Technical Report on Two-Step Estimation in RMPW Analysis

Edward Bein

Jonah Deutsch

Guanglei Hong

Kristin E. Porter

Xu Qin

Cheng Yang

Abstract

Causal mediation analysis decomposes the total effect of a treatment on the outcome into an indirect effect transmitted through a focal mediator and a direct effect. Ratio-of-mediator-probability weighting (RMPW) conveniently estimates these causal effects by adjusting for the confounding impact of a large number of pretreatment covariates through propensity score-based weighting. A weight is computed as a ratio of two estimated counterfactual probabilities for the mediator with respect to the two alternative treatment conditions. The causal effects of interest are then estimated when the estimated weight is applied to the sample data. Statistical inferences obtained from this two-step estimation procedure are potentially problematic if the estimated standard errors of the causal effect estimates do not reflect the sampling uncertainty in the estimation of the weights. This technical report extends to RMPW analysis a solution to the two-step estimation problem by stacking the score functions from both steps. We derive the asymptotic variance-covariance matrix for the indirect and direct effect two-step estimators, provide simulation results, and illustrate with an application study. Our simulation results indicate that the sampling uncertainty in the estimated weights should not be ignored. The standard error estimation using the stacking procedure offers a viable alternative to the bootstrapped standard error estimation.

Keywords

Direct effect, indirect effect, m-estimation, mediation analysis, method of moments, ratio-of-mediator-probability weighting

1. Introduction

Causal mediation analysis through ratio-of-mediator-probability weighting (RMPW), developed by Hong and others (Hong, 2010, 2015; Hong, Deutsch, & Hill, 2011, 2015; Hong & Nomi, 2012; Lange, Rasmussen, & Thygesen, 2013; Tchetgen Tchetgen & Shpister, 2012), is a weighting-based approach to identifying and estimating natural direct and indirect effects that compose the total effect of a treatment on the outcome. Huber (2014) and Tchetgen Tchetgen (2013) proposed related strategies employing weights that are mathematically equivalent to RMPW. All these weighting strategies are aimed at consistently estimating the population average potential outcomes that define the average natural direct effect and the average natural indirect effect. However, even in an experiment in which the treatment has been randomized, typically the mediator values are not randomized under each treatment condition. Therefore, the true ratio-of-mediator-probability weights are unknown and must be estimated from the sample data. It is well-known that the necessity of estimating weights often affects the variance of weighting-based estimators relative to what their variance would be if the true weights could somehow be employed (Cameron & Trivedi, 2005, pp.200-202; Wooldridge, 2010, pp. 409-413). This article contributes to the literature by deriving and estimating the variance of RMPW method of moments (MOM) estimators of natural direct and indirect effects, accounting for the use of estimated weights.

In the causal inference literature, there are several important examples of propensity score-based weighting in which, counterintuitively, standard errors of causal effect estimators are *smaller* when applying the estimated weights rather than the true (but typically unknown) weights. These include inverse-probability-of-treatment-weighted (IPTW) estimators of the average treatment effect and of marginal structural models more generally (see Wooldridge,

2010, pp. 922-923, and Robins, Hernan, & Brumback, 2000, p. 554, respectively). This seemingly counterintuitive result is in alignment with Rosenbaum's (1987) observation that, in comparison with the true propensity score, the estimated propensity score produces greater control of imbalances in the pretreatment covariates between the experimental group and the control group in a given sample. For such estimators, conservative statistical inference can be obtained by ignoring weight estimation. However, as we show later, this nice property is not shared by our RMPW estimators, and inference based on ignoring the reality that the RMPW weights must be estimated will sometimes produce downward-biased p-values and overly narrow confidence intervals.

Bootstrapping (Efron & Tibshirani, 1993) has often been recommended as a solution to the two-step estimation problem. Applied researchers resort to bootstrapping when analytic solutions are unfeasible or are not yet available. However, bootstrapping is computationally intensive and may generate unstable results when the sample size is relatively small. The statistical properties of bootstrapped standard errors are yet to be examined when applied to RMPW-based causal mediation analysis. In this study, we derive the correct large-sample standard errors given the necessity of weight estimation and explain how to consistently estimate them. For both large and small samples, we compare the inferential accuracy of these standard error estimators to the bootstrap standard error estimators as well as to the standard error estimators when the estimated weights are mistaken for the true weights.

We proceed as follows. Section 2 briefly reviews the definition of natural direct and indirect effects. Section 3 summarizes the RMPW estimation approach. In Section 4, we derive the asymptotic variance of the RMPW estimators, taking into account the need for weight estimation. Section 5 discusses variance estimation. We present in Section 6 a simulation study

comparing the performance of different methods of variance estimation. In Section 7, we apply the proposed variance estimators to the National Evaluation of Welfare-to-Work Strategies (NEWWS). Section 8 discusses implications for practice and next steps. Appendices provide technical details and describe Stata commands that implement our variance estimators.

2. Direct and Indirect Effects

In scientific research, the theoretical mechanism through which a treatment influences an outcome typically involves one or more mediators. A mediator is an intermediate outcome that is expected to be influenced by the initial treatment and will subsequently influence the final outcome. Direct and indirect effects of treatment relate to the extent to which the treatment effect on the outcome is mediated by a particular mediator or set of mediators. Consider the the simplest case in which there is a single mediator of interest. At one extreme, the effect of a treatment on the outcome is transmitted entirely through the mediator such that the treatment has no effect on the outcome via any other causal pathways. In this case, there is no direct effect. At the other extreme, either the treatment does not change the mediator value or any effect the treatment has on the mediator may have no subsequent effect on the outcome. In this case, the direct effect is the total effect. In social policy research, for any specified mediator or set of mediators, treatments tend to exhibit both direct and indirect effects on outcomes.

For example, analyzing the data collected from Riverside, California as a subset of the experimental data in the National Evaluation of Welfare-to-Work Strategies (NEWWS), Hong, Deutsch, and Hill (2011, 2015) examined the impact of a welfare-to-work strategy on the psychological well-being of welfare recipients with young children. The intervention emphasized and supported labor force participation and threatened sanctions if the participants failed to meet the requirements. This was sharply contrasted with the control condition that simply offered cash

assistance to welfare recipients. The researchers hypothesized that being assigned at random to the experimental condition would likely increase one's probability of employment. Becoming employed under the experimental condition would subsequently reduce depressive symptoms. Hence the treatment assignment is expected to generate a desired negative indirect effect on maternal depression. They further hypothesized that, without an increase in the probability of employment, being assigned to the experimental condition rather than the control condition would inadvertently heighten maternal depression, leading to an undesired positive direct effect of the treatment on maternal depression. The hypothesized counteracting indirect effect and direct effect may provide a theoretical explanation for the zero effect of the total treatment effect on maternal depression.

Below we use *T* to denote the treatment assignment, *M* for the mediator, and *Y* for the outcome. In the NEWWS application described above, *T* indicates whether a welfare recipient was assigned to the experimental condition or the control condition; *M* is employment during the period after randomization; and *Y* is maternal depression at the follow-up. Extending the Neyman-Rubin potential outcomes model of treatment effects (Holland, 1986; Neyman, 1923; Rubin, 1978), researchers have formally defined the direct and indirect effects that decompose the total effect of a treatment on an outcome (Pearl, 2001; Robins & Greenland, 1992). Let $T = 1$ if an individual was assigned to the experimental condition; and let $T = 0$ if the same individual was assigned to the control condition instead. Correspondingly, we posit that each individual in the population would have two potential intermediate outcomes denoted $M(0)$ and $M(1)$. In the NEWWS application, the former is a random variable taking value 0 if one is unemployed and 1 if employed under the control condition; the latter is another random variable representing the same individual's employment status if assigned to the experimental condition. $Y\big(1, M(1)\big)$ is the

potential outcome (i.e., the potential level of maternal depression) that the individual would display if assigned to the experimental condition; and $Y\big(0, M(0)\big)$ is the potential outcome that the same individual would display if assigned to the control condition. These are commonly written as $Y(1)$ and $Y(0)$, respectively, in other causal inference contexts. The difference between $Y\big(1, M(1)\big)$ and $Y\big(0, M(0)\big)$ defines the individual specific total effect of the treatment on maternal depression.

