Journal of Educational and Behavioral Statistics
2015, Vol. 40, No. 3, pp. 307–340
DOI: 10.3102/1076998615583902

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 Washington

Conventional methods for mediation analysis generate biased results when the mediator–outcome relationship depends on the treatment condition. This article shows how the ratio-of-mediator-probability weighting (RMPW) method can be used to decompose total effects into natural 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. Similar to other techniques for causal mediation analysis, RMPW generates causally valid results when the sequential ignorability assumptions hold. Yet unlike the model-based alternatives, including path analysis, structural equation modeling, and their latest extensions, RMPW requires relatively few assumptions about the distribution of the outcome, the distribution of the mediator, and the functional form of the outcome model. Correct specification of the propensity score models for the mediator remains crucial when parametric RMPW is applied. This article gives an intuitive explanation of the RMPW rationale, a mathematical proof, and simulation results for the parametric and nonparametric RMPW procedures. We apply the technique to identifying whether employment mediated the relationship between an experimental welfare-to-work program and maternal depression. A detailed delineation of the analytic procedures is accompanied by online Stata code as well as a stand-alone RMPW software program to facilitate users' analytic decision making.

Keywords: *causal inference; direct effect; indirect effect; mediation mechanism; potential outcome; propensity score*

Many important research questions in education, prevention science, and social sciences relate to how interventions work: What are the mechanisms through which a treatment exerts an impact on some outcome? To assess the role of a hypothesized mediator that could be affected by a treatment and could subsequently affect the outcome, researchers may 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. Moreover, conventional techniques for analyzing mediation rely on strong assumptions about the structural relationships among the treatment, the mediator, and the outcome. The analytic results are invalid when these model-based assumptions do not hold.

One of the assumptions is that there is no interaction between the treatment and the mediator in their influence on the outcome (Holland, 1988). However, 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. In other words, a treatment may alter the mediator–outcome relationship. 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). An intuitive 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. The number of study hours was speculated to be 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. The treatment-by-mediator interaction occurs, in this case, because the focal mediator (i.e., study hours) operates through its interactions with other unspecified mediators (e.g., attentiveness).

Treatment-by-mediator interactions may sometimes explain why an intervention fails to produce its intended effect on the outcome. As some researchers have argued or illustrated (Collins, Graham, & Flaherty, 1998; MacKinnon, Krull, & Lockwood, 2000; Preacher & Hayes, 2008; Sheets & Braver, 1999; 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 that heightens anxiety may increase study hours while reducing the amount of learning produced per hour, which may lead to a null effect of the encouragement treatment. 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 interactions (Baron & Kenny, 1986; Judd & Kenny, 1981). Importantly, whether the treatment alters the mediator–outcome relationship 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).

This article 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; Hong, 2010a). The RMPW strategy relaxes important constraining assumptions and is relatively straightforward to implement in common statistical packages. This propensity score-based weighting strategy adjusts for a large number of pretreatment covariates that may confound the mediator–outcome relationship. Moreover, it allows one to quantify the treatment effect on the outcome transmitted partly through a change in the mediator value and partly through a change in the mediator–outcome relationship. Yet, there is no need to explicitly include all the covariates and interaction terms in the outcome model. Hence, RMPW greatly simplifies the outcome model specification. We derive RMPW mathematically under specific identification assumptions. This article describes in detail the parametric and nonparametric RMPW analytic procedures in the context of an empirical application. We provide computer code in Stata and a free stand-alone RMPW software program in the online supplementary material along with the application data example. The performance of the RMPW method is assessed through a series of Monte Carlo simulations, from which we examine statistical properties of the estimation results and draw implications for practice.

The RMPW strategy overcomes some important limitations of a number of 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 popular techniques in social science and education 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). We have shown in Appendix A that omitting a nonzero treatment-by-mediator interaction will bias the estimation of direct and indirect effects. 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; VanderWeele, 2009). The IV method relies on the exclusion restriction, which implies that treatment assignment (e.g., encouragement), used as an instrument for the focal mediator (e.g., study hours), does not influence the outcome through other unspecified pathways (e.g., attentiveness) that would manifest in a treatment-by-mediator interaction. Marginal structural models take the same structural form as path analysis models in specifying the relationships among the treatment, the mediator, and the outcome, even though they avoid entering covariates directly into the structural models. As Coffman and Zhong (2012) acknowledged, however, without assuming that the treatment and the mediator additively affect the outcome, marginal structural models cannot be used to obtain an estimate of the natural indirect effect.

Recently, some new analytic strategies have emerged that relax the no-treatment-by-mediator-interaction assumption (see Hong [2015] for a review). These include modified regression approaches (Pearl, 2010; Petersen, Sinisi, & van der Laan, 2006; Preacher, Rucker, & Hayes, 2007; Valeri & VanderWeele, 2013; VanderWeele, 2013; VanderWeele & Vansteelandt, 2009, 2010), direct effect models (van der Laan & Peterson, 2008), conditional structural models (VanderWeele, 2009), and a resampling approach (Imai, Keele, & Tingley, 2010a; Imai, Keele, & Yamamoto, 2010b). While these methods are more flexible than the conventional approaches, correct specification of the outcome model is almost always crucial for generating unbiased estimates of the direct and indirect effects. An outcome model omitting multiway interactions between the treatment, the mediator, and the covariates can easily lead to biased estimation of the causal effects. The RMPW strategy is distinct from most of the previously mentioned approaches by minimizing the need to specify the outcome model.

Several alternative weighting methods (Huber, 2014; Tchetgen Tchetgen, 2013; Tchetgen Tchetgen & Shpitser, 2012) have similarly avoided these restrictions. A common theoretical rationale shared by RMPW and these alternative weighting methods is that the distribution of the mediator in the experimental group and that in the control group can be effectively equated through weighting under the identification assumptions with regard to the ignorability of the treatment and the mediator. We will explicate these assumptions in a later section. This transformation of mediator distribution makes possible the estimation of population average counterfactual outcomes essential to treatment effect decomposition. Yet the rationale is implemented differently by these different weighting methods. In particular, the inverse probability weight (IPW) proposed by Huber (2014) and the inverse odds ratio weight proposed by Tchetgen Tchetgen (2013) estimate the conditional probability of each treatment condition as a function of mediator values and covariate values. In contrast, the RMPW method estimates the

conditional probability of each mediator value under each treatment condition as a function of covariate values. In practice, applied researchers often have scientific knowledge about the selection mechanism to aid in modeling the latter (e.g., what type of students would study additional hours when encouraged). Modeling the treatment as a function of the mediator and covariates, however, does not have immediate substantive interpretations, given that the treatment causally precedes rather than succeeds the mediator. We will discuss later that, nonetheless, modeling the treatment may appear to be computationally convenient for continuous mediators.

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 nonparametric weighting procedures applied to binary mediators. After presenting the simulation results, we discuss the relative strengths and potential limitations as well as possible extensions of the RMPW strategy, and raise issues for future research.

Application Example

In the late 1990s, the U.S. 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). This change in federal policy was heavily influenced by experiments conducted earlier in the decade, some of which showed increased employment and earnings for welfare recipients as a result of employment-focused incentives and services (Grogger & Karoly, 2005). Since then, concerns have been raised about the impact of welfare-to-work programs on the long-term psychological well-being of welfare recipients, who tend to be low-income single mothers with young children, especially if they fail to secure employment (Cheng, 2007; Jagannathan, Camasso, & Sambamoorthi, 2010; Morris, 2008).

