HCEO
hceconomics.org

Human Capital and Economic Opportunity Global Working Group

Working Paper Series

Working Paper No. 2013-00ň

Ratio-of-Mediator-Probability Weighting for Causal Mediation Analysis in the Presence of Treatment-by-Mediator Interaction

$N9G@E>BǩHG@
'HG9Aǩ ! >NML<A
%>9MA>Kǩ ! ńǩBEEǩ

1> I M> F ;>K, 2013

Human Capital and Economic Opportunity Global Working Group
Economics Research Center
University of Chicago
1126 E. 59th Street
Chicago IL 60637
www.hceconomics.org

Ratio-of-Mediator-Probability Weighting for Causal Mediation Analysis in the Presence of Treatment-by-Mediator Interaction

Guanglei Hong

University of Chicago

Jonah Deutsch

Mathematica Policy Research

Heather D. Hill

University of Chicago

Author Notes

Acknowledgement:

This research was supported by a major research grant funded by the Spencer Foundation, a Scholars Award from the William T. Grant Foundation, a US Department of Education Institute of Education Sciences "Statistical and Research Methodology in Education" grant, and 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. The authors owe special thanks to Howard Bloom, Larry Hedges, Stephen Raudenbush, Patrick Shrout, and seminar participants at the University of Wisconsin-Madison, the University of Chicago, Northwestern University, and the University of California-Los Angeles for their comments on earlier versions of the manuscript.

Correspondence:

Guanglei Hong

The University of Chicago

Department of Comparative Human Development,

5736 S. Woodlawn Ave., Chicago, IL 60637

Phone: 773-702-9481

Email: ghong@uchicago.edu

Abstract

Conventional methods for mediation analysis generate biased results when the mediator-outcome relationship depends on the treatment condition. This article introduces a new technique, ratio-of-mediator-probability weighting (RMPW), for decomposing total effects into direct and indirect effects in the presence of treatment-by-mediator interactions. The indirect effect can be further decomposed into a pure indirect effect and a natural treatment-by-mediator interaction effect. The latter captures the treatment effect transmitted through a change in the mediational process. We illustrate how to apply the technique to identifying whether employment mediated the relationship between an experimental welfare program and maternal depression. In comparison with other techniques for mediation analysis, RMPW requires relatively few assumptions about the distribution of the outcome, the distribution of the mediator, and the functional form of the outcome model, and is easy to implement using standard statistical software. Simulation results reveal satisfactory performance of the parametric and non-parametric RMPW procedures under the identification assumptions and show a relatively higher level of robustness of the non-parametric procedure. We provide a tutorial and Stata code for implementing this technique.

Keywords:

Causal inference; direct effect; ignorability; indirect effect; mediation mechanism; potential outcome; propensity score.

Many important research questions in psychology, prevention science, and other social science fields relate to how interventions work: What are the mechanisms through which a treatment exerts an impact on some outcome? Great strides have been made in theory and methodology for identifying mediators that could have been affected by a treatment and could have subsequently affected the outcome. To assess the role of a hypothesized mediator, researchers typically attempt to decompose the total effect of a treatment into two pieces: an "indirect effect" that channels the treatment effect through the hypothesized mediator and a "direct effect" that works directly (or through other unspecified mechanisms). However, causal mediation analysis is challenging because, even in randomized controlled trials of interventions, participants are rarely randomized to different mediator values. Estimates of the indirect effect and the direct effect will be biased if the analyst ignores confounding variables that predict the mediator and the outcome. Moreover, conventional techniques for analyzing mediation rely on strong assumptions about the structural relationships among the treatment, the mediator, and the outcome that, ironically, are the relationships that the analyst sets about to investigate.

One of the assumptions is that there is no interaction between the treatment and the mediator in their influence on the outcome (Holland, 1988). Yet, as Judd and Kenny (1981) pointed out, a treatment may produce its effects not only through changing the mediator value but also in part by altering the mediational process that normally produces the outcome. Hence they emphasized that investigating treatment-by-mediator interactions should be an important component of mediation analysis, a point echoed in the more recent discussions (Kraemer, Wilson, Fairburn, Agras, 2002; Muller, Judd, & Yzerbyt, 2005; Spencer, Zanna, & Fong, 2005).

A straightforward example comes from Powers and Swinton's (1984) study, revisited by Holland (1988), in which students were assigned at random either to an experimental condition

that encouraged them to study for a test and provided study materials or to a control condition. Holland speculated that the amount a student studied was a response to the experimental condition and was a self-imposed treatment that might have an effect on test performance. Hence, the amount of study is a *mediator* of the effect of encouragement on test performance. Suppose that students in the experimental group, as a result of receiving encouragement along with the study materials, not only spent more time studying for the test but also studied more attentively and effectively than did the control students. The intervention might then exert its impact on test performance partly through increasing the number of study hours and partly through increasing the amount of learning produced by every additional hour of study. This would be a case in which the intervention alters not only the mediator value but also the relationship between the mediator and the outcome. Even though encouragement designs are widespread in social interventions and prevention studies, treatment effects on mediational processes have been largely overlooked in conceptualizations and data analyses.

Treatment-by-mediator interactions may sometimes provide an explanation for why an intervention fails to produce its intended effect on the outcome. As some researchers have argued (Collins, Graham, & Flaherty, 1998; MacKinnon, Krull, & Lockwood, 2000; Preacher & Hayes, 2008; Shrout & Bolger, 2002), mediation could occur when the total effect of the treatment on the outcome is zero. For example, an encouragement that comes with an undue amount of pressure may increase study hours yet at the same time may reduce the amount of learning produced per hour. Even though an increase in the amount of study is expected to increase learning, with a fixed amount of study, the student would learn less under the experimental condition than under the control condition, leading to a null effect of the encouragement treatment.

Importantly, whether the treatment alters the mediational process is distinct from another class of research questions about for whom and under what conditions the treatment works; the latter focuses on subpopulations and contextual features as pretreatment moderators (Kraemer, Kiernan, Essex, & Kupfer, 2008). Investigations of whether treatment effects differ across subpopulations or across contexts can be readily carried out through multiple regression or ANOVA. In comparison, causal mediation analysis is much more challenging. Even though analysts are advised to investigate the mediator-outcome relationship across the treatment conditions, they are generally not instructed how to decompose the treatment effect in the presence of treatment-by-mediator interaction (Baron & Kenny, 1986; Judd & Kenny, 1981).

This paper clarifies the concepts under the framework of potential outcomes (Holland, 1986, 1988; Pearl, 2001; Robins & Greenland, 1992; Rubin, 1978) and introduces a new strategy for mediation analysis using ratio-of-mediator probability weighting (RMPW). The RMPW strategy relaxes important constraining assumptions and is relatively straightforward to implement in common statistical packages. In particular, RMPW adjusts for the confounding of the mediator-outcome relationship and allows for the treatment-by-mediator interaction without having to explicitly include all the covariates and interaction terms in the outcome model (Hong, 2010a; Hong & Nomi, 2012). Moreover, RMPW allows one to quantify the treatment effect on the outcome transmitted through a change in the mediational process. Unlike most of the existing strategies for mediation analysis, RMPW minimizes the need for specifying the outcome model and simplifies the computation of standard errors. This analytic framework is broadly applicable to binary and multi-valued mediators and outcomes. The paper provides a tutorial and reveals statistical properties of the parametric and non-parametric RMPW results through a series of Monte Carlo simulations.

The RMPW strategy overcomes some important limitations of the existing alternatives. Path analysis (Alwin & Hauser, 1975; Duncan, 1966; Wright, 1934) and structural equation modeling (SEM) (Bollen, 1989; Jo, 2008; Jöreskog, 1970; MacKinnon, 2008) have been the most commonly used techniques in psychological research for analyzing mediation. They require a series of strong assumptions including the assumption that the mediator model and the outcome model are correctly specified and that there should be no treatment-by-mediator interaction (Bullock, Green, & Ha, 2010; Holland, 1988; Sobel, 2008). The assumption of no treatment-by-mediator interaction is also required by two additional approaches that have been extended to mediation analysis, the instrumental variable (IV) method widely used by economists (Heckman & Robb, 1985; Kling, Liebman & Katz, 2007; Raudenbush, Reardon, & Nomi, 2012) and marginal structural models well known to epidemiologists (Coffman & Zhong, 2012; Robins, 2003; Robins & Greenland, 1992). Treatment-by-mediator interactions will bias the estimates of direct and indirect effects produced by each of these techniques. Appendix A derives the bias term for path analysis models.

Recently, some new analytic strategies have emerged that relax the no-treatment-by-mediator assumption. These include modified regression approaches (Pearl, 2010; Petersen, Sinisi, & van der Lann, 2006; Preacher, Rucker, & Hayes, 2007; Valeri & VanderWeele, 2013; VanderWeele & Vansteelandt, 2009, 2010), direct effect models (van der Lann & Petersen, 2008), conditional structural models (VanderWeele, 2009), and a resampling approach (Imai, Keele, and Yamamoto, 2010; Imai, Keele, & Tingley, 2010). While these methods are more flexible than the conventional approaches, correct specification of the outcome model involving multi-way interactions among the treatment, the mediator, and the covariates is always crucial for generating unbiased estimates of the direct and indirect effects. Their implementation

typically requires extensive programming or specialized software. Most importantly, none of these methods estimates the treatment effect on the outcome transmitted through a change in the mediational process.

We illustrate the RMPW strategy with an analysis of the impact of a welfare-to-work program on maternal depression mediated by employment experience when there is evidence that employment (the mediator) affects depression (the outcome) differently under different policy conditions (the treatment). The application example is described in the next section, followed by definitions of the causal parameters, the theoretical rationale for using RMPW to identify the causal effects of interest, the identification assumptions, and the parametric and non-parametric weighting procedures applied to binary mediators. After presenting the simulation results, we show extensions to causal mediation moderated by pretreatment characteristics, to multi-category mediators, and to quasi-experimental data. The last section discusses the relative strengths and potential limitations of the RMPW strategy and raises issues for future research.

Application Example

In the late-1990s, the US government's six decade-long welfare cash assistance program (i.e., Aid to Families with Dependent Children, AFDC) was replaced nationwide by a new program (i.e., Temporary Assistance for Needy Families, TANF). This change in federal policy was heavily influenced by experiments conducted earlier in the decade, which showed increased employment and earnings for welfare recipients as a result of employment-focused incentives and services (Michalopoulos, Schwartz, & Adams-Ciardullo, 2001).

We use data from one such experiment, the National Evaluation of Welfare-to-Work Strategies Labor Force Attachment program (henceforth LFA) in Riverside, California. At the program orientation, all applicants to the AFDC program and current recipients who were not

working full time (defined as 30 or more hours per week) were randomly assigned to either the LFA program or the control condition. Individuals assigned to the control condition continued to receive public assistance from AFDC and eligibility-focused case management. The LFA program primarily targeted non-workers; individuals assigned to LFA who were already working 15 to 29 hours per week did not have to participate in program activities.

The LFA program included four key components: (1) *employment-focused case management*, including encouragement, support, and an emphasis on taking any job that became available; (2) *Job Club*, a class focused on skill building, resources, and support for job searching; (3) *job developers*, who worked with businesses and nonprofits in the community to identify jobs that might be filled by program participants; and (4) *sanctions* that penalized non-compliance in program activities or work by reducing LFA group members' welfare benefits. A key feature of LFA in Riverside is that it encouraged, but did not guarantee, employment among treatment group members. The four components of the program were designed to improve the likelihood that a welfare recipient would find employment, but the program did not provide jobs or strictly enforce employment.

As expected, the program increased employment and earnings and reduced welfare receipt (Hamilton, Greedman, Gennetian, Michalopoulos, Walter, Adams-Ciardullo, Gassman-Pines, McGorder, Zaslow, Ahluwalia, & Brooks, 2001). Yet an additional concern is that low-income single mothers with young children experience disproportionately high rates of depressive symptoms and clinical depression (Coiro, 2001; Moore, Zaslow, Coiro, Miller, & Magenheim, 1995; Siefert, Bowman, Heflin, Danziger, & Williams, 2000). Despite rhetoric and past evidence suggesting that welfare-to-work programs would benefit or harm the psychological well-being of welfare recipients (Cheng, 2007; Jagannathan, Camasso, & Sambamoorthi, 2010;

Knab, McLanahan, & Garfnikel, 2008; Morris, 2008), LFA in Riverside did not show a statistically significant total effect on maternal depression (Hamilton et al., 2001).