To understand the mediating role of employment, we need to introduce another potential outcome $Y\big(1, M(0)\big)$ that represents, in the current example, the level of maternal depression that one would display if assigned to the experimental condition yet counterfactually experiencing an employment status as one would under the control condition. For example, suppose that an individual would have a 0.7 probability of being employed if assigned to the experimental condition and a 0.4 probability of being employed under the control condition. $Y\big(1, M(0)\big)$ is the individual's potential level of depression if assigned to the experimental condition when her probability of employment would counterfactually be 0.4 rather than 0.7. The individual-specific natural direct effect is $Y\big(1, M(0)\big) - Y(0, M(0))$. This is the treatment effect on maternal depression should the treatment fail to change the individual's employment experience. The individual-specific natural indirect effect is $Y\big(1, M(1)\big) - Y(1, M(0))$. This is defined as the treatment effect on maternal depression solely attributable to the treatment-induced change in her employment experience when the individual is assigned to the experimental condition. In the earlier example, it is the change in maternal depression under the experimental condition when the individual's probability of employment is raised from 0.4 to 0.7.

Causal mediation analysis focuses on identifying and estimating the *population average natural direct effect*

$$NDE \equiv E[Y\big(1, M(0)\big) - Y(0, M(0))] \quad (1)$$

and the *population average natural indirect effect*

$$NIE \equiv E[Y\big(1, M(1)\big) - Y(1, M(0))] \quad (2)$$

Clearly,

$$NIE + NDE = E\big[Y\big(1, M(1)\big) - Y\big(0, M(0)\big)\big] = \text{Total Effect.}$$

According to Pearl (2001), these causal effects are "natural" rather than "controlled" because, under each treatment condition, the mediators are allowed to take random values that vary naturally across individuals rather than taking fixed values strictly controlled by the experimenter. In Robins and Greenland's (1992) terminology, *NDE* and *NIE* are called the "pure direct effect" and the "total indirect effect," respectively. The latter can be further decomposed when the researcher wants to know whether the treatment-induced change in the mediator (such as an increase in the probability of employment from .4 to .7) would influence the outcome differently between the experimental condition and the control condition.

The above definitions of the causal effects are provided under the simplifying assumption that an individual's potential mediators and potential outcomes are unaffected by other individuals' treatment assignments and mediator value assignments.[1] This is related to the *Stable Unit Treatment Value Assumption* (SUTVA) (Rubin, 1980). Yet this framework allows an individual's potential mediator under a given treatment condition to take random rather than fixed values. In the above example, the individual whose probability of employment is 0.7 when

[1] This assumption would not hold if there is a general equilibrium effect. For example, when employment opportunities are limited in a local job market, assigning a greater proportion of welfare applicants to the experimental condition will likely generate a greater demand for low-paying positions. As one's probability of employment diminishes, the depression level of an individual who remains unemployed under the experimental condition will likely become heightened. Although beyond the scope of the current study, possible spill-overs of treatment effects have emerged as an important topic in the causal inference literature (Hong, 2004, 2015; Hong & Raudenbush, 2006, 2013; Hudgens & Halloran, 2008; Sobel, 2006). Possible spill-overs of mediator effects are particularly relevant to causal mediation analysis (Hong, 2015).

assigned to the experimental condition may display a mediator value $M(1) = 1$ if an employment opportunity becomes available and may display an alternative mediator value $M(1) = 0$ if such an opportunity disappears in the market. Correspondingly, the individual assigned to the experimental condition has a 0.7 probability of displaying potential outcome $Y(1,1)$ and a 0.3 probability of displaying potential outcome $Y(1,0)$. Hence, when $T = 1$, the individual's potential outcome $Y(1)$ may not have a *stable* value; the same can be said of the individual's potential outcome under the control condition $Y(0)$.

3. RMPW Method of Moments Estimation of Natural Direct and Indirect Effects

Here we focus on decomposing the total effect of the treatment on the outcome into a natural direct effect and a natural indirect effect defined in (1) and (2). This section outlines the RMPW method-of-moments (MOM) estimation of *NDE* and *NIE*, which involves the estimation of the population average potential outcomes $E\left[Y\left(1, M(1)\right)\right]$, $E\left[Y\left(0, M(0)\right)\right]$, and $E\left[Y\left(1, M(0)\right)\right]$. Further decomposition of *NIE* is straightforward and will be addressed in the discussion section. For simplicity, we consider a randomized experiment with a binary treatment and a binary mediator, although the RMPW MOM strategy can be easily extended to non-randomized treatment assignments and multivalued mediators (Hong, 2015; Hong & Nomi, 2012).

A comprehensive presentation of the rationale of RMPW and the identification assumptions has appeared elsewhere (Hong, 2010, 2015; Hong, Deutsch, & Hill, 2011, 2015; Lange, Vansteelandt, & Bekaert, 2012). Similar to other related weighting strategies (Huber, 2014; Tchetgen Tchetgen, 2013; Tchetgen Tchetgen & Shpitser, 2012), the theoretical rationale is to equate the distribution of the mediator in the experimental group and that in the control group through weighting. Transforming the mediator distribution in the experimental group

through weighting then makes possible the identification of $E\left[Y\left(1, M(0)\right)\right]$ essential to treatment effect decomposition. The identification requires the sequential ignorability assumption (Imai et al, 2010a, 2010b). Namely, the treatment assignment and the mediator value assignment under each treatment can be viewed as randomized within levels of the observed pretreatment covariates. This section proceeds under the pretense that the RMPW weights are known. In the next section, we consider the realistic but more complicated case where these weights are unknown and must be estimated.

Suppose that a random sample of size *n* has been drawn from the population of interest. The i^{th} participant in the sample has observed data

$$O_i = (X_i, T_i, M_i, Y_i),$$

where X_i is a vector of baseline (i.e., pretreatment) covariates; T_i is the random treatment assignment indicator; M_i is an observed intermediate variable; and Y_i is an observed outcome.

The *ratio-of-mediator probability weight* for the i^{th} participant is

$$w_i \equiv w_i(M_i, X_i) = M_i \frac{P(M_i = 1|T_i = 0, X_i)}{P(M_i = 1|T_i = 1, X_i)} + (1 - M_i)\frac{P(M_i = 0|T_i = 0, X_i)}{P(M_i = 0|T_i = 1, X_i)}. \quad (3)$$

From (3),

$$E[w_i(M_i, X_i)|\, T_i = 1, X_i] = P(M_i = 1|T_i = 0, X_i) + P(M_i = 0|T_i = 0, X_i) = 1,$$

which implies that

$$E[w_i(M_i, X_i)|\, T_i = 1] = 1. \quad (4)$$

We define the shorthand for each population average potential outcome:

$$\mu_0 \equiv E\left[Y\left(0, M(0)\right)\right];$$

$$\mu_* \equiv E\left[Y\left(1, M(0)\right)\right];$$

$$\mu_1 \equiv E\big[Y\big(1, M(1)\big)\big].$$

These are useful because

$$NIE = \mu_1 - \mu_*,$$

$$NDE = \mu_* - \mu_0.$$

We now employ standard MOM technology to develop RMPW estimators of *NIE* and *NDE*.[2] MOM estimation is described in Cameron and Trivedi (2005, section pp. 166-168) and is a special case of the econometric technique of generalized method of moments (GMM) estimation (Cameron & Trivedi, 2005, chapter 6; Hansen, 1982), which it predates.[3] Consider the following estimating or score functions of observed data and parameters:

$$h_{0i}(O_i, \mu) = (Y_i - \mu)(1 - T_i),$$

$$h_{*i}(O_i, \mu) = (Y_i - \mu)w_i T_i,$$

$$h_{1i}(O_i, \mu) = (Y_i - \mu)T_i.$$

These estimating functions satisfy the population moment conditions at the true parameter values:

$$E\big[h_{0i}(O_i, \mu_0)\big] = 0,$$

$$E\big[h_{*i}(O_i, \mu_*)\big] = 0,$$

$$E\big[h_{1i}(O_i, \mu_1)\big] = 0.$$

The second moment condition relies on a key result from the past literature on RMPW (Hong, 2010, 2015; Hong, Deutsch, and Hill, 2011, 2015; Hong & Nomi, 2012), which we refer to as the RMPW theorem: $\mu_* = E[wY|T = 1]$. We then have

[2] Hong and colleagues (Hong, 2010, 2015; Hong, Deutsch, and Hill, 2015) presented weighted least squares estimators of the natural direct and indirect effects, using the RMPW weights. Appendix 1 shows that these estimators are equivalent to the MOM estimators developed in this section.

[3] Some authors refer to MOM estimation as "m-estimation" (e.g., Stefanski & Boos, 2002), while others reserve the latter term for a class of estimation approaches that include MOM (e.g., Wooldridge, 2010, chapter 12).

$$
\begin{aligned}
E\big[h_{*i}(O_i, \mu_*)\big] &= E\{E[(Y_i - \mu_*)w_i T_i | T_i]\} \\
&= P(T_i = 1)E[(Y_i - \mu_*)w_i | T_i = 1] \\
&= P(T_i = 1)(E[Y_i w_i | T_i = 1] - \mu_* E[w_i | T_i = 1]) \\
&= P(T_i = 1)\big(\mu_*(1 - E[w_i | T_i = 1])\big) \\
&= 0,
\end{aligned}
$$

where the next-to-last equality follows from the RMPW theorem and the last equality follows from equation (4).

