Parametric and Non-Parametric Weighting Methods for Mediation Analysis: An Application to the National Evaluation of Welfare-to-Work Strategies

Guanglei Hong, Jonah Deutsch, Heather Hill
University of Chicago, 5736 S. Woodlawn Ave., Chicago, IL 60637

Abstract

This study analyzes the impact of a welfare-to-work program on maternal depression mediated by employment experience in the context of a randomized experiment. We develop parametric and non-parametric weighting procedures as alternative strategies for removing selection bias in estimating the controlled direct effect of the policy and that of the mediator. Through ratio-of-mediator-probability weighting, we further decompose the total effect into a natural direct effect and a natural indirect effect. Simulation results suggest that the parametric and non-parametric weighting methods both have satisfactory performance when the propensity score models are correctly specified. The weighting results replicate those from path analysis and the Instrumental Variable method when the assumption of no treatment-by-mediator interaction and the exclusion restriction hold. The weighting methods show great advantages when these assumptions do not hold. Additionally, the weighting approach is exempted from assuming the distribution of the outcome, the distribution of the mediator, and the functional form of the outcome model. Hence it provides a viable alternative to the model-based approaches.

Keywords

Causal inference; direct effect; indirect effect; inverse-probability-of-treatment weighting; marginal mean weighting through stratification; mediation; moderation; ratio-of-mediator-probability weighting

1. Introduction

In the late-1990s, the decades-long welfare cash assistance program (i.e., Aid to Families with Dependent Children, AFDC) was replaced by a new program (i.e., Temporary Assistance for Needy Families, TANF) which required labor force participation. The National Evaluation of Welfare-to-Work Strategies (NEWWS) among other studies randomly assigned welfare recipients to work mandates and found that treatment group members had higher employment rates and earnings than comparison group members.

However, the political debate focused on the policy impacts on welfare recipients psychologically as well as financially. Low-income single mothers with young children typically face multiple barriers to securing employment and are disproportionately likely to work in jobs that offer erratic hours, low pay, and little stimulation. In this context, having a job may contribute minimal to psychological well-being when employment is not required; a successful experience in meeting the challenges of the work mandate under the welfare-to-work policies may possibly reduce depressive symptoms; while the failure to find work when it is required for welfare receipt may potentially heighten depressive symptoms. Therefore, questions about the mediation mechanism would be relevant even if the average policy effect on maternal depression is essentially zero.

The current study is focused on identifying the mediating role of policy-induced changes in employment in explaining the impact of a welfare-to-work program on maternal depression. Using data from the NEWWS experiment in Riverside, California, we develops analytic procedures that use weighting to estimate (1) the average effect of the welfare-to-work policy on maternal depression if one remains unemployed, (2) the average effect of gaining success in the job market under the new policy, (3) the average effect of the welfare-to-work policy on maternal depression if the policy counterfactually failed to change one's employment experience, and (4) the average amount of change in maternal depression attributable to policy-induced changes in employment.

This paper contributes to the literature by introducing parametric weighting and non-parametric weighting as alternative procedures for decomposing the controlled direct effects. Additionally, we use ratio-of-mediator-probability weighting to decompose the total effect into the natural direct effect and the natural indirect effect. In doing so, we clarify the assumptions required for identifying the controlled direct effects and those required for identifying the natural direct and indirect effects. We also assess the performance of these weighting procedures relative to conventional, model-based approaches such as path analysis and the instrumental variable (IV) method. We focus on investigating the mediator measured on a binary scale—never employed vs. ever employed during the two years after randomization—but also extend the methods to investigations of moderated effects and multi-valued mediators.

1.1 Sample and Data

Data are drawn from NEWWS Riverside site (Hamilton, 2002). Our sample includes participants with a child aged 3 to 5 who were assigned at random to either a Labor Force Attachment (LFA) program or a control condition. The LFA program provided services emphasizing immediate engagement in job search activities and required participation in work or work-related activities as a condition of welfare receipt. Participants assigned to the control condition were eligible for welfare without being subject to the employment requirement. The analytic sample includes 208 LFA group members and 486 control group members.

Unemployment Insurance records maintained by the State of California provide quarterly administrative data on *employment* for each study participant. All participants were surveyed shortly before the randomization and again at the two-year follow-up. The self-administered questionnaire at the two-year follow-up included twelve items measuring *depressive symptoms* (e.g., I could not get going) on a frequency scale from 1 (rarely, less than 1 day during the past week) to 4 (most of the time, 5-7 days during the past week). The sum score ranged from 0 to 34 with a mean equal to 7.49 and a standard deviation equal to 7.74. About 17% of the participants reported no depressive symptoms.

The baseline survey provided rich information about participant background. These include (a) maternal psychological well-being; (b) personal history of employment and history of welfare dependence; (c) human capital, employment status, earnings, and income; (d) personal attitudes toward employment including preference to work, willingness to accept a low-wage job, ashamed to be on welfare; (e) perceived social support and perceived barriers to work; (f) practical support and barriers to work including childcare arrangement and extra family burden; (g) household composition including number and age of children and marriage status; (h) ever been a teen mother; (i) public housing residence, and residential mobility; (j) demographic features.

1.2 Methodological Challenges and Weighting as an Innovative Solution

Even though participants in this study were assigned at random to either the LFA program or the control condition, a participant's employment experience during the two years after randomization is subject to the influence of stressors and personal resources mentioned above that may also relate to depressive symptoms. Researchers in the past have developed alternative analytic strategies for mediation analysis. The most prevalent are path analysis and the instrumental variable (IV) method, both involving specifying models for the mediator and the outcome. These conventional methods have received criticisms for being constrained by strong and often implausible assumptions such as no treatment-by-mediator interaction and the exclusion restriction. The treatment-by-mediator interaction and the natural direct effect are among the key parameters of interest in the current study. Recent attempts to relax these assumptions typically resort to alternative model-based assumptions such as by incorporating treatment-by-mediator interactions and covariates in the outcome model (Pearl, 2010; Petersen, Sinisi, van der Lann, 2006; VanderWeele, 2009). The functional form of the outcome model may have direct consequences for identification (Drake, 1993; Holland, 1988; Sobel, 2008), however, and the computation of standard error becomes cumbersome for each causal effect estimate represented as a function of multiple parameters. Viewing the counterfactual outcomes as missing data, van der Lann and Petersen (2008) outlined a series of methods for estimating the natural direct effect, all of which require a user-supplied parameterization of the natural direct effect and additional modeling assumptions to obtain estimators with good practical performance. More recently, Imai and colleagues (Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010) have developed a computationally intensive algorithm that requires fitting a mediator model and an outcome model followed by repeatedly simulating at least 1,000 times the potential values of the mediator and the potential outcomes given the simulated values of the mediator. The analysis nonetheless depends on correct specification of the functional form of both the outcome and mediator models.