Importantly, the null total effect does not rule out possible mediation. We propose two distinct scenarios in which the null total effect of the program on maternal depression would mask mediated effects. In both cases, the direct and indirect effects of the program on an individual could offset one another. First, program-induced employment might benefit a participant's mental health by boosting self-efficacy (a positive indirect effect due to a change in the mediator value), while other aspects of the program, such as the threat of sanctions, might be stressful and adversely affect the participant's mental health (a negative direct effect). If similar in size, these countervailing effects could result in a null total effect. Second, program expectations with regard to employment and the threat of sanctions could alter the relationship between employment and depression, such that employment would be more beneficial, and lack of employment more detrimental, to psychological well-bring if a mother was assigned to the LFA program than if she was assigned to the control condition. This second scenario, a classic case of treatment-by-mediator interaction, highlights a positive indirect effect due to a change in the mediational process, which again could be offset by a negative direct effect.

In this application we will investigate (1) whether the effect of employment on depression depended on treatment assignment, (2) whether through increasing employment, the program generated an indirect effect that either heightened or reduced depression, and (3) whether being assigned to the LFA program would have had a beneficial or detrimental direct effect had there been no change in employment.

Our sample includes 208 LFA group members and 486 control group members with a child aged 3 to 5 years. 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 from the Center for Epidemiology Studies—Depression Scale (CES-D; Radloff, 1977) 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 summary score ranged from 0 to 34 with a mean equal to 7.49 and a standard deviation equal to 7.74. The data had complete information on policy assignment and employment record.

The baseline survey given to NEWWS participants provided rich information about participant characteristics shown previously to be important predictors of both employment and depressive symptoms. These include measures of: (a) maternal psychological well-being; (b) history of employment and welfare use, employment status, earnings, and income in the quarter prior to randomization; (c) human capital; (d) personal attitudes toward employment, including the preference to work, willingness to accept a low-wage job, shame to be on welfare; (e) perceived social support and barriers to work; (f) practical support and barriers to work such as childcare arrangement and extra family burden; (g) household composition, including number and age of children and marital status; (h) teen parenthood; (i) public housing residence and residential mobility; and (j) demographic features including age and race/ethnicity.

Causal Parameters

Notation. Let A denote random assignment; Z, employment experience during the two years after randomization; and Y, 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. For simplicity, we start with mediators measured on a binary scale, that is, $Z = 1$ if ever

employed and $Z = 0$ if never employed during the two-year period. We will show later that our logic applies to multi-valued mediators as well.

Instead of using path coefficients to define the causal effects in mediation problems, we define the person-specific causal effects in terms of the counterfactual outcomes. The definitions correspond to our substantive research questions and involve no structural models because, as we argued earlier, the structural models are unknown to the researchers. Table 1 provides a glossary for all the causal effects defined below.

1. What is the treatment effect on the mediator?

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. Of these two potential outcomes, one is observed and the other is an unobserved counterfactual. The person-specific causal effect of being assigned to the LFA program versus control on a mother's employment is $Z_1 - Z_0$. The only assumptions implied by this definition are (1) that one's employment is affected only by one's own treatment assignment and is not affected by other individuals' treatment assignment and (2) that one's employment associated with a given treatment does not depend on whether the individual selected the treatment on her own or was assigned at random to the treatment (Rubin, 1986). Yet we allow each potential mediator value to be possibly altered by random events within or beyond the control of the experimenter. For example, a participant assigned to the LFA program may remain unemployed due to an economic downtown or an unexpected health problem of a family member.

2. What is the treatment effect on the outcome?

To define the LFA program effect on maternal depression, we use Y_1 to denote a mother's potential psychological outcome if assigned to the LFA program and Y_0 for the

potential outcome if assigned to the control condition. The person-specific treatment effect on a mother's depression is $Y_1 - Y_0$. Because each potential outcome in this case is also a function of the potential employment experience corresponding to the given treatment assignment, to be specific, we may write Y_1 and Y_0 as Y_{1Z_1} and Y_{0Z_0}, respectively. The first subscript "1" or "0" denotes the treatment that one could potentially be assigned to, and the second subscript "Z_1" or "Z_0" denotes the subsequent employment experience that one would potentially have in correspondence with the treatment.

3. What is the effect of the mediator on the outcome under each treatment condition?

As we have reasoned earlier, employment may affect depressive symptoms differently depending on whether the individual was assigned to the LFA program or the control condition. Let Y_{11} denote a mother's depression level if she was assigned to the LFA program and employed, and let Y_{10} denote her depression level if she was assigned to the LFA program and unemployed. Here the first subscript represents the assignment to the LFA program while the second subscript represents whether one is employed or not. The causal effect of employment on maternal depression if the mother was assigned to the LFA program is defined as $Y_{11} - Y_{10}$. In parallel, let Y_{01} denote the mother's depression level if she was assigned to the control condition and employed, and let Y_{00} denote her depression level if she was assigned to the control condition and unemployed. The causal effect of employment on maternal depression if she was assigned to the control condition is defined as $Y_{01} - Y_{00}$. The effect of employment on maternal depression depends on the treatment condition if $Y_{11} - Y_{10} \neq Y_{01} - Y_{00}$.

4. What is the direct effect of the treatment on the outcome?

We use Y_{1Z_0} to denote a mother's counterfactual outcome if assigned to the LFA program yet experiencing employment as she would have had under the control condition. The direct effect of the policy on depression is defined by $Y_{1Z_0} - Y_{0Z_0}$, representing the effect of the policy on maternal depression if the policy, perhaps counterfactually, failed to change one's employment experience. The direct effect is to be attributed to other unspecified mediational processes independent of employment. For example, the threat of sanctions under LFA might heighten depression while interactions with caseworkers in the LFA program might lead to improved access to community mental health services. In the latest literature on causal mediation, Pearl (2001) labeled this the "natural direct effect" because the mediator value under the control condition Z_0 is allowed to vary naturally across participants.

5. What is the indirect effect of the treatment on the outcome?

To determine whether employment mediates the treatment effect on depression, we ask whether a mother assigned to the LFA program would become more or less depressed should she counterfactually experience the same level of employment as she would under the control condition. Denoted by $Y_{1Z_1} - Y_{1Z_0}$, the indirect effect represents the change in a mother's depressive symptoms under LFA solely attributable to the policy-induced change in her employment experience (i.e., a change from Z_0 to Z_1). This has been called "the natural indirect effect" in Pearl's (2001) terminology and was instead called "the total indirect effect" according to Robins and Greenland (1992).

6. What is the indirect effect if the treatment changes the mediational process?

As we reasoned earlier, the LFA program relative to the control condition may affect maternal depression partly through increasing employment and partly through altering the mediational process, such that employment would be more beneficial under the LFA program

than under the control condition. In such cases, conceptually we may further decompose the indirect effect into two elements. The first element $Y_{0Z_1} - Y_{0Z_0}$ is the change in a mother's depressive symptoms *under the control condition* should her employment increase by an amount that the policy could induce. Robins and Greenland (1992) called this "the pure indirect effect." We hypothesize that, should the same increase in employment occur *under LFA*, there might be a greater change (positive or negative) in the mother's depressive symptoms. Hence the second element of the indirect effect, denoted by $\left(Y_{1Z_1} - Y_{1Z_0}\right) - \left(Y_{0Z_1} - Y_{0Z_0}\right)$, represents how the policy-induced change in employment would affect the mother's depression differently between the experimental condition and the control condition. We call this "the natural treatment-by-mediator interaction effect" because it is a function of the treatment effect on the mediational process in the natural world. The total effect of the treatment on the outcome $Y_{1Z_1} - Y_{0Z_0}$ is the sum of the direct effect and the indirect effect, and the latter is the sum of the pure indirect effect and the natural treatment-by-mediator interaction effect.

Table 2 illustrates the concepts with six participants, three of which were assigned to the LFA group, and three to the control group. For each participant, we list two potential mediator values corresponding to the two possible treatment conditions and three potential outcomes. For the first three participants, the only observables are Z_1 and Y_{1Z_1}; while for the second three, the only observables are Z_0 and Y_{0Z_0}. For example, the first participant would be employed when assigned to LFA ($Z_1 = 1$) and would also be employed if counterfactually assigned to the control condition instead ($Z_0 = 1$). The treatment effect on her employment is therefore zero ($Z_1 - Z_0 = 0$). Because employment does not serve as a mediator for this participant, the indirect effect of the treatment on depression is also zero ($Y_{1Z_1} - Y_{1Z_0} = Y_{11} - Y_{11} = 0$), while the direct

effect ($Y_{1Z_0} - Y_{0Z_0} = Y_{11} - Y_{01}$) equals the total effect ($Y_{1Z_1} - Y_{0Z_0} = Y_{11} - Y_{01}$). In contrast, the second participant would be employed when assigned to LFA ($Z_1 = 1$) but would be unemployed if counterfactually assigned to the control condition ($Z_0 = 0$). Hence the treatment has a non-zero effect on her employment ($Z_1 - Z_0 = 1$). In this case, the indirect effect of the treatment on her depression through the treatment-induced change in her employment is possibly nonzero ($Y_{1Z_1} - Y_{1Z_0} = Y_{11} - Y_{10}$).

Because researchers do not have observations of the counterfactual mediator values and the counterfactual outcome values, the individual-specific treatment effects can never be calculated. Yet research designs and analytic strategies can be employed, when certain identification assumptions are satisfied, to estimate the population average causal effects. For instance, taking an average of the individual-specific treatment effect on the mediator over all the individuals in a population, we obtain the population average treatment effect on the mediator and denote it by $E(Z_1 - Z_0)$. Here $E(\cdot)$, read as "the expected value of", represents the population mean of a random quantity. Because the equation $E(Z_1 - Z_0) = E(Z_1) - E(Z_0)$ always holds, the population average treatment effect on employment can be viewed as the difference between $E(Z_1)$, the population average employment rate when all individuals in the population are hypothetically assigned to LFA, and $E(Z_0)$, the population average employment rate when all individuals are hypothetically assigned to the control condition. Similarly, we define the population average treatment effect on the outcome $E(Y_1 - Y_0)$, the population average mediator effect on the outcome under the LFA program $E(Y_{11} - Y_{10})$ and that under the control condition $E(Y_{01} - Y_{00})$, the population average direct effect of the treatment on the outcome $E(Y_{1Z_0} - Y_{0Z_0})$,

the population average indirect effect of the treatment on the outcome $E\left(Y_{1Z_1} - Y_{1Z_0}\right)$, the population average pure direct effect $E\left(Y_{0Z_1} - Y_{0Z_0}\right)$, and the population average natural treatment-by-mediator interaction effect $E\left[\left(Y_{1Z_1} - Y_{1Z_0}\right) - \left(Y_{0Z_1} - Y_{0Z_0}\right)\right]$.

Hypothetical Experimental Designs for Causal Mediation Analysis

It is well known that random treatment assignment enables one to estimate without bias the treatment effect on the mediator and the treatment effect on the outcome. Yet such a design does not generate an unbiased estimate of the mediator effect on the outcome under each treatment condition. Nor does it provide an unbiased decomposition of the total effect into a direct effect and an indirect effect. There is an emerging literature on experimental designs for investigating causal mediation mechanisms (Hong, 2006, 2013; Imai, Tingley, & Yamamoto, 2013; Mattei & Mealli, 2011; Sobel & Stuart, 2012; Spencer, Zanna, & Fong, 2005). A review and comparison across these different designs are beyond the scope of this paper. Below we discuss several simplest hypothetical experimental designs in the context of the current application. Even though these designs are often infeasible in practice, they are foundational for developing our analytic framework for causal mediation analysis when only the treatment is randomly assigned.