The parameters μ_0, μ_*, and μ_1 are respectively estimated by MOM estimators $\hat{\mu}_0$, $\hat{\mu}_*$, and $\hat{\mu}_1$ (Cameron & Trivedi, 2005, p. 172) that satisfy the following sample moment conditions

$$\frac{1}{n}\sum_{i=1}^{n} h_{0i}(O_i, \hat{\mu}_0) = 0,$$

$$\frac{1}{n}\sum_{i=1}^{n} h_{*i}(O_i, \hat{\mu}_*) = 0,$$

$$\frac{1}{n}\sum_{i=1}^{n} h_{1i}(O_i, \hat{\mu}_1) = 0.$$

Let $n_1 = \sum_{i=1}^{n} T_i$ and $n_0 = \sum_{i=1}^{n}(1 - T_i)$ be the number of sampled participants in the experimental and control groups, respectively. Solving the estimating equations gives

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} Y_i\,(1 - T_i)}{n_0},$$

$$\hat{\mu}_* = \frac{\sum_{i=1}^{n} Y_i w_i T_i}{\sum_{i=1}^{n} w_i T_i},$$

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} Y_i\,T_i}{n_1}.$$

The variance matrix for these estimators follows from a standard result for MOM estimators (Cameron & Trivedi, 2005, p. 174). Let $\theta_0 = (\mu_0, \mu_*, \mu_1)'$, $\hat{\theta}$ be the corresponding vector of estimators, and $h^{(2)} = \left(h_0\ , h_*\ , h_1\ \right)'$. Then

$$\sqrt{n}\left(\hat{\theta} - \theta_o\right) \xrightarrow{d} N[0, G_o^{-1} S_o (G_o^{-1})'],$$

where, letting $p = P(T = 1)$ and $q = P(T = 0)$,

$$G_0 = E\left[\frac{\partial h^{(2)}}{\partial \theta'}|_{\theta_0}\right] = -\begin{bmatrix} q & 0 & 0 \\ 0 & p & 0 \\ 0 & 0 & p \end{bmatrix}, \tag{5}$$

and

$$S_0 = E_{\theta_0}\left[h^{(2)} h^{(2)'}\right]. \tag{6}$$

That is, in sufficiently large samples, $\hat{\theta}$ is approximately multivariate normally distributed with mean θ_0 and variance matrix

$$\frac{G_0^{-1} S_0 (G_0')^{-1}}{n}.$$

Let $\widehat{NIE} = \hat{\mu}_1 - \hat{\mu}_*$ and $\widehat{NDE} = \hat{\mu}_* - \hat{\mu}_0$ be estimators of the natural indirect and direct effects. Further let $v_{NIE} = (0, -1, 1)'$ and $v_{NDE} = (-1, 1, 0)'$. Then

$$\widehat{NIE} = v_{NIE}'\hat{\theta},$$

$$Var\left(\widehat{NIE}\right) = v_{NIE}' Var\left(\hat{\theta}\right) v_{NIE},$$

$$\widehat{NDE} = v_{NDE}'\hat{\theta},$$

$$Var\left(\widehat{NDE}\right) = v_{NDE}' Var\left(\hat{\theta}\right) v_{NDE}. \tag{7}$$

4. Variance of Two-step RMPW MOM Estimators

The NIE and NDE estimators presented above are infeasible, as they presuppose knowledge of the true RMPW weights. In practice, though, these weights are not known and must be estimated. The resulting estimators, utilizing estimated rather than true weights, are

commonly referred to as two-step estimators. In the first step, regression coefficients from logistic or probit regressions of binary mediator M on baseline covariates X are estimated and then used to estimate RPMW weights. In the second step, the NIE and NDE estimators given in the previous section are computed, but using the estimated weights from step one rather than the unknown true weights. The process of estimating the weights (via logistic regression in the discussion that follows) typically must be accounted for in determining the variance matrix of the two-step estimators. We do so, following the approach of Newey (1984), as presented in Cameron and Trivedi (2005, pp. 200-202); also see Hansen (1982), Pagan (1986), and Wooldridge (2010, pp. 409-413).

Two logistic regression models are used to estimate the RMPW weights; one model is fit to the control group to predict the potential mediator $M(0)$ and the other is fit to the experimental group to predict the potential mediator $M(1)$.[4] We refer to these as *mediator models*. In essence, these are propensity score models for the observed mediator under the two alternative treatment conditions. Here we assume that the mediator models are correctly specified and will discuss later the implications of one or both of the models being misspecified.

The logistic regression model with parameters α for predicting M under the control condition is

$$P(M_i = 1|T_i = 0, X_i) \equiv \mu_i^{\alpha} = \frac{1}{1 + \exp(-\alpha' X_i)}.$$

The logistic regression model with parameters β for predicting M under the treatment condition is

[4] Please refer back to (3), and recall that if $T = 0$, then $M = M(0)$ and if $T = 1$, then $M = M(1)$. Hence, predicting M on the control group is equivalent to predicting $M(0)$ on this group, and predicting M on the treatment group is equivalent to predicting $M(1)$ on this group. Since $M(0)$ and $M(1)$ are distinct variables, it is reasonable to fit a distinct logistic regression for each. This approach is equivalent to fitting a single logistic regression model for the observed mediator M that includes an interaction between T and every covariate.

$$P(M_i = 1|T_i = 1, X_i) \equiv \mu_i^{\beta} = \frac{1}{1 + \exp(-\beta' X_i)}.$$

The score function for estimating parameters α via maximum likelihood estimation (MLE) is

$$s_i^{\alpha} = (M_i - \mu_i^{\alpha}) X_i (1 - T_i).$$

The score function for parameters β via MLE is

$$s_i^{\beta} = \left(M_i - \mu_i^{\beta}\right) X_i T_i.$$

It is a standard result that these score functions have population mean zero at the true regression coefficient values α_0 and β_0, respectively. Further, the maximum likelihood estimates are obtained by solving

$$\frac{1}{n}\sum_{i=1}^{n} s_i^{\alpha} = 0,$$

$$\frac{1}{n}\sum_{i=1}^{n} s_i^{\beta} = 0.$$

We will estimate the variance matrix for the estimators of all the parameters $\vartheta_0 = (\alpha_0, \beta_0, \mu_0, \mu_*, \mu_1)'$, where the last three entries in this vector constitute θ_0 from the preceding section. Supposing the number of parameters for each logistic model is equal to *P*, we estimate (2*P* + 3) parameters in total, of which the 2*P* logistic regression coefficients are not of substantive interest and hence are nuisance parameters. Following Cameron and Trivedi (2005, pp. 200-202), we stack these score functions on top of the estimating functions from the previous section:

$$h_{2step,i} = \begin{pmatrix} h_i^{(1)} \\ h_i^{(2)} \end{pmatrix}$$

$$h_i^{(1)} = (s_i^{\alpha}, s_i^{\beta})'$$

$$h^{(2)} = \left(h_0\ , h_*\ , h_1\ \right)'$$

but where estimating function h_* is now defined in terms of the logistic regression-modeled RMPW weights. That is, the RMPW weights are now

$$w_i(M_i, X_i, \alpha, \beta) = M_i \frac{\mu_i^{\alpha}}{\mu_i^{\beta}} + (1 - M_i) \frac{1-\mu_i^{\alpha}}{1-\mu_i^{\beta}}, \tag{8}$$

and $w_i(M_i, X_i, \alpha_0, \beta_0)$ is the true RMPW weight. Then h_* is defined as

$$h_{*i}(O_i, \mu, \alpha, \beta) = (Y_i - \mu) w_i(M_i, X_i, \alpha, \beta) T_i,$$

and has population mean zero at $\mu = \mu_*, \alpha = \alpha_0, \beta = \beta_0$.

