmitted through the focal mediator as

$$E\left[Y\left(1, Q(1), M\left(1, Q(1)\right)\right)\right] - E\left[Y\left(1, Q(1), M\left(0, Q(0)\right)\right)\right].$$

This definition is consistent with the NIE defined in the absence of the posttreatment confounder Q because $Y\left(1, Q(1), M\left(1, Q(1)\right)\right)$ can be simplified as $Y\left(1, M(1)\right)$ and because $Y\left(1, Q(1), M\left(0, Q(0)\right)\right)$ can be simplified as $Y\left(1, M(0)\right)$. Using a weighting-based approach to sensitivity analysis, Hong, Qin, and Fan (2018) derived the effect size of bias in identifying the NIE due to the confounding effect of *Q*. Hong, Qin, and Fan (working paper) extended this result to the multisite causal mediation studies.

Based on an important result from Hong (2015), $E\left[Y_{ij}\left(1, Q_{ij}(1), M_{ij}\left(0, Q_{ij}(0)\right)\right)|S_{ij} = j\right]$ can be identified by the weighted average observed outcome of the experimental group in the site under a specific set of assumptions. Assumption 3.3 is modified such that strongly ignorable nonresponse is assumed to hold given **X** and *Q*, and Assumption 3.4 is modified such that it now assumes strongly ignorable mediator value assignment given **X** and *Q*. In addition to Assumptions 3.1 and 3.2 and the modified Assumptions 3.3 and 3.4, a new assumption is invoked that, within each site, the assignment to different values of the posttreatment confounder *Q* is strongly ignorable given **X**.

If *Q* is the only posttreatment covariate that confounds the response-mediator, response-outcome, or mediator-outcome relationship, $\beta_j^{(I)}(1)$ will be identified when both **X** and *Q* are adjusted for:

$$A_{Qj} = E\left[W_{Dij}W_{Tij}W_{Q.Rij}Y_{ij}\,|T_{ij} = 1, S_{ij} = j\right] - E\left[W_{Dij}W_{Tij}W_{Q.Rij}W_{Q.Mij}Y_{ij}\,|T_{ij} = 1, S_{ij} = j\right],$$

For individual *i* in site *j* displaying mediator value *m* and with pretreatment covariate values **x** and posttreatment covariate value *q*, under treatment randomization, the weights are

$$W_{Q.Rij} = \frac{Pr\left(R_{ij} = 1|T_{ij} = t, D_{ij} = 1, S_{ij} = j\right)}{Pr\left(R_{ij} = 1|\mathbf{X}_{ij} = \mathbf{x}, Q_{ij} = q, T_{ij} = t, D_{ij} = 1, S_{ij} = j\right)} \text{ for } t = 0, 1,$$

$$W_{Q.Mij} = \frac{pr\left(M_{ij} = m | Z_{ij} = 0, Q_{ij} = q, \mathbf{X}_{ij} = \mathbf{x}\right)}{pr\left(M_{ij} = m | Z_{ij} = 1, Q_{ij} = q, \mathbf{X}_{ij} = \mathbf{x}\right)}.$$

Adjusting for the Omitted Site-Specific Increment for a Posttreatment Confounder

The predictive relationship between *Q* and the response *R* or that between *X* and the mediator under either or both treatment conditions is not necessarily constant across the sites. When this is the case, a response model or a mediator model that includes *Q* but omits the site-specific increment for *Q* is misspecified. In order to assess the bias due to this model misspecification, I specify the nonresponse model as

$$\log\left[\frac{p_{Q^\star.Rtij}}{1-p_{Q^\star.Rtij}}\right] = \boldsymbol{X}_{ij}'\boldsymbol{\pi}_{Rt} + r_{Rtj} + \lambda_{Rt}Q + r_{1Rtj}Q, \left(r_{Rtj}, r_{1Rtj}\right)' \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\upsilon_{Rt}^2 & \upsilon_{01Rt}\\ \upsilon_{01Rt} & \upsilon_{1Rt}^2\end{pmatrix}\right),$$

and the mediator model as

$$\log\left[\frac{p_{Q^\star.Mtij}}{1-p_{Q^\star.Mtij}}\right] = \boldsymbol{X}_{ij}'\boldsymbol{\pi}_{Mt} + r_{Mtj} + \lambda_{Mt}Q + r_{1Mtj}Q, \left(r_{Mtj}, r_{1Mtj}\right)' \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\upsilon_{Mt}^2 & \upsilon_{01Mt}\\ \upsilon_{01Mt} & \upsilon_{1Mt}^2\end{pmatrix}\right).$$

The site-specific NIE is now identified by

$$A_{Q^\star j} = E\left[W_{Dij}W_{Tij}W_{Q^\star.Rij}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right] - E\left[W_{Dij}W_{Tij}W_{Q^\star.Rij}W_{Q^\star.Mij}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right],$$

where $W_{Q^\star.Rij}$ and $W_{Q^\star.Mij}$ are obtained from analyzing the modified logistic regression models.

Generic Form of Bias in Identifying the Site-Specific NIE

It is noteworthy that, across all four types of omissions, the identification of the site-specific NIE takes similar forms. Henceforth I use $A_j^{\#}$ as a general form of the adjusted identification result of NIE standing for A_{Pj}, $A_{X^\star j}$, $A_{P^\star j}$, A_{Qj}, or $A_{Q^\star j}$, use $W_{Rij}^{\#}$ as a generic form of the new weights standing for $W_{P.Rij}$, $W_{X^\star.Rij}$, $W_{P^\star.Rij}$, $W_{Q.Rij}$, or $W_{Q^\star.Rij}$ and use $W_{Mij}^{\#}$ as a generic form

of the new weights standing for $W_{P.Mij}$, $W_{X^{\star}.Mij}$, $W_{P^{\star}.Mij}$, $W_{Q.Mij}$, or $W_{Q^{\star}.Mij}$. Hence, the general form of the bias in identifying the site-specific NIE is

$$\begin{aligned}
A_j - A_j^{\#} = &\{E\left[W_{Dij}W_{Tij}W_{Rij}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right] \\
&- E\left[W_{Dij}W_{Tij}W_{Rij}W_{Mij}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right]\} \\
&- \{E\left[W_{Dij}W_{Tij}W_{Rij}^{\#}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right] \\
&- E\left[W_{Dij}W_{Tij}W_{Rij}^{\#}W_{Mij}^{\#}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right]\} \\
&= \{E\left[W_{Dij}W_{Tij}W_{Rij}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right] \\
&- E\left[W_{Dij}W_{Tij}W_{Rij}^{\#}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right]\} \\
&- \{E\left[W_{Dij}W_{Tij}W_{Rij}W_{Mij}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right] \\
&- E\left[W_{Dij}W_{Tij}W_{Rij}^{\#}W_{Mij}^{\#}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right]\},
\end{aligned}$$

in which A_j is the initial identification results without adjustment for the omitted pretreatment or posttreatment confounders or for the between-site variations of the selection mechanisms.

$$\begin{aligned}
&E\left[W_{Dij}W_{Tij}W_{Rij}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right] - E\left[W_{Dij}W_{Tij}W_{Rij}^{\#}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right] \\
&= E\left[\left(W_{Dij}W_{Tij}W_{Rij} - W_{Dij}W_{Tij}W_{Rij}^{\#}\right)Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right] \\
&= cov\left(W_{Dij}W_{Tij}W_{Rij} - W_{Dij}W_{Tij}W_{Rij}^{\#}, Y_{ij}|T_{ij} = 1, S_{ij} = j\right) \\
&+ E\left[W_{Dij}W_{Tij}W_{Rij} - W_{Dij}W_{Tij}W_{Rij}^{\#} \,|T_{ij} = 1, S_{ij} = j\right]E\left[Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right] \\
&= cov\left(W_{Dij}W_{Tij}W_{Rij} - W_{Dij}W_{Tij}W_{Rij}^{\#}, Y_{ij}|T_{ij} = 1, S_{ij} = j\right)
\end{aligned}$$

The last equation holds because $E\left[W_{Dij}W_{Tij}W_{Rij} \,|T_{ij} = 1, S_{ij} = j\right] = E\left[W_{Dij}W_{Tij}W_{Rij}^{\#} \,|T_{ij} = 1, S_{ij} = j\right] = 1$. Similarly,

$$\begin{aligned}
&E\left[W_{Dij}W_{Tij}W_{Rij}W_{Mij}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right] - E\left[W_{Dij}W_{Tij}W_{Rij}^{\#}W_{Mij}^{\#}Y_{ij} \,|T_{ij} = 1, S_{ij} = j\right] \\
&= cov\left(W_{Dij}W_{Tij}W_{Rij}W_{Mij} - W_{Dij}W_{Tij}W_{Rij}^{\#}W_{Mij}^{\#}, Y_{ij}|T_{ij} = 1, S_{ij} = j\right)
\end{aligned}$$

Generic Form of Sensitivity Parameters

The general form of the bias in identifying the site-specific NIE can be rewritten as

$$A_j - A_j^{\#} = \left(\sigma_{1j}^{\#}\rho_{1j}^{\#} - \sigma_{*j}^{\#}\rho_{*j}^{\#}\right)\sigma_{Y|T=1,j},$$

where $\sigma_{1j}^{\#} = \sqrt{var\left(W_{Dij}W_{Tij}W_{Rij} - W_{Dij}W_{Tij}W_{Rij}^{\#}|T_{ij} = 1, S_{ij} = j\right)}$ is the standard deviation of the weight discrepancy in the experimental group at site j and $\rho_{1j}^{\#} = corr\left(W_{Dij}W_{Tij}W_{Rij} - W_{Dij}W_{Tij}W_{Rij}^{\#}, Y_{ij}|T_{ij} = 1, S_{ij} = j\right)$ is the correlation between the weight discrepancy and the outcome in the same group at the same site for identifying the population mean outcome under the experimental condition; $\sigma_{*j}^{\#} =$

$\sqrt{var\left(W_{Dij}W_{Tij}W_{Rij}W_{Mij} - W_{Dij}W_{Tij}W_{Rij}^{\#}W_{Mij}^{\#}|T_{ij} = 1, S_{ij} = j\right)}$ and $\rho_{*j}^{\#} =$

$corr\left(W_{Dij}W_{Tij}W_{Rij} - W_{Dij}W_{Tij}W_{Rij}^{\#}, Y_{ij}|T_{ij} = 1, S_{ij} = j\right)$ are for identifying the population mean potential outcome under the experimental condition while the mediator takes the value under the counterfactual control condition; and $\sigma_{Y|T=1,j} = \sqrt{var\left(Y_{ij}|T_{ij} = 1, S_{ij} = j\right)}$ is the within-site standard deviation of the outcome in the experimental group. In an application in which more than one type of omissions exist, the analyst may compute their aggregate bias.