In this study, we develop both parametric and non-parametric weighting procedures for estimating the causal effects of interest. The parametric procedure involves (a) inverse-probability-of-treatment weighting (IPTW) (Robins, 1999; Rosenbaum, 1987) for estimating the controlled direct effects (VanderWeele, 2009) and (b) ratio-of-mediator-probability weighting (RMPW) for decomposing the total effect into the natural direct effect and the natural indirect effect (Hong, 2010a). The non-parametric procedure involves (a) marginal mean weighting through stratification (MMW-S) (Hong, 2010b, 2011) for estimating the controlled direct effects and (b) non-parametric ratio-of-mediator-probability weighting (NRMPW) for estimating the natural direct and indirect effects. We will show that employing either RMPW or NRMPW minimizes the need for specifying the outcome model and simplifies the computation of standard errors. In theory, the non-parametric MMW-S method and the NRMPW method reduce the reliance on correct specification of the functional form of the mediator model. Both the parametric and the non-parametric weighting approaches can be flexibly applied to multi-valued treatments and mediators.

This paper is organized as follows. Section 2 defines the causal estimands. Section 3 presents the general theoretical results of using weighting adjustment for bias reduction and clarifies the identification assumptions. Section 4 applies the parametric and non-parametric weighting procedures to the NEWWS data and estimates the causal effects of interest. Section 5 shows simulation results comparing the performance of parametric weighting with that of non-parametric weighting. Simulations are also used to reveal the relative strengths and limitations of the weighting procedures in comparison with path analysis and the IV method. Section 6 extends the weighting methods to investigations of moderated effects and multi-valued mediators. Section 7 concludes and discusses issues for future research.

2. Causal Estimands

We define the causal effects of interest in terms of the counterfactual outcomes (Holland, 1986, 1988; Pearl, 2001; Robins & Greenland, 1992; Rubin, 1978). Let A denote random assignment, Z for employment experience during the two years after randomization, and Y for depressive symptoms at the two-year follow-up. Let A = 1 if a welfare mother was assigned to the LFA program and A = 0 if assigned to the control condition. We use Z_1 to denote a mother's potential employment experience if assigned to the LFA program and Z_0 for the mother's potential employment experience if assigned to the control condition. We use Y_{1Z_1} to denote a mother's potential psychological outcome if assigned to the LFA program and Y_{0Z_0} for the potential outcome if assigned to the control condition. We use Y_{1Z_0} to denote a mother's counterfactual outcome if assigned to the LFA program yet counterfactually experiencing employment as she would have had under the control condition. For simplicity, below we define the causal effects when Z is measured on a binary scale, that is, Z = 1 if ever employed and Z = 0 if never employed during the two-year period.

The *controlled direct effect of the policy* if one remains unemployed regardless of the treatment assignment is defined as $Y_{10} - Y_{00}$; analogously, the controlled direct effect of the policy if one remains employed regardless of the treatment assignment is defined as $Y_{11} - Y_{01}$. The *controlled direct effect of employment* if assigned to the LFA program is defined as $Y_{11} - Y_{10}$; analogously, the controlled direct effect of employment under the control condition is defined as $Y_{01} - Y_{00}$.

The *natural direct effect of the policy* on maternal depression is defined by $Y_{1Z_0} - Y_{0Z_0}$. A positive natural direct effect would indicate that being assigned to the LFA program increased a mother's depressive symptoms two years later if the assignment to the LFA program counterfactually left her employment experience unchanged. The *natural indirect effect of the policy*, defined by $Y_{1Z_1} - Y_{1Z_0}$, represents the change in a mother's depressive symptoms attributable to the policy-induced change in her employment experience. The total effect of treatment assignment $Y_{1Z_1} - Y_{0Z_0}$ is the sum of the natural direct effect and the natural indirect effect.

3. Identification through Weighting Adjustment

This section presents the theoretical results clarifying the identification assumptions under which weighting removes selection bias in estimating the causal effects defined above. Following van der Lann and Petersen (2008), we represent the joint distribution of the observed data $O = \left(\mathbf{X}, A, Z_A, Y_{AZ_A}\right)$ in general as follows:

$$f^{(a,z)}(Y_{az}|A = a, Z_a = z, \mathbf{X}) \times q^{(a)}(Z_a = z|A = a, \mathbf{X}) \times p(A = a|\mathbf{X}) \times h(\mathbf{X}),$$

Where $\mathbf{X}$ denotes a vector of observed pretreatment covariates and where $f^{(a,z)}(\cdot)$, $q^{(a)}(\cdot)$, $p(\cdot)$, and $h(\cdot)$ are density functions. For simplicity, we use $f(\cdot)$ to represent $f^{(a,z)}(\cdot)$ in the discussion below. Theorem 1 summarizes the results from past research (Robins, 1999; Rosenbaum, 1987) showing the assumptions required for identifying $E\left(Y_{aZ_a}\right)$ through weighting.

THEOREM 1. $E(Y^*|A = a) \equiv E\left(W_{(aZ_a)}Y|A = a\right)$ is an observed data estimand for $E\left(Y_{aZ_a}\right)$, where

$$W_{(aZ_a)} = p(A = a)/p(A = a|\mathbf{X}) \quad (1)$$

for all possible values of a under the following assumptions:

Assumption 1 (Nonzero probability of treatment assignment). $0 < p(A = a \mid \mathbf{X}) < 1$.

Assumption 2 (Independence of treatment assignment and potential outcomes). $Y_{az} \coprod A \mid \mathbf{X}$.

Assumptions 1 and 2 are trivial in this experiment assigning welfare mothers at random to either the LFA program or the control condition. In this case, we have that $p(A = a|\mathbf{X}) = p(A = a)$ and hence $W_{(aZ_a)} = 1.0$ and $Y^* = Y$ for all units.

Theorem 2 summarizes the results for identifying $E(Y_{az})$ through weighting (VanderWeele, 2009).

THEOREM 2. $E(Y^*|A = a) \equiv E\big(W_{(az)}Y|A = a, Z = z\big)$ is an observed data estimand for $E(Y_{az})$, where

$$W_{(az)} = \big\{q^{(a)}(Z_a = z|A = a)/q^{(a)}(Z_a = z|A = a, \mathbf{X})\big\} \times \{p(A = a)/p(A = a|\mathbf{X})\} \quad (2)$$

for all possible values of a under the following additional assumptions:

Assumption 3 (Nonzero probability of mediator value assignment within a treatment). $0 < pr(Z_a = z \mid A, \mathbf{X}) < 1$.

