Running Head: Effect Heterogeneity with Time-varying Treatments and Moderators

ESTIMATING HETEROGENEOUS CAUSAL EFFECTS WITH TIME-VARYING TREATMENTS AND TIME-VARYING EFFECT MODERATORS: STRUCTURAL NESTED MEAN MODELS AND REGRESSION-WITH-RESIDUALS

Geoffrey T. Wodtke

University of Toronto

Daniel Almirall

University of Michigan

Corresponding Author

Geoffrey T. Wodtke, Department of Sociology, University of Toronto, 725 Spadina Avenue, Toronto, ON M5S SJ4, Canada. Email: geoffrey.wodtke@utoronto.ca

*This research was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 0718128, by the National Institute of Child Health and Human Development under Grant Nos. T32 HD007339 , R24 HD041028, and R01 HD074603 to the Population Studies Center and Survey Research Center at the University of Michigan, and by the National Institute of Mental Health under Grant No. R03 MH097954.

ABSTRACT

Individuals differ in how they respond to a particular treatment, intervention, or exposure, and social scientists are often interested in understanding how treatment effects are systematically moderated by observed characteristics of individuals. Effect moderation occurs when individual covariates dampen or amplify the effect of some exposure. This article focuses on conceptualizing and estimating moderated causal effects in longitudinal settings where both the treatment and effect moderator of interest vary over time. Effect moderation is typically examined using covariate by treatment interactions in conventional regression analyses, but in the longitudinal setting, this approach is problematic because time-varying moderators of future treatment may be affected by prior treatment—that is, moderators may also be mediators. Conditioning on a mediator of prior treatment in a conventional regression model can lead to bias from over-control of intermediate pathways and collider stratification. This article introduces moderated intermediate causal effects and the structural nested mean model for analyzing effect heterogeneity in the longitudinal setting. It discusses problems with conventional regression estimation, presents a new approach to estimation that avoids these problems (regression-with-residuals), and describes different ways to account for confounding with this approach. The method is illustrated using longitudinal data from the PSID to examine whether the effects of time-varying exposures to concentrated neighborhood poverty on the risk of adolescent childbearing are moderated by time-varying family income.

INTRODUCTION

In the social sciences, researchers are often interested in understanding how the effects of a particular treatment, intervention, or exposure vary by characteristics of the individuals, families, or households exposed. For example, it is commonly hypothesized that the developmental impact of living in a high-poverty neighborhood is more severe for children in poor families than for children in wealthier families (Jencks and Mayer 1990; Wilson 1987; Wilson 1996). The effects of marital dissolution on child outcomes are also thought to depend on the degree of parental conflict, where the impact of divorce may be less severe if a primary caregiver is leaving an abusive spousal relationship (Amato 2004; Wallerstein 1991). Finally, the effects of school, classroom, and teacher characteristics on student achievement are often assumed to be a function of student abilities, where future gains are built upon foundations laid down earlier (Heckman 2006; Sanders et al. 1997). In fact, treatment effect heterogeneity is endemic to nearly all social contexts (Xie 2007; Xie et al. 2012), and it has important implications for social theory, research, and policy (Brand and Xie 2010; Heckman et al. 2006; Manski 2007; Wodtke et al. 2012).

Another common feature of treatments and the individuals, families, and households exposed to them is that they change over time. In the examples mentioned previously, people move and neighborhoods change, and family income fluctuates throughout the life course. Similarly, household conflict intensifies and subsides, and spouses maintain or dissolve marriages after varying lengths of time. With respect to school, classroom, and teacher effects, students advance through grades and schools at scheduled intervals, and their abilities evolve at different rates as they learn. Although these time-dependent phenomena are often shoehorned into simplified point-in-time research questions (e.g., Brooks-Gunn et al. 1993; Parcel and Dufur 2001), the proliferation of rich longitudinal data in which treatments, covariates, and outcomes

are measured at multiple time points provides a valuable opportunity for social scientists to examine questions that more accurately reflect the causal processes unfolding in the real world.

In particular, longitudinal data allow for an analysis of how the effects of time-varying treatments are moderated by an individual's evolving covariate history. Analyzing moderated causal effects in the longitudinal setting can provide a more rigorous test of the social theories that motivate research on systematic effect heterogeneity, illuminate the developmental and life-course processes through which social exposures incrementally affect individual behavior, and help to identify which individuals will be more sensitive or resilient to additional treatments on the basis of their evolving characteristics, outcomes, and needs.