Three- and four-treatment arm experimental designs. To decompose the total effect into a direct effect and an indirect effect, Sobel and Stuart (2012) proposed a three-treatment arm experimental design for causal mediation analysis. Applying this design to the current example, one might assign welfare applicants at random to three treatment conditions: (1) the control condition, (2) the LFA program, and (3) the LFA program in which each participant's employment would take on a value associated with the counterfactual control condition. The observed mean outcomes of these three treatment groups would be unbiased estimates of the

respective population average potential outcomes $E(Y_{0Z_0})$, $E(Y_{1Z_1})$, and $E(Y_{1Z_0})$. The mean difference in the observed outcome between groups (2) and (3) estimates the indirect effect while that between groups (3) and (1) estimates the direct effect. To further decompose the indirect effect into the pure indirect effect and the natural treatment-by-mediator interaction effect, we would add a fourth treatment arm by assigning participants at random to (4) the control program in which each participant's employment would take on a value associated with the counterfactual LFA program. The mean difference in the observed outcome between groups (4) and (1) estimates the pure indirect effect while the difference between the (2)-(3) contrast and the (4)-(1) contrast estimates the natural treatment-by-mediator interaction effect.

Sequential randomized design. To estimate the mediator effect on the outcome under each treatment condition would require a different randomized design. When the treatment and the mediator are both binary, these causal effects involve four potential outcomes— $E(Y_{11})$, $E(Y_{10})$, $E(Y_{01})$, and $E(Y_{00})$. One could apply a sequential randomized experiment that, in the first step, would assign welfare applicants at random to either the LFA program or the control condition. In the second step, one would assign applicants within each treatment group at random to be either employed or unemployed. Suppose that according to earlier research, the employment rate would be 40% in the control group and 65% in the LFA group, the second step randomization would assign the control units to employment at random with a probability .4 and would assign the LFA units to employment with a probability .65. The mean observed outcome obtained from each of the four treatment-by-employment combinations would provide an unbiased estimate of the corresponding population average potential outcome. One would thereby obtain an unbiased estimate of the employment effect on depression under LFA and that

under the control condition and test whether the employment effect on depression depends on the treatment condition.

RMPW-Based Analytic Framework for Causal Mediation Analysis

Unfortunately, the sequential design and the three- or four-treatment arm design are both impractical in the context of welfare-to-work programs. Welfare agencies are not generally in a position of offering jobs to applicants or assigning them at random to employment. While it might be conceivable to randomize employment in "New Deal"-type public jobs programs, sequential randomization still would not allow one to decompose the total treatment effect. Although the three- or four-treatment arm experiment could allow for decomposition, it is even more challenging to implement because it requires that the experimenter be able to predict whether a participant would be employed under LFA and, additionally, whether the same person would be employed under the control condition. We will show that the RMPW technique can be applied to either a sequential randomized design or to a standard randomized experiment (where only the treatment is randomized) in order to approximate a three- or four-treatment arm design for decomposing the total effect. When only the treatment is randomized, RMPW adjusts for the potential bias due to the non-random mediator value assignment.

RMPW under Sequential Randomization

From the first step of a sequential design, the observed mean outcome of the control group and that of the LFA group are unbiased estimates of the population average potential outcomes associated with the control condition and with the LFA program, respectively. In the second step, once employment has been randomized under each treatment condition, through applying RMPW to the LFA group, one obtains an unbiased estimate of the population average

potential outcome associated with the LFA program when each participant's employment takes on a value associated with the counterfactual control condition.

The rationale for RMPW can be derived from the inherent connections between the two sets of population average potential outcomes listed in Table 2. First, the average potential outcome associated with LFA $E(Y_{1Z_1})$ is the average of the potential outcome of employment under LFA $E(Y_{11})$ and that of unemployment under LFA $E(Y_{10})$ proportionally weighted by the employment rate and the unemployment rate, respectively, under LFA:

$$E(Y_{1Z_1}) = E(Y_{11}) \times pr(Z_1 = 1) + E(Y_{10}) \times pr(Z_1 = 0).$$

Here $pr(Z_1 = 1)$ is the employment rate and $pr(Z_1 = 0)$ the unemployment rate if the entire population would be assigned to LFA. Second, the average potential outcome associated with the control condition $E(Y_{0Z_0})$ is the average of the potential outcome of employment under the control condition $E(Y_{01})$ and that of unemployment under the control condition $E(Y_{00})$ proportionally weighted by the employment rate and the unemployment rate, respectively, under the control condition:

$$E(Y_{0Z_0}) = E(Y_{01}) \times pr(Z_0 = 1) + E(Y_{00}) \times pr(Z_0 = 0).$$

Finally, the average potential outcome associated with LFA when each individual's employment would counterfactually remain the same as that under the control condition $E(Y_{1Z_0})$ is the average of the potential outcome of employment under LFA and that of unemployment under LFA proportionally weighted by the employment rate and the unemployment rate, respectively, under the *control* condition:

$$E(Y_{1Z_0}) = E(Y_{11}) \times pr(Z_0 = 1) + E(Y_{10}) \times pr(Z_0 = 0).$$

The above derivation has made clear that, to estimate $E(Y_{1Z_0})$ from the observed data in a sequential design, we may simply transform the employment rate and the unemployment rate in the LFA group to resemble those in the control group. The transformation can be done through weighting because

$$E(Y_{1Z_0}) = E(Y_{11}) \times pr(Z_0 = 1) + E(Y_{10}) \times pr(Z_0 = 0)$$

$$= E(Y_{11}) \times \frac{pr(Z_1 = 1)}{pr(Z_1 = 1)} \times pr(Z_0 = 1) + E(Y_{10}) \times \frac{pr(Z_1 = 0)}{pr(Z_1 = 0)} \times pr(Z_0 = 0)$$

$$= E\left(\frac{pr(Z_0 = 1)}{pr(Z_1 = 1)} \times Y_{11} \right) \times pr(Z_1 = 1) + E\left(\frac{pr(Z_0 = 0)}{pr(Z_1 = 0)} \times Y_{10} \right) \times pr(Z_1 = 0).$$

Here the potential outcome of employment under LFA Y_{11} is weighted by the ratio of the probability of employment under the control condition to that under LFA, $pr(Z_0 = 1)/pr(Z_1 = 1)$; in parallel, the potential outcome of unemployment under LFA Y_{10} is weighted by the ratio of the probability of unemployment under the control condition to that under LFA, $pr(Z_0 = 0)/pr(Z_1 = 0)$.

In analyzing data from a sequential design, we obtain the proportion of units employed in the control group, $pr(Z = 1 | A = 0)$, as an unbiased estimate of $pr(Z_0 = 1)$; the proportion of units employed in the LFA group, $pr(Z = 1 | A = 1)$, is an unbiased estimate of $pr(Z_1 = 1)$. Hence $pr(Z = 1 | A = 0)/pr(Z = 1 | A = 1)$ estimates the RMPW for the employed LFA units, while $pr(Z = 0 | A = 0)/pr(Z = 0 | A = 1)$ estimates the RMPW for the unemployed LFA units.

To estimate the direct effect and the indirect effect, we may combine the control group and the LFA group with a duplicate set of the LFA group. Let *D*1 be a dummy indicator that takes value 1 for the duplicate LFA units and 0 otherwise. The weight is 1.0 for the control units

(i.e., $A = 0$, $D1 = 0$), is $pr(Z_0 = 1)/pr(Z_1 = 1)$ for the employed LFA units (i.e., $A = 1$, $D1 = 0$, $Z = 1$) and $pr(Z_0 = 0)/pr(Z_1 = 0)$ for the unemployed LFA units (i.e., $A = 1$, $D1 = 0$, $Z = 0$), and is 1.0 for the duplicate LFA units (i.e., $A = 1$, $D1 = 1$). These three weighted groups approximate data from a three-treatment arm design. We then regress the outcome Y on the treatment indicator A and the indicator for LFA duplicate $D1$ in a weighted model:

$$Y = \gamma^{(0)} + \gamma^{(DE)}A + \gamma^{(IE)}D1 + e. \quad (1)$$

Here $\gamma^{(0)}$ estimates $E(Y_{0Z_0})$; $\gamma^{(0)} + \gamma^{(DE)}$ estimates $E(Y_{1Z_0})$; and $\gamma^{(0)} + \gamma^{(DE)} + \gamma^{(IE)}$ estimates $E(Y_{1Z_1})$. Hence $\gamma^{(DE)}$ estimates the average direct effect $E(Y_{1Z_0} - Y_{0Z_0})$ and $\gamma^{(IE)}$ estimates the average indirect effect $E(Y_{1Z_1} - Y_{1Z_0})$. To account for the duplication of every LFA unit, we identify individual units as clusters and obtain cluster-robust standard errors.

If the research interest also lies in estimating the pure indirect effect and the natural treatment-by-mediator interaction effect, it will become necessary to estimate $E(Y_{0Z_1})$ from the observed data. In a sequential design, we may simply transform the employment rate and the unemployment rate in the control group to resemble those in the experimental group.

$$E(Y_{0Z_1}) = E(Y_{01}) \times pr(Z_1 = 1) + E(Y_{00}) \times pr(Z_1 = 0)$$

$$= E\left(\frac{pr(Z_1 = 1)}{pr(Z_0 = 1)} \times Y_{01}\right) \times pr(Z_0 = 1) + E\left(\frac{pr(Z_1 = 0)}{pr(Z_0 = 0)} \times Y_{00}\right) \times pr(Z_0 = 0).$$

To implement, we may additionally create a duplicate set of the control group to approximate the fourth treatment arm in a hypothetical four-treatment arm design. Let $D0$ be a dummy indicator that takes value 1 for the duplicate control units and 0 otherwise. The weight is $pr(Z_1 = 1)/pr(Z_0 = 1)$ for the duplicate set of the employed control units (i.e., $A = 0$, $D0 = 1$, $Z =$

1) and is $pr(Z_1 = 0)/pr(Z_0 = 0)$ for the duplicate set of the unemployed control units (i.e., $A = 0$, $D0 = 1$, $Z = 0$). The weighting scheme is summarized in Table 3a.

We then conduct a weighted analysis regressing the outcome Y on the treatment indicator A, the indicator for LFA duplicates $D1$, and the indicator for control duplicates $D0$:

$$Y = \gamma^{(0)} + \gamma^{(DE)}A + \gamma^{(IE.1)}D1 + \gamma^{(IE.0)}D0 + e. \quad (2)$$

Here $\gamma^{(IE.1)}$ estimates the total indirect effect $E(Y_{1Z_1} - Y_{1Z_0})$, $\gamma^{(IE.0)}$ estimates the pure indirect effect $E(Y_{0Z_1} - Y_{0Z_0})$, and hence $\gamma^{(IE.1)} - \gamma^{(IE.0)}$ estimates the natural treatment-by-mediator interaction effect. Its standard error is $Var(\hat{\gamma}^{(IE.1)}) + Var(\hat{\gamma}^{(IE.0)}) - 2Cov(\hat{\gamma}^{(IE.1)}, \hat{\gamma}^{(IE.0)})$.

RMPW under Random Treatment Assignment

The NEWWS data are representative of many applications in which only the treatment is randomized. Within each treatment group, some individuals might have a higher likelihood of employment than others due to their prior education and training, personal predispositions, past employment experience, and family situations. Suppose that an individual's probability of employment under a given treatment is a function of the observed pretreatment characteristics **X**. We may envision that the data approximate a sequential randomized block design in which individuals with homogeneous pretreatment characteristics **X** = **x** constitute blocks. Those in the same block are hypothetically randomized first to LFA or the control condition and subsequently to employment or unemployment. In the current study, hypothetical randomization to employment within each block could be a result of unpredictable events in the nature. Hence in a given block, the observed mediator values of individuals assigned to the control condition provide counterfactual information of the mediator values that the LFA units would likely display should they be assigned to the control condition instead; Similarly, the LFA unit's

observed mediator values are what the control units if assigned to LFA instead would counterfactually display. Hence we apply RMPW to the "as-if" sequential randomized data within each block and summarize the results over all blocks.

We estimate RMPW as functions of the pretreatment characteristics that determine one's block membership. To be specific, for estimating $E(Y_{1Z_0})$, an LFA unit displaying pretreatment characteristics **x** would be weighted by $pr(Z_0 = 1 | \mathbf{X} = \mathbf{x}) / pr(Z_1 = 1 | \mathbf{X} = \mathbf{x})$ if the unit were employed. Here $pr(Z_0 = 1 | \mathbf{X} = \mathbf{x})$ can be estimated by the proportion of control units employed in this particular block, denoted by $pr(Z = 1 | A = 0, \mathbf{X} = \mathbf{x})$; while $pr(Z_1 = 1 | \mathbf{X} = \mathbf{x})$ can be estimated by the proportion of LFA units employed in the same block, denoted by $pr(Z = 1 | A = 1, \mathbf{X} = \mathbf{x})$. If the LFA unit were unemployed, the weight would be $pr(Z_0 = 1 | \mathbf{X} = \mathbf{x}) / pr(Z_1 = 1 | \mathbf{X} = \mathbf{x}) = pr(Z = 1 | A = 0, \mathbf{X} = \mathbf{x}) / pr(Z = 1 | A = 1, \mathbf{X} = \mathbf{x})$. After generating a duplicate set of the LFA units, a weighted analysis of model (1) estimates the direct effect and the indirect effect.