Assumption 4 (No confounding of mediator-outcome relationship within the treatment assigned). $Y_{az} \coprod Z_a \mid A = a, \mathbf{X}$.

Assumption 3 implies that the controlled direct effects cannot be identified for units who are "always employed under either LFA or control" or are "never employed under either LFA or control." Assumption 4 requires that the observed covariates adequately account for all the potential confounding of the mediator-outcome relationship under each treatment condition.

Theorem 3 summarizes the results for identifying $E\big(Y_{aZ_{a'}}\big)$ through weighting (Hong, 2010a).

THEOREM 3. $E(Y^*|A = a) \equiv E\big(W_{(aZ_{a'})}Y|A = a\big)$ is an observed data estimand for $E\big(Y_{aZ_{a'}}\big)$, where

$$W_{(aZ_{a'})} = \big\{q^{(a')}(Z_{a'} = z|A = a', \mathbf{X})/q^{(a)}(Z_a = z|A = a, \mathbf{X})\big\} \times \{p(A = a)/p(A = a|\mathbf{X})\}. \quad (3)$$

for all possible values of a under the following additional assumptions:

Assumption 5 (No confounding of treatment-mediator relationship). $Z_a \coprod A \mid \mathbf{X}$.

Assumption 6 (No confounding of mediator-outcome relationship across treatment conditions). $Y_{az} \coprod Z_{a'} \mid A = a, \mathbf{X}$.

Assumption 5 is trivial when A is randomized. Assumption 6 implies that there are no post-treatment covariates given the observed pretreatment covariates $\mathbf{X}$ (Pearl, 2001). It is important to note that identifying the natural direct effect and the natural indirect effect does not require Assumption 3. Instead, it only requires precluding the possibility that $pr(Z_a = z|A = a, \mathbf{X}) = 0$ when $pr(Z_{a'} = z|A = a', \mathbf{X}) > 0$. Let $\Omega_{Z_a|\mathbf{X}}$ be the support for mediator values under the target treatment $A = a$ conditioning on pretreatment covariates $\mathbf{X}$; Let $\Omega_{Z_{a'}|\mathbf{X}}$ be the corresponding conditional support for mediator values under the reference treatment $A = a'$. We replace Assumption 3 with the following:

*Assumption 3** (Conditional support for mediator values under the reference treatment not exceeding that under the target treatment). $\Omega_{Z_{a'}|\mathbf{X}} \leq \Omega_{Z_a|\mathbf{X}}$.

Assumption 3* implies that when decomposing the total effect into the natural direct effect and the natural indirect effect, there is no need to exclude welfare mothers who would be "always employed regardless of treatment" or "never employed regardless of treatment." This is because for such units, $pr(Z_1 = z|A = 1, \mathbf{X}) = pr(Z_0 = z|A = 0, \mathbf{X})$ for $z = 0,1$. Hence their ratio-of-mediator-probability weight would be equal to 1.0. However, the natural direct and indirect effects cannot be identified for those who would be "always employed only if treated," that is, $pr(Z_1 = 0|A = 1, \mathbf{X}) = 0$ and $pr(Z_0 = 0|A = 0, \mathbf{X}) > 0$, or would be "never employed only if treated," that is, $pr(Z_1 = 1|A = 1, \mathbf{X}) = 0$ and $pr(Z_0 = 1|A = 0, \mathbf{X}) > 0$.

4. Application of Parametric and Non-Parametric Weighting to NEWWS

Applying the above theoretical results to an analysis of the NEWWS data, we present the analytic procedures and report the results. The data had complete information on policy assignment and employment. We identified 86 pretreatment covariates that predict either maternal depression or employment experience. After creating a missing category for those with missing information in each categorical covariate and assuming missing at random, we conducted maximum likelihood-based missing imputation in the outcome and in the continuous covariates (Little & Rubin, 2002).

According to the results of intent-to-treat analysis, assignment to the LFA program increased the employment rate from 39.5% to 65.4%. However, the average policy effect on maternal depression cannot be statistically distinguished from zero (coefficient = 0.11, SE = 0.64, t = 0.18, p = 0.86).

4.1 Mediator on a Binary Scale: Parametric Approach

In the NEWWS data, policy assignment A is binary. When the mediator is also binary, to estimate the controlled direct effects of the policy and the controlled direct effects of employment, we need to estimate the marginal mean outcomes $E(Y_{00})$, $E(Y_{10})$, $E(Y_{01})$, and $E(Y_{11})$. To estimate the natural direct effect and the natural indirect effect requires obtaining sample estimates of each of the three marginal mean outcomes $E\left(Y_{0Z_0}\right)$, $E\left(Y_{1Z_1}\right)$, and $E\left(Y_{1Z_0}\right)$.

4.1.1 Estimation of the Controlled Direct Effects though Parametric Weighting

Because the policy assignment was randomized, Equation (2) is simplified as $W_{(az)} = q^{(a)}(Z_a = z|A = a)/q^{(a)}(Z_a = z|A = a, \mathbf{X})$. We predict the conditional probability of being employed if assigned to the control condition and the conditional probability of being employed if assigned to the LFA program each as a function of $\mathbf{X}$ denoted by $\theta_{Z_0}(\mathbf{X})$ and $\theta_{Z_1}(\mathbf{X})$, respectively. We identify, within the LFA group, individuals who had no counterfactual information on the basis of the estimated $\theta_{Z_1}(\mathbf{X})$. Similarly, we identify individuals in the control group who had no counterfactual information on the basis of the estimated $\theta_{Z_0}(\mathbf{X})$. See online Table 1 for the exclusion criteria. A predefined propensity score caliper of width no more than 0.2 standard deviation of the logit propensity score is often acceptable (Austin, 2011). To be conservative, we choose the caliper width to be 10% of a standard deviation of the logit propensity score. Applying the same exclusion criteria to both groups, we exclude 18 LFA members and 39 control group members from the analytic sample.

Following Equation (2), we obtain the parametric weight as follows for $a = 0,1$:

If $A = a$ and $Z = 1$, then $IPTW = pr(Z = 1|A = a)/\theta_{Z_a}(\mathbf{X})$;

If $A = a$ and $Z = 0$, then $IPTW = pr(Z = 0|A = a)/\left(1 - \theta_{Z_a}(\mathbf{X})\right)$.