Let $\hat{\vartheta} = (\hat{\alpha}, \hat{\beta}, \hat{\mu}_0, \hat{\mu}_*, \hat{\mu}_1)'$, where the last three entries in this vector constitute $\hat{\theta}$. By the same MOM result presented in the preceding section[5], we have

$$\sqrt{n}\left(\hat{\vartheta} - \vartheta_o\right) \xrightarrow{d} N[0, A_o^{-1} B_o (A_o^{-1})'],$$

in which

$$B_o = E\left[h_{2step,i} h'_{2step,i}\right] = E\begin{bmatrix} h_i^{(1)} h_i^{(1)\prime} & h_i^{(1)} h_i^{(2)\prime} \\ h_i^{(2)} h_i^{(1)\prime} & h_i^{(2)} h_i^{(2)\prime} \end{bmatrix} \equiv \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_0 \end{bmatrix},$$

$$A_o = E\left[\frac{\partial h_{2step,i}}{\partial \vartheta'}\Big|_{\vartheta_0}\right] = E\begin{bmatrix} \dfrac{\partial h_i^{(1)}}{\partial(\alpha, \beta)} & \dfrac{\partial h_i^{(1)}}{\partial \theta} \\ \dfrac{\partial h_i^{(2)}}{\partial(\alpha, \beta)} & \dfrac{\partial h_i^{(2)}}{\partial \theta} \end{bmatrix} \equiv \begin{bmatrix} G_{11} & 0 \\ G_{21} & G_0 \end{bmatrix},$$

where S_0 and G_0, as defined in (5) and (6), are elements of the variance matrix for $\hat{\theta}$ given in the preceding section.

$$G_{11} = E\left[\frac{\partial h_i^{(1)}}{\partial(\alpha, \beta)}\right] = E\begin{bmatrix} \dfrac{\partial s_i^{\alpha}}{\partial \alpha} & \dfrac{\partial s_i^{\alpha}}{\partial \beta} \\ \dfrac{\partial s_i^{\beta}}{\partial \alpha} & \dfrac{\partial s_i^{\beta}}{\partial \beta} \end{bmatrix} = E\begin{bmatrix} \dfrac{\partial s_i^{\alpha}}{\partial \alpha} & 0 \\ 0 & \dfrac{\partial s_i^{\beta}}{\partial \beta} \end{bmatrix},$$

[5] Using the Cameron and Trivedi (2005) stacking approach, θ_0 is estimated simultaneously with the nuisance parameters α_0 and β_0. Nonetheless, $\hat{\theta}$ is still referred to as a two-step estimator. Stacking estimating functions was also discussed by Stefanski and Boos (2002).

$$G_{21} = E\left[\frac{\partial h_i^{(2)}}{\partial(\alpha,\beta)}\right] = E\left[\begin{matrix}0\\ \frac{\partial h_{*i}}{\partial(\alpha,\beta)}\\ 0\end{matrix}\right].^{6}$$

Importantly, G_{21} relates the estimation of θ to the coefficients in the mediator models.

Because ϑ_0 includes nuisance parameters, we are not interested in the large-sample variance of $\hat{\vartheta}$ per se, but rather in the large-sample variance of $\hat{\theta}$. Recall that when true RMPW weights are (infeasibly) used, this variance matrix is

$$\frac{G_0^{-1}S_0(G_0')^{-1}}{n}.$$

For the feasible two-step estimator $\hat{\theta}$, the large-sample variance matrix is (Cameron & Trivedi, 2005, equation 6.65)

$$\frac{G_0^{-1}S_0(G_0')^{-1}}{n} + \frac{G_0^{-1}\{G_{21}[G_{11}^{-1}S_{11}G_{11}^{-1}]G_{21}' - G_{21}G_{11}^{-1}S_{12} - S_{21}G_{11}^{-1}G_{21}'\}(G_0')^{-1}}{n}. \quad (9)$$

The second term in (9) gives the adjustment to the variance of the infeasible estimator to account for the two-step nature of the feasible estimator. The second term is equal to zero if $G_{21} = 0$, which holds when the true rather than estimated RMPW weights are used. Pre- and post-multiplying the variance matrix given in (9) by v_{NIE} yields the large-sample variance of the two-step estimator of the natural indirect effect; pre- and post-multiplying the variance matrix given in (9) by v_{NDE} yields the large-sample variance of the two-step estimator of the natural direct effect.

5. Estimating the Variance of the Two-step Estimators

The entries in the G and S matrices in (9) are unknown population means, and hence the variance matrix given in (9) cannot be used directly. To estimate the variance matrix, the

[6] The value of G_{21} is given in Appendix 3.

population mean entries in the matrices in (9) can be replaced by their sample mean analogues. The estimating functions, score functions, and their partial derivatives are defined in terms of population parameters (e.g., h_{0i} is defined in terms of μ_0), and hence any sample means involving estimating or score functions or partial derivatives will replace the population parameters with their estimates (e.g., μ_0 is replaced by $\hat{\mu}_0$ in h_{0i}). Per Wooldridge (2010, Lemma 12.1), this approach yields a consistent variance estimator. Since the first term in (9) is the variance matrix of the infeasible one-step estimator, this also can be consistently estimated via the just-described approach. Generalized method of moments commands such as the Stata *GMM* command can be used to estimate stacked MOM estimators (recall MOM is a special case of GMM), and will automatically estimate the variance matrix given in (5) (StataCorp, 2013). See Appendix 2 for an example of the use of this Stata command.

These variance estimation procedures can be contrasted with the nonparametric bootstrap (Efron & Tibshirani, 1993). For each bootstrap sample, the two logistic regressions predicting the mediator probabilities from baseline covariates are fit via MLE, the RMPW weights are obtained as functions of the estimated logistic regression coefficients, and the estimates of the natural direct and indirect effects are computed using the estimated weights. The sample variances of the collection of bootstrap estimates of the direct and indirect effects serve as estimates of the variances of the two-step estimators.

6. Simulation Study

We turned to Monte Carlo simulations to address three remaining issues regarding the two-step estimator: (1) the performance of the two-step estimators of $SE(\widehat{NDE})$ and $SE(\widehat{IDE})$ with relatively small sample sizes; (2) the direction and magnitude of the bias associated with the estimators of $SE(\widehat{NDE})$ and $SE(\widehat{IDE})$ that ignore the uncertainty in the RMPW weights; and (3)

the performance of the bootstrap method when compared to the two-step method for estimating $SE(\widehat{NDE})$ and $SE(\widehat{IDE})$.

The derivation in this paper invoked the asymptotic theory to show that the two-step method produces consistent estimators of the variances and thus of standard errors of the direct and indirect effect estimates. For applied research, it is important to know how reliant these estimators' robustness is on large sample sizes. We generated samples of size 100 and 1,000 and estimated bias of the standard error estimators using eight distinct data-generating processes described in Appendix 4. In each case, we varied one feature of the model assumptions at a time, which led to a different variance of the RMPW weights. For both sample sizes, and for both direct and indirect effects, we found no evidence of bias in the two-step standard error estimators in any of the simulated scenarios.

The standard error estimators that ignore the uncertainty in the RMPW weights were biased in some cases. When bias occurred, it led to attenuated estimates in some circumstances but inflated estimates in others. In particular, we found scenarios in which ignoring the uncertainty in the RMPW weights led to severely attenuated standard error estimates and an increased type I error rate. These downward biases could be as high as 16% of the true standard error for $\widehat{SE}(\widehat{NDE})$ and 86% for $\widehat{SE}(\widehat{IDE})$, when the sample size was 1000. This led to true coverage rates of nominal 95% confidence intervals as low as 90.5% and 19.8%, respectively. Appendix 4 tabulates the parameters that we used for data generation (see Tables 4.1 and 4.2) and presents the results that illustrate cases of conservative and liberal standard error estimates and coverage (see Tables 4.3 and 4.4).

We also compared our results to those obtained from a nonparametric bootstrap, which is frequently used to incorporate the uncertainty of weight estimation in other causal effect

estimators that rely on weights such as IPTW (e.g. Little & Rubin, 2002; Huber, 2014). The bootstrap estimator uses 1,000 replications (each re-sampled with replacement), and takes the standard deviation of the estimated coefficients across these replications as the standard error of the associated estimate. We give Stata code for implementing bootstrap standard errors in Appendix 2. As for the two-step estimators of $SE(\widehat{NDE})$ and $SE(\widehat{IDE})$, the bootstrap standard error estimators are not biased in any of the eight data-generating distributions. However, according to our simulation results, the bootstrap standard error estimators on the basis of 1,000 replications are often less efficient than the two-step standard error estimators. In general, the precision of a bootstrap standard error estimator tends to improve as the number of replications increases. Yet in practice, the researcher would not have a clear sense, in every unique case, how many bootstrap samples are needed for the bootstrap standard error estimator to perform comparably to a two-step standard error estimator. The main advantage of the two-step method is that it has a closed-form expression, while bootstrapping tends to be much more computationally intensive.

7. Application Study

We applied the two-step estimator to the NEWWS data introduced in section 2 that had previously been analyzed by Hong, Deutsch, and Hill (2015). We repeated the analysis in this article, using the same data set and variables, but using the two-step variance estimator from equation (9) and a bootstrap variance estimator to account for the uncertainty in the RMPW weights.