To illustrate what is meant by effect moderation in the longitudinal setting, consider our motivating empirical example: neighborhood effects on teen childbearing. With measures of exposure to neighborhood poverty over different time intervals, of family income over different time intervals, and of our primary end-of-study outcome—whether a subject has a child during adolescence—we can investigate the following types of questions: "what is the impact of living in a high-poverty (versus low-poverty) neighborhood during childhood and then a low-poverty neighborhood during adolescence on the risk of adolescent childbearing among subjects whose families were poor during childhood" and "what is the impact of living in a low poverty neighborhood during childhood and then a high-poverty (versus low-poverty) neighborhood during adolescence on the risk of adolescent childbearing among subjects whose families were poor during adolescence?" These questions address the distal, proximal, and incremental effects of exposures to neighborhood poverty conditional on the evolving economic position of the family.

Analyzing effect moderation in the longitudinal setting is difficult because time-varying moderators of future treatment may be affected by prior treatment. For example, if a family is exposed to a high-poverty neighborhood during a subject's childhood, this may affect whether his or her family is poor in the future. Time-varying moderators affected by prior treatment present both conceptual and methodological challenges (Almirall et al. 2010; Elwert 2013; Robins et al. 2007; Robins 1994; VanderWeele and Robins 2007a). At the conceptual level, seemingly reasonable questions, such as "what is the effect of living in a high-poverty (versus low-poverty) neighborhood throughout childhood and adolescence among subjects whose families were poor throughout these periods," do not translate into well-defined causal contrasts. Causal contrasts compare the same subjects in two counterfactual states, but this question compares different subjects because the children in families who would stay poor had they continuously been exposed to high-poverty neighborhoods and the children in families who would stay poor had they been continuously exposed to low-poverty neighborhoods are almost certainly not the same set of subjects. When prior treatments help to create the subgroup of interest in the future, crafting coherent estimands for moderated causal effects is a difficult conceptual challenge.

At the methodological level, time-varying moderators affected by prior treatment complicate conventional estimation strategies. With cross-sectional data, point-in-time treatments, and pre-treatment moderators, effect moderation is typically examined using covariate by treatment interactions in conventional regression analyses (Wooldridge 2010) or using propensity score stratification methods (Xie et al. 2012). However, with longitudinal data, time-varying treatments, and time-varying moderators, regression and propensity score stratification methods that naively condition on time-varying moderators affected by prior

treatment cannot estimate moderated causal effects without bias. Even with a well-defined set of causal estimands that can be identified from observational or experimental data, conventional estimation strategies are biased because the moderator of interest is also a mediator of the treatment-outcome relationship.

To overcome these challenges, this article (1) introduces moderated intermediate causal effects and the structural nested mean model for analyzing effect heterogeneity in the longitudinal setting, (2) presents a simple regression-with-residuals estimation strategy that avoids the problems associated with conventional methods and can be implemented with off-the-shelf software, (3) describes several different ways that this estimation strategy can be extended to account for confounding, and (4) illustrates an application of these methods in the context of neighborhood-effects research. We begin with a brief review of effect moderation in the point-in-time setting. Next, we extend these ideas to a simplified longitudinal setting with only two time points, a binary treatment, and a single binary moderator. Then, we discuss further extensions of these methods for more complex scenarios that involve multivalued treatments and moderators, and confounding. Finally, we demonstrate an application of these methods using longitudinal data from the Panel Study of Income Dynamics (PSID) to investigate whether the effects of childhood and adolescent exposures to concentrated neighborhood poverty on the risk of teen childbearing are moderated by time-varying family income.

EFFECT MODERATION IN THE POINT-IN-TIME SETTING

A moderator is a pre-treatment variable which systematically modifies the form, direction, or strength of the effect of a treatment, A, on a response variable of interest, Y. Analyses of effect moderation typically compare some measure of the treatment-outcome association across levels

of third variable, M. If the measure of association differs significantly across levels of the third variable, then M is said to moderate the effect of A on Y.

Effect moderation is metric dependent. Under the same data-generating process, one measure of the treatment-outcome relationship may differ across levels of a putative moderator while another measure does not. The most common effect metrics are the difference in outcomes linked to different treatments and the ratio of outcomes linked to different treatments. We focus on the difference metric because it is the easiest to interpret and is widely considered to have the greatest relevance for public policy (Rothman et al. 1980; VanderWeele and Robins 2007b).