We use data from the National Evaluation of Welfare-to-Work Strategies (NEWWS) Labor Force Attachment program (henceforth LFA) in Riverside, California. Rather than focusing on employment as the outcome of primary interest, we examine whether and how employment mediated the program impact on maternal depression in the long run. 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. 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 noncompliance in program activities or work by reducing LFA group members' welfare benefits. A key feature of LFA in Riverside is that it encouraged and increased the likelihood of, but did not guarantee, employment among treatment group members.

As expected, the program increased employment and earnings and reduced welfare receipt *during* the 2 years after randomization. LFA in Riverside did not show a statistically significant effect on maternal depression *at the end* of those 2 years (Hamilton et al., 2001). Importantly, the null effect on depression does not rule out possible mediation by the participants' intermediate experience with employment. We hypothesize two distinct scenarios in which the null total effect of the program on depression would mask mediated effects. First, program-induced employment might eventually benefit a participant's mental health (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 on depression in the long run. Second, program expectations with regard to employment and the threat of sanctions could alter the relationship between employment and subsequent mental health, such that employment during the study period would be more beneficial, and unemployment during the same period more detrimental, to long-term psychological well-being if a mother was assigned to LFA than if she was assigned to the control condition. This second scenario, a classic case of treatment-by-mediator interaction, highlights an indirect effect due to a change in the mediator–outcome relationship, which again could be offset by a direct effect. In this application, we will investigate (1) whether the effect of employment *during* the 2 years after randomization on depression *at the end* of those 2 years depended on treatment assignment, (2) whether through increasing employment, the program generated an indirect effect that reduced depression in the end, and (3) whether being assigned to LFA would have had a 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 years to 5 years. Unemployment Insurance records maintained by the State of California provide quarterly administrative data on *employment* for each participant. We summarize the employment records over the 2 years after randomization in a binary measure indicating whether a participant was ever employed during the 2-year period. All participants were surveyed shortly before the randomization and again at the 2-year follow-up. The self-administered questionnaire at the 2-year follow-up included 12 items (Center for

Epidemiologic Studies–Depression Scale; Radloff, 1977) measuring *depressive symptoms* during the past week (e.g., I could not get going) on a frequency scale from 0 (*rarely*) to 3 (*most of the time*). The summary score ranged from 0 to 34 with a mean equal to 7.49 and a standard deviation equal to 7.74.

The baseline survey provided rich information about participant characteristics shown previously to be important predictors of 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) education credentials and academic skills; (d) personal attitudes toward employment, including the preference to work, willingness to accept a low-wage job, and shame to be on welfare; (e) perceived social support and barriers to work; (f) practical support and barriers to work such as child care 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 treatment assignment; Z, employment experience during the 2 years after randomization; and Y, depressive symptoms at the 2-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. Let $Z = 1$ if a welfare mother was ever employed and $Z = 0$ if never employed during the 2-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 counterfactual outcomes. Table 1 provides a glossary for all the causal effects defined subsequently.

1. What is the treatment effect on the mediator?

We use Z_1 to denote a mother's potential employment experience if assigned to LFA and Z_0 for the mother's potential employment experience if assigned to the control condition. Of these two potential intermediate outcomes, one is observed and the other is an unobserved counterfactual. The person-specific causal effect of treatment assignment on employment is $Z_1 - Z_0$. This definition implies that one's employment is affected only by one's own treatment assignment and is not affected by other individuals' treatment assignment (Rubin, 1986). Yet we allow each potential mediator value to be possibly altered by random events often beyond the control of the experimenter. For example, a participant assigned to LFA who otherwise would have become employed might

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 relative to unemployment 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 relative to unemployment on maternal depression if the mother is assigned to the control condition
Controlled treatment-by-mediator interaction effect	$(Y_{11} - Y_{10}) - (Y_{01} - Y_{00})$	The difference between the LFA program and the control condition in the effect of employment on maternal depression
Treatment effect decomposition		
Natural 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; also called "the pure direct effect"
Natural 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; also called "the total indirect effect"
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

(continued)

Table 1. *(continued)*

Label	Notation	Definition
Natural treatment-by-mediator interaction effect	$(Y_{1Z_1} - Y_{1Z_0}) - (Y_{0Z_1} - Y_{0Z_0})$	The difference between the LFA program and the control condition in how the policy-induced change in employment affects maternal depression

Note. LFA = Labor Force Attachment.

remain unemployed due to an economic downturn or an unexpected health problem of a family member.

2. What is the treatment effect on the outcome?

To define the treatment effect on maternal depression at the 2-year follow-up, we use Y_1 to denote a mother's potential psychological outcome if assigned to LFA and Y_0 for the mother's potential outcome if assigned to the control condition. The person-specific treatment effect on 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, 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 LFA or the control condition. Let Y_{11} denote a mother's depression level if she was assigned to LFA and employed, and let Y_{10} denote her depression level if she was assigned to LFA and unemployed. Here, the first subscript represents the assignment to LFA, while the second represents whether one is employed. The causal effect of employment relative to unemployment on maternal depression if the mother was assigned to LFA 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 relative to unemployment 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 $(Y_{11} - Y_{10}) - (Y_{01} - Y_{00}) \neq 0$, which, according to

Pearl's (2001) terminology, is called the controlled treatment-by-mediator interaction effect.

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 LFA yet experiencing employment as she would have had under the control condition. The direct effect, defined by $Y_{1Z_0} - Y_{0Z_0}$, represents the effect of treatment assignment on maternal depression if the treatment, perhaps counterfactually, failed to change one's employment experience. The direct effect can be attributed to other unspecified pathways independent of employment. For example, the threat of sanctions under LFA might heighten depression, while interactions with LFA caseworkers might lead to improved access to mental health services. 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. Robins and Greenland (1992) called this the pure direct effect instead.

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 LFA would become more or less depressed at the 2-year follow-up should she counterfactually experience the same level of employment as she would under the control condition. Defined 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 treatment-induced change in her employment experience (i.e., a change from Z_0 to Z_1). This has been called the total indirect effect by Robins and Greenland (1992) and the natural indirect effect by Pearl (2001).

6. What is the indirect effect if the treatment changes the mediator–outcome relationship?

As we reasoned earlier, the LFA program relative to the control condition might affect maternal depression partly through increasing employment and partly through altering the mediator–outcome relationship, such that employment would be more beneficial under LFA 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 treatment 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, defined by $(Y_{1Z_1} - Y_{1Z_0}) - (Y_{0Z_1} - Y_{0Z_0})$, reflects how the treatment-induced

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}	Y_{0Z_1}
1	1	1	1	Y_{11}	Y_{11}	Y_{01}	Y_{01}
2	1	1	0	Y_{11}	Y_{10}	Y_{00}	Y_{01}
3	1	0	0	Y_{10}	Y_{10}	Y_{00}	Y_{00}
4	0	1	1	Y_{11}	Y_{11}	Y_{01}	Y_{01}
5	0	1	0	Y_{11}	Y_{10}	Y_{00}	Y_{01}
6	0	0	0	Y_{10}	Y_{10}	Y_{00}	Y_{00}
Population average		$E(Z_1)$	$E(Z_0)$	$E(Y_{1Z_1})$	$E(Y_{1Z_0})$	$E(Y_{0Z_0})$	$E(Y_{0Z_1})$

change in employment would affect the mother's depression differently between LFA and the control condition. This "natural treatment-by-mediator interaction effect" is the treatment effect on the mediator–outcome relationship in the natural world and is different from $(Y_{11} - Y_{10}) - (Y_{01} - Y_{00})$ for participants whose employment experience is unchanged by the treatment (i.e., $Z_1 = Z_0$). The total effect of the treatment on the outcome 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. That is,

$$\begin{aligned} Y_{1Z_1} - Y_{0Z_0} &= (Y_{1Z_0} - Y_{0Z_0}) + (Y_{1Z_1} - Y_{1Z_0}) \\ &= (Y_{1Z_0} - Y_{0Z_0}) + (Y_{0Z_1} - Y_{0Z_0}) + [(Y_{1Z_1} - Y_{1Z_0}) - (Y_{0Z_1} - Y_{0Z_0})]. \end{aligned}$$