Analyzing data from the NEWWS Labor Force Attachment program (LFA) in Riverside, California, Hong, Deutsch, and Hill (2015) examined whether and how employment mediated the program impact on depression in the long run for mothers with young children. Individuals

who were assigned at random to the control condition continued to receive public assistance from AFDC. Those who were assigned to the LFA program received support and training for job searching and were threatened with sanctions for non-compliance in program activities or work. The LFA program, however, did not guarantee employment.

The sample includes 208 LFA group members and 486 control group members with a child aged 3 to 5 years. Unemployment Insurance records maintained by the State of California provide quarterly administrative data on employment for each participant. All participants were surveyed shortly before the randomization and again at the two-year follow-up. The data contain a rich set of pre-treatment covariates essential to satisfying the assumptions employed by the RMPW theorem.

Table 1 displays the results from our re-analysis of the NEWWS Riverside data. We compare across three different estimation methods: (a) ignoring the uncertainty in the estimated weight, (b) applying the proposed two-step estimation procedure, and (c) bootstrapping with 10,000 replications. The estimated direct effect is about 17% of a standard deviation of the outcome. The direct effect estimate indicates that, if the treatment had counterfactually generated no impact on employment, maternal depression would have increased. Using the two-step estimation procedure, we find the standard error to be 0.86, slightly smaller than the standard error estimate 0.87 when the estimation uncertainty in the weight is ignored, and considerably smaller than the bootstrapped standard error estimate 0.95. The effect size of the estimated indirect effect is about -0.12, which indicates that, if all individuals were hypothetically assigned to LFA, the LFA-induced change in employment would produce a reduction in maternal depression on average. The standard error estimate obtained from the two-step estimation is 0.48, slightly higher than the standard error estimate 0.47 when one ignores the estimation uncertainty

in the weight, yet much smaller than the bootstrapped standard error 0.61. In this application, the true standard errors are unknown to the researcher. Yet the notable discrepancies between the two-step standard error estimates and their corresponding bootstrap standard error estimates again raise concerns that these two methods for standard error estimation may not generate comparable results.

Table 1. Standard Errors of Direct and Indirect Effects from NEWWS Analysis Using Different Estimators

	Ignoring uncertainty	Two-step method (eq. 9)	Bootstrap
Direct effect			
Coefficient	1.257	1.257	1.257
Standard error	0.871	0.862	0.959
t-statistic	1.443	1.458	1.311
p-value	0.149	0.145	0.190
Indirect effect			
Coefficient	-0.879	-0.879	-0.879
Standard error	0.462	0.472	0.601
t-statistic	-1.903	-1.862	-1.463
p-value	0.057	0.063	0.143

Note: The standard deviation of the outcome in the control group is 7.666.

8. Discussion

The RMPW method decomposes a total treatment effect into a natural direct effect and a natural indirect effect through propensity score-based adjustment for mediator value selection. This manuscript has focused on how to accurately adjust for the estimation of RMPW weights when drawing inferences about natural direct and indirect effects. As we have noted, in other

causal inference contexts where weighted estimators are used (e.g., IPTW estimator of the average treatment effect), lack of adjustment typically leads to conservative inference. In the present context, however, our simulation studies have shown that ignoring the uncertainty in weight estimation sometimes leads to conservative inference and at other times liberal inference and therefore inflated Type I errors. Because we have not found any definitive way to distinguish the data generating mechanisms that lead to conservative vs. liberal inference, we believe that all use of RMPW estimators should adjust for weight estimation.

We have extended to RMPW applications the two-step estimation method that uses the "stacking trick." By stacking the score functions from the two steps of analysis (i.e., propensity score estimation in Step 1 and causal effect estimation in Step 2), the two-step estimation procedure generates a consistent variance estimator for each causal effect estimator that captures the sampling variability in both steps. We derived the asymptotic variance-covariance matrix for the two-step estimators. This method is based on large sample theories. Our simulation studies have examined the accuracy of the variance estimators obtained from the two-step procedure in comparison with the bootstrapped variance estimators for both small and moderate sized samples that are commonly seen in real applications. We have shown that the two-step estimation method works well in both. Comparing the two-step estimation procedure with the bootstrapping procedure that uses 1,000 replications, we find many cases in which the two-step estimator appears to be more precise than the bootstrapping estimator (the former often generates a relatively narrower range of 95% plausible values and a smaller mean square error). Most importantly, the two-step procedure involves considerably less computation than the bootstrapping.

In many applications, one may prefer the RMPW method to path analysis (Alwin & Hauser, 1975; Baron & Kenny, 1986; Duncan, 1966; Wright, 1934) and structural equation modeling (SEM) (Bollen, 1989; Jo, 2008; Jöreskog, 1970; MacKinnon, 2008) and to marginal structural models (Coffman & Zhong, 2012; Robins, 2003; Robins & Greenland, 1992; VanderWeele, 2009) because it avoids possible misspecifications of the mediator-outcome relationship and because it can flexibly accommodate treatment-by-mediator interactions. In the presence of such an interaction, one may further decompose the natural indirect effect into a population average pure indirect effect $E[Y\big(0,M(1)\big)-Y(0,M(0))]$ and a population average natural treatment-by-mediator interaction effect $E\{\big[Y\big(1,M(1)\big)-Y\big(1,M(0)\big)\big]-[Y\big(0,M(1)\big)-Y(0,M(0))]\}$. This further decomposition involves a fourth population average potential outcome $E\big[Y\big(0,M(1)\big)\big]$ that can be similarly identified under the sequential ignorability assumption and estimated through RMPW MOM estimation. We have implemented the two-step estimation procedure for decomposing the total treatment effect into a pure direct effect, a pure indirect effect, and a natural treatment-by-mediator interaction effect in a stand-alone RMPW software program freely available online at http://hlmsoft.net/ghong/. The same method can be extended to multivalued and continuous mediators.

We conclude with a brief discussion of two future research directions. The first involves extending the RMPW estimation and the adjustment for the estimation of the weights to the context of multisite studies. Here, a site might be a school, a job training center, a hospital, or a community. Each site houses a mini-population in its own right. Of interest are not only the population average causal effects but also the between-site variance of each causal effect. Qin and Hong (2014, 2015) offer one translation; and we are currently evaluating this approach as well as exploring others.

Second, we are concerned with the issue of model selection for weight estimation. Our "stacking trick" approach presupposes that the logistic mediation models that one specifies for estimating RMPW weights are determined a priori on a theoretical basis. An alternative approach would be to employ a model selection procedure to choose the mediator models, with the intent of finding models that will yield RMPW estimators with relatively small bias and mean squared error (given that the ideal of employing the true mediator models is unlikely in practice).[7] Brookhart and van der Laan (2006) and Vansteelandt, Bekaert, and Claeskens (2012) offer two possible model selection algorithms, though Rotnitzky and Vansteelandt (2015) note reservations. Use of model selection would then require care with post-selection inference (Claeskens & Hjort, 2008, chapter 7). Hong (2015) and colleagues (Hong, Deutsch, & Hill, 2011, 2015) proposed a semiparametric approach to estimating the propensity scores and the weights that are relatively robust to mediator model misspecifications. Others have recommended generalized boosted models for propensity score estimation (McCaffrey, Ridgeway, & Morral, 2004). We plan to explore the relative merits of these and other approaches and will extend the two-step estimation method to these alternative approaches.

[7] Wooldridge (2010, chapter 13) contains a general discussion of the consequences of model misspecification for two-step estimators. Even if the mediator models are misspecified, their MLEs $\hat{\alpha}$ and $\hat{\beta}$ still have probability limits, call them α^* and β^*, respectively. Let $\mu_{**} \equiv \frac{E[Y_i w_i(M_i,X_i,\alpha^*,\beta^*)|T_i=1]}{E[w_i(M_i,X_i,\alpha^*,\beta^*)|T_i=1]}$. It can be shown that, under model misspecification, the asymptotic bias of the two-step estimator of NIE is $E\big[Y\big(1,M(0)\big)\big] - \mu_{**}$, and the analogous asymptotic bias for the NDE estimator is $\mu_{**} - E\big[Y\big(1,M(0)\big)\big]$.