Our goal is to approximate a second-stage randomization under Assumptions 3 and 4. We then analyze the following regression model through weighted least squares.

$$Y = \delta^{(0)} + \delta^{(A)}A + \delta^{(Z)}Z + \delta^{(AZ)}AZ + e. \tag{4}$$

The estimated controlled direct effect of the policy if unemployed is 1.73 (SE = 1.12, t = 1.55, p = .12); the estimated controlled direct effect of the policy if employed is -2.20 (SE = 0.95, t = -2.31, p < .05). Clearly, the controlled direct effect of the policy if unemployed is significantly different from that if employed (coefficient = 3.93, SE = 1.47, t = 2.68, p < .01). The estimated controlled direct effect of employment under LFA (coefficient = -2.64, SE = 1.24, t = 2.14, p < .05) is also significantly different from that under the control condition (coefficient = 1.29, SE = 0.79, t = 1.63, p = .10). Even though employment does not seem to affect maternal depression by a significant amount had all welfare mothers continued to be covered by the old policy, it appears that once employment becomes one of the primary qualifications for welfare receipt, the impact of employment success on one's psychological well-being becomes salient.

4.1.2 Estimation of the Natural Direct Effect and the Natural Indirect Effect through Parametric Weighting

As stated under Theorem 3, to estimate the natural direct effect and the natural indirect effect requires Assumption 3* instead of Assumption 3. In practice, we define the "always employed regardless of treatment" as those whose estimated θ_{Z_0} and θ_{Z_1} were both higher than the respective maximum values displayed by those who were unemployed; we define the "never employed regardless of treatment" as those whose estimated θ_{Z_0} and θ_{Z_1} were both lower than the respective minimum value displayed by those who were employed. Online Table 2 summarizes the inclusion criteria for identifying the common support for estimating the natural direct and indirect effects. We do not find "always employed regardless of treatment" or "never employed regardless of treatment" in this data set.

According to Theorem 1, the sample mean outcome of the control group members provides an unbiased estimate of $E\left(Y_{0Z_0}\right)$ while the sample mean outcome of the LFA participants provide an unbiased estimate of $E\left(Y_{1Z_1}\right)$. According to Theorem 3, the ratio-of-mediator-probability weighted sample mean outcome of the LFA participants provide an unbiased estimate of $E\left(Y_{1Z_0}\right)$. To proceed, we reconstruct the data set to include the sampled control group members, the sampled LFA members, and a duplicate set of the sampled LFA members. Let D be a dummy indicator that takes value 1 for the duplicate LFA members and 0 otherwise. We assign the weight as follows to units in the analytic sample:

If $A = 0$, and $D = 0$, then $RMPW = 1.0$;
If $A = 1$, and $D = 1$, then $RMPW = 1.0$;
If $A = 1, D = 0,$ and $Z = 1$, then $RMPW = \theta_{Z_0}/\theta_{Z_1}$;
If $A = 1, D = 0,$ and $Z = 0$, then $RMPW = \left(1 - \theta_{Z_0}\right)/\left(1 - \theta_{Z_0}\right)$.

The control group and the RMPW adjusted LFA group display the same distribution of employment. We then analyze an outcome model through generalized least squares and obtain robust standard errors.

$$Y = \gamma^{(0)} + \gamma^{(ND)}A + \gamma^{(NI)}AD + e. \qquad (5)$$

Here $\gamma^{(0)}$ estimates $E\left(Y_{0Z_0}\right)$; $\gamma^{(ND)}$ estimates the natural direct effect $E\left(Y_{1Z_0} - Y_{0Z_0}\right)$; and $\gamma^{(NI)}$ estimates the natural indirect effect $E\left(Y_{1Z_1} - Y_{1Z_0}\right)$. The estimated natural direct effect is 1.02 (SE = 0.93; *Wald* χ^2 = 1.20, p = 0.27), about 14% of a standard deviation of the outcome; the estimated natural indirect effect is -0.70 (SE = 1.04; *Wald* χ^2 = 0.46, p = 0.50). On one hand, even if the employment experience of all welfare mothers counterfactually remained unchanged by the welfare-to-work policy, maternal depression would increase only by an insignificant amount; on the other hand, apparently the policy induced change in employment is not great enough to produce a significant amount of reduction in maternal depression on average.

4.2 Mediator on a Binary Scale: Non-Parametric Approach

4.2.1 Non-Parametric Estimation of the Controlled Direct Effects

We rank order the LFA members by θ_{Z_1} and then divide the LFA sample into six equal portions denoted by $S_{Z_1} = 1, \ldots, 6$. Let N denote the analytic sample size. Let $I_i\left(S_{Z_1} = s\right) = 1$ if LFA member i is in stratum s and 0 otherwise. Let $I_i(Z_1 = z)$ be the

indicator for employment status under LFA for $z = 0{,}1$. The marginal mean weight through stratification for the LFA members is computed as follows.
If $A = 1$, $Z = z$, and $S_{Z_1} = s$, then

$$MMWS = \frac{\sum_{i=1}^{N} A_i\, I_i(S_{Z_1} = s)}{\sum_{i=1}^{N} A_i\, I_i(Z_1 = z) I_i(S_{Z_1} = s)} \times \frac{\sum_{i=1}^{N} A_i\, I_i(Z_1 = z)}{\sum_{i=1}^{N} A_i}.$$

Similarly, for those assigned to the control group, we rank order them by θ_{Z_0} and then divide the control sample into six equal portions denoted by $S_{Z_0} = 1, \dots, 6$.
If $A = 0$, $Z = z$, and $S_{Z_0} = s$,

$$MMWS = \frac{\sum_{i=1}^{N}(1 - A_i)\, I_i(S_{Z_0} = s)}{\sum_{i=1}^{N}(1 - A_i)\, I_i(Z_0 = z) I_i(S_{Z_0} = s)} \times \frac{\sum_{i=1}^{N}(1 - A_i)\, I_i(Z_0 = z)}{\sum_{i=1}^{N}(1 - A_i)}.$$

Analyzing Model (4) weighted by $MMWS$ through weighted least squares, we obtain an estimate of the controlled direct effect of the policy for the unemployed (coefficient = 1.62, *SE* = 1.09, *t* = 1.49, *p* = 0.14), and an estimate of the controlled direct effect of the policy for the employed (coefficient = -1.61, *SE* = 0.92, *t* = -1.75, *p* = 0.08). There is clear indication that the controlled direct effect for the unemployed is significantly different from that for the employed. The estimated interaction effect is 3.23 (*SE* = 1.42, *t* = 2.27, *p* < 0.05). The estimated controlled direct effect of employment for the control group members is 0.74 (*SE* = 0.76, *t* = 0.97, *p* = 0.33); while that for the LFA participants is -2.49 (*SE* = 1.20, *t* = -2.07, *p* < 0.05). Similar to the parametric results, successful experience with employment under the welfare-to-work policy apparently reduces maternal depression by a significant amount while showing no significant impact under the control condition.