In this section, we formally define causal effect moderation in the point-in-time setting using the potential outcomes framework and briefly discuss several estimation strategies. Although the potential outcomes framework in the point-in-time setting is now well-understood and frequently used in the social sciences (Manski 2007; Morgan and Winship 2007), the review of these concepts and methods is intended to facilitate our exposition of effect moderation in the time-varying setting, which is considerably more complex.

A Counterfactual Model

Let A_i indicate exposure to a dichotomous point-in-time treatment for subject i. That is, $A_i = 1$ if subject i is exposed to treatment, and $A_i = 0$ if the subject is not exposed. In addition, let $Y_i(a)$ be the end-of-study outcome for subject i had she received the treatment $A_i = a$, possibly contrary to fact. The set $\{Y_i(0), Y_i(1)\}$ is known as the set of potential outcomes for subject i.

In the counterfactual framework, causal effects are defined as contrasts between potential outcomes. The *individual causal effect* for subject i is given by

$$ICE_i = Y_i(1) - Y_i(0), \quad (1)$$

which is the difference between a subject's outcome had she received treatment and the same subject's outcome had she not received treatment. For each subject, we only ever observe the single potential outcome that corresponds to the treatment actually received, and the other potential outcomes are counterfactuals. Thus, individual causal effects can never be observed in reality. This is known as the fundamental problem of causal inference (Holland 1986; Rubin 1974). Under certain assumptions, such as the homogeneity of subjects or temporal invariance, the ICE_i can be computed, but these assumptions are generally indefensible in the social sciences (Holland 1986). To overcome these difficulties, researchers typically focus on the *average causal effect*, which is given by

$$ACE = E(ICE_i) = E\big(Y_i(1) - Y_i(0)\big). \tag{2}$$

This quantity describes how treatment affects subjects in the population of interest on average. It can be estimated under weaker assumptions (discussed below) than those required to compute individual causal effects.

Based on these definitions of individual and average causal effects, *moderated average causal effects* can be defined as

$$\mu(M_i, a) = E(Y_i(a) - Y_i(0)|M_i) = a \times E(Y_i(1) - Y_i(0)|M_i), \tag{3}$$

where M_i is a pre-treatment variable. Formally, M_i is a moderator for the causal effect of treatment if $\mu(M_i, a)$ is non-constant in M_i—that is, if M_i helps to summarize variability in the individual causal effects across the population of interest. Note that this definition specifies that the moderator occurs before treatment, which in turn occurs before the outcome. This temporal ordering is implicit in the notation because M_i is not indexed by a as a potential outcome of treatment. In addition, this definition neither requires nor prohibits that M_i has a causal effect on

the outcome, but it does specify that any such effect is not the primary effect of interest in the current analysis.

Moderated average causal effects can be linked to the conditional mean of the potential outcomes through the following additive decomposition:

$$E(Y_i(a)|M_i) = \beta_0 + \varepsilon(M_i) + \mu(M_i, a), \quad (4)$$

where $\beta_0 = E(Y_i(0)|M_i = 0)$, $\varepsilon(M_i) = E(Y_i(0)|M_i = m) - E(Y_i(0)|M_i = 0)$, and $u(M_i, a)$ is defined as above. The intercept, β_0, gives the mean outcome value had individuals in the subgroup for which $M_i = 0$ not received treatment; the function $\varepsilon(M_i)$ is the associational effect of the moderator on the outcome had subjects not received treatment; and the function $u(M_i, a)$ describes the moderated causal effects of treatment. Because our primary interest is in the causal function, $\mu(M_i, a)$, the associational effect of the moderator, $\varepsilon(M_i)$, is called a nuisance function.

We consider linear parametric models for the causal and nuisance functions in Equation 4. Any parameterization of the causal function $\mu(M_i, a)$ must satisfy the constraint that it equals zero when $a = 0$. This constraint motivates the common use of interaction terms to model effect moderation. For example, when both treatment and the moderator are binary, a saturated model for $\mu(M_i, a)$ is

$$\mu(M_i, a; \beta) = a(\beta_1 + \beta_2 M_i) = \beta_1 a + \beta_2 M_i a, \quad (5)$$

where $\beta_1$is the average causal effect of treatment among subjects in group $M_i = 0$ and β_2 increments this effect for subjects with $M_i = 1$. If $\beta_2 = 0$, then M_i is not a moderator.