Appendix 1

Equivalence of RMPW Weighted Least Squares and Method of Moments Estimators

We show that the weighted least squares-style (WLS) estimators of natural effects presented by Hong and colleagues (Hong, 2010, 2015; Hong, Deutsch, and Hill, 2015) are identical to the method of moments (MOM) estimators presented in section 3. The WLS approach adds a duplicate of each treatment group observation to the data set, where $D = 1$ indicates that an observation is a duplicate and $D = 0$ otherwise. If N_0 denotes the number of participants in the control group and N_1 denotes the number of participants in the treatment group. The outcome vector $\tilde{Y}$ is of length $N_0 + 2N_1$ due to the presence of the duplicate outcome observations. The design matrix L is $(N_0 + 2N_1) \times 3$, where the first column is a vector of ones, the second is the vector of T values, and the third is the vector of D values. There is also a diagonal weight matrix W, where outcome observations with $D = 1$ or $T = 0$ have weights of 1, and outcome observations with $D = 0$ and $T = 1$ have weight $w = w(M, X)$; that is, the RMPW weight is used. The WLS estimator is then

$$\hat{\delta} = (L' WL)^{-1} L' W\tilde{Y} \tag{1}$$

We may order the observations so that the control group observations are followed by the treatment group observations, which are followed by the duplicate treatment group observations:

$$L = \begin{bmatrix} L_a \\ L_b \\ L_c \end{bmatrix}, W = \begin{bmatrix} W_a & 0 & 0 \\ 0 & W_b & 0 \\ 0 & 0 & W_c \end{bmatrix}, \tilde{Y} = \begin{bmatrix} Y_0 \\ Y_1 \\ Y_1 \end{bmatrix} \tag{2}$$

where each of the N_0 rows in L_a is $(1\ 0\ 0)$, each of the N_1 rows in L_b is $(1\ 1\ 0)$, and each of the N_1 rows in L_c is $(1\ 1\ 1)$; W_a and W_c are identity matrices of dimension N_0 and N_1, respectively, and W_b is a diagonal matrix with the N_1 w values; Y_0 is the vector of N_0 control group outcomes and Y_1 is the vector of N_1 treatment group outcomes. It follows that

$$L' WL = L_a' W_a L_a + L_b' W_b L_b + L_c' W_c L_c = \begin{bmatrix} N_0 + N_1 + \sum w & N_1 + \sum w & N_1 \\ N_1 + \sum w & N_1 + \sum w & N_1 \\ N_1 & N_1 & N_1 \end{bmatrix}$$

where $\sum w$ is the sum of the N_1 RMPW weights, and the inverse matrix is

$$(L' WL)^{-1} = \frac{1}{N_0 N_1 \sum w} \begin{bmatrix} N_1 \sum w & -N_1 \sum w & 0 \\ -N_1 \sum w & N_0 N_1 + N_1 \sum w & -N_0 N_1 \\ 0 & -N_0 N_1 & N_0 N_1 + N_0 \sum w \end{bmatrix}. \tag{3}$$

Further,

$$L' W\tilde{Y} = L_a' W_a Y_0 + L_b' W_b Y_1 + L_c' W_c Y_1 = \begin{bmatrix} \sum y_0 + \sum w y_1 + \sum y_1 \\ \sum w y_1 + \sum y_1 \\ \sum y_1 \end{bmatrix} \tag{4}$$

where $\sum y_0$ is the sum of the N_0 control group outcomes, $\sum y_1$ is the sum of the N_1 treatment group outcomes, and $\sum w y_1$ is the sum of the N_1 products of the RMPW weights and the treatment group outcomes. From (3) and (4) we have

$$\hat{\delta} = (L' WL)^{-1} L' W\tilde{Y} = \begin{bmatrix} \frac{\sum y_0}{N_0} \\ \frac{\sum w y_1}{\sum w} - \frac{\sum y_0}{N_0} \\ \frac{\sum y_1}{N_1} - \frac{\sum w y_1}{\sum w} \end{bmatrix} \tag{5}$$

The second entry in $\hat{\delta}$ is the MOM estimator of the natural direct effect and the third entry is the MOM estimator of the natural indirect effect.

Appendix 2

Software for Two-step RPMW Estimation

The method of moments estimator described in this article is a type of generalized method of moments estimator, introduced by Hansen (1982). Stata contains a command, "GMM," that generates such estimates for these models (StataCorp, 2013a). The Stata code below, divided into three steps, illustrates how to implement the stacked method of moments estimator. See the Stata manual for a full description of the command (StataCorp, 2013b).

1. Generate a constant to be used in the GMM command

   generate cons = 1

2. Execute the GMM command

   *gmm (eq1: (M - ((1 / (1 + exp(-{xb1: cons X1 … Xp }))))) * T) ///*

   *(eq2: (M - ((1 / (1 + exp(-{xb2: cons X1 … Xp }))))) * (1 - T)) ///*

   *(eq3: (Y - {Mu0}) * (1 - T)) ///*

   *(eq4: (Y - {Mu1}) * T) ///*

   *(eq5: (Y - {MuStar}) * ///*

   *(((M * ((1 / (1 + exp(-{xb2:})))/(1 / (1 + exp(-{xb1:}))))) + ((1 - M) * ///*

   *(exp(-{xb2:}) / (1 + (exp(-{xb2:})))) / (exp(-{xb1:}) / (1 + (exp(-{xb1:}))))))) * T),*

   /// instruments(eq1: X1 … Xp) instruments(eq2: X1 … Xp) instruments(eq3:) ///

   instruments(eq4:) instruments(eq5:) winitial(identity) onestep

Here *X1 … Xp* represent *P* covariates entered in each logistic regression model. This step produces coefficient estimates and standard errors for the covariates entered in the logistic regression models as well as for the sample means, *Mu0* (i.e., μ_0) and *Mu1* (i.e., μ_1), and the

weighted sample mean, *MuStar* (i.e., μ_*). The "///" allows the code to be written on separate lines for clarity while indicating that each line pertains to one command.

3. Use the "lincom" command to estimate natural indirect effect. This calculates the linear combination of coefficients estimated in the prior step, taking into account the covariance term in the standard error. The variance and covariance estimates used in this step come directly from the GMM estimation procedure in step 2, and are saved in memory.

Estimate Natural Indirect Effect

lincom _b[/Mu1] - _b[/MuStar]

Estimate Natural Direct Effect

lincom _b[/MuStar] - _b[/Mu0]

Some programming is required to obtain bootstrapped standard errors in Stata. We used the following code in our estimation.

```
* define program to compute estimates of NDE and NIE
program nat_effects, rclass
* compute RMPW weights
logistic M X1 … Xp if T==0
predict p_T0, pr
logistic M X1 … Xp if T==1
predict p_T1, pr
gen w = M*(p_T0 / p_T1) + (1-M)*( (1-p_T0) / (1-p_T1) )
* estimate Mu0, Mu1, MuStar
summ T
local n1 = r(sum)
```

```
local n0 = _N - `n1'
gen num0 = Y*(1-T)
summ num0
local Mu0 = r(sum) / `n0'
gen num1 = Y*T
summ num1
local Mu01= r(sum) / `n1'
gen numstar = Y*w*T
gen denomstar = w*T
summ numstar
local sumstar = r(sum)
summ denomstar
local MuStar = `sumstar' / r(sum)
drop num0 num1 numstar denomstar
* estimate NDE and NIE
return scalar nde_est = `MuStar' - `Mu0'
return scalar nie_est = `Mu1' - `MuStar'
end
* get bootstrapped SEs
bootstrap nde=r(nde_est) nie=r(nie_est), reps(1000): nat_effects
```

We have similarly implemented the two-step estimation method in R (code available upon request) and in a stand-alone RMPW software program freely available online.

Appendix 3

Derivation of the Two-Step Estimator

We derive the elements in $G_{21} = E\begin{bmatrix} 0 \\ \frac{\partial h_{*i}}{\partial(\alpha,\beta)} \\ 0 \end{bmatrix}$ that contribute to the adjustment in equation (9). Because only h_{*i} is dependent on $w_i(M_i, X_i, \alpha, \beta)$, hence dependent on (α, β), only the second row of this $3 \times 2P$ matrix is non-zero. It takes some work to specify $\frac{\partial h_{*i}}{\partial(\alpha,\beta)}$.

$$\frac{\partial h_{*i}}{\partial \alpha} = (Y_i - \mu_*)T_i \frac{\partial w_i}{\partial \alpha};$$

$$\frac{\partial h_{*i}}{\partial \beta} = (Y_i - \mu_*)T_i \frac{\partial w_i}{\partial \beta}.$$

We will make use of standard results for logistic regression models:

$$\frac{\partial \mu_i^\alpha}{\partial \alpha} = \mu_i^\alpha(1 - \mu_i^\alpha)X_i,$$

$$\frac{\partial \mu_i^\beta}{\partial \beta} = \mu_i^\beta\left(1 - \mu_i^\beta\right)X_i,$$

$$\frac{\partial s_i^\alpha}{\partial \alpha} = -\mu_i^\alpha(1 - \mu_i^\alpha)X_i X_i'(1 - T_i),$$

$$\frac{\partial s_i^\beta}{\partial \beta} = -\mu_i^\beta\left(1 - \mu_i^\beta\right)X_i X_i' T_i.$$