Similarly, in order to estimate $E(Y_{0Z_1})$, RMPW transforms the employment rate in the control group within each block to resemble that in the LFA group. The weight is $pr(Z_1 = 1 | \mathbf{X} = \mathbf{x}) / pr(Z_0 = 1 | \mathbf{X} = \mathbf{x}) = pr(Z = 1 | A = 1, \mathbf{X} = \mathbf{x}) / pr(Z = 1 | A = 0, \mathbf{X} = \mathbf{x})$ for the employed control units and is $pr(Z_1 = 0 | \mathbf{X} = \mathbf{x}) / pr(Z_0 = 0 | \mathbf{X} = \mathbf{x}) = pr(Z = 0 | A = 1, \mathbf{X} = \mathbf{x}) / pr(Z = 0 | A = 0, \mathbf{X} = \mathbf{x})$ for the unemployed control units. Applying the weight to a duplicate set of the control units and analyzing model (2), we obtain estimates of the pure indirect effect and the natural treatment-by-mediator interaction effect. These theoretical results are summarized in Table 3(b).

RMPW for Multivalued Mediators

This framework can be extended easily to multivalued mediators. We will show in a later section an example in which employment is measured on a 3-point scale, denoted by $z = 0, 1, 2$ for unemployment, low employment, and high employment, respectively. To estimate $E(Y_{1Z_0})$, we apply to LFA units the following weight:

$$\omega = \frac{pr(Z_0 = z \mid \mathbf{X} = \mathbf{x})}{pr(Z_1 = z \mid \mathbf{X} = \mathbf{x})} = \frac{pr(Z = z \mid A = 0, \mathbf{X} = \mathbf{x})}{pr(Z = z \mid A = 1, \mathbf{X} = \mathbf{x})}.$$

Here $pr(Z = z \mid A = 0, \mathbf{X} = \mathbf{x})$ is the proportion of control units with pretreatment characteristics $\mathbf{X} = \mathbf{x}$ who experienced employment level *z*; and $pr(Z = z \mid A = 1, \mathbf{X} = \mathbf{x})$ is the proportion of LFA units with pretreatment characteristics $\mathbf{X} = \mathbf{x}$ who experienced employment level *z*. To estimate $E(Y_{0Z_1})$, we apply to control units the weight:

$$\omega = \frac{pr(Z_1 = z \mid \mathbf{X} = \mathbf{x})}{pr(Z_0 = z \mid \mathbf{X} = \mathbf{x})} = \frac{pr(Z = z \mid A = 1, \mathbf{X} = \mathbf{x})}{pr(Z = z \mid A = 0, \mathbf{X} = \mathbf{x})}.$$

Identification Assumptions

Here we summarize specific identification assumptions under which one may employ RMPW when only the treatment assignment is randomized. In essence, we require that the data resemble what one would obtain from a sequential randomized block design. That is, individuals who share the same observed pretreatment characteristics would have the same probability of employment. Data from such a hypothetical design would satisfy the following assumptions:

Assumption 1: Nonzero probability of treatment assignment. Every unit in the population has a nonzero probability of being assigned to each treatment condition.

Assumption 2: No confounding of treatment-outcome relationship. Treatment assignment is independent of the potential outcomes.

In the NEWWS study, because of the randomized treatment assignment, these two assumptions are both satisfied. Hence the mean observed outcome of the control units provides an unbiased estimate of $E(Y_{0Z_0})$ while the mean observed outcome of the LFA units provides an unbiased estimate of $E(Y_{1Z_1})$.

Assumption 3: No confounding of treatment-mediator relationship. Treatment assignment is independent of the potential intermediate outcomes.

Assumption 4: Nonzero probability of mediator value assignment. Within levels of the observed pretreatment characteristics, every unit has a nonzero probability of being assigned to each mediator value under each treatment condition.

Assumption 5: No confounding of mediator-outcome relationship within a treatment. Within levels of the observed pretreatment characteristics and under a given treatment condition, mediator value assignment is independent of the potential outcomes.

Assumption 6: No confounding of mediator-outcome relationship across treatment conditions. Within levels of the observed pretreatment characteristics, mediator value assignment under a given treatment condition is independent of potential outcomes associated with an alternative treatment condition.

The treatment randomization guaranteed that Assumption 3 is satisfied. Assumption 4 represents a *probabilistic* view of mediator value assignment. That is, given that the job market was at least partially governed by uncertainty, many of those who were unemployed under the LFA condition may have had a nonzero probability of being employed; similarly, many of those who were employed under the control condition may have had a nonzero probability of becoming unemployed. Assumptions 5 and 6 together require that the observed pretreatment covariates adequately account for all the potential confounding of the mediator-outcome

relationships within and across the treatment conditions. Robins (2003) argued that it is unlikely that one would accept Assumption 6 unless one believed that the mediator value assignment was randomized by nature within levels of the pretreatment covariates. The above six assumptions constitute the "sequential ignorability" (Imai, Keele, and Yamamoto, 2010; Imai, Keele, & Tingley, 2010), that is, the treatment assignment and the mediator value assignment under each treatment can be viewed as randomized within levels of the observed pretreatment covariates. Under all six assumptions, the RMPW adjusted mean observed outcome of the LFA units provides an unbiased estimate of $E\left(Y_{1Z_0}\right)$ while the RMPW adjusted mean observed outcome of the control units provides an unbiased estimate of $E\left(Y_{0Z_1}\right)$. Unlike many of the existing methods, the RMPW strategy does not require the no treatment-by-mediator interaction assumption.

Parametric RMPW Procedure

We describe a parametric procedure for estimating RMPW in this section and a nonparametric procedure in the next section for binary mediators. The parametric approach estimates RMPW directly as a ratio of the estimated conditional probability of mediator value assignment under the control condition to that under the LFA condition.

Step 1: Select and prepare the pretreatment covariates. We have selected 86 pretreatment covariates that are theoretically associated with maternal depression or with employment experience. After creating a missing category for each categorical covariate with missing information, we impute the missing data in the outcome and the continuous covariates and generate five impute data sets (Little & Rubin, 2002). We then carry out steps 2 through 7 with each imputed data set one at a time and, at the end, combine the estimated causal effects over the five imputed data sets. For simplicity, below we discuss the analytic procedure with the first imputed data set.

Step 2: Specify the propensity score model for the mediator under each treatment condition. Analyzing data from the LFA group, we predict an LFA unit's conditional probability of (i.e., the propensity score for) employment under LFA, denoted by $\theta_{Z_1} = pr(Z_1 = 1 | \mathbf{X} = \mathbf{x}) = pr(Z = 1 | A = 1, \mathbf{X} = \mathbf{x})$, as a function of the unit's observed pretreatment characteristics. After stepwise selection of the outcome and mediator predictors, the propensity score model is analyzed through logistic regression (Rosenbaum & Rubin, 1984). Similarly, using data from the control group, we predict a control unit's conditional probability of employment under the control condition, denoted by $\theta_{Z_0} = pr(Z_0 = 1 | \mathbf{X} = \mathbf{x}) = pr(Z = 1 | A = 0, \mathbf{X} = \mathbf{x})$. By virtue of the random treatment assignment, the propensity score model specified under the control condition would apply to the LFA units had they been counterfactually assigned to the control condition instead. Hence applying the coefficient estimates obtained from this second propensity score model, we can predict each LFA unit's θ_{Z_0}, that is, the unit's propensity score for employment under the counterfactual control condition. Similarly, applying the coefficient estimates obtained from the LFA group, we predict each control unit's propensity score for employment under the counterfactual LFA condition θ_{Z_1}.

Step 3: Identify the common support for mediation analysis in each treatment group. Among those who display the same propensity score for employment given the treatment, the employed units are expected to have their unemployed counterparts and vice versa. Units who do not have counterparts are excluded from the subsequent mediation analysis due to their lack of counterfactual information. To implement, we compare the distribution of the logit of θ_{Z_1} and that of θ_{Z_0} across the employed LFA units, the unemployed LFA units, the employed control units, and the unemployed control units, and identify cases in which the distribution of either

propensity score does not overlap across all four groups. One may add 20% of a standard deviation of the logit of each propensity score at each end to expand the range of the common support (Austin, 2011). In this application, only two individuals from the control group are excluded. See Appendix B5 for the Stata code for identifying the common support.

Step 4: Check balance in covariate distribution across the treatment-by-mediator combinations. Even though the identification assumptions cannot be empirically verified, if after propensity score adjustment, a considerable proportion of the observed pretreatment covariates remains predictive of the mediator, we view this as evidence that the adjustment fails to approximate data from a sequential randomized block design. Specifically, applying inverse-probability-of-treatment weighting (IPTW) (Robins, 1999) to the current example, we assign the weight $pr(Z=1 \mid A=1)/\theta_{Z_1}$ to the employed LFA units, $pr(Z=0 \mid A=1)/(1-\theta_{Z_1})$ to the unemployed LFA units, $pr(Z=1 \mid A=0)/\theta_{Z_0}$ to the employed control units, and $pr(Z=0 \mid A=0)/(1-\theta_{Z_0})$ to the unemployed control units. We expect that, 95% of the time, a categorical covariate will show equal proportion distribution and that a continuous covariate will show equal mean and variance across these four groups. One may improve the balance through modifying the propensity score models.

Step 5: Estimate the mediator effect on the outcome under each treatment condition. This step produces useful evidence with regard to whether the mediator-outcome relationship differs by treatment. Applying IPTW to the data, we simply regress the outcome on the binary treatment, the binary mediator, and their interaction.

Step 6: Create a duplicate and compute the parametric weight. We then reconstruct the data within common support to include the control units and their duplicate and the LFA units and their duplicate. The rest of this step has been summarized in Table 3(b). Most importantly, to

estimate $E(Y_{1Z_0})$, the weight for the employed LFA units is $\theta_{Z_0}/\theta_{Z_1}$ and that for the unemployed LFA units is $(1-\theta_{Z_0})/(1-\theta_{Z_1})$; to estimate $E(Y_{0Z_1})$, the weight for the employed control units is $\theta_{Z_1}/\theta_{Z_0}$ and that for the unemployed control units is $(1-\theta_{Z_1})/(1-\theta_{Z_0})$.

Step 7: Estimate the causal effects. Finally, conducting a weighted analysis of model (1), we obtain estimates of the direct effect and the indirect effect along with a cluster-robust standard error for each estimate. Analyzing model (2), we additionally obtain estimates of the pure indirect effect and the natural treatment-by-mediator interaction effect. One may improve precision by making additional covariance adjustment for strong predictors of the outcome. Appendix B1 shows the Stata code for the parametric analysis.

Analyzing the NEWWS data, we first generate estimates of the treatment effect on the mediator and the treatment effect on the outcome. An intention-to-treat (ITT) analysis shows that assignment to LFA increased the employment rate from 39.5% to 65.4%. Another ITT analysis shows that the average treatment effect on depression cannot be statistically distinguished from zero (coefficient = 0.11, SE = 0.64, t = 0.18, p = 0.86).

According to the results from Step 5, the employment effect on depression differed by treatment. Specifically, employment reduced depressive symptoms under LFA (coefficient = -2.49, *SE* = 1.20, t = -2.07, $p < 0.05$) but not under the control condition (coefficient = 0.74, *SE* = 0.76, t = 0.97, p = 0.33). The treatment-by-mediator interaction is statistically significant (coefficient = 3.23, *SE* = 1.42, t = 2.27, $p < 0.05$). According to these results, had all welfare mothers continued to be covered by the old policy, employment would not have affected maternal depression by a significant amount. However, once employment became one of the primary qualifications for welfare receipt, employment success apparently would lead to a reduction in depressive symptoms.