Table 2 illustrates the concepts with six participants, three of whom 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 four potential outcomes. For the first three participants, the only observables are Z_1 and Y_{1Z_1}; for the second three, the only observables are Z_0 and Y_{0Z_0}. Yet research designs and analytic strategies can be employed, when certain identification assumptions are satisfied, to estimate the population average causal effects. These include the population average treatment effect on the mediator denoted by $E(Z_1 - Z_0)$, 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 $E(Y_{1Z_1} - Y_{1Z_0})$, the population average pure indirect effect $E(Y_{0Z_1} - Y_{0Z_0})$, and the population average natural treatment-by-mediator interaction effect $E[(Y_{1Z_1} - Y_{1Z_0}) - (Y_{0Z_1} - Y_{0Z_0})]$.

RMPW-Based Analytic Framework for Causal Mediation Analysis

RMPW Under Sequential Randomization

In a hypothetical sequential randomized experiment, after assigning welfare applicants at random to either LFA or the control condition, the experimenter would subsequently assign applicants within each treatment group at random to employment. The mean observed outcomes obtained from the four treatment-by-employment combinations would provide unbiased estimates of the first set of population average potential outcomes $E(Y_{11})$, $E(Y_{10})$, $E(Y_{01})$, and $E(Y_{00})$. Yet the natural direct effect and the natural indirect effect are defined in terms of the second set of population average potential outcomes $E(Y_{1Z_1})$, $E(Y_{1Z_0})$, and $E(Y_{0Z_0})$. The rationale for RMPW can be derived from the inherent connections between these two sets of potential outcomes, as shown 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:

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

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

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

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

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

We may simply transform the employment rate in the LFA group to resemble that in the control group. The transformation can be done through weighting because the previously mentioned equation is equal to

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

Here, Y_{11} is weighted by the ratio of the probability of employment under the control condition to that under LFA, $\frac{\text{pr}(Z_0=1)}{\text{pr}(Z_1=1)}$; in parallel, Y_{10} is weighted by the

ratio of the probability of unemployment under the control condition to that under LFA, $\frac{\text{pr}(Z_0=0)}{\text{pr}(Z_1=0)}$. In a sequential randomized design, $\frac{\text{pr}(Z=1|A=0)}{\text{pr}(Z=1|A=1)}$ estimates the RMPW for the employed LFA units, while $\frac{\text{pr}(Z=0|A=0)}{\text{pr}(Z=0|A=1)}$ 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. The duplication allows for estimating $E(Y_{1Z_0})$ with the RMPW adjusted mean observed outcome of the LFA group and, at the same time, estimating $E(Y_{1Z_1})$ with the mean observed outcome of the same group. Let $D1$ 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., $A = 0$ and $D1 = 0$), is $\frac{\text{pr}(Z=1|A=0)}{\text{pr}(Z=1|A=1)}$ for the employed LFA units (i.e., $A = 1$, $D1 = 0$, and $Z = 1$) and $\frac{\text{pr}(Z=0|A=0)}{\text{pr}(Z=0|A=1)}$ for the unemployed LFA units (i.e., $A = 1$, $D1 = 0$, and $Z = 0$), and is 1.0 for the duplicate LFA units (i.e., $A = 1$ and $D1 = 1$). We then regress Y on the treatment indicator A and the indicator for LFA duplicate $D1$ in a weighted model:

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

Here, $\gamma^{(0)}$ estimates $E(Y_{0Z_0})$; $\gamma^{(0)} + \gamma^{(DE)}$ estimates $E(Y_{1Z_0})$; and $\gamma^{(0)} + \gamma^{(DE)} + \gamma^{(IE.1)}$ estimates $E(Y_{1Z_1})$. Hence $\gamma^{(DE)}$ estimates the average direct effect $E(Y_{1Z_0} - Y_{0Z_0})$, and $\gamma^{(IE.1)}$ 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 (*SE*s).

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 $E(Y_{0Z_1})$. We may simply transform the employment rate and the unemployment rate in the control group to resemble those in the experimental group.

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

To implement, we may additionally create a duplicate set of the control group indicated by $D0$. This is because the mean observed outcome of the control group estimates $E(Y_{0Z_0})$, while the RMPW adjusted mean observed outcome of the same group estimates $E(Y_{0Z_1})$. The weight is $\frac{\text{pr}(Z=1|A=1)}{\text{pr}(Z=1|A=0)}$ for the duplicate set of the employed control units (i.e., $A = 0$, $D0 = 1$, $D1 = 0$, and $Z = 1$) and is $\frac{\text{pr}(Z=0|A=1)}{\text{pr}(Z=0|A=0)}$ for the duplicate set of the unemployed control units (i.e., $A = 0$, $D0 = 1$, $D1 = 0$, and $Z = 0$). We then conduct a weighted analysis regressing Y on A, $D1$, and $D0$:

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

Here, $\gamma^{(IE.1)}$ estimates the average 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 *SE* is $\mathrm{Var}\left(\hat{\gamma}^{(IE.1)}\right) + \mathrm{Var}\left(\hat{\gamma}^{(IE.0)}\right) - 2\mathrm{Cov}\left(\hat{\gamma}^{(IE.1)}, \hat{\gamma}^{(IE.0)}\right)$.

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 $\mathbf{X}$. We may envision that the data approximate a sequential randomized block design in which individuals with homogeneous pretreatment characteristics $\mathbf{X} = \mathbf{x}$ constitute blocks. Those in the same block are hypothetically randomized first to LFA or the control condition and subsequently to employment or unemployment. Hypothetical randomization to employment within each block could be a result of unpredictable events, such as fluctuations in the local job market or a major change in one's household. In a given block, the observed mediator distribution of individuals assigned to the control condition provides counterfactual information of the mediator distribution that the LFA units would likely display should they be assigned to the control condition instead, and vice versa. We apply RMPW to the "as-if" sequential randomized data within each block and summarize the results over all blocks.

RMPW can be estimated as functions of $\mathbf{x}$ that determine one's block membership. To be specific, for estimating $E(Y_{1Z_0})$, an individual in the LFA group displaying pretreatment characteristics $\mathbf{x}$, if employed, would be weighted by $\frac{\mathrm{pr}(Z=1|A=0,\mathbf{X}=\mathbf{x})}{\mathrm{pr}(Z=1|A=1,\mathbf{X}=\mathbf{x})}$. If the individual were unemployed, the weight would be $\frac{\mathrm{pr}(Z=0|A=0,\mathbf{X}=\mathbf{x})}{\mathrm{pr}(Z=0|A=1,\mathbf{X}=\mathbf{x})}$. A weighted analysis of Model 1 estimates the direct effect and the indirect effect. Similarly, 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 $\frac{\mathrm{pr}(Z=1|A=1,\mathbf{X}=\mathbf{x})}{\mathrm{pr}(Z=1|A=0,\mathbf{X}=\mathbf{x})}$ for the employed control units and is $\frac{\mathrm{pr}(Z=0|A=1,\mathbf{X}=\mathbf{x})}{\mathrm{pr}(Z=0|A=0,\mathbf{X}=\mathbf{x})}$ for the unemployed control units. Analyzing Model 2, we estimate the pure indirect effect and the natural treatment-by-mediator interaction effect. Table 3 summarizes the weighting scheme.