Using the logistic regression results and (8), we have that

$$\frac{\partial w_i}{\partial \alpha} = \left[M_i \frac{1}{\mu_i^\beta} - (1 - M_i)\frac{1}{1 - \mu_i^\beta}\right]\frac{\partial \mu_i^\alpha}{\partial \alpha} = \left[M_i \frac{1}{\mu_i^\beta} - (1 - M_i)\frac{1}{1 - \mu_i^\beta}\right]\mu_i^\alpha(1 - \mu_i^\alpha)X_i;$$

$$\frac{\partial w_i}{\partial \beta} = \left[-M_i \mu_i^{\alpha} \frac{1}{\left(\mu_i^{\beta}\right)^2} + (1 - M_i)(1 - \mu_i^{\alpha}) \frac{1}{\left(1 - \mu_i^{\beta}\right)^2} \right] \frac{\partial \mu_i^{\beta}}{\partial \beta}$$

$$= \left[-M_i \mu_i^{\alpha} \frac{1}{\left(\mu_i^{\beta}\right)^2} + (1 - M_i)(1 - \mu_i^{\alpha}) \frac{1}{\left(1 - \mu_i^{\beta}\right)^2} \right] \mu_i^{\beta}\left(1 - \mu_i^{\beta}\right) X_i$$

$$= \left[-M_i \mu_i^{\alpha} \frac{\left(1 - \mu_i^{\beta}\right)}{\mu_i^{\beta}} + (1 - M_i)(1 - \mu_i^{\alpha}) \frac{\mu_i^{\beta}}{1 - \mu_i^{\beta}} \right] X_i.$$

Appendix 4

Simulation Study Specifications and Results

In this appendix we provide details about our simulation study. First we describe the data generating model and the parameter values we specified for our eight distinct simulated scenarios. Second, we provide results from two of the eight simulation scenarios for illustration. The first is a simulated scenario that has led to overestimation of standard errors of direct and indirect effect estimators when ignoring uncertainty in the estimated RMPW, and thus has resulted in conservative inference; the second simulated scenario has led to underestimation and thus liberal inference.

(1) Data Generating Model and Parameter Specifications

We generate three independent baseline covariates X_1, X_2, and X_3 with identical distributions $N(0,1)$. We randomly assign individuals to treatment, T_i, such that $P(T_i = 1) = 0.5$ for all i. Next, we generate a binary mediator from the following models under each treatment condition:

$$logit\{P(M_i = 1|T_i = 0, \boldsymbol{X_i})\} = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i}$$

$$logit\{P(M_i = 1|T_i = 1, \boldsymbol{X_i})\} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

where M_i is a binary mediator. Each individual in the control group then obtains a mediator probability, $\mu_i^{\alpha} = P(M_i = 1|T_i = 0, \boldsymbol{X_i})$, with which we generate the mediator value for each individual in the control group from $Bernoulli(\mu_i^{\alpha})$. Similarly, each individual in the treatment group obtains a mediator probability, $\mu_i^{\beta} = P(M_i = 1|T_i = 1, \boldsymbol{X_i})$, with which we generate the mediator value for each treated individual from $Bernoulli\left(\mu_i^{\beta}\right)$.

Finally, we generate a continuous outcome from the following model, allowing for an interaction between the treatment and the mediator:

$$Y_i = \theta_0 + \theta_1 T_i + \theta_2 M_i + \theta_3 T_i M_i + \theta^{(1)} X_{1i} + \theta^{(2)} X_{2i} + \theta^{(3)} X_{3i} + \varepsilon_i$$

in which $\varepsilon_i \sim N(0, \sigma^2)$. Based on expressions derived by Valeri & VanderWeele (2013) for the natural direct effect and natural indirect effect, θ_1, θ_2 and θ_3 can be computed as follows:

$$\theta_2 = \text{NIE} / \left[\mu^\beta - \mu^\alpha\right] - \theta_3$$

$$\theta_1 = \text{NDE} - \theta_3 \mu^\alpha$$

where $\mu^\beta = E\big(P(M = 1|T = 1, \boldsymbol{X})\big)$ and $\mu^\alpha = E\big(P(M = 1|T = 0, \boldsymbol{X})\big)$. We specify NDE to be three times NIE and compute θ_1, θ_2 and θ_3 as follows:

$$\theta_2 = 3/4 \times \text{NIE} / \left[\, \mu^\beta - \, \mu^\alpha\right],$$

$$\theta_3 = 1/3 \times \theta_2,$$

$$\theta_1 = \text{NDE} - \theta_3 \mu^\alpha,$$

in which

$$\mu^\beta = E\big(P(M = 1|T = 1, \boldsymbol{X})\big)$$

$$= \iiint \frac{exp(logit\{P(M = 1|T = 1, \boldsymbol{X})\})}{1 + exp(logit\{P(M = 1|T = 1, \boldsymbol{X})\})} f(X_1) f(X_2) f(X_3) dX_1 dX_2 X_3,$$

$$\mu^\alpha = E\big(P(M = 1|T = 0, \boldsymbol{X})\big)$$

$$= \iiint \frac{exp(logit\{P(M = 1|T = 0, \boldsymbol{X})\})}{1 + exp(logit\{P(M = 1|T = 0, \boldsymbol{X})\})} f(X_1) f(X_2) f(X_3) dX_1 dX_2 X_3.$$

We select two different sample sizes: *N* = 1000 representing a relatively big sample size and *N* = 100 representing a relatively small sample size. For each sample size, we design eight simulations defined by eight sets of parameters, which create a range of scenarios that applied researchers may experience in practice and several others that are more extreme so that we can understand the robustness of the estimators under even unusual conditions.

Each of the eight simulations varies one feature of the data-generating distribution at a time. The changes of the parameter values in the propensity score models lead to changes in the

magnitude and the variance of the RMPW weights and correspondingly the standard errors of the estimates of the direct effect and the indirect effect. In this way, we can evaluate the influence of each data generation feature on the estimation results, and assess the stability of performance for each estimation procedure.

(2) Results from Two Illustrative Simulations

Through Monte Carlo simulations, we compare the standard error estimator when the estimation uncertainty in the weights are ignored, the two-step standard error estimator, and the bootstrapped standard error estimator against the true standard error approximate obtained from 1000 replications.

For reference, we also compare with the results when using the true weights. The true weights are directly calculated with the parameters that we specify under each scenario, and thus no sampling variability is involved. In this way, we should almost always obtain an unbiased estimate of the standard error of a causal effect estimator.

For the sake of illustration, we present the details of simulations 4 and 8. The parameter specifications for these simulations are given in Tables 4.1 and 4.2, and the simulation results are presented in Tables 4.3 and 4.4.

Table 4.1. Parameter Specifications in the Mediator Models

	α_0	α_1	α_2	α_3	β_0	β_1	β_2	β_3
Simulation 4	-1	0.5	0.5	-0.5	1	0.5	0.5	-0.5
Simulation 8	-0.1	0.5	0.5	-0.5	0.1	0.5	0.5	-0.5

Table 4.2. Parameter Specifications in the Outcome Model

	θ_0	$\theta^{(1)}$	$\theta^{(2)}$	$\theta^{(3)}$	σ^2	NDE	NIE	Var(Y)	σ^2/Var(Y)	NDE	NIE
Simulation 4	20	0.4	0.6	0.9	0.36	0.39	0.13	1.69	21%	0.3	0.1
Simulation 8	20	0.4	0.6	0.9	0.36	0.39	0.13	1.69	21%	0.3	0.1

Note: NDE and NIE are presented as effect sizes.

For both the direct effect and the indirect effect, Tables 4.3 and 4.4 show (1) the bias of the estimators of the natural direct effect and natural indirect effect, $\widehat{NDE}$ and, $\widehat{NIE}$ (2) the standard deviation of the sampling distribution of each estimator as the Monte Carlo approximation of the true standard error of each estimator, $SE\left(\widehat{\mathrm{NDE}}\right)$ and $SE\left(\widehat{\mathrm{NIE}}\right)$, (3) the mean estimated standard error of each causal effect estimator, $\widehat{SE}\left(\widehat{\mathrm{NDE}}\right)$ and $\widehat{SE}\left(\widehat{\mathrm{NIE}}\right)$, (4) the mean square error (MSE) of each standard error estimator, $\mathrm{E}\{\left[\widehat{SE}\left(\widehat{\mathrm{NDE}}\right) - SE\left(\widehat{\mathrm{NDE}}\right)\right]^2\}$ and $\mathrm{E}\{\left[\widehat{SE}\left(\widehat{\mathrm{NIE}}\right) - SE\left(\widehat{\mathrm{NIE}}\right)\right]^2\}$, (5) the 2.5th and 97.5th percentiles of the standard error estimates across 1000 replications, which we refer to as the 95% plausible range, and (6) the true coverage rates for nominal 95% CIs for each of the causal effects. By comparing (2) and (3), we assess the bias of the standard error estimators, while (4) and (5) provide information about the precision of the standard error estimators, and (6) provides information about the implications of standard error estimation for accurate statistical inference.