By convention, the standard deviation of the outcome in the control group denoted by $\sigma_{Y|T=0}$ is usually the scaling unit for calculating effect sizes. Hence, the effect size of the bias in identifying $\beta_j^{(I)}(1)$ is

$$\frac{A_j - A_j^{\#}}{\sigma_{Y|T=0}} = c_j\left(\sigma_{1j}^{\#}\rho_{1j}^{\#} - \sigma_{*j}^{\#}\rho_{*j}^{\#}\right) \text{ where } c_j = \frac{\sigma_{Y|T=1,j}}{\sigma_{Y|T=0}}.$$

Here c_j is simply a conversion coefficient that converts the effect size to the conventional scale. This derivation makes clear that $\sigma_{1j}^{\#}, \rho_{1j}^{\#}, \sigma_{*j}^{\#}$, and $\rho_{*j}^{\#}$ are the key sensitivity parameters that

determine the size of the hidden bias at site *j* due to omissions of pretreatment or posttreatment confounders or between-site variations in the selection mechanism. Following the derivations in Hong, Qin, and Yang (2018), we are able to show that, in a single site *j*, $\sigma_{1j}^{\#}$ is associated with the degree to which the omissions predict nonresponse under either treatment condition, $\rho_{1j}^{\#}$ is related to the degree to which the omissions predict the outcome within levels of the response status in the experimental group, $\sigma_{*j}^{\#}$ is associated with the degree to which the omissions predict both nonresponse and mediator value assignment under either treatment condition, and $\rho_{*j}^{\#}$ is related to the degree to which the omissions predict the outcome within levels of the response status and mediator in the experimental group.

Bias in Identifying the Population Average NIE

The effect size of bias in identifying the population average NIE can be represented as

$$E\left[c_j\right]E\left[\sigma_{1j}^{\#}\rho_{1j}^{\#} - \sigma_{*j}^{\#}\rho_{*j}^{\#}\right] + cov\left(c_j, \sigma_{1j}^{\#}\rho_{1j}^{\#} - \sigma_{*j}^{\#}\rho_{*j}^{\#}\right)$$
$$= c\left[\sigma_1^{\#}\rho_1^{\#} + cov\left(\sigma_{1j}^{\#}, \rho_{1j}^{\#}\right) - \sigma_*^{\#}\rho_*^{\#} - cov(\sigma_{*j}^{\#}, \rho_{*j}^{\#})\right].$$

where $c = E\left[c_j\right] = E\left[\sigma_{Y|T=1,j}\right]/\sigma_{Y|T=0}$ is the average conversion coefficient, $\sigma_1^{\#} = E[\sigma_{1j}^{\#}]$, $\rho_1^{\#} = E[\rho_{1j}^{\#}]$, $\sigma_*^{\#} = E[\sigma_{*j}^{\#}]$, and $\rho_*^{\#} = E[\rho_{*j}^{\#}]$. The last equation holds under the assumption that $cov\left(c_j, \sigma_{1j}^{\#}\rho_{1j}^{\#} - \sigma_{*j}^{\#}\rho_{*j}^{\#}\right) = 0$. It seems reasonable, in general, to assume that the conversion coefficient c_j is independent of $\sigma_{1j}^{\#}\rho_{1j}^{\#}$ and $\sigma_{*j}^{\#}\rho_{*j}^{\#}$ and hence $cov\left(c_j, \sigma_{1j}^{\#}\rho_{1j}^{\#} - \sigma_{*j}^{\#}\rho_{*j}^{\#}\right) = 0$. This assumption would be violated if the confounding effect is greater in sites where the standard deviation of the outcome in the experimental group is greater (or smaller).

Bias in Identifying the Between-Site Variance of NIE

In identifying the between-site variance of NIE denoted by $\sigma^2_{I(1)}$, the bias is the difference in results between only adjusting for **X** and additionally adjusting for the omitted confounders and/or their site-specific increments, that is,

$$var(A_j) - var(A_j^{\#})$$

$$= var(A_j^{\#} + A_j - A_j^{\#}) - var(A_j^{\#})$$

$$= var(A_j^{\#}) + var(A_j - A_j^{\#}) + 2cov(A_j^{\#}, A_j - A_j^{\#}) - var(A_j^{\#})$$

$$= var(A_j - A_j^{\#}) + 2cov(A_j^{\#}, A_j - A_j^{\#}).$$

Suppose that Assumptions 3.3 and 3.4 hold after one makes the additional adjustment. When this is true, $A_j^{\#}$ will identify the site-specific NIE $\beta_j^{(I)}(1)$. Let $b_j^{(I)}(1)$ denote the corresponding effect size. The effect size of bias in identifying the between-site variance of NIE can be written as

$$var\left(c_j(\sigma_{1j}^{\#}\rho_{1j}^{\#} - \sigma_{*j}^{\#}\rho_{*j}^{\#})\right) + 2cov\left(b_j^{(I)}(1), c_j(\sigma_{1j}^{\#}\rho_{1j}^{\#} - \sigma_{*j}^{\#}\rho_{*j}^{\#})\right).$$

This bias involves two new sensitivity parameters:

(i) $var\left(c_j(\sigma_{1j}^{\#}\rho_{1j}^{\#} - \sigma_{*j}^{\#}\rho_{*j}^{\#})\right)$ is the between-site variance of the effect size of bias for site-specific NIE;

(ii) $cov\left(b_j^{(I)}(1), c_j(\sigma_{1j}^{\#}\rho_{1j}^{\#} - \sigma_{*j}^{\#}\rho_{*j}^{\#})\right)$ is the covariance between the effect size of site-specific NIE and the effect size of bias for site-specific NIE.

3.C.2 Sensitivity Analysis for the Other Causal Parameters

Following the same logic as above, I obtain the effect size of the bias in identifying each of the other causal parameters as follows.

The effect size of the bias in identifying the population average NDE $\gamma^{(D)}(0)$ is

$$c\left[\sigma_*^{\#}\rho_*^{\#} + cov\left(\sigma_{*j}^{\#}, \rho_{*j}^{\#}\right)\right] - c'\left[\sigma_0^{\#}\rho_0^{\#} + cov(\sigma_{0j}^{\#}, \rho_{0j}^{\#}\right].$$

in which $c' = E\left[c_j'\right] = E\left[\sigma_{Y|T=0,j}\right]/\sigma_{Y|T=0}$, $\sigma_{0j}^{\#} =$

$\sqrt{var\left(W_{Dij}W_{Tij}W_{Rij} - W_{Dij}W_{Tij}W_{Rij}^{\#}|T_{ij} = 0, S_{ij} = j\right)}$ is the standard deviation of the weight discrepancy in the control group at site j; $\rho_{0j}^{\#} = corr\left(W_{Dij}W_{Tij}W_{Rij} - W_{Dij}W_{Tij}W_{Rij}^{\#}, Y_{ij}|T_{ij} = 0, S_{ij} = j\right)$ is the correlation between the weight discrepancy and the outcome in the same group at the same site for identifying the population mean outcome under the control condition; and $\sigma_0^{\#}$ and $\rho_0^{\#}$ are the corresponding population average sensitivity parameters.

The effect size of the bias in identifying the population average ITT effect on the outcome $\gamma^{(T)}$ is

$$c\left[\sigma_1^{\#}\rho_1^{\#} + cov\left(\sigma_{1j}^{\#}, \rho_{1j}^{\#}\right)\right] - c'\left[\sigma_0^{\#}\rho_0^{\#} + cov(\sigma_{0j}^{\#}, \rho_{0j}^{\#})\right].$$

Clearly, the sum of the bias in identifying the population average NIE and that in identifying the population average NDE is equal to the bias in identifying the population average ITT effect on the outcome.

The effect size of the bias in identifying the population average PIE $\gamma^{(I)}(0)$ is

$$c'\left[\sigma_*^{\#\prime}\rho_*^{\#\prime} + cov\left(\sigma_{*j}^{\#\prime}, \rho_{*j}^{\#\prime}\right) - \sigma_0^{\#}\rho_0^{\#} - cov(\sigma_{0j}^{\#}, \rho_{0j}^{\#})\right].$$

in which $\sigma_{*j}^{\#\prime} = \sqrt{var\left(W_{Dij}W_{Tij}W_{Rij}W_{Mij} - W_{Dij}W_{Tij}W_{Rij}^{\#}W_{Mij}^{\#}|T_{ij} = 0, S_{ij} = j\right)}$ is the standard deviation of the weight discrepancy in the control group at site j; $\rho_{*j}^{\#\prime} =$ $corr\left(W_{Dij}W_{Tij}W_{Rij}W_{Mij} - W_{Dij}W_{Tij}W_{Rij}^{\#}W_{Mij}^{\#}, Y_{ij}|T_{ij} = 0, S_{ij} = j\right)$ is the correlation between the weight discrepancy and the outcome in the same group at the same site for identifying the population mean outcome under the control condition while the mediator taking the value under

the counterfactual experimental condition; and $\sigma_*^{\#\prime}$ and $\rho_*^{\#\prime}$ are the corresponding population average sensitivity parameters.