RMPW for Multivalued Mediators

This framework can be extended easily to multivalued mediators. To estimate $E(Y_{1Z_0})$, we apply to LFA units the following weight:

TABLE 3.
RMPW Application to NEWWS Data

	$E(Y_{0Z_0})$	$E(Y_{1Z_1})$	$E(Y_{1Z_0})$		$E(Y_{0Z_1})$	
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	$\frac{\text{pr}(Z=0\|A=0,\mathbf{X}=\mathbf{x})}{\text{pr}(Z=0\|A=1,\mathbf{X}=\mathbf{x})}$	$\frac{\text{pr}(Z=1\|A=0,\mathbf{X}=\mathbf{x})}{\text{pr}(Z=1\|A=1,\mathbf{X}=\mathbf{x})}$	$\frac{\text{pr}(Z=0\|A=1,\mathbf{X}=\mathbf{x})}{\text{pr}(Z=0\|A=0,\mathbf{X}=\mathbf{x})}$	$\frac{\text{pr}(Z=1\|A=1,\mathbf{X}=\mathbf{x})}{\text{pr}(Z=1\|A=0,\mathbf{X}=\mathbf{x})}$

Note. RMPW = ratio-of-mediator-probability weighting; NEWWS = National Evaluation of Welfare-to-Work Strategies.

$$\omega = \frac{\text{pr}(Z_0 = z|\mathbf{X} = \mathbf{x})}{\text{pr}(Z_1 = z|\mathbf{X} = \mathbf{x})} = \frac{\text{pr}(Z = z|A = 0, \mathbf{X} = \mathbf{x})}{\text{pr}(Z = z|A = 1, \mathbf{X} = \mathbf{x})}. \quad (3)$$

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

$$\omega = \frac{\text{pr}(Z_1 = z|\mathbf{X} = \mathbf{x})}{\text{pr}(Z_0 = z|\mathbf{X} = \mathbf{x})} = \frac{\text{pr}(Z = z|A = 1, \mathbf{X} = \mathbf{x})}{\text{pr}(Z = z|A = 0, \mathbf{X} = \mathbf{x})}. \quad (4)$$

Identification Assumptions

This section presents the theoretical results clarifying the identification assumptions under which RMPW removes selection bias in estimating the causal effects defined previously.

Assumption 1: Nonzero probability of treatment assignment. Within levels of $\mathbf{X}$, every unit has a nonzero probability of being assigned to each treatment condition. That is, for $a = 0, 1$,

$$0 < \text{pr}(A = a|\mathbf{X}) < 1.$$

Assumption 2: No confounding of treatment–outcome relationship. Treatment assignment is independent of the potential outcomes, given the observed pretreatment covariates $\mathbf{X}$. That is, for all possible values of a, a', and z,

$$Y_{aZ_a}, Y_{aZ_{a'}}, Y_{az} \coprod A|\mathbf{X}.$$

Here, $\coprod$ denotes independence in causal inference.

Assumption 3: No confounding of treatment–mediator relationship. Treatment assignment is independent of the potential intermediate outcomes given **X**. For all possible values of a,

$$Z_a \coprod A|\mathbf{X}.$$

Assumption 4: Nonzero probability of mediator value assignment. Within levels of **X**, every unit has a nonzero probability of being assigned to each mediator value under each treatment condition. That is, for all possible values of a and z,

$$0 < \text{pr}(Z_a = z|A = a, \mathbf{X}) < 1.$$

Assumption 5: No confounding of mediator–outcome relationship within a treatment. Within levels of **X** and under a given treatment condition, mediator value assignment is independent of the potential outcomes. That is, for all possible values of a and z,

$$Y_{az} \coprod Z_a|A = a, \mathbf{X}.$$

Assumption 6: No confounding of mediator–outcome relationship across treatment conditions. Within levels of **X**, mediator value assignment under a given treatment condition is independent of the potential outcomes associated with an alternative treatment condition. That is, for all possible values of a, a', and z,

$$Y_{az} \coprod Z_{a'}|A = a, \mathbf{X}.$$

The previously mentioned six assumptions constitute the sequential ignorability (Imai et al., 2010a, 2010b); 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. Assumptions 5 and 6 imply that the mediator–outcome relationships are not confounded by any posttreatment covariates (Pearl, 2001; Robins, 2003).

Subsequently, we derive the identification results under these assumptions. Following van der Laan and Petersen (2008), we represent the joint distribution of the observed data $O = (\mathbf{X}, A, Z_a, Y_{aZ_a})$ in general as follows:

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

where **X** denotes a vector of observed pretreatment covariates and where $f^{(a,z)}(\cdot)$, $p^{(a)}(\cdot)$, $q(\cdot)$, and $h(\cdot)$ are density functions. For simplicity, we use $f(\cdot)$ to represent $f^{(a,z)}(\cdot)$ henceforth. Theorem 1 summarizes the results from past research (Robins, 1999; Rosenbaum, 1987) showing that $E[Y_{aZ_a}]$ is identified; Theorem 2 summarizes the results for identifying $E[Y_{az}]$ (VanderWeele, 2009) through

an analysis of marginal structural models. Theorem 3 summarizes the results for identifying $E\left[Y_{aZ_{a'}}\right]$ through RMPW analysis (Hong, 2010a).

Theorem 1: $E(Y^*|A=a) \equiv E\left(\omega^{(a,Z_a)}Y|A=a\right)$ is unbiased for $E[Y_{aZ_a}]$ under Assumptions 1 and 2, where the inverse-probability-of-treatment weight

$$\omega^{(a,Z_a)} = \frac{q(A=a)}{q(A=a|\mathbf{X})},$$

for all possible values of a removes treatment selection. When the treatment is randomized, we have that $q(A=a|\mathbf{X}) = q(A=a)$ and hence $\omega^{(a,Z_a)} = 1.0$ and $Y^* = Y$ for all units.

Theorem 2: $E(Y^*|A=a) \equiv E\left(\omega^{(a,z)}Y|A=a,Z_a=z\right)$ is unbiased for $E[Y(a,z)]$ under Assumptions 1 through 5, where the product of the inverse-probability-of-mediator weight and the inverse-probability-of-treatment weight

$$\omega^{(a,z)} = \frac{p^{(a)}(Z_a=z|A=a)}{p^{(a)}(Z_a=z|A=a,\mathbf{X})} \times \frac{q(A=a)}{q(A=a|\mathbf{X})},$$

for all possible values of a and z removes treatment selection and mediator value selection within each treatment group. This weight, routinely applied in marginal structural models, does not allow for treatment effect decomposition in the presence of treatment-by-mediator interaction.