We present results from Simulation 4 because, when N = 1000, ignoring the uncertainty of the estimated RMPW leads to overestimation of standard errors of direct and indirect effect estimators and thus conservative inference. We present results from Simulation 8 because, when N = 1000 and when N = 100, ignoring the uncertainty of the estimated RMPW leads to underestimation of the standard errors of direct and indirect effect estimators and thus liberal inference. The findings across all eight simulations have been summarized in the main text.

Table 4.3. Conservative Statistical Inference When Ignoring Uncertainty in Estimated RMPW

	N=1000				*N*=100			
		est. weight				est. weight		
	true weight	ignoring uncertainty	two-step	bootstrap	true weight	ignoring uncertainty	two-step	bootstrap
Bias of $\widehat{\text{NDE}}$	0.0076		0.0053		0.0207		0.0176	
$SE(\widehat{\text{NDE}})$	0.1018		0.09		0.3167		0.3212	
$\widehat{SE}(\widehat{\text{NDE}})$ mean	0.1007	0.1021	0.089	0.0903	0.3081	0.3255	0.2931	0.3403
MSE of $\widehat{SE}(\widehat{\text{NDE}})$	2.39×10^{-5}	1.90×10^{-4}	1.75×10^{-5}	2.50×10^{-5}	0.0019	0.0067	0.0066	0.0081
95% plausible range for $\widehat{SE}(\widehat{\text{NDE}})$	[0.092,0.111]	[0.091,0.117]	[0.083,0.098]	[0.082,0.101]	[0.239,0.408]	[0.233,0.532]	[0.223,0.479]	[0.249,0.579]
coverage rate for 95% CI for NDE	95.80%	97.80%	94.40%	94.10%	92.90%	94.60%	92.20%	95.70%
Bias of $\widehat{\text{NIE}}$	-0.0039		-0.0016		-0.0078		-0.0047	
$SE(\widehat{\text{NIE}})$	0.0583		0.0376		0.1814		0.2011	
$\widehat{SE}(\widehat{\text{NIE}})$ mean	0.0577	0.0598	0.0363	0.0393	0.1707	0.2001	0.1718	0.2368
MSE of $\widehat{SE}(\widehat{\text{NIE}})$	1.96×10^{-5}	5.63×10^{-5}	4.65×10^{-5}	6.02×10^{-5}	0.0016	0.0082	0.0109	0.0118
95% plausible range for $\widehat{SE}(\widehat{\text{NIE}})$	[0.05,0.067]	[0.047,0.078]	[0.027,0.052]	[0.029,0.058]	[0.113,0.265]	[0.099,0.442]	[0.085,0.434]	[0.124,0.522]
coverage rate for 95% CI for NIE	94.30%	100.00%	93.80%	94.40%	92.60%	97.10%	96.20%	99.10%

Note. NDE=0.39; NIE=0.13.

Table 4.4. Liberal Statistical Inference When Ignoring Uncertainty in Estimated RMPW

	N=1000				*N*=100			
		est. weight				est. weight		
	true weight	ignoring uncertainty	two-step	bootstrap	true weight	ignoring uncertainty	two-step	bootstrap
Bias of $\widehat{\text{NDE}}$	0.0029		0.0029		-0.0005		0.0043	
$SE(\widehat{\text{NDE}})$	0.1229		0.1507		0.4043		0.5044	
$\widehat{SE}(\widehat{\text{NDE}})$ mean	0.126	0.1267	0.1542	0.1546	0.3969	0.4138	0.4935	0.5197
MSE of $\widehat{SE}(\widehat{\text{NDE}})$	1.49×10^{-5}	5.84×10^{-4}	1.91×10^{-5}	3.26×10^{-5}	0.0006	0.0096	0.0016	0.0025
95% plausible range for $\widehat{SE}(\widehat{\text{NDE}})$	[0.121,0.13]	[0.122,0.132]	[0.149,0.159]	[0.147,0.163]	[0.35,0.444]	[0.36,0.501]	[0.434,0.567]	[0.45,0.642]
coverage rate for 95% CI for NDE	95.00%	90.50%	95.40%	95.50%	93.90%	89.30%	94.50%	95.90%
Bias of $\widehat{\text{NIE}}$	-0.0004		-0.0003		-0.002		-0.0068	
$SE(\widehat{\text{NIE}})$	0.0058		0.0886		0.0184		0.313	
$\widehat{SE}(\widehat{\text{NIE}})$ mean	0.0058	0.0124	0.0896	0.0903	0.0186	0.1119	0.3139	0.3523
MSE of $\widehat{SE}(\widehat{\text{NIE}})$	6.43×10^{-8}	5.83×10^{-3}	6.32×10^{-6}	1.22×10^{-5}	0.0000	0.0436	0.0024	0.0047
95% plausible range for $\widehat{SE}(\widehat{\text{NIE}})$	[0.005,0.006]	[0.005,0.023]	[0.085,0.094]	[0.084,0.096]	[0.014,0.024]	[0.038,0.249]	[0.253,0.433]	[0.279,0.516]
coverage rate for 95% CI for NIE	94.80%	19.80%	95.90%	95.80%	94.60%	48.30%	94.80%	96.30%

Note. NDE=0.39; NIE=0.13.

References

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

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51,* 1173-1182.

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

Brookhart, A. M., & van der Laan, M. J. (2006). A semiparametric model selection criterion with applications to the marginal structural model. Computational Statistics and Data Analysis, 50, 475-498.

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

Claeskens, G., & Hjort, N. L. (2008). Model selection and model averaging. New York: Cambridge University Press.

Coffman, D. L. & Zhong, W. (2012). Assessing mediation using marginal structural models in the presence of confounding and moderation. *Psychological Methods, 17*(4), 642-664.

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

Efron, B., & Tibshirani, J. (1993). An introduction to the bootstrap. London: Chapman and Hall.

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

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

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

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

Hong, G., Deutsch, J., & Hill, H. (2011). Parametric and non-parametric weighting methods for estimating mediation effects: An application to the National Evaluation of Welfare-

to-Work Strategies. In *JSMProceedings*, Social Statistics Section. Alexandria, VA: American Statistical Association, pp.3215-3229.

Hong, G., Deutsch, J., & Hill, H. D. (2015). Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-by-mediator interaction. *Journal of Educational and Behavioral Statistics*, *40*(3), 307-340.

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

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

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

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

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.

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

Lange, T., Rasmussen, M., & Thygesen, L. C. (2013). Assessing natural direct and indirect effects through multiple pathways. *American Journal of Epidemiology*, *179*, 513-518.

Little, R. J. A. & Rubin, D. B. (2002) *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley and Sons Inc.

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

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*(4), 403-425.

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

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9, translated in Statistical Science, (with discussion), Vol 5, No 4, 465–480, 1990.

Pagan, A. (1986). Two stage and related estimators and their applications. *The Review of Economic Studies*, *53*(4), 517-538.

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

Qin, X., & Hong, G. (2014). Causal mediation analysis in multi-site trials: An application of ratio-of-mediator-probability weighting to the Head Start Impact Study. In *JSM Proceedings*, Social Statistics Section. Alexandria, VA: American Statistical Association, pp.912-926.

Qin, X., & Hong, G. (2015). Weighting methods for assessing mediation effect variation in multisite trials: An application to the National Job Corps Study. Paper presented at the 2015 Joint Statistical Meetings in Seattle, WA.

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

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

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology, 11, 550-560.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(398), 387-394.

Rotnitzky, A., & Vansteelandt, S. (2015). Double-robust methods. In G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis, & G. Verbeke (eds.), Handbook of Missing Data Methodology. Boca Raton, FL: CRC Press.

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

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, *75*(371), 591-593.

StataCorp. (2013a). Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.

StataCorp, (2013b). Stata 13 Base Reference Manual. College Station, TX: Stata Press.

Stefanski, L. A., & Boos, D. D. (2002). The calculus of m-estimation. American Statistician, 56, 29-38.

Tchetgen Tchetgen, E. J. (2013). Inverse odds ratio-weighted estimation for causal media-tion analysis. Statistics in medicine 32 (26), 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*(3), 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*(2), 137.

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

Vansteelandt, S., Bekaert, M. & Claeskens, G. (2012). On model selection and model misspecification in causal inference. Statistical Methods in Medical Research, 21, 7-30.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data (2nd ed.). Cambridge, MA: MIT Press.

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