The effect size of the bias in identifying the population average natural treatment-by-mediator interaction effect $\gamma^{(T\times M)}$ is

$$c\left[\sigma_1^{\#}\rho_1^{\#} + cov\left(\sigma_{1j}^{\#}, \rho_{1j}^{\#}\right) - \sigma_*^{\#}\rho_*^{\#} - cov\left(\sigma_{*j}^{\#}, \rho_{*j}^{\#}\right)\right]$$
$$- c'\left[\sigma_*^{\#\prime}\rho_*^{\#\prime} + cov\left(\sigma_{*j}^{\#\prime}, \rho_{*j}^{\#\prime}\right) - \sigma_0^{\#}\rho_0^{\#} - cov(\sigma_{0j}^{\#}, \rho_{0j}^{\#})\right].$$

The effect size of bias in identifying the between-site variance of NDE is

$$var\left(c_j\sigma_{*j}^{\#}\rho_{*j}^{\#} - c_j'\sigma_{0j}^{\#}\rho_{0j}^{\#}\right) + 2cov\left(b_j^{(D)}(0), c_j\sigma_{*j}^{\#}\rho_{*j}^{\#} - c_j'\sigma_{0j}^{\#}\rho_{0j}^{\#}\right),$$

where $b_j^{(D)}(0)$ denotes the effect size of NDE at site *j*.

To derive the effect size of bias in identifying the covariance between site-specific NIE and NDE, I use B_j to denote the original identification result of NDE and $B_j^{\#}$ as a general form of the adjusted identification result of NDE. Therefore, in identifying the covariance between site-specific NIE and NDE, the bias is

$$cov\left(A_j, B_j\right) - cov\left(A_j^{\#}, B_j^{\#}\right)$$
$$= cov\left(A_j^{\#} + A_j - A_j^{\#}, B_j^{\#} + B_j - B_j^{\#}\right) - cov\left(A_j^{\#}, B_j^{\#}\right)$$
$$= cov\left(A_j^{\#}, B_j - B_j^{\#}\right) + cov\left(A_j - A_j^{\#}, B_j^{\#}\right) + cov\left(A_j - A_j^{\#}, B_j - B_j^{\#}\right)$$
$$= cov\left(b_j^{(I)}(1), c_j\sigma_{*j}^{\#}\rho_{*j}^{\#} - c_j'\sigma_{0j}^{\#}\rho_{0j}^{\#}\right) + cov\left(b_j^{(D)}(0), c_j\left(\sigma_{1j}^{\#}\rho_{1j}^{\#} - \sigma_{*j}^{\#}\rho_{*j}^{\#}\right)\right)$$
$$+ cov\left(c_j\sigma_{*j}^{\#}\rho_{*j}^{\#} - c_j'\sigma_{0j}^{\#}\rho_{0j}^{\#}, c_j\left(\sigma_{1j}^{\#}\rho_{1j}^{\#} - \sigma_{*j}^{\#}\rho_{*j}^{\#}\right)\right).$$

The effect size of bias in identifying the between-site variance of ITT is

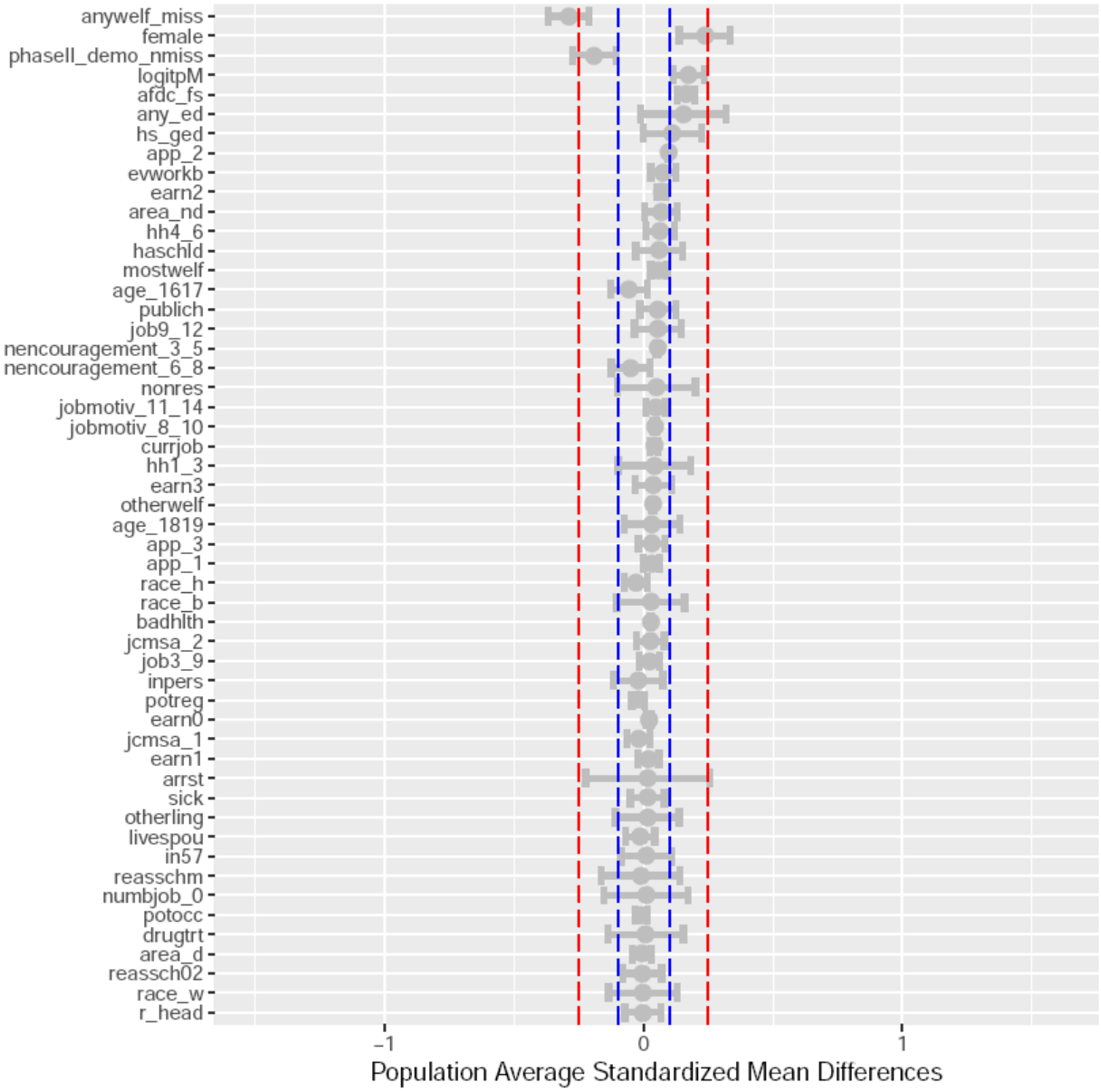
$$var\left(c_j\sigma_{1j}^{\#}\rho_{1j}^{\#} - c_j'\sigma_{0j}^{\#}\rho_{0j}^{\#}\right) + 2cov\left(b_j^{(T)}, c_j\sigma_{1j}^{\#}\rho_{1j}^{\#} - c_j'\sigma_{0j}^{\#}\rho_{0j}^{\#}\right),$$

where $b_j^{(T)}$ denotes the effect size of ITT at site *j*.

APPENDIX 3.D Balance Checking Results

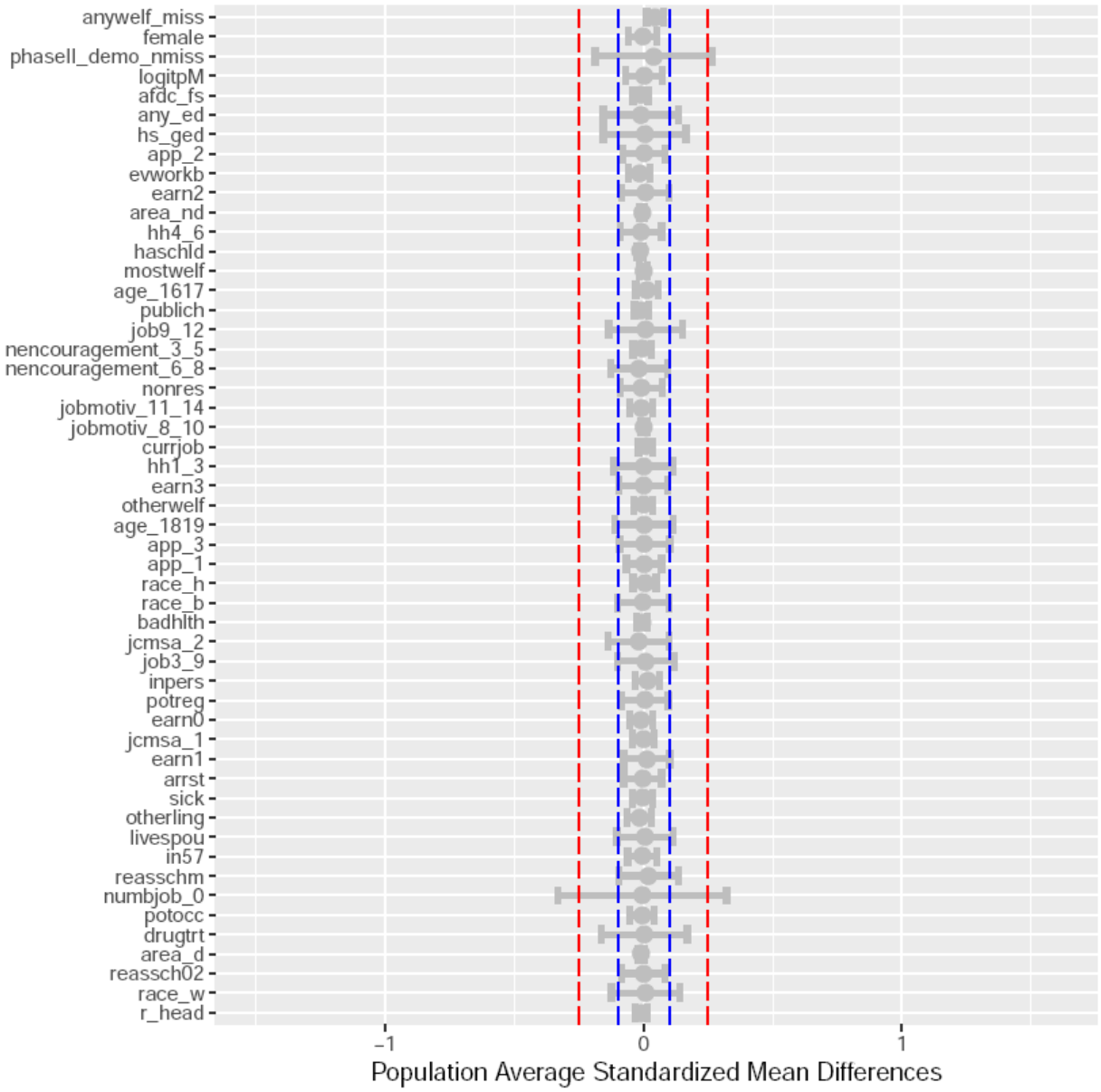
As a supplement to Section 3.3.3, this appendix displays balance checking results before and after propensity score adjustments.