4.2.2 Non-Parametric Estimation of the Natural Direct Effect and the Natural Indirect Effect

We rank order the units in the analytic sample by θ_{Z_1} and then divide the sample into three even portions. Within each of these three subclasses, we then rank order and subdivide again by θ_{Z_0}. Let $S = 1, \dots, 9$ denote the nine subclasses. We generate a duplicate set of the LFA group as before and use $D = 1$ to denote the duplicates and 0 otherwise. The non-parametric weight is computed as follows:

If $A = 0$, and $D = 0$, then $NRMPW = 1.0$;
If $A = 1$, and $D = 1$, then $NRMPW = 1.0$;
If $A = 1, D = 0,$ and $Z = z$, then

$$NRMPW = \frac{\sum_{i=1}^{N}(1 - A_i)\, I_i(Z_0 = z) I_i(S = s)}{\sum_{i=1}^{N}(1 - A_i)\, I_i(S = s)} \times \frac{\sum_{i=1}^{N} A_i\, I_i(S = s)}{\sum_{i=1}^{N} A_i\, I_i(Z_1 = z) I_i(S = s)}.$$

Applying the above non-parametric weight, we analyze Equation (5) through generalized least squares and obtain robust standard errors. The estimated natural direct effect is 1.07 (*SE* = 0.88, *Wald* χ^2 = 1.50, *p* = 0.22); the estimated natural indirect effect is -0.76 (*SE* = 0.99, *Wald* χ^2 = 0.58, *p* = 0.45).

5. Simulations

5.1 Comparisons between Parametric and Non-Parametric Weighting Strategies

We conduct a series of Monte Carlo simulations to assess the performance of the parametric weighting procedure relative to the non-parametric weighting procedure in the case of a binary randomized treatment indicator, a binary mediator, and a continuous outcome. Because nonlinear non-additive propensity score models are typically misspecified as a linear additive one, we also compare between the parametric and the non-parametric procedures the sensitivity of results to such model misspecifications. The simulated data resemble the structure of the NEWWS Riverside data.

In our baseline model, potential outcomes Y_{az} for $a = 0,1$ and $z = 0,1$ are each a linear additive function of three standard normal independent covariates X_1, X_2, and X_3. Let the logit of propensity for employment under each treatment be a linear additive function of these same covariates. The controlled direct effects are determined by three parameters $\delta^{(A)}$, $\delta^{(Z)}$, and $\delta^{(AZ)}$. Specifically, $E(Y_{10} - Y_{00}) = \delta^{(A)}$, $E(Y_{01} - Y_{00}) = \delta^{(Z)}$, $E(Y_{11} - Y_{01}) = \delta^{(A)} + \delta^{(AZ)}$ and $E(Y_{11} - Y_{10}) = \delta^{(Z)} + \delta^{(AZ)}$. The above three parameters together with $E\left(\theta_{Z_0}\right)$ and $E\left(\theta_{Z_1}\right)$ determine the natural direct effect and the natural indirect effect. Specifically, $E\left(Y_{1Z_0} - Y_{0Z_0}\right) = \gamma^{(ND)} = \delta_A + \delta_{AZ}E\left(\theta_{Z_0}\right)$, and $E\left(Y_{1Z_1} - Y_{1Z_0}\right) = \gamma^{(NI)} = (\delta_Z + \delta_{AZ})\left[E\left(\theta_{Z_1}\right) - E\left(\theta_{Z_0}\right)\right]$. We compare across four sets of parameter value specifications:

(a) $E\left(\theta_{Z_0}\right) = E\left(\theta_{Z_1}\right) = .5$, $\delta^{(A)} = \delta^{(Z)} = \delta^{(AZ)} = 0$, and hence $\gamma^{(ND)} = \gamma^{(NI)} = 0$;

(b) $E\left(\theta_{Z_0}\right) = .4$, $E\left(\theta_{Z_1}\right) = .65$, $\delta^{(A)} = \delta^{(Z)} = \delta^{(AZ)} = 0$, and hence $\gamma^{(ND)} = \gamma^{(NI)} = 0$;

(c) $E\left(\theta_{Z_0}\right) = .4$, $E\left(\theta_{Z_1}\right) = .65$, $\delta^{(A)} = 2.55$, $\delta^{(Z)} = 1.5$, $\delta^{(AZ)} = -4.5$, hence $\gamma^{(ND)} = 0.75$ and $\gamma^{(NI)} = -0.75$;

(d) $E\left(\theta_{Z_0}\right) = .2$, $E\left(\theta_{Z_1}\right) = .8$, $\delta^{(A)} = 1.5$, $\delta^{(Z)} = 2.5$, $\delta^{(AZ)} = -3.74$, hence $\gamma^{(ND)} = 0.75$ and $\gamma^{(NI)} = -0.75$.

The evaluation criteria for causal effect estimate $\hat{\lambda}$ include the following: (1) bias in the point estimate: $E\left(\hat{\lambda}\right) - \lambda$; (2) sampling variability of the point estimate: $\sigma^2\left(\hat{\lambda}\right) = E\left[\hat{\lambda} - E\left(\hat{\lambda}\right)\right]^2$; (3) mean square error: $E\left[\left(\hat{\lambda} - \lambda\right)^2\right] = \sigma^2\left(\hat{\lambda}\right) + \left[E\left(\hat{\lambda}\right) - \lambda\right]^2$; (4) approximate bias in the standard error estimate: $E\left[\widehat{\sigma\left(\hat{\lambda}\right)}\right] - \sigma\left(\hat{\lambda}\right)$. We select two different sample sizes. *N* = 800 represents a relatively small sample size similar to the NEWWS Riverside data; *N* = 5,000 represents a large sample size seen in some other national evaluations. For each given sample size, we generate 1,000 random samples. The tabulated results of all the simulations are available on the first author's web site (http://home.uchicago.edu/~ghong/JSM2011SimulationResults).

Across the four sets of parameter values, when the propensity score model is correctly specified, parametric weighting removes more initial bias than does non-parametric weighting. In scenarios (b), (c), and (d) in which the treatment has a nonzero effect on the mediator, both parametric and non-parametric weighting methods remove at least 90% of the bias in all causal effect estimates. We compute the initial bias with no adjustment for the covariates. The bias in standard error estimates never exceeds 4.5% of a standard deviation of a potential outcome. Non-parametric weighting is generally more efficient than parametric weighting. As expected, sampling variability increases as the sample size decreases and as $E\left(\theta_{Z_0}\right)$ and $E\left(\theta_{Z_1}\right)$ move away from .5.