After executing Step 6, we find that the weighted employment rates in the original control group, the duplicate LFA group, and the original LFA group are .395, .654, and .376, respectively. Step 7 then decomposes the total treatment effect. The estimated direct effect is 1.29 (*SE* = 0.87; *t* = 1.48, *p* = 0.14), about 17% of a standard deviation of the outcome; the estimated indirect effect is -0.87 (*SE* = 0.47; *t* = -1.87, *p* = 0.06). The direct effect estimate indicates that, if the assignment to LFA rather than to the control condition had counterfactually generated no impact on employment, maternal depression would not have increased by a statistically significant amount, on average. According to the indirect effect estimate, if all individuals were hypothetically assigned to LFA, the LFA-induced change in employment (i.e., the increase in employment rate from 39.5% to 65.4%) was almost great enough to produce a significant reduction in maternal depression, on average. Further decomposing the indirect effect into a "pure indirect effect" and a "natural treatment-by-mediator interaction effect," we find that, if all individuals were hypothetically assigned to the control condition instead, the same amount of change in employment as reported earlier would not have a statistically significant impact on the average level of depression (Coefficient = 0.32, *SE* = 0.27; *t* = 1.48, *p* = 0.14). The estimated natural treatment-by-mediator interaction effect is -1.19 (*SE* = 0.53; *t* = -2.26, $p < 0.05$), providing evidence that the LFA-induced increase in employment reduced depression under the LFA condition in a way that did not happen under the control condition.

Non-Parametric RMPW Procedure

In general, non-parametric analyses are more robust than their parametric counterparts because the former is less reliant on model-based assumptions. For example, past research has shown that, in evaluating the relative effectiveness of different treatments, parametric IPTW often generates biased results especially when the propensity score models are misspecified in

their functional forms (Hong, 2010b; Shaffer & Kang, 2008). In contrast, non-parametric weighting methods such as marginal mean weighting through stratification (MMW-S) produce robust results despite the misspecification of the propensity score models (Hong, 2010b, 2012). IPTW and MMW-S, however, are not suitable for decomposing the total effect into a direct effect and an indirect effect in the presence of treatment-by-mediator interaction. We develop a non-parametric RMPW procedure for mediation analysis and evaluate its performance in comparison with that of the parametric RMPW procedure through simulations.

In essence, the non-parametric RMPW procedure re-computes the conditional probability of employment under each treatment condition on the basis of propensity score stratification. It differs from the parametric RMPW procedure only in Steps 4, 5, and 6.

Step 4: Check balance in covariate distribution across the treatment-by-mediator combinations. Instead of checking balance in the data adjusted by IPTW, we apply the non-parametric MMW-S procedure to adjust for employment selection associated with the observed pretreatment covariates. We first rank the sampled units by θ_{Z_1} and divide the sample into three even portions. Within each of these three subclasses, we then rank and subdivide the units again by θ_{Z_0}. Let s = 1, …, 9 denote the nine strata. With a relatively large sample size, one may increase the number of strata along each propensity score dimension. Within stratum s, we then assign the weight $pr(Z=1 \mid A=1)/pr(Z=1 \mid A=1, S=s)$ to the employed LFA units,

$pr(Z=0 \mid A=1)/pr(Z=0 \mid A=1, S=s)$ to the unemployed LFA units,

$pr(Z=1 \mid A=0)/pr(Z=1 \mid A=0, S=s)$ to the employed control units, and

$pr(Z=0 \mid A=0)/pr(Z=0 \mid A=0, S=s)$ to the unemployed control units.

Here $pr(Z=1|A=1,S=s)$ is the proportion of LFA units in stratum *s* who were employed; $pr(Z=0|A=1,S=s)$ is the proportion of LFA units in stratum *s* who were unemployed; $pr(Z=1|A=0,S=s)$ and $pr(Z=0|A=0,S=s)$ represent, in stratum *s*, the respective proportions of control units who were employed and unemployed.

Step 5: Estimate the mediator effect on the outcome under each treatment condition. Applying MMW-S to the data, we regress the outcome on the treatment, the mediator, and their interaction, and test whether the mediator-outcome relationship depends on the treatment condition.

Step 6: Create a duplicate and compute the non-parametric weight. We estimate RMPW non-parametrically under the stratification described in Step 4. Within stratum *s*, we re-compute the conditional probability of employment under the control condition as the proportion of the control units in that stratum who were employed, denoted by $pr(Z=1|A=0,S=s)$. Similarly, we re-compute the conditional probability of employment under LFA in the same stratum as the proportion of the LFA units who were employed, denoted by $pr(Z=1|A=1,S=s)$. It is easy to show that the conditional probability of unemployment under the control condition in stratum *s* is $pr(Z=0|A=0,S=s)$ while the conditional probability of unemployment under LFA in the same stratum is $pr(Z=0|A=1,S=s)$.To estimate $E(Y_{1Z_0})$, the non-parametric weight for the employed LFA units is:

$$\omega = \frac{pr(Z=1|A=0,S=s)}{pr(Z=1|A=1,S=s)}.$$

The weight for the unemployed LFA units is

$$\omega = \frac{pr(Z=0|A=0,S=s)}{pr(Z=0|A=1,S=s)}.$$

To estimate $E(Y_{0Z_1})$, the non-parametric weight for the employed control units is:

$$\omega = \frac{pr(Z=1 \mid A=1, S=s)}{pr(Z=1 \mid A=0, S=s)}.$$

The weight for the unemployed control units is

$$\omega = \frac{pr(Z=0 \mid A=1, S=s)}{pr(Z=0 \mid A=0, S=s)}.$$

Applying the non-parametric RMPW to the regression model specified in Equation (1) , we estimate the direct effect and the indirect effect. Appendix B2 presents the Stata code for the non-parametric analysis. Under three-by-three stratification, the direct effect estimate is 0.57 (*SE* = 0.75, *t* = 0.75, *p* = 0.45); the indirect effect estimate is -0.19 (*SE* = 0.35, *t* = -0.53, *p* = 0.60). We then increase to four-by-four stratification for a higher percentage of bias removal. The direct effect estimate is 1.34 (*SE* = 0.79, *t* = 1.70, *p* = 0.09), and the indirect effect estimate is -0.93 (*SE* = 0.38, *t* = -2.43, $p < 0.05$). Further decomposing the indirect effect, we obtain an estimate of the pure indirect effect (Coefficient = 0.45, *SE* = 0.30, *t* = 1.50, *p* = 0.13) and an estimate of the natural treatment-by-mediator interaction effect (Coefficient = -1.38, *SE* = 0.49, *t* = -2.85, $p < 0.01$). The point estimates obtained under the four-by-four stratification are converging to the parametric weighting results. Yet the estimation with non-parametric weighting appears to be relatively more efficient. Hence we are able to detect a statistically significant negative indirect effect of the treatment. According to these results, the LFA-induced increase in employment rate would reduce maternal depression should the entire population be assigned to LFA. Additionally, there is clear evidence that the treatment changed the mediational process: the LFA-induced increase in employment would become beneficial to participants' mental health only under the LFA program while showing no benefit under the control condition.

Simulations

We conduct a series of Monte Carlo simulations to assess the performance of the non-parametric RMPW procedure relative to the parametric RMPW procedure in estimating the direct and indirect effects in the case of a binary randomized treatment, a binary mediator, and a continuous outcome. With non-parametric RMPW, we also compare three-by-three strata with four-by-four strata. Additionally, we compare the robustness of estimation between the parametric and the non-parametric procedures when the propensity score models are misspecified in their functional forms. The simulated data resemble the structure of the NEWWS Riverside data. 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 sample size, we generate 1,000 random samples.

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. We compare across three sets of parameter value specifications shown in Table 4. In our first contrast, the direct effect and the indirect effect are both set to be zero in simulation (a) and are nonzero in simulations (b) and (c). Secondly, we change the magnitude of the treatment effect on the mediator. Simulations (a) and (b) set the employment rates similar to those in the NEWWS data, while simulation (c) increases the employment rate in the LFA program and decreases that under the control condition, which essentially reduces the statistical power under the same total sample size.

The evaluation criteria for causal effect estimate $\hat{\gamma}$ (that is, either $\hat{\gamma}^{(DE)}$ or $\hat{\gamma}^{(IE)}$) include the following: (1) bias in the point estimate: $E(\hat{\gamma})-\gamma$; (2) sampling variability of the point

estimate: $Var(\hat{\gamma}) = E[\hat{\gamma} - E(\hat{\gamma})]^2$; (3) mean square error (MSE): $E[(\hat{\gamma} - \gamma)^2] = Var(\hat{\gamma}) + [E(\hat{\gamma}) - \gamma]^2$; and (4) bias in the standard error estimate: $E[\hat{\sigma}(\hat{\gamma})] - \sigma(\hat{\gamma})$. Results from a naïve analysis serve as the baseline for assessing the performance of the RMPW procedures. A naïve RMPW procedure is employed as if the data were from a sequential experimental design with both treatment randomization and employment randomization. We assess the extent to which the parametric and non-parametric RMPW procedures successfully removes bias associated with the pretreatment covariates.

Simulation Results

Correctly Specified Propensity Score Models. Table 5 summarizes the key results corresponding to the three sets of baseline parameter values when the propensity score models are correctly specified. The parametric RMPW and the non-parametric RMPW procedures both perform generally well in all three cases. The parametric procedure removes nearly 100% of the bias; the non-parametric procedure with three-by-three strata removes 85% or more of the initial bias while that with four-by-four strata removes 90% or more of the bias when the sample size is relatively large. However, in a relatively small sample, the advantage of increasing strata apparently disappears when $E(\theta_{Z_0})$ and $E(\theta_{Z_1})$ are shifting away from 0.5. With a relatively large sample size, the non-parametric estimates often show a higher efficiency and a smaller MSE when compared with the parametric estimates. However, with a relatively small sample size, an increase in the number of strata seems to result in a loss of efficiency. Finally, comparing the standard error estimates with the corresponding sampling standard deviations approximated on the basis of 1,000 samples, we find the average discrepancy close to zero across all cases and never exceeding 0.047 standard deviations of a potential outcome in any single case.

Misspecified Propensity Score Models. We then modify the data generation plan to allow for a comparison between the parametric and the non-parametric RMPW procedures when a nonlinear, non-additive propensity score model is misspecified as a linear additive one. According to our results (not tabulated here), regardless of sample size, the parametric RMPW procedure generates estimates that are increasingly biased as the degree of nonlinearity or non-additivity increases. In contrast, the non-parametric RMPW results remain robust in all cases.

Extensions of the RMPW Approach

The RMPW approach can be easily extended to an analysis of causal mediation mechanisms that may vary across subpopulations of units and to multi-category mediators. We also briefly discuss extensions to quasi-experimental data. Because the non-parametric RMPW strategy is sometimes constrained by sample size, in this section we present the parametric RMPW approach only and focus on the estimation of the direct effect and the indirect effect.

RMPW Procedure for Analyzing Moderated Mediation

A moderator defines subpopulations across which the treatment effect and the mediation mechanisms may differ. For example, comparing welfare recipients who had been teen parents in the past with those who had never become teen parents, we find that the assignment to LFA increased non-teen mothers' depressive symptoms while showing a zero effect for the teen mothers. We test whether the direct effect and the indirect effect of the treatment on maternal depression depend on teen parenthood status.

Let $V = 1$ represent a teen parent and $V = 0$ for a non-teen parent. To investigate teen parenthood as a potential moderator, conventionally one would conduct two-group comparisons in SEM, a strategy that shows limitations when the treatment and the mediator interact in either or both subpopulations. We apply the first six steps of the parametric RMPW procedure within

each subpopulation. In Step 7, we modify Equation (1) to include two sub-models, one for teen mothers and the other for non-teen mothers:

$$Y = V\left(\gamma_{V1}^{(0)} + \gamma_{V1}^{(DE)}A + \gamma_{V1}^{(IE)}D\right) + (1-V)\left[\gamma_{V0}^{(0)} + \gamma_{V0}^{(DE)}A + \gamma_{V0}^{(IE)}D\right] + e.$$

Here $\gamma_{V1}^{(DE)}$ and $\gamma_{V1}^{(IE)}$ estimate the direct effect and the indirect effect, respectively, for teen mothers; $\gamma_{V0}^{(DE)}$ and $\gamma_{V0}^{(IE)}$ estimate the direct effect and the indirect effect, respectively, for non-teen mothers. Additionally, we test whether $\gamma_{V1}^{(DE)} = \gamma_{V0}^{(DE)}$ and $\gamma_{V1}^{(IE)} = \gamma_{V0}^{(IE)}$ by simply re-parameterizing the above model as follows:

$$Y = \gamma_{V0}^{(0)} + \gamma_{V0}^{(DE)}A + \gamma_{V0}^{(IE)}D + V\gamma_{V1-0}^{(0)} + \gamma_{V1-0}^{(DE)}VA + \gamma_{V1-0}^{(IE)}VD + e.$$

where $\gamma_{V1-0}^{(DE)} = \gamma_{V1}^{(DE)} - \gamma_{V0}^{(DE)}$ and $\gamma_{V1-0}^{(IE)} = \gamma_{V1}^{(IE)} - \gamma_{V0}^{(IE)}$. Appendix B3 shows the Stata code for the moderated mediation analysis.