Figure 3.D.1 Imbalance between Response Levels before Weighting in the Job Corps Group



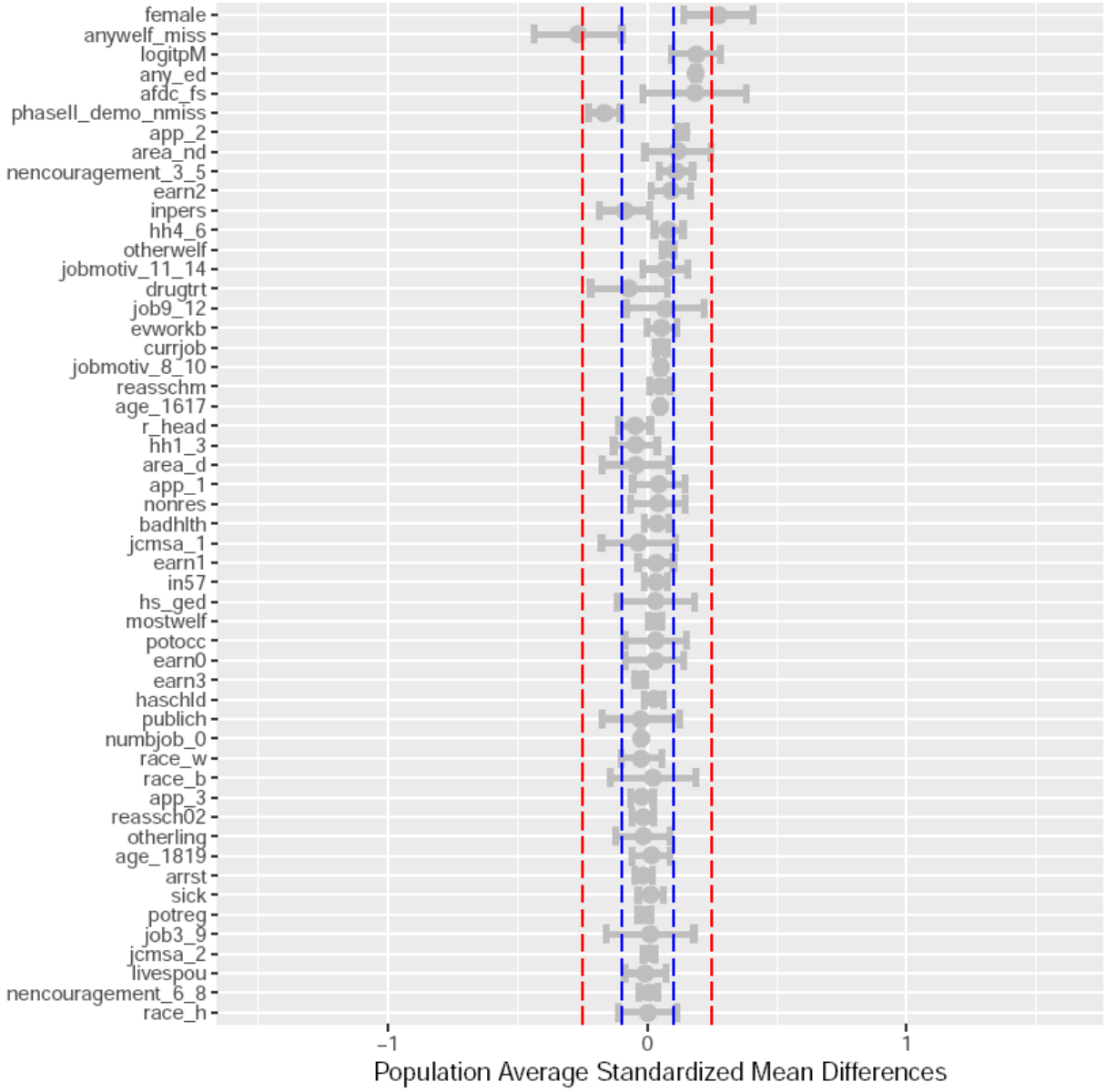
Note: logitpR is the logit of the propensity score of the response status. Each grey dot indicates the overall mean difference in the corresponding covariate between response levels in the Job Corps group, divided by the pooled standard deviation of the covariate in the Job Corps group. Each grey interval indicates the 95% plausible value range of the site-specific mean differences. The red dotted lines indicate the threshold of ±0.25. The blue dotted lines indicate the threshold of ±0.1.

Figure 3.D.2 Imbalance between Response Levels after Weighting in the Job Corps Group



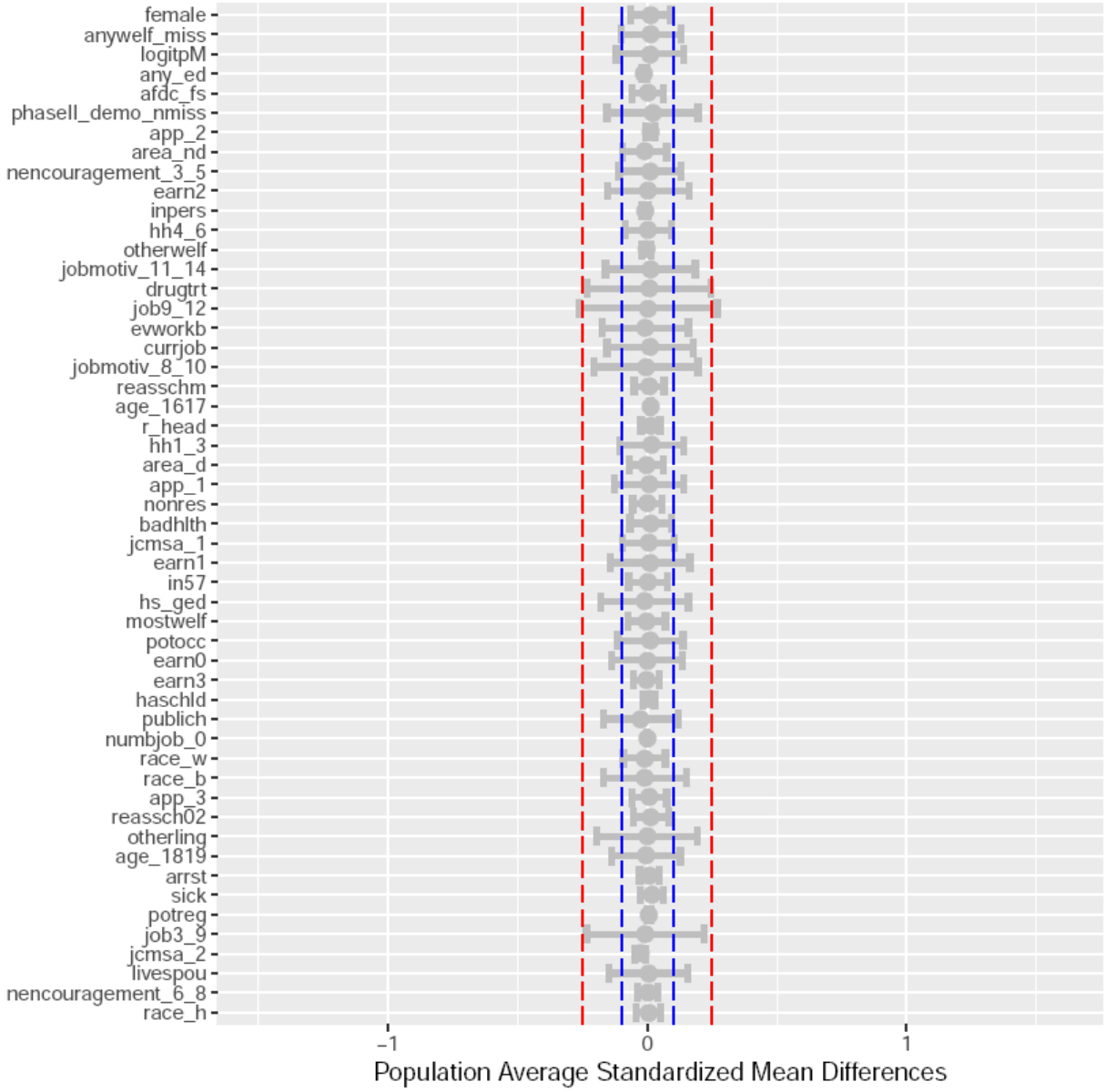
Note: logitpR is the logit of the propensity score of the response status. Each grey dot indicates the overall weighted mean difference in the corresponding covariate between response levels in the Job Corps group, divided by the pooled standard deviation of the covariate in the Job Corps group. Each grey interval indicates the 95% plausible value range of the site-specific weighted mean differences. The red dotted lines indicate the threshold of ±0.25. The blue dotted lines indicate the threshold of ±0.1.