Earlier research (Hong, 2010) has shown that, when a nonlinear non-additive propensity score model is misspecified, non-parametric weighting estimates of average treatment effects are typically more robust than parametric weighting estimates. To investigate whether the same pattern exists in mediation analysis, we let the true propensity score models include either X_3^2 or X_2X_3 each having a relatively weak yet realistic association (coefficient = 0.2) with the logit of propensity. We find that the non-parametric weighting results remain robust. The parametric weighting results are generally comparable to those of non-parametric weighting except for the estimation of $\delta^{(AZ)}$ in which case the parametric weighting results contain a considerably larger amount of bias.

5.2. Comparisons with Conventional Approaches to Mediation Analysis

5.2.1 Comparisons with Path Analysis

We first analyze naïve path analysis models with no statistical adjustment for any covariates and with no treatment-by-mediator interaction. The amount of bias in the naïve path analysis results now serves as the basis for computing bias removal in comparing across the adjustment methods. In scenarios (b), (c), and (d), the weighting methods consistently remove at least 90% of the bias. In scenario (a) in which the treatment has zero effect on the mediator, the naïve path analysis outperforms the weighting methods in estimating the controlled direct effect of the treatment and the natural direct effect.

We then conduct path analysis with linear covariance adjustment for all three covariates yet with no treatment-by-mediator interaction. The weighting results replicate the adjusted path analysis results when there is a nonzero treatment effect on the mediator and when there is no treatment-by-mediator interaction. The weighting methods outperform the adjusted path analysis when the treatment effect on the mediator is nonzero and when there is treatment-by-mediator interaction.

5.2.2 Comparisons with the IV Method

To compare the weighting methods with the IV method, we assume that the exclusion restriction holds and therefore let $\delta_A = \delta_{AZ} = 0$ while keeping all other parameter values the same as before. The natural direct effect is now zero; and the natural indirect effect is equal to $\delta_Z\left[E\left(\theta_{Z_1}\right) - E\left(\theta_{Z_0}\right)\right]$. The parameter value specifications are modified as follows:

(c*) $E\left(\theta_{Z_0}\right) = .4$, $E\left(\theta_{Z_1}\right) = .65$, $\delta^{(A)} = 0$, $\delta^{(Z)} = 1.5$, $\delta^{(AZ)} = 0$, hence $\gamma^{(ND)} = 0$ and $\gamma^{(NI)} = 0.375$;

(d*) $E\left(\theta_{Z_0}\right) = .2$, $E\left(\theta_{Z_1}\right) = .8$, $\delta^{(A)} = 0$, $\delta^{(Z)} = 2.5$, $\delta^{(AZ)} = 0$, hence $\gamma^{(ND)} = 0$ and $\gamma^{(NI)} = 1.5$.

Specifically, we compare the weighting adjusted estimate of the natural indirect effect with the IV adjusted estimate of the natural indirect effect. The latter is equivalent to the ITT effect of the treatment on the outcome under the exclusion restriction. For additional comparisons, we also estimate the ITT effect with covariance adjustment for all the covariates. The parametric weighting estimates are almost identical to the ITT estimates and to the adjusted ITT estimates when the sample size is relatively large. The non-parametric weighting results show a slightly larger bias as expected. In general, the weighting results replicate the IV results when the identification assumptions hold.

6. Extensions of the Weighting Methods

The weighting strategies can be extended to studies of mediational relationships moderated by subpopulation status and to studies of multi-valued mediators. Because the parametric and the non-parametric weighting strategies have shown comparable performance, we will present the parametric results for simplicity.

6.1 Extensions to Moderated Effects

Our preliminary examination of the NEWWS data has revealed that welfare recipients who had been teen parents in the past appeared to respond differently to the policy than those who had never become teen parents. Non-teen mothers assigned to the LFA program appeared to have heightened depressive symptoms two years after randomization in comparison with their counterparts assigned to the control group. In contrast, although teen mothers started with a higher average level of depression at the baseline, assignment to the LFA program did not seem to increase their depressive symptoms more than assignment to the control group two years later. It is also noteworthy that, while non-teen mothers assigned to the LFA program showed a higher employment rate (63%) than those assigned to the control group (41%), an even larger difference in employment rate among teen mothers emerged between the LFA group (70%) and the control group (37%).

Here we extend the analytic procedures to consider teen parenthood as a moderator. Let V be a dummy indicator for teen parent status, that is, $V = 1$ for a teen mother and $V = 0$ for a non-teen mother. We conduct the same set of analysis as before within each subpopulation. Equation (4) is modified to include two submodels, one for teen mothers and the other for non-teen mothers:

$$Y = V\left(\delta_{V1}^{(0)} + \delta_{V1}^{(A)}A + \delta_{V1}^{(Z)}Z + \delta_{V1}^{(AZ)}AZ\right) + (1 - V)\left(\delta_{V0}^{(0)} + \delta_{V0}^{(A)}A + \delta_{V0}^{(Z)}Z + \delta_{V0}^{(AZ)}AZ\right) + e.$$

A reparameterization of the above model enables one to test whether any of the controlled direct effects significantly differ between teen mothers and non-teen mothers.

We find that under the LFA program, unemployment would increase the depressive symptoms among non-teen mothers by a significant amount (coefficient = 3.09, *SE* = 1.52, *t* = 2.03, *p* < .05). For teen mothers, the impact of unemployment on depressive symptoms under the LFA program appears to be smaller (coefficient = 1.16, *SE* = 2.04, *t* = 0.57, *p* = .57). The difference between teen mothers and non-teen mothers, however, cannot be distinguished from zero. In general, the controlled direct effect of employment and that of program assignment do not differ between teen mothers and non-teen mothers.

Similarly, we modify Equation (5) as follows:

$$Y = V\left(\gamma_{V1}^{(0)} + \gamma_{V1}^{(ND)}A + \gamma_{V1}^{(NI)}AD\right) + (1 - V)\left(\gamma_{V0}^{(0)} + \gamma_{V0}^{(ND)}A + \gamma_{V0}^{(NI)}AD\right) + e.$$

Here $\gamma_{V1}^{(ND)}$ and $\gamma_{V1}^{(NI)}$ are estimates of the natural direct effect and the natural indirect effect, respectively, for teen mothers; $\gamma_{V0}^{(ND)}$ and $\gamma_{V0}^{(NI)}$ are the corresponding estimates for non-teen mothers. We find that, for teen mothers, the natural direct effect is -0.53 (*SE* = 1.23, *Wald* χ^2 = 0.18, *p* = .67) and the natural indirect effect is -0.42 (*SE* = 1.35, *Wald* χ^2 = 0.10, *p* = .76); while for non-teen mothers, the natural direct effect is 1.90 (*SE* = 1.33, *Wald* χ^2 = 2.04, *p* = .15) and the natural indirect effect is -0.81 (*SE* = 1.47, *Wald* χ^2 = 0.30, *p* = .58). We find no evidence that the natural direct and indirect effects differ between teen mothers and non-teen mothers.