Theorem 3: $E(Y^*|Z=z) \equiv E\left(\omega^{(a,Z_{a'})}Y|A=a\right)$ is unbiased for $E\left[Y_{aZ_{a'}}\right]$ under Assumptions 1 through 6, where the ratio-of-mediator-probability weight in combination with the inverse-probability-of-treatment weight for removing treatment selection is given by:

$$\omega^{(a,Z_{a'})} = \frac{p^{(a')}(Z_{a'}=z|A=a',\mathbf{X})}{p^{(a)}(Z_a=z|A=a,\mathbf{X})} \times \frac{q(A=a)}{q(A=a|\mathbf{X})},$$

for all possible values of a and z. Appendix B presents a proof of Theorem 3.

Parametric RMPW Procedure

Applying the previously mentioned theoretical results to an analysis of the NEWWS data, we describe a parametric procedure for estimating RMPW in this section and a nonparametric procedure in the next section for binary mediators. These procedures can be carried out in standard statistical programs. We provide Stata code in the online supplementary material for all the analyses presented in this article. Additionally, a free stand-alone RMPW software program provides user-friendly interfaces designed not only to ease computation but also to assist the applied user with analytic decision making. The software can be accessed through the website: http://hlmsoft.net/ghong/.

The parametric approach estimates RMPW as a ratio of the estimated propensity score of being assigned to a mediator value under one treatment to that under the alternative treatment.

Step 1: Select and Prepare the Pretreatment Covariates

We have selected 86 pretreatment covariates that are theoretically associated with maternal depression or with employment. After creating a missing category for each categorical covariate with missing information, we impute the missing data in the outcome and in the continuous covariates and generate five imputed 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, subsequently, we discuss the analytic procedure with one imputed data set. We first estimate the treatment effects on the mediator and on the outcome. Assignment to LFA increased employment rate from 39.5% to 65.4%. The average treatment effect on depression cannot be statistically distinguished from zero (coefficient = 0.11, $SE = 0.64$, $t = 0.18$, $p = .86$).

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 propensity score for employment under LFA, denoted by $\theta_{Z_1} = \theta_{Z_1}(\mathbf{x}) = \text{pr}(Z = 1|A = 1, \mathbf{X} = \mathbf{x})$, 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, we predict a control unit's propensity score for employment under the control condition, denoted by $\theta_{Z_0} = \theta_{Z_0}(\mathbf{x}) = \text{pr}(Z = 1|A = 0, \mathbf{X} = \mathbf{x})$ using data from the control group. Under Assumption 1, "nonzero probability of treatment assignment" and Assumption 3, "no confounding of treatment–mediator relationship," the propensity score model specified under the control condition can be applied to the LFA units had they been counterfactually assigned to the control condition instead. Hence, applying the coefficient estimates obtained from the control group, 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}. We will discuss alternative model fitting approaches in the last section of this article.

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. To approximate data from a sequential randomized block design, 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 joint distribution of θ_{Z_1} and θ_{Z_0} across the four treatment-by-mediator groups and identify cases in which the distribution 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 of the observed distribution to expand the range of the common support (Austin, 2011).

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 portion 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 weighting (Robins, 1999; VanderWeele, 2009) to the current example, we assign the weight $\frac{\text{pr}(Z=1|A=1)}{\theta_{Z_1}}$ to the employed LFA units, $\frac{\text{pr}(Z=0|A=1)}{\left(1-\theta_{Z_1}\right)}$ to the unemployed LFA units, $\frac{\text{pr}(Z=1|A=0)}{\theta_{Z_0}}$ to the employed control units, and $\frac{\text{pr}(Z=0|A=0)}{\left(1-\theta_{Z_0}\right)}$ to the unemployed control units. If the weighting adjustment has been successful, we expect that, 95% of the time, a categorical covariate will show equal frequency distribution and that a continuous covariate will show equal mean and variance across the 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

By applying the marginal structural models (VanderWeele, 2009), this step produces useful evidence with regard to whether the mediator–outcome relationship differs by treatment. We simply regress Y on A, Z, and A-by-Z interaction under the inverse-probability weighting. The results show that the employment effect on depression differed by treatment. Specifically, having all participants employed as opposed to having none employed would reduce depressive symptoms under LFA (coefficient $= -2.49$, $SE = 1.20$, $t = -2.07$, $p < .05$) but not under the control condition (coefficient $= 0.74$, $SE = 0.76$, $t = 0.97$, $p = .33$). The treatment-by-mediator interaction is statistically significant (coefficient $= -3.23$, $SE = 1.42$, $t = -2.27$, $p < .05$). However, the analysis in Step 5 does not decompose the total effect to reveal the mediation mechanism.

Step 6: Create a Duplicate and Compute the Parametric RMPW

We then reconstruct the data within common support to include a duplicate for each control unit and one for each LFA unit. The rest of this step has been summarized in Table 3. To estimate $E(Y_{1Z_0})$, the weight for the employed LFA units is $\frac{\theta_{Z_0}}{\theta_{Z_1}}$ and that for the unemployed LFA units is $\frac{(1-\theta_{Z_0})}{(1-\theta_{Z_1})}$; to estimate $E(Y_{0Z_1})$, the weight for the employed control units is $\frac{\theta_{Z_1}}{\theta_{Z_0}}$ and that for the unemployed control units is $\frac{(1-\theta_{Z_1})}{(1-\theta_{Z_0})}$. The employment rate in the RMPW-adjusted LFA group (38.1%) approximates that in the control group (39.5%), while that in the RMPW-adjusted control group (66.9%) approximates that in the LFA group (65.4%).

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 *SE* for each estimate. Analyzing weighted 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 continuous outcome. The estimated direct effect is 1.29 ($SE = 0.87$; $t = 1.48$, $p = .14$), about 17% of a standard deviation of the outcome; the estimated indirect effect is −0.87 ($SE = 0.47$; $t = -1.87$, $p = .06$). The direct effect estimate indicates that, if the treatment had counterfactually generated no impact on employment (i.e., if the employment rate had remained at 39.5% rather than increasing to 65.4%), maternal depression would have increased, but not 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 previously would not have a statistically significant impact on the average level of depression (coefficient = 0.32, $SE = 0.27$; $t = 1.48$, $p = .14$). The estimated natural treatment-by-mediator interaction effect is −1.19 ($SE = 0.53$; $t = -2.26$, $p < .05$), 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. Because the treatment assignment changed some but not all participants' status from being unemployed to being employed, unsurprisingly, the magnitude of "the natural treatment-by-mediator interaction effect" is considerably smaller than that of "the controlled treatment-by-mediator interaction

effect." The sum of the estimated natural direct effect, the pure indirect effect, and the natural treatment-by-mediator interaction effect is 0.42 and is equal to the total treatment effect on depression in the analytic sample.

Nonparametric RMPW Procedure

In general, nonparametric analyses are relatively more robust than their parametric counterparts because the former are less reliant on model-based assumptions. For example, past research has shown that, in evaluating the relative effectiveness of different treatments, parametric inverse-probability-of-treatment weighting (IPTW) often generates biased results, especially when the propensity score models are misspecified in their functional forms (Hong, 2010b; Kang & Schafer, 2007; Schafer & Kang, 2008; Waernbaum, 2012). In contrast, nonparametric weighting methods, such as marginal mean weighting through stratification (MMWS), produce robust results despite such misspecifications (Hong, 2010b, 2012). IPTW and MMWS, however, are not suitable for decomposing the total effect into a direct effect and an indirect effect in the presence of treatment-by-mediator interactions. We develop a nonparametric RMPW procedure for mediation analysis and evaluate its performance in comparison with that of the parametric RMPW procedure through simulations.