Any parameterization of the nuisance function, $\varepsilon(M_i)$, must satisfy the constraint that it equals zero when the moderator is equal to zero. A saturated model for $\varepsilon(M_i)$ with a binary moderator is

$$e(M_i; \lambda) = \lambda_1 M_i, \quad (6)$$

where λ_1 gives the associational effect of the moderator on the outcome. Combining the models for the causal and nuisance functions yields a saturated linear model for the conditional mean of $Y_i(a)$ given M_i:

$$E(Y_i(a)|M_i) = \beta_0 + \lambda_1 M_i + a(\beta_1 + \beta_2 M_i), \quad (7)$$

which is the familiar linear model with an intercept term, "main effects" for treatment and the moderator, and a treatment by moderator interaction.

Estimation

The moderated causal effects defined previously can be identified from observed data under the assumption of ignorability of treatment assignment (Holland 1986; Morgan and Winship 2007; Rubin 1974). One version of this assumption can be written as

$$Y_i(a) \perp A_i|M_i \; \forall \; a \quad (8)$$

where $\perp$ denotes statistical independence. Substantively, this condition states that there are not any pre-treatment variables other than the moderator that directly affect selection into treatment and the outcome. The ignorability assumption is met by design in experimental studies where treatment is randomly assigned.[1] Figure 1 displays two directed acyclic graphs (DAGs) that describe simple causal systems where this assumption is satisfied. Panel A describes the scenario where treatment assignment is random, while Panel B depicts the situation where selection into treatment is determined solely on the basis of the moderator.

When the ignorability assumption is satisfied, the causal parameters defined previously can be estimated with the following observed data model:

$$E(Y_i|A_i, M_i) = \beta_0^* + \lambda_1^* M_i + A_i(\beta_1^* + \beta_2^* M_i). \quad (9)$$

[1] In fact, experimental studies typically meet stronger ignorability assumptions—for example, $Y_i(a) \perp A_i \; \forall \; a$.

The asterisks on these parameters denote the distinction between the fundamentally unobservable mean differences between potential outcomes in Equation 7 and the observed differences between conditional means in Equation 9, which are equivalent only under the ignorability assumption. In this situation, ordinary least squares estimates of the regression of Y_i on M_i, A_i, and M_iA_i can be used to estimate the moderated causal effects of interest.

Adjustment for Confounding

It is often the case in the social sciences that the ignorability assumption defined in Equation 8 does not hold because randomization is not possible or there are pre-treatment variables other than the moderator that affect both treatment selection and the outcome. These variables are called confounders, and they lead to bias if they are not properly accounted for. Figure 2 contains a DAG that graphically depicts the problem of confounding bias. It shows that a fourth variable, C, directly affects both treatment and the outcome. Under a slightly modified version of the ignorability assumption, the moderated causal effects of interest can still be identified from observed data in the presence of confounders, but more complicated estimation methods are required (Holland 1986; Morgan and Winship 2007; Rubin 1974). Formally, this assumption can be written as

$$Y_i(a) \perp A_i | M_i, C_i \; \forall \; a. \tag{10}$$

Substantively, it states that there are no unobserved determinants of both treatment and the outcome other than M_i and C_i. Thus, in the special case where all confounders of treatment are observed by the researcher, unbiased estimation of moderated causal effects remains possible.

There are two approaches to adjusting for observed confounders. With the first approach, all observed confounders are included along with A_i and M_i in the observed data regression

model for Y_i. This covariate-adjusted regression approach requires correct specification of how the outcome relates to the treatment, moderator, and confounders. Typically, only "main effects" for observed confounders are included in the model, but higher-order terms, including interactions between confounders and treatment, are also possible. In the simple case with just a single binary confounder, this model can be expressed as

$$E(Y_i|A_i, M_i, C_i) = \beta_0^* + \eta_1^* C_i + \lambda_1^* M_i + A_i(\beta_1^* + \beta_2^* M_i). \quad (11)$$

It assumes that the effect of treatment on the outcome varies only by levels of M_i. This approach to estimation becomes problematic if C_i is also a moderator for the treatment-outcome relationship or if the number of observed confounders is large. In these situations, covariate-adjusted regression estimation becomes heavily reliant on functional form, and the analyst may have to move away from a model for effect moderation by M_i toward a model for effect moderation by multiple covariates, including C_i, which may not be of scientific interest.