RMPW Procedure for A Multi-Category Mediator

To distinguish among participants who were employed to varying degrees, we examine a three-category measure: Never employed ($z = 0$), low employment ($z = 1$) (i.e., employed for no more than 50% of the two-year period), and high employment ($z = 2$) (i.e., employed for more than 50% of the two-year period). The direct effect and the indirect effect of treatment assignment on maternal depression are defined the same as before. Below we highlight the modifications in Steps 2~6 of the parametric RMPW procedure for analyzing a three-category mediator. The same procedure applies when the mediator contains more than three categories.

Step 2. Each unit now has three propensity scores under a given treatment condition. Specifically, $\theta_{Z_1=0}$, $\theta_{Z_1=1}$, and $\theta_{Z_1=2}$ represent the conditional probability of having zero employment, low employment, and high employment under LFA; $\theta_{Z_0=0}$, $\theta_{Z_0=1}$, and $\theta_{Z_0=2}$ represent the conditional probabilities of having these three levels of employment under the

control condition. A comparison between a multinomial logistic regression model and an ordinal model shows that, in this case, the latter fits the data as adequately as the former.

Step 3. We identify the range of the logit of $\theta_{Z_1=z}$ and that of $\theta_{Z_0=z}$ for *z* = 0, 1, 2 within which the distributions of the three employment groups under LFA and the three groups under the control condition overlap. Appendix B5 provides the Stata code for identifying the common support in the case of a multi-category mediator.

Step 4. Again we employ IPTW in balance checking by assigning weight $pr(Z=z \mid A=1)/\theta_{Z_1=z}$ to the LFA units displaying employment level *z* and assigning weight $pr(Z=z \mid A=0)/\theta_{Z_0=z}$ to the control units displaying employment level *z* for *z* = 0, 1, 2.

Step 5. Applying IPTW to the data, we regress the outcome on the binary treatment, dummy indicators for two of the three treatment levels, and their interactions.

Step 6. As before, we reconstruct the data set to include a duplicate set. The weight is again 1.0 for the original control units and the duplicate LFA units. For the original LFA units, the parametric RMPW is simply $\omega = \theta_{Z_0=z}/\theta_{Z_1=z}$.

Regardless of the distribution of the multi-category mediator, the outcome model is specified the same as that in model (1). The estimated direct effect is 1.24 (*SE* = 0.81, *t* = 1.54, *p* = 0.13); the estimated indirect effect is -.86 (*SE* = 0.37, *t* = -2.31, *p* = 0.02). The magnitude of these results is similar to the decomposition of the total effect when employment is measured on a binary scale. However, by considering employment on a three-category scale, the treatment effect decomposition becomes more precise. We are again able to detect a negative indirect effect that reaches the statistical significance level. The Stata code is shown in Appendix B4.

Extensions to Quasi-Experimental Data

A randomized experiment, especially those with a longitudinal design, often suffers from nonrandom attrition such that, among those in the remaining sample, the experimental group and the control group may become systematically different. When randomization is unfeasible, researchers typically analyze quasi-experimental data for evaluating treatment effects and for investigating mediation mechanisms. In all these cases, a wide array of statistical techniques is now available for reducing selection bias in treatment effect estimation. For example, IPTW and MMW-S can be employed to equate the observed pretreatment composition between the experimental group and the control group. To proceed with mediation analysis, one may carry out Steps 2 and 3 with the data already adjusted for treatment selection through such weighting. In Steps 4 and 5, the weight for adjusting treatment selection can be multiplied by the weight for adjusting mediator selection. And finally, the weight for adjusting treatment selection is multiplied by RMPW and then applied to Equation (1) in Step 7.

Conclusion and Discussion

When a treatment changes not only the distribution of a mediator but also how the mediator influences the outcome, the treatment-by-mediator interaction becomes an important component of the causal mediation mechanism. However, such data pose an analytic challenge when one attempts to decompose the total effect. Additionally, when only the treatment is randomized, it is necessary to remove potential confounding of the mediator-outcome relationship.

This paper has presented a new approach to causal mediation analysis that addresses these challenges. We have contributed to the conceptualization of causal mediation, first of all, by defining “the natural treatment-by-mediator interaction effect.” In the NEWWS application, we have found this interaction effect to be an essential component of the indirect effect reflecting

how the treatment-induced change in the mediational process transmitted the treatment effect on the outcome.

Conventional analysis typically ignores the interaction effect and therefore generates biased estimates of the indirect effect and the direct effect. The RMPW strategy reconstructs the data to estimate (1) the population average potential outcome should all the units be assigned to the control condition, (2) the population average potential outcome should all the units be assigned to the experimental condition, and (3) the population average potential outcome should all the units be assigned to the experimental condition yet the mediator values would counterfactually remain the same as that under the control condition. The above third hypothetical treatment arm is constructed by transforming the mediator distribution of the experimental group to resemble that of the control group. Contrasting the mean outcome between the groups, the outcome model generates a direct effect estimate and an indirect effect estimate along with their standard errors. To adjust for the selection of mediator values, the transformation of mediator distribution is conducted within subgroups of individuals who would respond similarly at the intermediate stage to the treatment given their pretreatment characteristics. To estimate the natural treatment-by-mediator interaction effect requires the estimation of (4) the population average potential outcome should all the units be assigned to the control condition yet the mediator would counterfactually take the same values as those under the experimental condition. We simply add a fourth hypothetical treatment arm by transforming the mediator distribution of the control group to resemble that of the experimental group.

This paper has delineated the analytic steps for implementing the RMPW strategy. According to the simulation results, the parametric and the non-parametric RMPW procedures both demonstrate satisfactory performance under the identification assumptions. While the

parametric RMPW results are sensitive to possible misspecifications of the functional form of the propensity score models, the nonparametric RMPW results are generally robust. We have shown RMPW extensions to analyses of moderated mediation effects, to multi-category mediators, as well as to quasi-experimental data.

Advantages of the RMPW Strategy

The RMPW strategy shows its strengths in comparison with the existing methods because it relies on relatively fewer identification assumptions and model-based assumptions, and because it can be implemented fairly easily with standard statistical software. In addition to assuming sequential ignorability, the conventional path analysis/SEM approach and the marginal structural models require the assumption that there is no treatment-by-mediator interaction. Latest advancements in causal mediation analysis all require "sequential ignorability" and, at the same time, accommodate treatment-by-mediator interactions typically by resorting to model-based assumptions with regard to how the treatment, the mediator, and the covariates interact in the outcome model. It is well-known that misspecifications of the outcome model would often bias causal effect estimation (Drake, 1993). In contrast, the RMPW strategy not only relaxes the no treatment-by-mediator interaction assumption but also greatly simplifies the specification of the outcome model. Hence the RMPW strategy has broad applications regardless of the distribution of the outcome, the distribution of the mediator, or the functional relationship between the outcome and the mediator.

Most existing methods estimate the indirect effect and sometimes the direct effect each as a function of the sample estimates of multiple parameters. Extra programming using the delta method therefore is required for estimating the asymptotic standard errors. In contrast, the RMPW strategy generates cluster-robust standard errors for the causal effect estimates and

provides immediate tests of the null hypotheses. This method can be readily implemented with standard software such as Stata, SPSS, SAS, and R.

Limitations of the RMPW strategy

RMPW identifies the causal effects of interest under the untestable assumption of sequential ignorability. Even though the ignorability of treatment assignment can be warranted by treatment randomization, mediator value assignment is typically not randomized. Similar to most existing methods described above, RMPW removes selection bias associated with the observed pretreatment covariates. The result will be biased if, for those who share the same observed pretreatment characteristics, the mediator-outcome relationship is confounded by omitted pretreatment or post-treatment covariates. Post-treatment covariates can be viewed as other potential mediators that are not independent of the focal mediator. For example, immediately after the randomization of treatment assignment, suppose that some participants' depressive symptoms would be heightened if assigned to LFA but not if assigned to the control condition instead. The post-randomization depressive symptoms at a heightened level under LFA would likely impede one's ability to secure employment and also predict depression at the two-year follow-up. In causal mediation analyses that allow for treatment-by-mediator interactions, the potential confounding effect of post-treatment covariates cannot be adjusted for directly (Imai, Keele, Tingley, & Yamamoto, 2011) but only indirectly through the adjustment for the related pretreatment covariates such as baseline depressive symptoms in the current example. Sensitivity analysis may be employed to assess the consequence of a possible omission (Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010; VanderWeele, 2010). However, in the case that the observed pretreatment covariates have explained nearly all the systematic variation

in the outcome, the remaining potential bias associated with the omitted pretreatment and post-treatment covariates may become negligible.

Future research may extend the RMPW approach to studies of multiple concurrent mediators, multiple consecutive mediators, time-varying mediators, and mediation problems in multi-level data such as data from cluster randomized trials or multisite trials.

References

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review, 40,* 37– 47.

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.

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,* 1173-1182.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

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.

Cheng, T. C. (2007). Impact of work requirements on the psychological wellbeing of TANF recipients. *Health and Social Work, 32*(1), 41-48.

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.

Coiro, M. J. (2001). Depressive symptoms among women receiving welfare. In M.C. Lennon (Ed.), *Welfare, work, and wellbeing* (pp.1-23). Bingham, NY: The Haworth Medical Press.

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

Duncan, O. D. (1966). “Path analysis: Sociological examples,” *American Journal of Sociology*, *72*, 1-16.

Hamilton, G., Freedman, S., Gennetian, L., Michalopoulos, C., Walter, J., Adams-Ciardullo, D., Gassman-Pines, A., McGroder, S., Zaslow, M., Ahluwalia, S., & Brooks, J. (2001). *How effective are different welfare-to-work approaches: Five-year adult and child impacts for eleven programs*. New York, NY: MDRC

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of an intervention. *Journal of Econometrics*, *30*, 239-267.

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. (2006). “Multilevel experimental designs and quasi-experimental approximations for studying intervention implementation as a mediator.” Paper presented at the 2006 Annual Meeting of the American Educational Research Association, San Francisco, CA.

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. (2012). Marginal mean weighting through stratification: A generalized method for evaluating multi-valued and multiple treatments with non-experimental data. *Psychological Methods*, *17*(1), 44-60.

Hong, G. (2013). Covariate-informed parallel design. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, *176*(1), 35.

Hong, G., & Nomi, T. (2012). Weighting methods for assessing policy effects mediated by peer change. *Journal of Research on Educational Effectiveness* special issue on the statistical approaches to studying mediator effects in education research, *5*(3), 261-289.

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

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*(4), 765-789.

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

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mech-anisms (with discussions). *Journal of the Royal Statistical Society, Series A (Statistics in Society). 176*(1), 5-51.

Jagannathan, R., Camasso, M. J., & Sambamoorthi, U. (2010). Experimental evidence of welfare reform impact on clinical anxiety and depression levels among poor women. *Social Science & Medicine, 71*(1), 152-160.

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

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 evaluation. *Evaluation Review*, *5*, 602–619.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of Neighborhood effects. *Econometrica, 75*(1), 83-119.

Knab, J., McLanahan, S., &Garfinkel, I.(2008).The effects of welfare and child support policies on maternal health and wellbeing. In R. Schoeni, J. House, G. Kaplan, & G. Pollack (Eds.), *Making Americans healthier: Social and economic policy as health policy* (pp. 281-305). New York: Russell Sage.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, *59*(10), 877-883.

Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacAuthor approaches. *Health Psychology*, *27*(2 Suppl), S101-S108.