In essence, the nonparametric RMPW procedure recomputes the conditional probability of mediator value assignment 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. We apply the nonparametric MMWS 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 four even portions. Within each of these four subclasses, we then rank and subdivide the units by θ_{Z_0}. Let $s = 1, \ldots, 16$ denote the 16 strata. With a relatively large sample size, one may increase the number of strata. Within stratum s, we assign the weight $\frac{\text{pr}(Z=z|A=a)}{\text{pr}(Z=z|A=a,S=s)}$ to the individuals displaying mediator value z in treatment group a for $a = 0, 1$ and $z = 0, 1$. We then examine covariate balance across the four treatment-by-mediator categories in the MMWS-adjusted sample.

Step 5: Estimate the mediator effect on the outcome under each treatment condition. Applying MMWS 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.

Step 6: Create a duplicate and compute the nonparametric weight. We estimate RMPW nonparametrically under the stratification described in Step 4. To estimate $E(Y_{1Z_0})$, the nonparametric weight is

$$\text{RMPW} = \frac{\text{pr}(Z = z|A = 0, S = s)}{\text{pr}(Z = z|A = 1, S = s)},$$

for an LFA unit in stratum s displaying mediator value z; to estimate $E(Y_{0Z_1})$, the weight is given by:

$$\text{RMPW} = \frac{\text{pr}(Z = z|A = 1, S = s)}{\text{pr}(Z = z|A = 0, S = s)},$$

for a control unit in stratum s displaying mediator value z.

The nonparametric RMPW is then applied to the outcome models specified in Equations 1 and 2. Under a four-by-four stratification, the direct effect estimate is 1.34 ($SE = 0.79$, $t = 1.70$, $p = .09$), and the indirect effect estimate is -0.93 ($SE = 0.38$, $t = -2.43$, $p < .05$). Further decomposing the indirect effect, we estimate the pure indirect effect (coefficient $= 0.45$, $SE = 0.30$, $t = 1.50$, $p = .13$) and the natural treatment-by-mediator interaction effect (coefficient $= -1.38$, $SE = 0.49$, $t = -2.85$, $p < .01$). These point estimates are similar to the parametric weighting results. Yet the estimation with nonparametric weighting appears to be relatively efficient, which allows us to detect a statistically significant negative indirect effect of the treatment.

Simulations

We conduct a series of Monte Carlo simulations to assess the performance of the nonparametric 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 nonparametric RMPW, we also compare 3×3 strata with 4×4 strata. Additionally, we compare the robustness of estimation between the parametric and the nonparametric procedures when the propensity score models are misspecified in their functional forms. 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. The direct effect and the indirect effect are both set to be zero in Simulation a and are nonzero in Simulations b and c. Simulations a and b set the employment rates similar to those in the NEWWS data, while Simulation c increases the employment rate under LFA and decreases that under the control condition, which essentially reduces the statistical power under the same total sample size.

TABLE 4.
Summary of Simulation Results Under Correct Specification of the Propensity Score Models

		$N = 5{,}000$			$N = 800$		
	Model	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 ($\hat{\gamma}^{(IE.1)}$)							
% 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

Note. MSE = mean square error; RMPW = ratio-of-mediator-probability weighting; NRMPW = nonparametric ratio-of-mediator-probability weighting.

The evaluation criteria for causal effect estimate $\hat{\gamma}$ (either $\hat{\gamma}^{(DE)}$ for the direct effect or $\hat{\gamma}^{(IE)}$ for the indirect effect) include the following: (1) bias in the point estimate: $E(\hat{\gamma}) - \gamma$; (2) sampling variability of the point estimate: $\mathrm{Var}(\hat{\gamma}) = E[\hat{\gamma} - E(\hat{\gamma})]^2$; (3) mean square error (MSE): $E\left[(\hat{\gamma} - \gamma)^2\right] = \mathrm{Var}(\hat{\gamma}) + [E(\hat{\gamma}) - \gamma]^2$; and (4) bias in the *SE* estimate: $E[\hat{\sigma}(\hat{\gamma})] - \sigma(\hat{\gamma})$. A naïve RMPW procedure is employed as if the data were generated from a sequential experimental design. Results from this naïve analysis serve as the baseline for assessing the extent to which the parametric and nonparametric RMPW procedures successfully remove bias associated with the pretreatment covariates.

Table 4 summarizes the key results corresponding to the three sets of parameter values when the propensity score models are correctly specified. The parametric and nonparametric RMPW procedures both perform generally well

in all three cases. The parametric procedure removes nearly 100% of the bias; the nonparametric procedure with 3×3 strata removes 85% or more of the initial bias, while that with 4×4 strata removes 90% or more of the bias when the sample size is relatively large. The nonparametric estimates often show a higher efficiency and a smaller MSE when compared with the parametric estimates. However, in a relatively small sample, an increase in the number of strata seems to result in a loss of efficiency without further reducing bias, especially when $E(\theta_{Z_0})$ and $E(\theta_{Z_1})$ are shifting away from 0.5. Finally, comparing the average *SE* estimates with the corresponding sampling standard deviations approximated on the basis of 1,000 samples, we find the discrepancy close to zero across all cases and never exceeding 0.047 standard deviations of a potential outcome in any single case.

We then modify the data generation plan to allow for a comparison between the parametric and the nonparametric RMPW procedures when nonlinear, nonadditive propensity score models are misspecified as linear additive. According to our results (please see supplementary material online), regardless of sample size, the parametric RMPW procedure generates estimates that are increasingly biased as the degree of nonlinearity or nonadditivity increases. In contrast, the nonparametric RMPW results remain robust in all cases.

We have conducted additional simulations (results available upon request) showing that, when there is no treatment-by-mediator interaction in the simulated data, the RMPW results replicate those from path analysis and IV results. As expected, RMPW outperforms these conventional methods, especially in bias correction, when the assumption of no treatment-by-mediator interaction does not hold.

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. Conventional analysis typically ignores the interaction effect and therefore generates biased estimates of the indirect effect and the direct effect. When only the treatment is randomized, how to adjust for a large number of pretreatment covariates that confound the mediator–outcome relationship is another major concern.

This article has described a relatively new approach to causal mediation analysis that addresses these challenges. In addition to estimating the population average potential outcome should all the units be assigned to the control condition and the population average potential outcome should all the units be assigned to the experimental condition, the RMPW strategy reconstructs the data to estimate 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 weighting transforms 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 *SE*s. 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 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. Hence, we additionally transform the mediator distribution of the control group to resemble that of the experimental group.

This article has provided details of the analytic steps for implementing the RMPW strategy. According to the simulation results, the parametric and nonparametric RMPW procedures both demonstrate satisfactory performance under the identification assumptions. As anticipated, the parametric RMPW results are sensitive to possible misspecifications of the functional form of the propensity score models. In contrast, the nonparametric RMPW results are relatively robust and efficient. There are also nonparametric approaches to propensity score estimation, including generalized boosted models, which reduce model misspecification errors (McCaffrey, Ridgeway, & Morral, 2004). Future research may investigate the application of these approaches to RMPW analysis.