Figure 3.D.3 Imbalance between Response Levels before Weighting in the Control Group



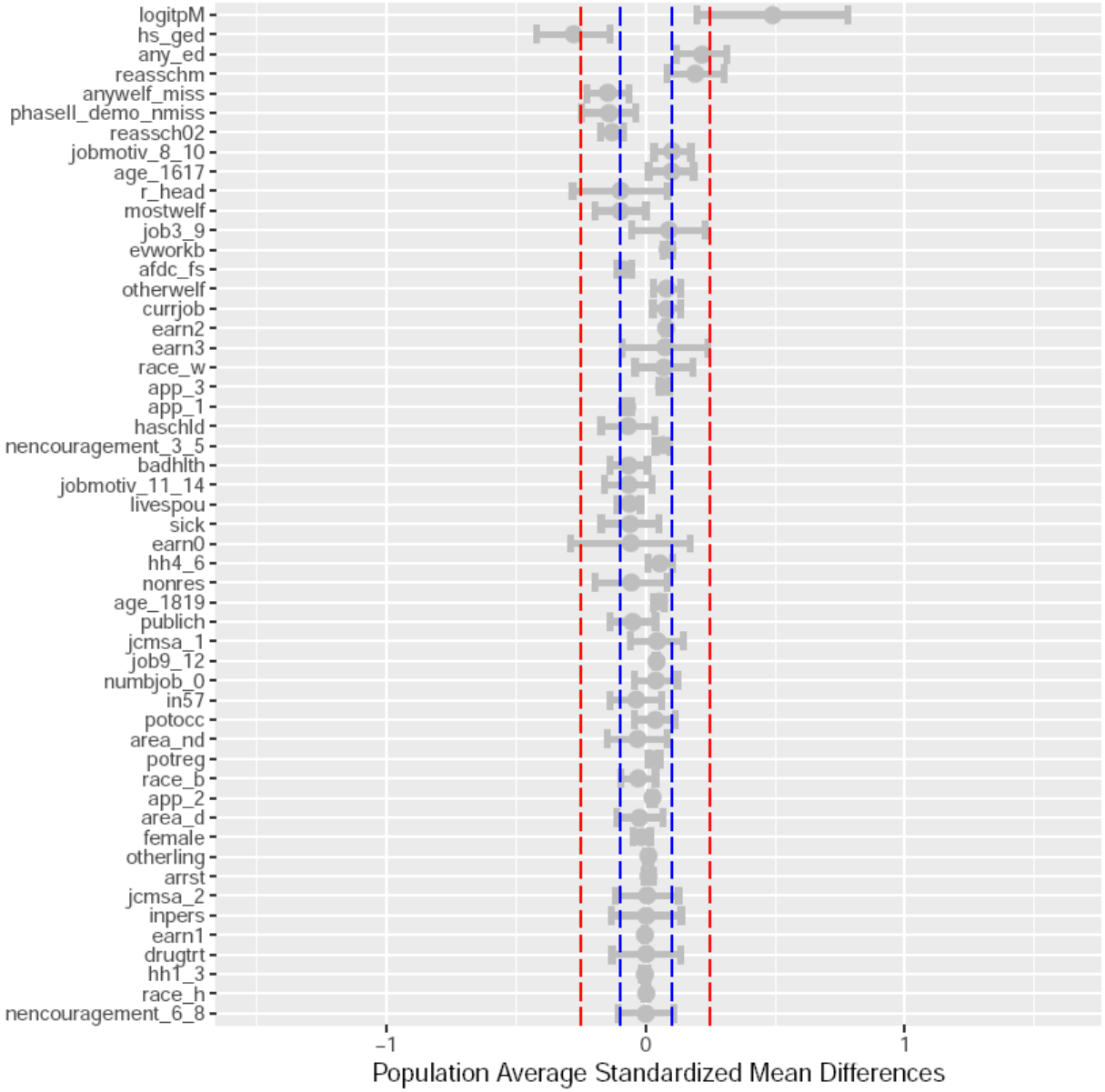
Note: logitpR is the logit of the propensity score of the response status. Each grey dot indicates the overall mean difference in the corresponding covariate between response levels in the control group, divided by the pooled standard deviation of the covariate in the control group. Each grey interval indicates the 95% plausible value range of the site-specific mean differences. The red dotted lines indicate the threshold of ±0.25. The blue dotted lines indicate the threshold of ±0.1.

Figure 3.D.4 Imbalance between Response Levels after Weighting in the Control Group



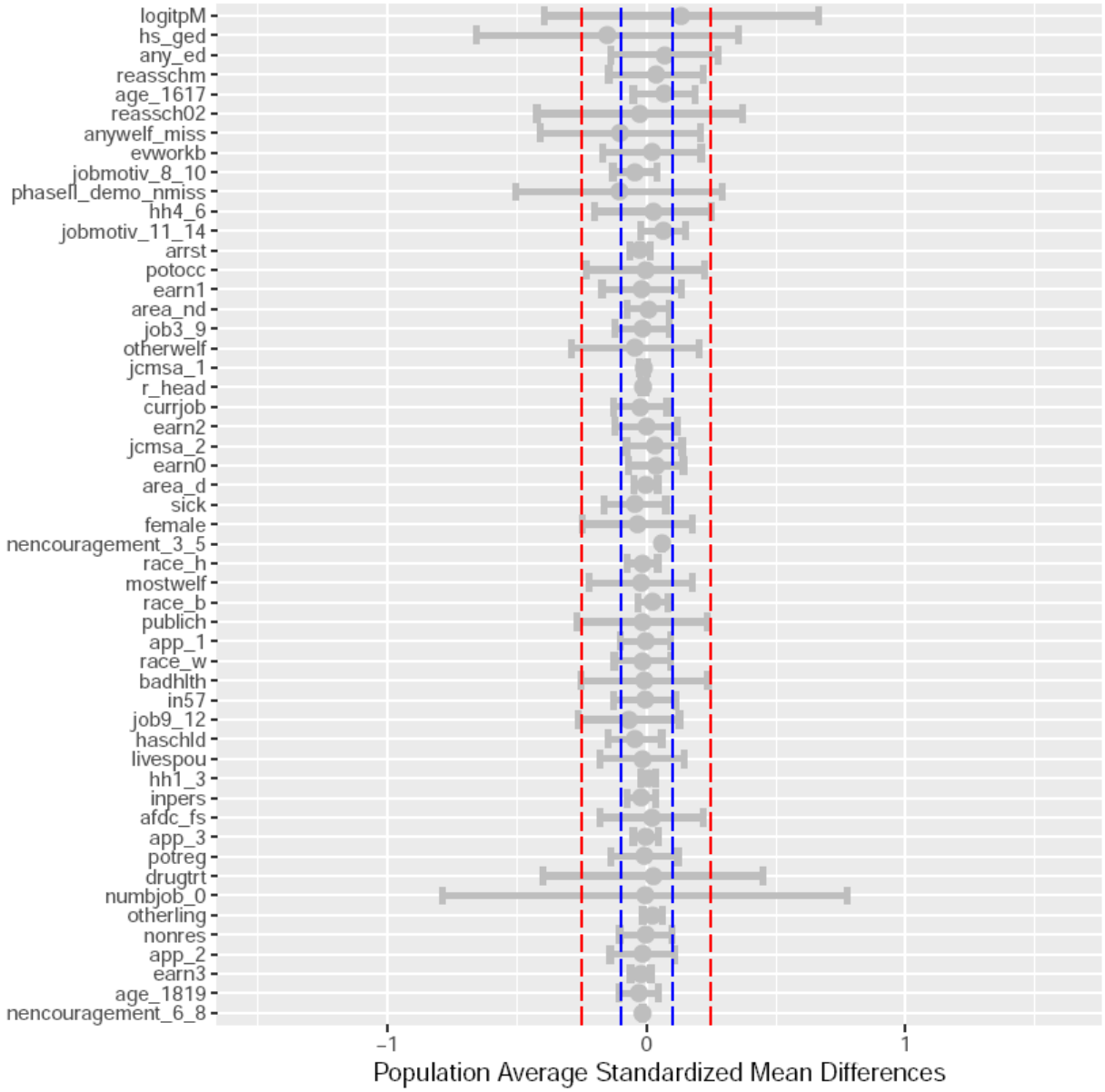
Note: logitpR is the logit of the propensity score of the response status. Each grey dot indicates the overall weighted mean difference in the corresponding covariate between response levels in the control group, divided by the pooled standard deviation of the covariate in the control group. Each grey interval indicates the 95% plausible value range of the site-specific weighted mean differences. The red dotted lines indicate the threshold of ±0.25. The blue dotted lines indicate the threshold of ±0.1.