6.2 Extensions to Mediators on a Multi-Valued Scale

The binary measure of employment does not distinguish among people who were employed to varying degrees. To overcome this limitation, we examine a three-category measure of employment: Never employed, employed for no more than 50% of the two-year period, and employed for more than 50% of the two-year period. In the NEWWS data, the proportion of participants employed for no more than 50% of the two-year period was higher in the LFA group (34.6%) than in the control group (23.3%). An even larger difference in the proportion of those employed for more than 50% of the time lies between the LFA group (30.8%) and the control group (16.3%). We hypothesize that employment for more than 50% of the time may reduce maternal depression more effectively than employment for no more than 50% of the time, and that the benefit of employment for reducing depression is expected to be more evident under the LFA program than under the control condition.

Below we highlight the modifications in the analytic procedure for accommodating a multi-valued mediator. Let $z = 0, 1, 2$ denote zero employment, low employment, and high employment, respectively. The controlled direct effect of the policy dependent on the level of employment is $E(Y_{1z} - Y_{0z})$ for $z = 0, 1, 2$; the controlled direct effect of employment dependent on the policy is $E(Y_{az} - Y_{az'})$ for $a = 0, 1$ and $z = 0, 1, 2$. The above definitions involve the marginal mean outcomes $E(Y_{00})$, $E(Y_{01})$, $E(Y_{02})$, $E(Y_{10})$, $E(Y_{11})$, and $E(Y_{12})$. The natural direct effect and the natural indirect effect are defined the same as before regardless of the change in mediator distributions.

For each individual unit, we estimate the conditional probabilities of having zero employment, low employment, and high employment if assigned to the LFA program as a function of pretreatment covariates **X**, denoted by $\theta_{Z_1=0}$, $\theta_{Z_1=1}$, and $\theta_{Z_1=2}$, respectively; we also estimate the conditional probabilities of having zero employment, low employment, and high employment if assigned to the control group, denoted by $\theta_{Z_0=0}$, $\theta_{Z_0=1}$, and $\theta_{Z_0=2}$, respectively. A comparison between a multinomial logistic regression model and an ordinal logistical regression model shows that, in this case, the latter fits the data as adequately as the former.

6.2.1 Estimation of the Controlled Direct Effects

To identify common support as required by Assumption 3, we choose the analytic sample by excluding units who do not have counterfactual information under an alternative treatment level when the treatment is given. Online Table 3 summarizes the exclusion criteria. The analytic sample for estimating the controlled direct effects includes 166 LFA participants and 402 control units.

Applying Equation (2) to units whose $A = a$ and $Z = z$, the parametric weight is simply $IPTW = pr(Z = z|A = a)/\theta_{Z_a=z}$. Under Assumptions 3 and 4, the weighted data approximate a two-stage randomized experiment in which individuals assigned to each treatment would then be assigned at random to zero employment, low employment, or high employment. We modify the model in Equation (4) to include two dummy indicators $I(Z = 1)$ and $I(Z = 2)$ for low employment and high employment, respectively, and then analyze through weighted least squares.

$$Y = \delta^{(0)} + \delta^{(A)}A + \delta^{(Z=1)}I(Z = 1) + \delta^{(Z=2)}I(Z = 2) + \delta^{(A,Z=1)}AI(Z = 1) \\ + \delta^{(A,Z=2)}AI(Z = 2) + e.$$

The controlled direct effect of the policy remains insignificant regardless of whether the employment level is set at zero, low, or high. Even though a large contrast exists in the controlled direct effect of the policy between zero employment and high employment (coefficient = 3.30, *SE* = 1.79, *t* = 1.84, *p* = .07), the difference fails to reach statistical significance. Under the control condition, the controlled direct effects of zero vs. low, of low vs. high, and of zero vs. high employment are all indistinguishable from zero. However, under the welfare-to-work policy, the controlled direct effect of zero vs. high is noteworthy (coefficient = 3.92, *SE* = 1.47, *t* = 2.67, *p* < .01). In comparison with zero employment, high employment is expected to reduce maternal depression by as much as 50% of a standard deviation.

6.2.2 Estimation of the Natural Direct Effect and the Natural Indirect Effect

To satisfy Assumption 3*, the analytic sample includes not only units who have a nonzero probability of having each level of employment under each treatment condition but also those who would always have zero employment, low employment, or high employment regardless of treatment assignment. Online Table 4 shows the empirical criteria for identifying these subpopulations. In this study, apparently none of the welfare mothers would always have zero employment, low employment, or high employment regardless of treatment assignment.

As before, we reconstruct the data to include the sampled control group members, the sampled LFA members, and a duplicate set of the sampled LFA members. Let D be a dummy indicator that takes value 1 for the duplicate LFA members and 0 otherwise. Applying Equation (3) to the analytic sample, the parametric RMPW can be computed as follows:

If $A = 0$, and $D = 0$, then $RMPW = 1.0$;

If $A = 1$, and $D = 1$, then $RMPW = 1.0$;

If $A = 1, D = 0,$ and $Z = z$, then $RMPW = \theta_{Z_0=z}/\theta_{Z_1=z}$.

Regardless of the distribution of the multi-valued mediator, the outcome model is specified the same as that in Equation (5) and is analyzed through generalized least squares with robust standard errors. The estimated natural direct effect is 1.19 (*SE* = 0.90, $Wald\ \chi^2$ = 1.76, *p* = 0.19); the estimated natural indirect effect is -.62 (*SE* = 1.03, $Wald\ \chi^2$ = 0.37, *p* = 0.55). These results are similar to the decomposition of the total effect when employment is measured on a binary scale.

7. Conclusion and Discussion

This paper has presented a series of weighting approaches to mediation analysis. Applications to the NEWWS Riverside data have generated interesting empirical results suggesting that the welfare-to-work policy could potentially heighten maternal depression if one is unemployed, but that a positive change in a welfare mother’s employment status could potentially reduce depression when she is subject to the new policy requirements. However, given that 34.6% of the welfare mothers were already employed even without the policy and given that as many as 40% of the welfare mothers remained unemployed under the new policy, the policy-induced change in employment was only modest. Hence on average, there was a minimal policy impact on maternal depression mediated by employment. The LFA program under study was a precursor to the current welfare-to-work regulations. Our empirical findings therefore have immediate implications for today’s policy discussion.