The RMPW strategy shows its strengths in comparison with many existing methods that similarly require the sequential ignorability. The conventional path analysis/SEM approach and the marginal structural models additionally require the assumption that there is no treatment-by-mediator interaction. The latest advancements in causal mediation analysis accommodate treatment-by-mediator interactions, often 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 tend to bias causal effect estimation (Drake, 1993). The RMPW strategy and several other closely related weighting strategies relax the no-treatment-by-mediator interaction assumption; the weighted outcome models simply provide mean contrasts between the potential outcomes defined earlier and therefore are nonparametric in nature.

Moreover, the RMPW strategy and other related weighting strategies have broad applications regardless of the distribution of the outcome, the distribution of the mediator, or the functional relationship between the outcome and the mediator. The weights as specified in their general forms in Equations 3 and 4 can be applied to multivalued mediators without changing the outcome model

specifications in Equations 1 and 2. One may also test whether the causal mediation mechanism differs across subpopulations. We provide additional Stata code in the online supplementary material for RMPW analysis with multivalued mediators and for moderated mediation analysis. Finally, in analyses of quasi-experimental data, RMPW can be easily combined with IPTW or MMWS, as shown in Theorem 3, to further remove treatment selection bias (Hong and Nomi, 2012). In the case of continuous mediators, the ratio of probabilities of mediator values may be replaced by the ratio of mediator densities. The IPW strategy (Huber, 2014) alternatively estimates the ratio of probabilities of treatment, given the mediator and the pretreatment covariates, and appears convenient for accommodating continuous mediators. The performance of these alternative weighting strategies needs to be compared in future research.

In most existing methods for causal mediation analysis, the indirect effect and sometimes the direct effect are each represented as a function of multiple parameter values. Extra programming using the delta method or bootstrapping is required for estimating the asymptotic or empirical *SE*s of the sample estimates. Using the off-the-shelf statistical packages, the RMPW strategy presented in this article generates cluster-robust *SE*s for the causal effect estimates and provides immediate tests of the null hypotheses. In general, the analyst needs to consider the statistical uncertainty in the two-step estimation—that is, estimating the propensity score model coefficients to construct the weight followed by estimating the causal effects—in computing asymptotic standard errors (Bein et al., 2015; Cameron & Trivedi, 2005; Wooldridge, 2012). This is implemented in the stand-alone free RMPW software program and can also be carried out with the generalized method-of-moments procedure in Stata.

It is crucial to emphasize that RMPW identifies the causal effects of interest under the untestable assumptions of sequential ignorability. Even though the ignorability of treatment assignment can be warranted by treatment randomization, mediator value assignment is typically not randomized. Therefore, similar to most existing methods described previously, the causal validity of an RMPW analysis depends critically on the quality of the baseline data in terms of the extent to which they predict the mediator and the outcome. Cook and Steiner (2010) highlighted the special role of pretest measures relative to all other covariates. In the NEWWS application, the pretest measures include, most importantly, baseline employment record and baseline depression score. Schochet and Burghardt (2007) suggested collecting baseline predictions by program staff on the likely program experiences of program-eligible individuals (e.g., whether one would likely be employed under LFA). The RMPW approach is recommended, in the end, only if there are credible baseline covariates that can remove a large portion of selection bias.

The mediator–outcome relationship may be additionally confounded by posttreatment covariates. 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 might also independently predict depression at the 2-year follow-up. In causal mediation analyses that allow for treatment-by-mediator interactions, the potential confounding effect of observed posttreatment covariates cannot be adjusted for directly (Avin, Shpitser, & Pearl, 2005; 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. Viewing an important posttreatment covariate as a mediator temporally precedent to the focal mediator, we may extend the RMPW strategy to a causal mediation analysis involving two consecutive mediators (Hong, 2015; Huber, 2014). If such a posttreatment covariate is unobserved, sensitivity analysis may be employed to assess the consequence of the possible omission (Imai et al., 2010a, 2010b; VanderWeele, 2010). In the cases in which the observed pretreatment covariates have explained nearly all the systematic variation in the outcome, however, the remaining potential bias associated with the omitted pretreatment and posttreatment covariates may become negligible.

Finally, the problem of overfitting the propensity score models is a potential concern when the sample size is small relative to the number of parameters in the prediction model (Hawkins, 2004). For example, when the propensity score model for employment tailored to the control sample is applied to the LFA sample, the prediction error may become inflated. Future studies may incorporate either cross-validation or leave-one-out bootstrapping to avoid model overfitting and assess their relative effectiveness in causal mediation analysis (Abadie, Chingos, & West, 2013; Hastie, Tibshirani, & Friedman, 2009). The cross-validation strategy sets apart a training sample to which a prediction model is fitted and is then applied to a cross-validation sample. The leave-one-out bootstrap strategy fits a model to $N - 1$ cases in a sample of N and then uses the fitted model to make a prediction for the case that has been left out. The model is then refitted each time when a different case is left out. Peck (2003, 2007) and Schochet and Burghardt (2007) provided applications of the former to propensity score-based subgroup analysis.

The RMPW strategy is also applicable to cluster randomized designs, which are common in education. In such a design, schools or classrooms are randomized to the experimental or the control condition. While individual-level outcomes are typically of interest, the theorized mediator could be either at the cluster or the individual level. For a cluster-level mediator, the propensity score model under each treatment condition will be analyzed at the cluster level. RMPW will be computed subsequently for each cluster. The outcome model will include a cluster-level treatment indicator and a cluster-level duplicate indicator. However, standard multilevel software programs do not decompose the variance appropriately when there is duplication, which would complicate the estimation

of model-based *SE*s. One may obtain robust *SE*s for the causal effect estimates and may use bootstrapping to construct a confidence interval for each causal effect. For an individual-level mediator, the RMPW procedure is similar except that the propensity score model analysis and the computation of RMPW will be conducted at the individual level instead. The RMPW strategy, when applied to cluster randomized designs, assumes intact clusters, no interference between clusters and between individuals within a cluster, as well as the sequential ignorability.

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. We will show that, when the mediator–outcome relationship depends on the treatment, the bias in the path analysis estimate of the direct effect is a product of three elements: the treatment-by-mediator interaction effect, the treatment effect on the mediator, and the proportion of units assigned to the control group. The bias in the indirect effect estimate takes the opposite sign.

Let $A = 1$ if a unit is treated and 0 if the unit is assigned to the control condition. As a mediator, Z_a is a function of treatment assignment a for $a = 0, 1$ and can be generated by $Z_a = \beta_0 + \beta_1 a + \varepsilon_z$ where ε_z is a random error. In other words, we have that $\text{pr}(Z_0 = 1) = \beta_0$ and that $\text{pr}(Z_1 = 1) = \beta_0 + \beta_1$. 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 $Y = \gamma_0 + \gamma_1 A + \gamma_2 Z + \text{e}$. We can show that $\gamma_2 = \theta_2 + \theta_3 \times \text{pr}(A = 1)$. The path analysis model represents the indirect effect as $\gamma_2\beta_1 = \theta_2\beta_1 + \theta_3\beta_1 \times \text{pr}(A = 1) = (\theta_2 + \theta_3)\beta_1 - \theta_3\beta_1 \times \text{pr}(A = 0)$, which contains the bias $-\theta_3\beta_1 \times \text{pr}(A = 0)$. This is equivalent to $-E\{(Y_{11} - Y_{01}) - (Y_{10} - Y_{00})\} \times \{\text{pr}(Z_1 = 1) - \text{pr}(Z_0 = 1)\} \times \text{pr}(A = 0)$.