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

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Erlbaum.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. Prevention Science, 1(4), 173-181.

Mattei, A. and Mealli, F. (2011). Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society, Series B (Methodological), 73*(5), 729-752.

Michalopoulos, C., Schwartz, C., & Adams-Ciardullo, D. (2001). *National evaluation of the welfare-to-work strategies: What works best for whom. Impacts of 20 welfare- to-work programs by subgroup*. New York: Manpower Demonstration Research Corporation.

Moore, K. A., Zaslow, M. J., Coiro, M. J., Miller, S. M., & Magenheim, E. B. (1995). *How well are they faring? AFDC families with preschool-aged children in Atlanta at the outset of the JOBS evaluation.* Washington, DC: U.S. Department of Health and Human Services.

Morris, P. A. (2008). Welfare program implementation and parents' depression. *Social Service Review*, 82(4), 579-614.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89*(6), 852-863.

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.

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology*, *76*(2), 266-278.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879-891.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42*(1), 185-227.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385-401.

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* Special Issue on the Statistical Approaches to Studying mediator Effects in Education Research, *5*(3), 303-332.

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. (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*(2), 143-155.

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. (1986). Statistics and causal inference: Comments: Which ifs have causal answers. *Journal of the American Statistical Association, 81*(396), 961-962.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13*(4), 279-313.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological Methods*, *4*, 422–445.

Siefert, K. P., Bowman, P. S., Heflin, C. M., Danziger, S., & Williams, D. R. (2000). Social and environmental predictors of maternal depression in current and recent welfare recipients. *American Journal of Orthopsychiatry, 70*(4), 510-522.

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

Sobel, M. E., & Stuart, E. A. (2012). Comments. *Journal of Research on Educational Effectiveness* Special Issue on the Statistical Approaches to Studying mediator Effects in Education Research, *5*(3), 290-292.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analysis in examining psychological processes. *Journal of Personality and Social Psychology*, *89*(6), 845-851.

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), 137-150.

van der Laan, M. J., & Peterson, M. L. (2008). Direct effect models. *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.

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology, 21*, 540-551.

VanderWeele, T. J., & Vansteelandt, S. (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.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5*, 161-215.

Appendix A

Bias in Path Analysis Estimation due to the Omission of Treatment-by-Mediator Interaction

For simplicity, suppose that the treatment and the mediator are both binary. Also suppose that treatment assignment and mediator value assignment under each treatment are both randomized. Let *A* = 1 if a unit is treated and 0 if the unit is assigned to the control condition. We use *Z* to denote the binary mediator and use *Y* to denote the outcome. If a unit is treated, the potential mediator is Z_1 that can take values 1 or 0. If the unit is assigned to the control condition instead, the potential mediator is denoted by Z_0 that again can take values 1 or 0. As a mediator, Z_a is a function of treatment assignment *a* and can be generated by $Z_a = \beta_0 + \beta_1 a + \varepsilon_z$ $Z_a = \beta_0 + \beta_1 a + \varepsilon_z$ where ε_z is a random error. In other words, we have that $pr(Z_0 = 1) = \beta_0$ and that $pr(Z_1 = 1) = \beta_0 + \beta_1$. We denote the potential outcome by Y_{az} if a unit is assigned to treatment *a* and displays mediator value *z*. Suppose that the data generation function for the potential outcomes is $Y_{az} = \theta_0 + \theta_1 a + \theta_2 z + \theta_3 az + \varepsilon_Y$. Hence the total effect is $\theta_1 + \theta_3\beta_0 + (\theta_2 + \theta_3)\beta_1$, the direct effect is $\theta_1 + \theta_3\beta_0$, and the indirect effect is $(\theta_2 + \theta_3)\beta_1$. Path analysis invokes the assumption of linearity and additivity (Holland, 1988) and specifies the observed outcome model as $Y = \gamma_0 + \gamma_1 A + \gamma_2 Z + e$. We can show that $\gamma_2 = \theta_2 + \theta_3 \times pr(A = 1)$. The indirect effect estimate is $\gamma_2\beta_1 = \theta_2\beta_1 + \theta_3\beta_1 \times pr(A = 1) = (\theta_2 + \theta_3)\beta_1 - \theta_3\beta_1 \times pr(A = 0)$. Hence the bias in the indirect effect estimate is $-\theta_3\beta_1 \times pr(A = 0)$, which is equivalent to $-E\{(Y_{11} - Y_{01}) - (Y_{10} - Y_{00})\} \times \{pr(Z_1 = 1) - pr(Z_0 = 1)\} \times pr(A = 0)$. The bias in the direct effect estimate takes the opposite sign.

Appendix B

B1. Stata Code for Parametric RMPW Analysis with a Binary Mediator

```
***  Run binary logit for the control group
logit Z X1 X2 X3 X4 X5 X6 X7 X8 X9 if A==0

*** Generate predicted probability, that is P(Z=1|X,A=0), for both experimental and control
groups.
* Note that this will be an in-sample prediction for those in the control group and an
* out-of-sample prediction for those in the experimental group.
predict p0, pr

*** Run binary logit for the experimental group
logit Z X1 X2 X3 X4 X5 X6 X7 X8 X9 if A==1

*** Generate predicted probability, that is P(Z=1|X,A=1).  Note that this will be an
* in-sample prediction for those in the experimental group and an out-of-sample prediction for
* those in the control group.
predict p1, pr

*** Generate weight
gen rmpw=1
replace rmpw=p0/p1 if A==1 & Z==1
replace rmpw=(1-p0)/(1-p1) if A==1 & Z==0

*** Generate a unique identifier, called "obs", for each person.  This will allow
* duplicates to have the same identifier, which will be necessary for obtaining the correct
* standard errors.
gen obs=_n

*** Generate duplicate observations for the experimental group, where D1 is the indicator
* for duplicate. D1=0 for all control group observations and original experimental group
* observations.
expand 2 if A==1, gen(D1)

*** Make sure duplicates get a weight=1. Notice that this means that duplicate
* observations receive a different weight than their original.
replace rmpw=1 if D1==1

*** Outcome model.
*Command weights each observation and clusters standard errors at the person level,
* adjusting for correlation in errors within each set of duplicates.
*** Adjustment for covariate X1, centered at its sample mean, for improving precision.
```

```
reg Y A D1 X1 [pweight=rmpw], vce(cluster obs)

*** To decompose natural indirect effect into the pure indirect effect and the natural
*treatment-by-mediator interaction effect
*Create a duplicate set of the control group, which will be weighted
expand 2 if A==0, gen(D0)

* Generate a new set of weights for the duplicate control group
replace rmpw = p1/p0   if  A==0 & Z==1 & D0==1
replace rmpw = (1-p1)/(1-p0)   if  A==0 & Z==0 & D0==1

*** Outcome model to estimate the pure indirect effect and the natural treatment-by-mediator
* interaction effect
*** Adjustment for covariate X1, centered at its sample mean, for improving precision.
reg Y A D1 D0 X1 [pweight=rmpw], vce(cluster obs)
lincom D1 – D0

*** The coefficient for D0 represents the pure indirect effect. The coefficient for D1 represents
* the total indirect effect.
*** The post-estimation command estimates, and does a significance test on, the natural
* treatment-by-mediator interaction effect, which is the total indirect effect less the pure indirect
* effect.
```

B2. Stata Code for Non-Parametric RMPW Analysis with a Binary Mediator

```
*** Run binary logit for the control group.
logit Z X1 X2 X3 X4 X5 X6 X7 X8 X9 if A==0

*** Generate logit score (not probability), which will be used later to create a categorical
* variable used in creating non-parametric weights.
predict xb0, xb

*** Run binary logit for the experimental  group.
logit Z X1 X2 X3 X4 X5 X6 X7 X8 X9 if A==1

***  Generate logit score, which will be used later to create a categorical variable used in
* creating non-parametric weights.
predict xb1, xb

*** Generate weight
*** Place all observations into three equal-sized categories based on their logit score
* from the experimental group model.
* Generate categorical variable h1=(0, 1, 2) based on terciles in xb1.
egen h1=cut(xb1), group(3)
```

```
*** Within each category of h1, generate categorical variables (h00, h01, h02)= (0, 1, 2)
* based on terciles in xb0.
* Loop through each category of h1.
forvalues j=0(1)2  {
egen h0`j'=cut(xb0) if h1==`j', group(3)
}

***  Generate a strata variable to place each observation into one of 9 strata,
* based on joint distribution of xb0 and xb1.
gen strata=.
replace strata=0 if h1==0 & h00==0
replace strata=1 if h1==0 & h00==1
replace strata=2 if h1==0 & h00==2
replace strata=3 if h1==1 & h01==0
replace strata=4 if h1==1 & h01==1
replace strata=5 if h1==1 & h01==2
replace strata=6 if h1==2 & h02==0
replace strata=7 if h1==2 & h02==1
replace strata=8 if h1==2 & h02==2

***  Calculate probabilities P(Z=1| A, strata). “Prij” is the probability that Z=1 in treatment
* group i and strata j. Loop over treatment groups (i = 0, 1) and strata (j = 0, 1, . . . , 8).
forvalues i=0(1)1  {
forvalues j=0(1)8  {
qui sum Z if A==`i' & strata==`j'
sca pr`i'`j'=r(mean)
}
}

***  Generate non-parametric RMPW weights based on these calculated probabilities,
* treatment group membership, strata membership, and Z.
gen nrmpw=1
* Loop over strata categories (j= 0, 1, . . . , 8).
forvalues j=0(1)8  {
replace nrmpw = pr0`j'/pr1`j' if A==1 & strata==`j' & Z==1
replace nrmpw = (1-pr0`j')/(1-pr1`j') if A==1 & strata==`j' & Z==0
}

***  Use the same process here as in the parametric case to create a person-specific identifier
* and generate duplicate observations.
gen obs=_n
expand 2 if A==1, gen(D1)
```

```
* Ensure duplicates receive a weight equal to 1.
replace nrmpw=1 if D1==1

*** Outcome model.
*Command weights each observation and clusters standard errors at the person level,
* adjusting for correlation in errors within each set of duplicates.
*** Adjustment for covariate X1, centered at its sample mean, for improving precision.
reg Y A D1 X1 [weight=nrmpw], vce(cluster obs)

*** To decompose the indirect effect into the pure indirect effect and the natural
*treatment-by-mediator interaction effect
*Create a duplicate set of the control group, which will be weighted
expand 2 if A==0, gen(D0)

* Generate new set of weights for the duplicate control group
* Loop over strata categories (j= 0, 1, . . . , 8).
forvalues j=0(1)8  {
replace nrmpw = pr1`j'/pr0`j' if A==0 & strata==`j' & Z==1 & D0==1
replace nrmpw = (1-pr1`j')/(1-pr0`j') if A==0 & strata==`j' & Z==0 & D0==1
}

*** Outcome model to estimate the pure indirect effect and the natural treatment-by-mediator
* interaction effect
*** Adjustment for covariate X1, centered at its sample mean, for improving precision.
reg Y A D1 D0 X1 [pweight=nrmpw], vce(cluster obs)
lincom D1 – D0
```

B3. Stata Code for RMPW Analysis of Moderated Mediation Effects

```
*****  Run binary logit  ***
***  Loop over treatment groups A (j= 0, 1) and moderator V (k= 0, 1).  There will be 4
* models run and 4 predicted probabilities for every observation.  For each observation,
* one of these predicted probabilities will be in-sample while the other 3 will be
* out-of-sample predictions.
forvalues j=0(1)1  {
forvalues k=0(1)1  {
logit Z X1 X2 X3 X4 X5 X6 X7 X8 X9 if A==`j' & V==`k'

* Generate predicted probabilities.
predict pt`j'`k', pr
}
}
```

```
*** Generate weight
gen rmpwt=1

*** Loop over values of moderator V (k= 0, 1)
forvalues k=0(1)1  {
replace rmpwt=pt0`k'/pt1`k' if A==1 & Z==1 & V==`k'
replace rmpwt=(1-pt0`k')/(1-pt1`k') if A==1 & Z==0 & V==`k'
}

*** Create a person-specific identifier and generate duplicate observations.
gen obs=_n
expand 2 if A==1, gen(D)
replace rmpwt=1 if D==1

*** Generate interactions
gen V_A=V*A
gen V_D=V*D
gen V_X1=V*X1
gen V0= (-1*V) + 1
gen V0_A=V0*A
gen V0_D=V0*D
gen V0_X1=V0*X1

*** Outcome models
* Each model is a "fully-interacted" model, and is run in two separate specifications
* to allow for direct hypothesis testing of each difference, without having to
* run post-estimation tests of linear combinations of coefficients in separate commands.
reg Y A D V V_A V_D X1 V_X1 [weight=rmpwt], vce(cluster obs)
reg Y A D V0 V0_A V0_D X1 V0_X1 [weight=rmpwt], vce(cluster obs)
```