Identifying the causal effects requires the assumptions of no confounding of mediator-outcome relationship within the treatment assigned and across treatment conditions. These are strong assumptions yet are not entirely implausible given the rich set of pretreatment measures. In fact, temporal fluctuation in a local job market that is completely exogenous to the potential outcomes often contributes to the variation in employment. There could be potential concerns that the assignment to the LFA program might immediately heighten some welfare mothers’ depression level and might subsequently undermine their ability to seek and maintain employment. We reason that the random assignment to LFA by itself is unlikely to be a strong trigger of depression. The adjustment for baseline depressive symptoms through weighting is expected to remove most if not all bias associated with post-treatment depression.

The study has clarified identification assumptions required for the causal inference. In particular, we have pointed out that the common support needed for identifying the natural direct and indirect effects is different than that needed for identifying the controlled direct effects. This clarification fills a gap in the causal inference literature.

The parametric and non-parametric ratio-of-mediator-probability weighting methods applied in this study have provided a viable alternative to the conventional methods such as path analysis and the IV method. The RMPW methods show strengths especially when the assumptions of no treatment-mediator interaction and the exclusion restriction do not hold. There is clear evidence that neither assumption seems plausible in the current application. In contrast with the regression-based approaches, the RMPW adjusted outcome is specified as a function of the natural direct effect and the natural indirect effect with no model-based assumptions. It does not require combining multiple parametric models as has been the case in most other existing methods. Hence as Hong

(2010a) pointed out, the RMPW strategy applies regardless of the distribution of the outcome, the distribution of the mediator, or the functional relationship between the outcome and the mediator. The parametric approach is flexible for handling even continuous mediators. The weighting strategies are suitable for handling a large number of pretreatment covariates without a need to specify multi-way interactions among the treatment, the mediator, and the covariates. Future research may extend this approach to studies of multi-valued treatments, multiple mediators, time-varying treatments, time-varying moderators, time-varying mediators, and mediation problems in multi-level data.

Acknowledgement

This research was supported by a major research grant entitled "Improving Research on Instruction: Models, Designs, and Analytic Methods" funded by the Spencer Foundation, a Scholars Award from the William T. Grant Foundation and the start-up funds from the University of Chicago for the first author. The National Opinion Research Center (NORC) at the University of Chicago has provided research facilities.

References

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, **10**(2), 150-161.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatments effects. *Biometrics*, **49**, 1231-1236.

Hamilton, G. (2002). *Moving people from welfare to work: Lessons from the National Evaluation of Welfare-to-Work Strategies*. Manpower Demonstration Research Corporation.

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

Holland, P. (1988). "Causal inference, path analysis, and recursive structural equations models," *Sociological methodology*, **18**, 449-484.

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

Hong, G. (2010b). Marginal mean weighting through stratification: Adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics*, **35**(5), 499-531.

Hong, G. (2011). Marginal mean weighting through stratification: A generalized method for evaluating multi-valued and multiple treatments with non-experimental data. *Psychological Methods*. Advanced online publication. Doi:10.1037/a0024918.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, **15**(4), 309-334.

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

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the American Statistical Association Joint Statistical Meetings*. Minn, MN: MIRA Digital Publishing, 1572-1581, August 2005.

Pearl, J. (2010). The mediation formula: A guide to the assessment of causal pathways in non-linear models. Los Angeles, CA: University of California, Los Angeles. Technical report R-363, July 2010.

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

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In M. Elizabeth Halloran and Donald Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (pp.95-134). New York: Springer.

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

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

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

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

van der Lann, M. J., & Peterson, M. L. (2008). Direct effect models. *The International Journal of Biostatistics*, **4**(1), Article 23.

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

Appendix

Proof of Theorem 3

Theorem 3 requires that we derive a weight $W_{(aZ_{a'})}$ such that $E\left(Y_{aZ_{a'}}\right)$ can be consistently estimated by $E\left(W_{(aZ_{a'})}Y|A=a\right)$.

$$E\left(Y_{aZ_{a'}}\right) \equiv E\left\{E\left(Y_{aZ_{a'}}|\mathbf{X}\right)\right\}.$$

By Assumptions 1 and 2, the above is equal to

$$E\left\{E\left(Y_{aZ_{a'}}|A=a,\mathbf{X}\right)\right\}$$

$$\equiv \iiint\limits_{\mathbf{x},z,y} y \times f(Y_{az}=y|A=a,Z_{a'}=z,\mathbf{X}=\mathbf{x}) \times q^{(a')}(Z_{a'}=z|A=a,\mathbf{X}=\mathbf{x}) \times \square(\mathbf{X}=\mathbf{x})dydzd\mathbf{x},$$

which, by Assumptions 4, 5, and 6, is equal to

$$\iiint\limits_{\mathbf{x},z,y} y \times f(Y_{az}=y|A=a,Z_{a}=z,\mathbf{X}=\mathbf{x}) \times q^{(a')}(Z_{a'}=z|A=a',\mathbf{X}=\mathbf{x}) \times \square(\mathbf{X}=\mathbf{x})dydzd\mathbf{x}$$

which, by Bayes Theorem and by Assumptions 1, 2, and 3* is equal to

$$\iiint\limits_{\mathbf{x},z,y} y \times f(Y_{az}=y|A=a,Z_{a}=z,\mathbf{X}=\mathbf{x}) \times q^{(a)}(Z_{a}=z|A=a,\mathbf{X}=\mathbf{x}) \times \square(\mathbf{X}=\mathbf{x}|A=a) \times \frac{q^{(a')}(Z_{a'}=z|A=a',\mathbf{X}=\mathbf{x})}{q^{(a)}(Z_{a}=z|A=a,\mathbf{X}=\mathbf{x})} \times \frac{p(A=a)}{p(A=a|\mathbf{X}=\mathbf{x})}dydzd\mathbf{x} = E(Y^*|A=a),$$

where $Y^* = W_{(aZ_{a'})}Y$ and

$W_{(aZ_{a'})} = \left\{q^{(a')}(Z_{a'}=z|A=a',\mathbf{X})/q^{(a)}(Z_{a}=z|A=a,\mathbf{X})\right\} \times \{p(A=a)/p(A=a|\mathbf{X})\}$.

This concludes the proof. □