Figure 3.D.5 Imbalance between Mediator Levels before Weighting in the Job Corps Group



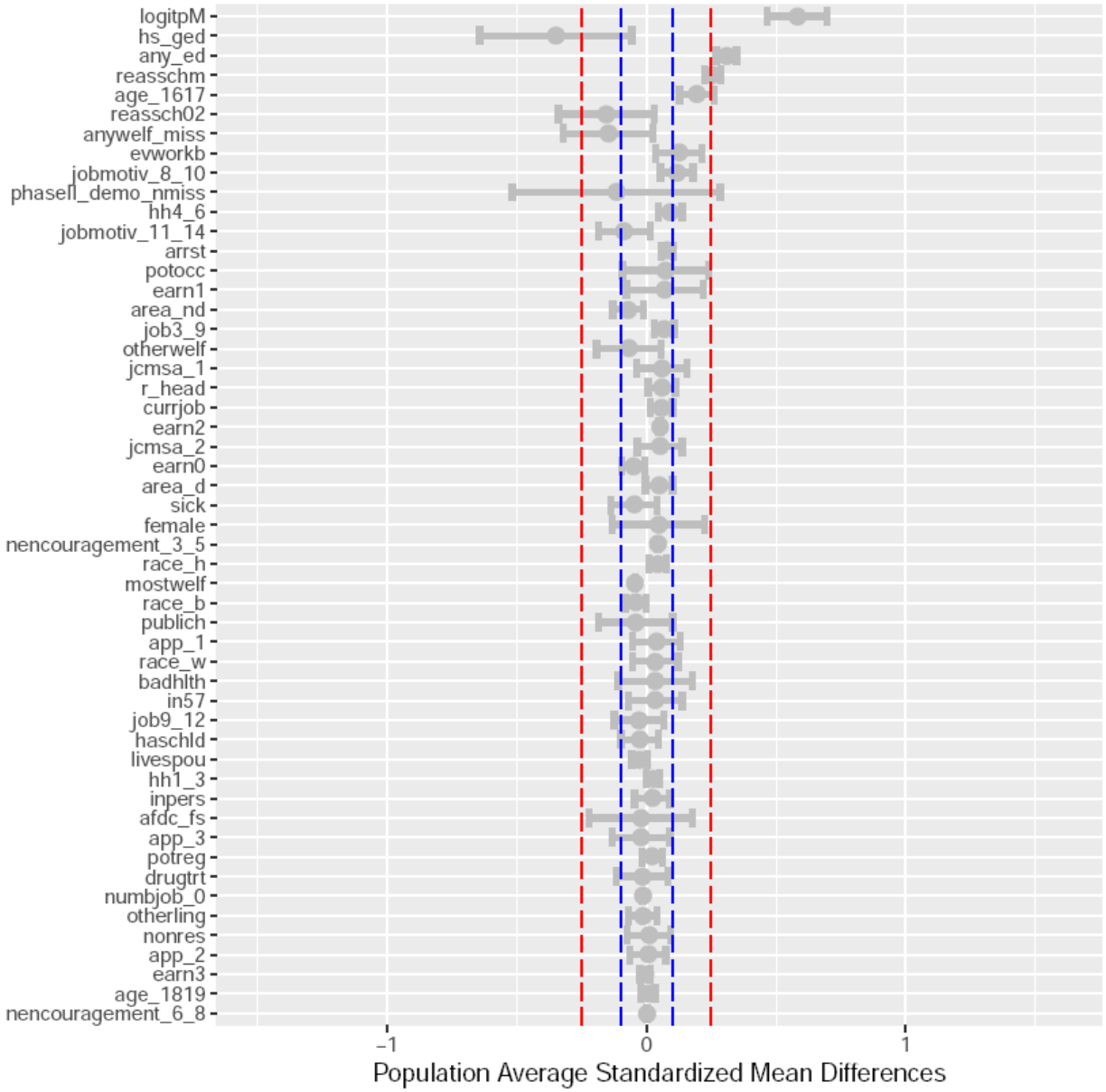
Note: logitpM is the logit of the propensity score of the mediator. Each grey dot indicates the overall mean difference in the corresponding covariate between mediator levels in the Job Corps group, divided by the pooled standard deviation of the covariate in the Job Corps group. Each grey interval indicates the 95% plausible value range of the site-specific mean differences. The red dotted lines indicate the threshold of ±0.25. The blue dotted lines indicate the threshold of ±0.1.

Figure 3.D.6 Imbalance between Mediator Levels after Weighting in the Job Corps Group



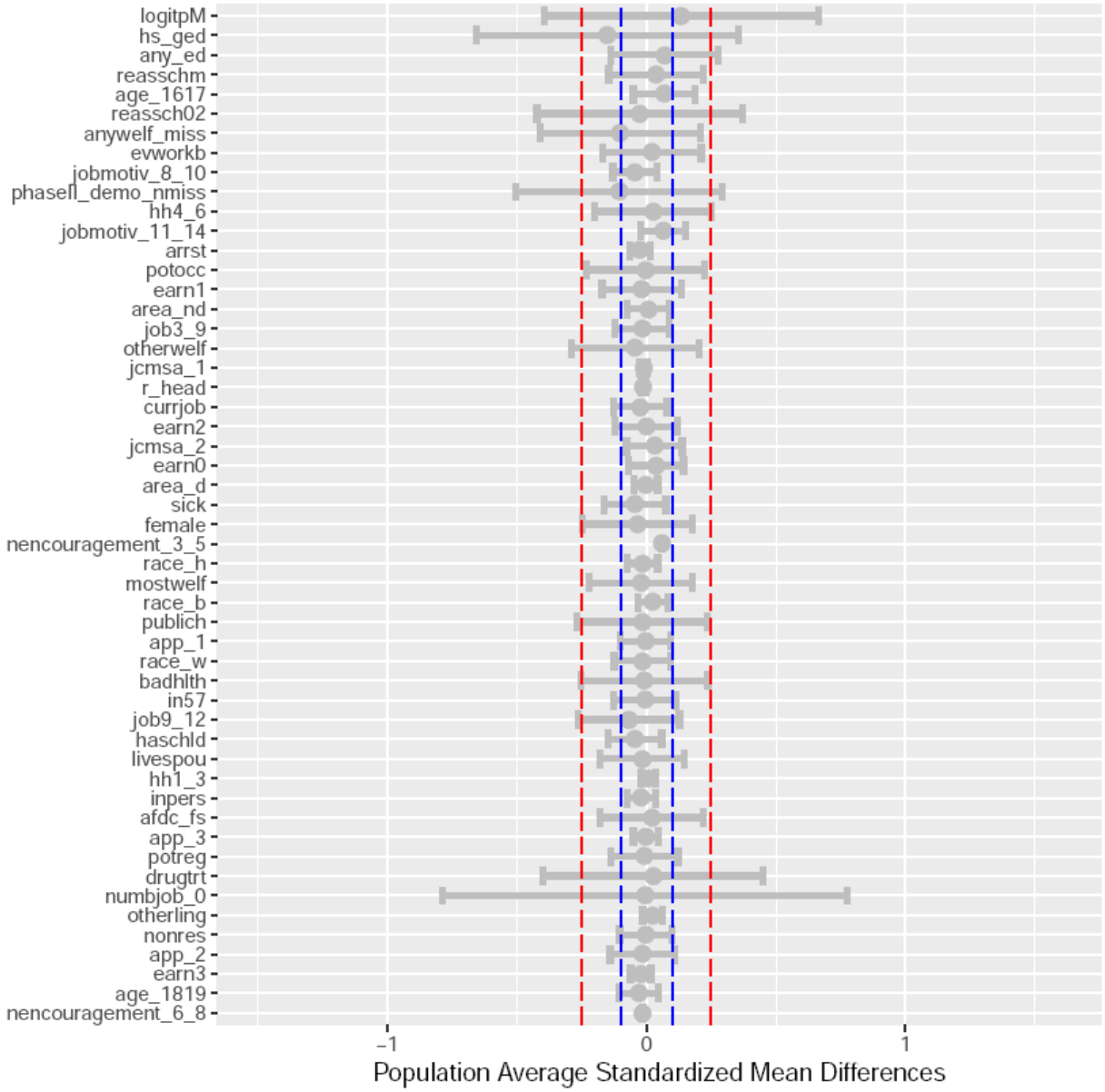
Note: logitpM is the logit of the propensity score of the mediator. Each grey dot indicates the overall weighted mean difference in the corresponding covariate between mediator levels in the Job Corps group, divided by the pooled standard deviation of the covariate in the Job Corps group. Each grey interval indicates the 95% plausible value range of the site-specific weighted mean differences. The red dotted lines indicate the threshold of ±0.25. The blue dotted lines indicate the threshold of ±0.1.

Figure 3.D.7 Imbalance between Mediator Levels before Weighting in the Control Group



Note: logitpM is the logit of the propensity score of the mediator. Each grey dot indicates the overall mean difference in the corresponding covariate between mediator levels in the control group, divided by the pooled standard deviation of the covariate in the control group. Each grey interval indicates the 95% plausible value range of the site-specific mean differences. The red dotted lines indicate the threshold of ±0.25. The blue dotted lines indicate the threshold of ±0.1.

Figure 3.D.8 Imbalance between Mediator Levels after Weighting in the Control Group



Note: logitpM is the logit of the propensity score of the mediator. Each grey dot indicates the overall weighted mean difference in the corresponding covariate between mediator levels in the control group, divided by the pooled standard deviation of the covariate in the control group. Each grey interval indicates the 95% plausible value range of the site-specific weighted mean differences. The red dotted lines indicate the threshold of ±0.25. The blue dotted lines indicate the threshold of ±0.1.