Appendix B

Proof of Theorem 3

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

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

By Assumptions 1 and 2, the previously mentioned equation is equal to

$$\begin{aligned}
&E\left\{E\left[Y_{aZ_{a'}}|A=a,\mathbf{X}\right]\right\}\\
&\equiv \int_{\mathbf{x}}\int_{z}\int_{y} y \times f(Y_{az}=y|A=a,Z_{a'}=z,\mathbf{X}=\mathbf{x}) \times p^{(a')}(Z_{a'}=z|A=a,\mathbf{X}=\mathbf{x})\\
&\times h(\mathbf{X}=\mathbf{x})dydzd\mathbf{x},
\end{aligned}$$

which, by Assumptions 3, 5, and 6, is equal to

$$\begin{aligned}
&\int_{\mathbf{x}}\int_{z}\int_{y} y \times f(Y_{az}=y|A=a,Z_{a}=z,\mathbf{X}=\mathbf{x}) \times p^{(a')}(Z_{a'}=z|A=a',\mathbf{X}=\mathbf{x})\\
&\times h(\mathbf{X}=\mathbf{x})dydzd\mathbf{x},
\end{aligned}$$

which, by Bayes Theorem and by Assumptions 1, 2, and 4 is equal to

$$\begin{aligned}
&\int_{\mathbf{x}}\int_{z}\int_{y} y \times f(Y_{az}=y|A=a,Z_{a}=z,\mathbf{X}=\mathbf{x}) \times p^{(a)}(Z_{a}=z|A=a,\mathbf{X}=\mathbf{x})\\
&\times h(\mathbf{X}=\mathbf{x}|A=a) \times \frac{p^{(a')}(Z_{a'}=z|A=a',\mathbf{X}=\mathbf{x})}{p^{(a)}(Z_{a}=z|A=a,\mathbf{X}=\mathbf{x})} \times \frac{q(A=a)}{q(A=a|\mathbf{X}=\mathbf{x})}dydzd\mathbf{x}\\
&= E(Y^*|A=a),
\end{aligned}$$

where $Y^* = \omega^{(a,Z_{a'})}Y$ and $\omega^{(a,Z_{a'})} = \frac{p^{(a')}(Z_{a'}=z|A=a',\mathbf{X}=\mathbf{x})}{p^{(a)}(Z_{a}=z|A=a,\mathbf{X}=\mathbf{x})} \times \frac{q(A=a)}{q(A=a|\mathbf{X}=\mathbf{x})}$.

This concludes the proof.

Acknowledgments

The authors owe special thanks to Howard Bloom, Larry Hedges, Stephen Raudenbush, Patrick Shrout, seminar participants at the University of Wisconsin-Madison, the University of Chicago, Northwestern University, the University of California-Los Angeles, Carnegie Mellon University, the Prevention Science and Methodology Group, participants at the William T. Grant Foundation "Learning from variation in program effects" conference, and four anonymous reviewers for their comments on earlier versions of the article. Daniel McCaffrey provided extremely important editorial guidance that led to critical improvements in the final manuscript. Richard Congdon deserves major credit for designing and programming the ratio-of-mediator-probability weighting (RMPW) software.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a major research grant entitled "Improving Research on Instruction: Models, Designs, and Analytic Methods" funded by the Spencer Foundation, a Scholars Award from the William T. Grant Foundation, and the start-up funds from the University of Chicago for the first author. Additional support came from the Institute of Education Sciences, U.S. Department of Education, through Grant R305D120020 to the National Opinion Research Center. The opinions expressed are those of the authors and do not represent views of the funding agencies listed here.

Supplementary Material

The online appendices are available at http://jeb.sagepub.com/supplemental.

References

Abadie, A., Chingos, M. M., & West, M. R. (2013). *Endogenous stratification in randomized experiments* (Working Paper 19742). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w19742

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*, 150–161.

Avin, C., Shpitser, I., & Pearl, J. (2005). Identifiability of path-specific effects. *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence* (pp. 357–363). Edinburgh, Scotland, UK.

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.

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

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

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge, England: Cambridge University Press.

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

Coffman, D. L., & Zhong, W. (2012). Assessing mediation using marginal structural models in the presence of confounding and moderation. *Psychological Methods*, *17*, 642–664.

Collins, L., Graham, J., & Flaherty, B. (1998). An alternative framework for defining mediation. *Multivariate Behavioral Research*, *33*, 295–312.

Cook, T., & Steiner, P. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods*, *15*, 56–68.

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

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

Grogger, J., & Karoly, L. A. (2005). *Welfare reform: Effects of a decade of change*. Cambridge, MA: Harvard University Press.

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *44*, 1–12.

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. (2010a). Ratio of mediator probability weighting for estimating natural direct and indirect effects. In *JSM Proceedings* (pp. 2401–2415), Biometrics Section. 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*, 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*, 44–60.

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

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*, 261–289.

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

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

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

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*, 765–789.

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*, 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.

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, *22*, 523–539.

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

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*, 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 MacArthur approaches. *Health Psychology*, *27*, S101–S108.

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

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

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

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*, 403–425.

Morris, P. A. (2008). Welfare program implementation and parents' depression. *Social Service Review*, *82*, 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*, 852–863.

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

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

Peck, L. R. (2003). Subgroup analysis in social experiments: Measuring program impacts based on posttreatment choice. *American Journal of Evaluation*, *24*, 157–187.

Peck, L. R. (2007). What are the effects of welfare sanction policies? Or, using propensity scores as a subgroup indicator to learn more from social experiments. *American Journal of Evaluation*, *28*, 256–274.

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

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology*, *76*, 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*, 879–891.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, *42*, 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*, 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*, 303–332.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In M. Elizabeth Halloran & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp. 95–134). New York, NY: Springer.

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

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

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

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 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*, 961–962.

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

Schochet, P. Z., & Burghardt, J. (2007). Using propensity scoring to estimate program-related subgroup impacts in experimental program evaluations. *Evaluation Review*, *31*, 95–120.

Sheets, V. L., & Braver, S. L. (1999). Organizational status and perceived sexual harassment: Detecting the mediators of a null effect. *Personality and Social Psychology Bulletin*, *25*, 1159–1171.

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

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

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*, 845–851.

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

Tchetgen Tchetgen, E. J., & Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *The Annals of Statistics*, *40*, 1816–1845.

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*, 137–150.

van der Laan, M. J., & Petersen, M. L. (2008). Direct effect models. *The International Journal of Biostatistics*, *4*, 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. (2013). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, *24*, 224–232.

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.

Waernbaum, I. (2012). Model misspecification and robustness in causal inference: Comparing matching with doubly robust estimation. *Statistics in Medicine*, *31*, 1572–1581.

Wooldridge, J. M. (2012). *Introductory econometrics: A modern approach* (5th ed.). Mason, OH: Cengage Learning.

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

Authors

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

JONAH DEUTSCH is a researcher at Mathematica Policy Research, 111 East Wacker Drive, Suite 920, Chicago, IL 60601; e-mail: jdeutsch@mathematica-mpr.com. His research interests are program evaluation methods, value-added modeling, and education policy.

HEATHER D. HILL is an associate professor at the Daniel J. Evans School of Public Affairs at the University of Washington, 323 Parrington Hall Box 353055, Seattle, WA 98195; e-mail: hdhill@uw.edu. Her research interests are social policy, family economic circumstances, and child development.

Manuscript received April 6, 2014
First revision received September 28, 2014
Second revision received January 6, 2015
Accepted March 12, 2015