B4. Stata Code for Parametric RMPW Analysis with a Three-Category Mediator

```
*****  Run ordered logit  ***
*** Same as binary parametric analysis except that we have an ordered logit, where Z= 0, 1, 2
* with three predicted probabilities under each treatment.
ologit Z X1 X2 X3 X4 X5 X6 X7 X8 X9 if A==0
predict p00 p01 p02, pr
ologit Z X1 X2 X3 X4 X5 X6 X7 X8 X9 if A==1
predict p10 p11 p12, pr

*** Generate weight
gen rmpw3=1
forvalues i=0(1)2  {
```

```
replace rmpw3=p0`i'/p1`i' if A==1 & Z==`i'
}

***  Create a person-specific identifier,
* generate duplicate observations, and give duplicates a weight equal to 1.
gen obs=_n
expand 2 if A==1, gen(D)
replace rmpw3=1 if D==1

*** Outcome model
reg Y A D X1 [weight=rmpw3], vce(cluster obs)
```

B5. Stata Code for Identifying Common Support

```
**** For a binary mediator

*** Variables xb0 and xb1 are the logit scores of the respective propensity models for
* the experimental and control groups.  See section B2.
* Loop over the logit scores (a= 0, 1) and the treatment categories (b= 0, 1)
forvalues a=0(1)1  {
* Calculate the standard deviation of each logit score, to be used below.
qui sum xb`a'
sca sd`a'=r(sd)
forvalues b=0(1)1  {
qui sum xb`a' if A==`b'
* Calculate the "minimum" and maximum values of each logit score for each treatment group.
* Where the "maximum"("minimum") is actually 20% of a standard deviation of the logit score
* above (below) the actual maximum (minimum).
sca xb`a'`b'max=r(max) + .2*sd`a'
sca xb`a'`b'min=r(min) - .2*sd`a'
}
}

*** Generate an "exclude" indicator
* Loop over each combination of logit score and treatment category.
gen exclude=0
forvalues a=0(1)1  {
forvalues b=0(1)1  {
replace exclude=1 if xb`a'<max(xb`a'0min,xb`a'1min) | xb`a'>min(xb`a'0max,xb`a'1max)
}
}

*** For a binary mediator with a moderator V
```

```
*** Run binary logit
***  Loop over treatment groups A (j= 0, 1) and moderator V (k= 0, 1).
forvalues j=0(1)1  {
forvalues k=0(1)1  {
logit Z X1 X2 X3 X4 X5 X6 X7 X8 X9 if A==`j' & V==`k'
* Generate logit scores.
predict xbt`j'`k', xb
}
}

* Loop over logit score treatment categories (a= 0, 1), logit score moderator categories (c= 0, 1),
* and actual treatment categories (b= 0, 1).
forvalues c=0(1)1  {
forvalues a=0(1)1  {
* Calculate the standard deviation of logit for each moderator group.
qui sum xbt`a'`c' if V==`c'
sca sdt`a'`c'=r(sd)
forvalues b=0(1)1  {
* Calculate "minimums" and "maximums" as above.
qui sum xbt`a"c' if A==`b' & V==`c'
sca xbt`a'`b'`c'max=r(max) + .2*sdt`a'`c'
sca xbt`a'`b'`c'min=r(min) - .2*sdt`a'`c'
}
}
}

*** Generate an "exclude" indicator
* Loop over each combination of logit score and moderator categories.
gen excludet=0
forvalues c=0(1)1  {
forvalues k=0(1)1  {
replace excludet=1 if V==`k' & (xbt`c'`k'<max(xbt`c'0`k'min,xbt`c'1`k'min) |
xbt`c'`k'>min(xbt`c'0`k'max,xbt`c'1`k'max))
}
}

*** For a three-category mediator

*** Run ordered logit for each treatment group and generate logit scores.
ologit Z X1 X2 X3 X4 X5 X6 X7 X8 X9 if A==0
predict xbo0, xb
ologit Z X1 X2 X3 X4 X5 X6 X7 X8 X9 if A==1
```

```
predict xbo1, xb

* Loop over logit score treatment categories (a= 0, 1), logit score moderator categories (c= 0, 1),
* and actual treatment categories (b= 0, 1).
forvalues i=0(1)1  {
* Calculate the standard deviation of each logit score.
qui sum xbo`i'
sca sdo`i'=r(sd)
}
forvalues i=0(1)1  {
forvalues j=0(1)2  {
* Calculate "minimums" and "maximums" for each treatment-by-mediator group as above.
qui sum xbo0 if A==`i' & Z==`j'
sca max0`i'`j'=r(max) + .2*sdo0
sca min0`i'`j'=r(min) - .2*sdo0
qui sum xbo1 if A==`i' & Z==`j'
sca max1`i'`j'=r(max) + .2*sdo1
sca min1`i'`j'=r(min) - .2*sdo1
}
}

*** Generate an "exclude" indicator
* Loop over each combination of logit score and moderator categories.
gen exclude3=0
forvalues i=0(1)1  {
forvalues k=0(1)1  {
replace exclude3=1 if xbo`i'>min(scalar(max`i'`k'0),scalar(max`i'`k'1),scalar(max`i'`k'2))
replace exclude3=1 if xbo`i'<max(scalar(min`i'`k'0),scalar(min`i'`k'1),scalar(min`i'`k'2))
}
}
```

Table 1

Glossary of Causal Effects in Mediation Analysis

Label	Notation	Definition
Treatment effect on the mediator	$Z_1 - Z_0$	The effect of being assigned to the LFA program versus control on a mother's employment
Treatment effect on the outcome	$Y_1 - Y_0$	The effect of being assigned to the LFA program versus control on a mother's depressive symptoms
Mediator effect on the outcome under the experimental condition	$Y_{11} - Y_{10}$	The effect of employment on maternal depression if the mother is assigned to the LFA program
Mediator effect on the outcome under the control condition	$Y_{01} - Y_{00}$	The effect of employment on maternal depression if the mother is assigned to the control condition
Direct effect of the treatment on the outcome	$Y_{1Z_0} - Y_{0Z_0}$	The effect of the policy on maternal depression if the policy fails to change one's employment experience
(Total) indirect effect of the treatment on the outcome	$Y_{1Z_1} - Y_{1Z_0}$	The effect of the policy on maternal depression under the LFA program solely attributable to the policy-induced change in her employment experience
Pure indirect effect of the treatment on the outcome	$Y_{0Z_1} - Y_{0Z_0}$	The effect of the policy on maternal depression under the control condition solely attributable to the policy-induced change in her employment experience
Natural treatment-by-mediator interaction effect	$\left(Y_{1Z_1} - Y_{1Z_0}\right) - \left(Y_{0Z_1} - Y_{0Z_0}\right)$	The difference between the LFA program and the control condition in how the policy-induced change in employment affects maternal depression

Table 2

Potential Mediators and Potential Outcomes

	Treatment	Potential Mediators		Potential Outcomes		
Individual Unit	A	Z_1	Z_0	Y_{1Z_1}	Y_{1Z_0}	Y_{0Z_0}
1	1	1	1	Y_{11}	Y_{11}	Y_{01}
2	1	1	0	Y_{11}	Y_{10}	Y_{00}
3	1	0	0	Y_{10}	Y_{10}	Y_{00}
4	0	1	1	Y_{11}	Y_{11}	Y_{01}
5	0	1	0	Y_{11}	Y_{10}	Y_{00}
6	0	0	0	Y_{10}	Y_{10}	Y_{00}
Population Average		$E(Z_1)$	$E(Z_0)$	$E(Y_{1Z_1})$	$E(Y_{1Z_0})$	$E(Y_{0Z_0})$

Table 3

(a) RMPW Applied to Data from a Sequential Randomized Design for Approximating a Three- or Four-Treatment Arm Design

	$E(Y_{0Z_0})$	$E(Y_{1Z_1})$	$E(Y_{1Z_0})$		$E(Y_{0Z_1})$	
A	0	1	1		0	
$D1$	0	1	0		0	
$D0$	0	0	0		1	
Z	0 or 1	0 or 1	0	1	0	1
ω	1.0	1.0	$\frac{pr(Z=0\mid A=0)}{pr(Z=0\mid A=1)}$	$\frac{pr(Z=1\mid A=0)}{pr(Z=1\mid A=1)}$	$\frac{pr(Z=0\mid A=1)}{pr(Z=0\mid A=0)}$	$\frac{pr(Z=1\mid A=1)}{pr(Z=1\mid A=0)}$

(b) RMPW Applied to Data from a Sequential Randomized Block Design for Approximating a Three- or Four-Treatment Arm Design

	$E(Y_{0Z_0})$	$E(Y_{1Z_1})$	$E(Y_{1Z_0})$		$E(Y_{0Z_1})$	
A	0	1	1		0	
$D1$	0	1	0		0	
$D0$	0	0	0		1	
Z	0 or 1	0 or 1	0	1	0	1
ω	1.0	1.0	$\frac{pr(Z=0\mid A=0,\mathbf{X}=\mathbf{x})}{pr(Z=0\mid A=1,,\mathbf{X}=\mathbf{x})}$	$\frac{pr(Z=1\mid A=0,\mathbf{X}=\mathbf{x})}{pr(Z=1\mid A=1,,\mathbf{X}=\mathbf{x})}$	$\frac{pr(Z=0\mid A=1,\mathbf{X}=\mathbf{x})}{pr(Z=0\mid A=0,,\mathbf{X}=\mathbf{x})}$	$\frac{pr(Z=1\mid A=1,\mathbf{X}=\mathbf{x})}{pr(Z=1\mid A=0,,\mathbf{X}=\mathbf{x})}$

Table 4

Parameter Values for Three Sets of Simulations

	$E(\theta_{Z_0})$	$E(\theta_{Z_1})$	$\gamma^{(DE)}$	$\gamma^{(IE)}$
(a)	.3918	.6609	0	0
(b)	.3918	.6609	0.7869	-0.8073
(c)	.1938	.8066	0.77325	-0.7660

Table 5

Summary of Simulation Results under Correct Specification of the Propensity Score Models

	Model	N = 5,000			N = 800		
		RMPW	*NRMPW 3×3*	*NRMPW 4×4*	*RMPW*	*NRMPW 3×3*	*NRMPW 4×4*
Direct Effect Estimate ($\hat{\gamma}^{(DE)}$)							
% Bias removal	(a)	0.999	0.857	0.905	0.984	0.843	0.872
	(b)	0.980	0.875	0.923	0.999	0.854	0.864
	(c)	0.995	0.859	0.908	0.988	0.818	0.771
Relative efficiency	(a)	0.960	0.964	0.980	0.885	0.918	0.888
	(b)	0.990	1.048	1.062	0.943	0.993	0.938
	(c)	0.856	0.941	0.949	0.656	0.794	0.774
MSE	(a)	0.004	0.003	0.002	0.011	0.012	0.012
	(b)	0.006	0.004	0.004	0.022	0.022	0.023
	(c)	0.008	0.010	0.007	0.037	0.040	0.046
Indirect Effect Estimate ($\hat{\gamma}^{(IE)}$)							
% Bias removal	(a)	0.998	0.856	0.904	0.985	0.856	0.885
	(b)	0.991	0.865	0.913	0.990	0.864	0.874
	(c)	0.999	0.856	0.904	0.994	0.823	0.776
Relative efficiency	(a)	1.145	1.872	1.749	0.985	1.337	1.102
	(b)	0.778	0.693	0.688	0.563	0.669	0.613
	(c)	0.752	0.973	0.933	0.490	0.708	0.680
MSE	(a)	0.002	0.001	0.001	0.002	0.003	0.003
	(b)	0.004	0.003	0.002	0.012	0.011	0.012
	(c)	0.006	0.008	0.005	0.022	0.024	0.030