APPENDIX 4.A

Proof of Theorems 4.1 and 4.2

This Appendix provides a supplement to Section 4.2. The derivation below proves that, under Assumptions 4.1 $\sim$ 4.5, the expectation of each potential outcome at site j, $E\left[Y(t, M_V(t'), M_E(t''))|S = j\right]$, for $t, t', t'' = 0, 1$, can be identified with a weighted average of the observed outcome at that site. Let $\mathbf{X} = \{\mathbf{X}_D \cup \mathbf{X}_T \cup \mathbf{X}_R \cup \mathbf{X}_V \cup \mathbf{X}_E\}$ be the union of all the observed pretreatment confounders. To simplify notations, I drop the subscript ij of each variable.

$$
\begin{aligned}
&E[Y(t, M_V(t'), M_E(t''))|S = j] \\
&= E\{E[Y(t, M_V(t'), M_E(t''))|\mathbf{X} = \mathbf{x}, S = j]\} \\
&= \int_{\mathbf{x}} \int_{m_V} \int_{m_E} \int_{y} y \times f(Y(t, m_V, m_E) = y|M_V(t') = m_V, M_E(t'') = m_E, \mathbf{X} = \mathbf{x}, S = j) \\
&\times Pr(M_V(t') = m_V|M_E(t'') = m_E, \mathbf{X} = \mathbf{x}, S = j) \times Pr(M_E(t'') = m_E|\mathbf{X} = \mathbf{x}, S = j) \\
&\times g(\mathbf{X} = \mathbf{x}|S = j) dy dm_V dm_E d\mathbf{x}.
\end{aligned}
$$

Under Assumption 4.1, $\{Y(t, m_V, m_E), M_V(t), M_E(t)\} \perp\!\!\!\perp D|\mathbf{X}_D = \mathbf{x}_D, S = j$. Because $\mathbf{X}_D \subset \mathbf{X}$, $\{Y(t, m_V, m_E), M_V(t), M_E(t)\} \perp\!\!\!\perp D|\mathbf{X} = \mathbf{x}, S = j$ also holds. Hence, the above equation is equal to

$$
\begin{aligned}
&\int_{\mathbf{x}} \int_{m_V} \int_{m_E} \int_{y} y \times f(Y(t, m_V, m_E) = y|D = 1, M_V(t') = m_V, M_E(t'') = m_E, \mathbf{X} = \mathbf{x}, S = j) \\
&\times Pr(M_V(t') = m_V|D = 1, M_E(t'') = m_E, \mathbf{X} = \mathbf{x}, S = j) \\
&\times Pr(M_E(t'') = m_E|D = 1, \mathbf{X} = \mathbf{x}, S = j) \times g(\mathbf{X} = \mathbf{x}|S = j) dy dm_V dm_E d\mathbf{x}.
\end{aligned}
$$

By Bayes theorem,

$$
g(\mathbf{X} = \mathbf{x}|S = j) = g(\mathbf{X} = \mathbf{x}|D = 1, S = j) \times \frac{Pr(D = 1|S = j)}{Pr(D = 1|\mathbf{X} = \mathbf{x}, S = j)}.
$$

where $0 < Pr(D = 1|\mathbf{X} = \mathbf{x}, S = j) < 1$. When the strongly ignorable sampling mechanism (Assumption 4.1) holds, controlling for $\mathbf{X}_D$ removes sampling selection. Because $\mathbf{X}_D \subset \mathbf{X}$, $Pr(D = 1|\mathbf{X} = \mathbf{x}, S = j) = Pr(D = 1|\mathbf{X}_D = \mathbf{x}_D, S = j)$. Hence, it is equivalanet to assuming that $0 < Pr(D = 1|\mathbf{X}_D = \mathbf{x}_D, S = j) < 1$. This is known as positivity assumption. Let $W_D = \frac{Pr(D=1|S=j)}{Pr(D=1|\mathbf{X}=\mathbf{x},S=j)} = \frac{Pr(D=1|S=j)}{Pr(D=1|\mathbf{X}_D=\mathbf{x}_D,S=j)}$, and thus

$$
\begin{aligned}
&E[Y(t, M_V(t'), M_E(t''))|S = j] \\
&= \int_{\mathbf{x}} \int_{m_V} \int_{m_E} \int_{y} W_D \times y \\
&\times f(Y(t, m_V, m_E) = y|D = 1, M_V(t') = m_V, M_E(t'') = m_E, \mathbf{X} = \mathbf{x}, S = j) \\
&\times Pr(M_V(t') = m_V|D = 1, M_E(t'') = m_E, \mathbf{X} = \mathbf{x}, S = j) \\
&\times Pr(M_E(t'') = m_E|D = 1, \mathbf{X} = \mathbf{x}, S = j) \times g(\mathbf{X} = \mathbf{x}|D = 1, S = j) dy dm_V dm_E d\mathbf{x}.
\end{aligned}
$$

Similarly, under the assumption that $0 < Pr(T = 1|\mathbf{X}_T = \mathbf{x}_T, D = 1, S = j) < 1$, let $W_T = \frac{Pr(T=t|D=1,S=j)}{Pr(T=t|\mathbf{X}_T=\mathbf{x}_T,D=1,S=j)}$. When Assumption 4.2 holds, i.e. $\{Y(t, m_V, m_E), M_V(t), M_E(t)\} \perp\!\!\!\perp T|D = 1, \mathbf{X}_T = \mathbf{x}_T, S = j$, and by Bayes theorem, the above equation is equal to

$$
\begin{aligned}
&\int_{\mathbf{x}} \int_{m_V} \int_{m_E} \int_{y} W_D W_T \times y \\
&\times f(Y(t, m_V, m_E) = y|T = t, D = 1, M_V(t') = m_V, M_E(t'') = m_E, \mathbf{X} = \mathbf{x}, S = j) \\
&\times Pr(M_V(t') = m_V|T = t', D = 1, M_E(t'') = m_E, \mathbf{X} = \mathbf{x}, S = j) \\
&\times Pr(M_E(t'') = m_E|T = t'', D = 1, \mathbf{X} = \mathbf{x}, S = j) \\
&\times g(\mathbf{X} = \mathbf{x}|T = t, D = 1, S = j) dy dm_V dm_E d\mathbf{x}.
\end{aligned}
$$

Under the assumption that $0 < Pr(R = 1|\mathbf{X}_R = \mathbf{x}_R, T = t, D = 1, S = j) < 1$, let $W_R = \frac{Pr(R=1|T=t,D=1,S=j)}{Pr(R=1|\mathbf{X}_R=\mathbf{x}_R,T=t,D=1,S=j)}$. When Assumption 4.3 holds, i.e. $\{Y(t, m_V, m_E), M_V(t), M_E(t)\} \perp\!\!\!\perp R|T = t, D = 1, \mathbf{X}_R = \mathbf{x}_R, S = j$, and by Bayes theorem, the above equation is

equal to

$$\int_{\mathbf{x}}\int_{m_V}\int_{m_E}\int_{y} W_D W_T W_R \times y$$
$$\times f(Y(t, m_V, m_E) = y | R = 1, T = t, D = 1, M_V(t') = m_V, M_E(t'') = m_E, \mathbf{X} = \mathbf{x}, S = j)$$
$$\times Pr(M_V(t') = m_V | R = 1, T = t', D = 1, M_E(t'') = m_E, \mathbf{X} = \mathbf{x}, S = j)$$
$$\times Pr(M_E(t'') = m_E | R = 1, T = t'', D = 1, \mathbf{X} = \mathbf{x}, S = j)$$
$$\times g(\mathbf{X} = \mathbf{x} | R = 1, T = t, D = 1, S = j) dy dm_V dm_E d\mathbf{x}.$$

When $t = t' = t''$, it is easy to obtain the following identification result,

$$E[Y(t, M_V(t), M_E(t)) | S = j] = E[W_D W_T W_R Y | R = 1, T = t, D = 1, S = j].$$

When $t' \neq t$ or $t'' \neq t$ and if Assumption 4.5 holds, i.e. $M_V(t) \perp\!\!\!\perp M_E(t')\} | R = 1, T = t, D = 1, \mathbf{X}_V = \mathbf{x}_V, \mathbf{X}_E = \mathbf{x}_E, S = j$, because $\{\mathbf{X}_V, \mathbf{X}_E\} \subset \mathbf{X}$, $M_V(t) \perp\!\!\!\perp M_E(t')\} | R = 1, T = t, D = 1, \mathbf{X} = \mathbf{x}, S = j$ also holds. Hence,

$$E[Y(t, M_V(t'), M_E(t'')) | S = j]$$
$$= \int_{\mathbf{x}}\int_{m_V}\int_{m_E}\int_{y} W_D W_T W_R \times y$$
$$\times f(Y(t, m_V, m_E) = y | R = 1, T = t, D = 1, M_V(t') = m_V, M_E(t'') = m_E, \mathbf{X} = \mathbf{x}, S = j)$$
$$\times Pr(M_V(t') = m_V | R = 1, T = t', D = 1, \mathbf{X} = \mathbf{x}, S = j)$$
$$\times Pr(M_E(t'') = m_E | R = 1, T = t'', D = 1, \mathbf{X} = \mathbf{x}, S = j)$$
$$\times g(\mathbf{X} = \mathbf{x} | R = 1, T = t, D = 1, S = j) dy dm_V dm_E d\mathbf{x}.$$

If Assumption 4.4 holds, because $\mathbf{X}_V \subset \mathbf{X}$ and $\mathbf{X}_E \subset \mathbf{X}$, $Y(t, m_V, m_E) \perp\!\!\!\perp \{M_V(t), M_V(t')\} | R = 1, T = t, D = 1, \mathbf{X} = \mathbf{x}, S = j$ and $Y(t, m_V, m_E) \perp\!\!\!\perp \{M_E(t), M_E(t')\} | R = 1, T = t, D =$

$1, \mathbf{X} = \mathbf{x}, S = j$ also hold. Hence,

$$
\begin{aligned}
&E[Y(t, M_V(t'), M_E(t''))|S = j] \\
&= \int_{\mathbf{x}} \int_{m_V} \int_{m_E} \int_{y} W_D W_T W_R \times y \\
&\times f(Y(t, m_V, m_E) = y|R = 1, T = t, D = 1, M_V(t) = m_V, M_E(t) = m_E, \mathbf{X} = \mathbf{x}, S = j) \\
&\times Pr(M_V(t') = m_V|R = 1, T = t', D = 1, \mathbf{X} = \mathbf{x}, S = j) \\
&\times Pr(M_E(t'') = m_E|R = 1, T = t'', D = 1, \mathbf{X} = \mathbf{x}, S = j) \\
&\times g(\mathbf{X} = \mathbf{x}|R = 1, T = t, D = 1, S = j) dy dm_V dm_E d\mathbf{x}.
\end{aligned}
$$

let $W_{Vt'} = \frac{Pr(M_V(t')=m_V|\mathbf{X}=\mathbf{x},R=1,T=t',D=1,S=j)}{Pr(M_V(t)=m_V|\mathbf{X}=\mathbf{x},R=1,T=t,D=1,S=j)} = \frac{Pr(M_V=m_V|\mathbf{X}=\mathbf{x},R=1,T=t',D=1,S=j)}{Pr(M_V=m_V|\mathbf{X}=\mathbf{x},R=1,T=t,D=1,S=j)}$ and $W_{Et''} = \frac{Pr(M_E(t'')=m_E|\mathbf{X}=\mathbf{x},R=1,T=t'',D=1,S=j)}{Pr(M_E(t)=m_E|\mathbf{X}=\mathbf{x},R=1,T=t,D=1,S=j)} = \frac{Pr(M_E=m_E|\mathbf{X}=\mathbf{x},R=1,T=t'',D=1,S=j)}{Pr(M_E=m_E|\mathbf{X}=\mathbf{x},R=1,T=t,D=1,S=j)}$, because $M_V(t) = M_V$ and $M_E(t) = M_E$ when $T = t$ and, similarly, $M_V(t') = M_V$ when $T = t'$, and $M_E(t'') = M_E$ when $T = t''$. This is based on the assumption that $0 < Pr(M_V = m_V|\mathbf{X} = \mathbf{x}, R = 1, T = t, D = 1, S = j) < 1$ and $0 < Pr(M_E = m_E|\mathbf{X} = \mathbf{x}, R = 1, T = t, D = 1, S = j) < 1$. When the strongly ignorable mediator selection mechanism (Assumption 4.4) holds, controlling for $\mathbf{X}_V$ removes selection of mediator M_V and controlling for $\mathbf{X}_E$ removes selection of mediator M_E. Because $\mathbf{X}_V \subset \mathbf{X}$ and $\mathbf{X}_E \subset \mathbf{X}$, the positivity assumptions can be simplified as $0 < Pr(M_V = m_V|\mathbf{X}_V = \mathbf{x}_V, R = 1, T = t, D = 1, S = j) < 1$ and $0 < Pr(M_E = m_E|\mathbf{X}_E = \mathbf{x}_E, R = 1, T = t, D = 1, S = j) < 1$, and the weights are equal to $W_{Vt'} = \frac{Pr(M_V=m_V|\mathbf{X}_V=\mathbf{x}_V,R=1,T=t',D=1,S=j)}{Pr(M_V=m_V|\mathbf{X}_V=\mathbf{x}_V,R=1,T=t,D=1,S=j)}$ and $W_{Et''} = \frac{Pr(M_E=m_E|\mathbf{X}_E=\mathbf{x}_E,R=1,T=t'',D=1,S=j)}{Pr(M_E=m_E|\mathbf{X}_E=\mathbf{x}_E,R=1,T=t,D=1,S=j)}$. Then

$$
\begin{aligned}
&E[Y(t, M_V(t'), M_E(t''))|S = j] \\
&= \int_{\mathbf{x}} \int_{m_V} \int_{m_E} \int_{y} W_D W_T W_R W_{Vt'} W_{Et''} \times y \\
&\times f(Y(t, m_V, m_E) = y|R = 1, T = t, D = 1, M_V(t) = m_V, M_E(t) = m_E, \mathbf{X} = \mathbf{x}, S = j) \\
&\times Pr(M_V(t) = m_V|R = 1, T = t, D = 1, \mathbf{X} = \mathbf{x}, S = j) \\
&\times Pr(M_E(t) = m_E|R = 1, T = t, D = 1, \mathbf{X} = \mathbf{x}, S = j) \\
&\times g(\mathbf{X} = \mathbf{x}|R = 1, T = t, D = 1, S = j) dy dm_V dm_E d\mathbf{x} \\
&= E[W_D W_T W_R W_{Vt'} W_{Et''} Y|R = 1, T = t, D = 1, S = j].
\end{aligned}
$$

If $t' = t$, then $W_{Vt'} = 1$; if $t'' = t$, then $W_{Et''} = 1$.

REFERENCES

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

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American sociological review*, 37-47.

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

Baron, R. M., & 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.

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

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

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 rmpw analysis. *Statistics in Medicine, 37*(8), 1304-1324.

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

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

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 817-842.

Blundell, R., Dearden, L., Meghir, C., & Sianesi, B. (1999). Human capital investment: the returns from education and training to the individual, the firm and the economy. *Fiscal studies*, *20*(1), 1-23.

Bollen, K. A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological methodology*, 37-69.

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, *104*(3), 396.

Bullock, J.G., Green, D.P., & Ha, S.E. (2010). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology* , 98, 550–558.

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.

Coffman, D. L., & Zhong, W. (2012). Assessing Mediation Using Marginal Structural Models in the Presence of Confounding and Moderation. *Psychological Methods*, *17*(4), 642-664.

Daniel, R., De Stavola, B., Cousens, S., & Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, *71* (1), 1–14.

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

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

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

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

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

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., & Neumann, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: The case of Job Corps. *The Review of Economics and Statistics*, *94*(1), 153-171.

Flores-Lagunes, A., Gonzalez, A., & Neumann, T. C. (2010). Learning but not earning? The impact of Job Corps training on Hispanic youth. *Economic Inquiry*, 48(3), 651-667.

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). Chichester, England: John Wiley.

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

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.

Heckman, J. J., Hsse, J., & Rubinstein, Y. (2000). The GED is a mixed signal: The effect of cognitive and non-cognitive skills on human capital and labor market outcomes. *University of Chicago xerox*.

Heckman, J. J., & Robb Jr, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of econometrics*, *30*(1-2), 239-267.

Heckman, J. J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. *American Economic Review*, *91*(2), 145-149.

Heckman, J. J., Smith, J., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, *64*(4), 487-535.

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

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

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Holland, P. W. (1988). Causal Inference, Path Analysis, and Recursive Structural Equations Models. *Sociological Methodology*, 449-484.

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

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

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

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

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

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

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

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

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

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

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

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

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods*, 13(4), 314-336.

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

Jöreskog, K.G. (1970). A general method for analysis of covariance structures. *Biometrika* , 57, 239–251.

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

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 523–539.

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

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

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

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

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

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

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

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.

MacKinnon, D.P. (2008). *Introduction to Statistical Mediation Analysis* , Erlbaum, Mahwah, NJ.

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

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

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

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

Olsen, R. B. (2017). Evaluating educational interventions when impacts may vary across sites. *Journal of Research on Educational Effectiveness*, *10*(4), 907-911.

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

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

Pouncy, H. (2000). New directions in job training strategies for the disadvantaged. In S. Danziger and J. waldfogel (eds.), *Securing the future: Investing in children from birth to college*. (pp.264-282). New York: Russel Sage Foundation.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007) Addressing moderated mediation hypotheses: Theory, Methods, and Prescriptions. *Multivariate Behavioral Research*, *42*, 185–227.

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

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

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

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

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

Reardon, S. F., Unlu, F., Zhu, P., & Bloom, H. S. (2014). Bias and bias correction in multisite instrumental variables analysis of heterogeneous mediator effects. *Journal of Educational and Behavioral Statistics, 39*(1), 53-86.

Robins, J.M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials*. New York: Springer.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In P. J. Green, N. L. Hjort, & S. Richardson (Eds.), *Highly structured stochastic systems* (pp. 70–81). New York, NY: Oxford University Press.

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

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

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, 34–58.

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

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

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

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, *13* (4), 279.

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., & 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., & McConnell, S. (2008) Does Job Corps Work? Impact Findings from the National Job Corps Study. *The American Economic Review*, 98(5), 1864-1886.

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

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.

Sobel, M.E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics* , 33(2), 230–251.

Spencer, M. B. (2006). Phenomenology and ecological systems theory: Development of diverse groups. In R. M. Lerner & W. Damon (Eds.), *Handbook of child psychology, vol. 1: Theoretical models of human development, 6th ed.* (pp. 829-893). New York: Wiley Publishers.

Spencer, M. B. (2008). Lessons learned and opportunities ignored since Brown v. Board of Education: Youth development and the myth of a color-blind society. *Educational Researcher*, *37*(5), 253.

Spencer, M. B., & Swanson, D. P. (2013). Opportunities and challenges to the development of healthy children and youth living in diverse communities. *Development and Psychopathology, 25*, 1551-1566.

Spencer, M. B., Swanson, D. P., & Harpalani, V. (2015). Development of the self. *Handbook of child psychology and developmental science*.

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

Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas* (Vol. 39). Englewood Cliffs, NJ: Prentice-Hall.

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

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

Valeri, L., & 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.

van der Laan, M. J., & Petersen, M. L. (2008). Direct effect models. *The international journal of biostatistics*, *4*(1), 1-27.

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology, 20*, 18–26)

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

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

VanderWeele, T.J. (2013). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology, 24*, 224-232.

VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

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

VanderWeele, T. J., & Vansteelandt (2009) Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2, 457-468.

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

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

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

Zimmermann, K. F., Biavaschi, C., Eichhorst, W., Giulietti, C., Kendzia, M. J., Muravyev, A., . . . others (2013). Youth unemployment and vocational training. *Foundations and Trends* R *in Microeconomics, 9* (1–2), 1–157.