With the second approach, inverse-probability-of-treatment (IPT) weighting is used to balance observed confounders, C_i, across levels of the treatment (Hirano and Imbens 2001; Robins et al. 2000; Rosenbaum and Rubin 1983), and the linear regression model is reserved for examining effect moderation by M_i. This method involves reweighting the observed data by a function of the propensity score to generate a pseudo-sample in which treatment is no longer confounded by observed covariates. The IPT-weight for subject i is given by

$$w_i = \frac{P(A_i = a_i|M_i)}{P(A_i = a_i|M_i, C_i)}. \quad (12)$$

It is the ratio of the conditional probability that a subject is exposed to the actual treatment she received given all her pre-treatment covariates and the conditional probability of treatment given only the moderator. IPT-weighting balances confounders across treatment levels by giving more weight to subjects with confounders that are underrepresented in their treatment group and less

weight to subjects with confounders that overrepresented in their treatment group. The true IPT weights are unknown and must be estimated from data, which requires a correctly specified model for the conditional probability of treatment (Hirano and Imbens 2001; Robins et al. 2000; Rosenbaum and Rubin 1983). After estimates of the IPT-weights are computed, a weighted least squares regression of Y_i on M_i, A_i, and M_iA_i with weights equal to $\widehat{w}_i$ provides estimates of the moderated causal effects of interest.

EFFECT MODERATION IN THE LONGITUDINAL SETTING

This section transitions to the situation in which both the treatment and putative moderator vary over time. For expositional simplicity, we focus on a simple example with a binary treatment measured at two points in time, a single binary moderator measured at two points in time, and an end-of-study outcome variable.

In the longitudinal setting, formulating coherent causal questions can be conceptually challenging. The challenge arises from the possibility that future moderators may be affected by prior treatment, and thus prior treatment creates, at least in part, the time-varying subgroups of interest. This complication precludes intuitively appealing questions about how the effects of long-term treatment trajectories vary across subgroups defined in terms of long-term moderator trajectories, such as "what is the effect of being *always* (versus *never*) treated among subjects who were *always* in a particular subgroup?" These types of questions are inappropriate because they cannot be translated into counterfactuals that compare the same set of subjects: had a subject been always (versus never) treated, she may not have been always in the one particular subgroup of interest.

In this section, we resolve these conceptual difficulties by introducing *moderated intermediate causal effects*, which isolate the average causal effects of one additional wave of treatment (versus no additional waves) conditional on treatments and moderators prior to that wave. Moderated intermediate causal effects address questions like "what is the effect of being treated only at wave 1 (versus not being treated at all) among subjects who were in a particular subgroup prior to wave 1" and "what is the effect of being treated at wave 2 (versus not being treated at wave 2) among subjects who were in a particular subgroup prior to wave 2 and exposed to a particular treatment at wave 1?" By carefully attending to the temporal ordering of treatments, moderators, and the outcome, this approach allows for the formulation of coherent causal questions about effect moderation in the longitudinal setting.

Moderated Intermediate Causal Effects

Let A_{it} indicate exposure to a dichotomous treatment for subject i at time $t = 1,2$. That is, $A_{it} = 1$ if subject i is exposed to treatment at time t, and $A_{it} = 0$ otherwise. In addition, let $Y_i(a_1, a_2)$ be the end-of-study outcome for subject i had she received the treatment sequence $(A_{i1} = a_1, A_{i2} = a_2)$, possibly contrary to fact. By "end-of-study," we mean occurring after a_2. The set $\{Y_i(0,0), Y_i(1,0), Y_i(0,1)Y_i(1,1)\}$ gives all possible potential end-of-study outcomes for subject i. Now, let M_{i1} be the binary moderator of interest for subject i at time $t = 1$, which is measured just prior to treatment at time 1. Similarly, let $M_{i2}(a_1)$ be the binary moderator for subject i at time $t = 2$ had she been exposed to the prior treatment a_1. This measure of the moderator occurs after treatment at time 1 but before treatment at time 2. It is therefore indexed as a potential outcome of treatment at time 1. The set $\{M_{i2}(0), M_{i2}(1)\}$ gives the potential

outcomes of the moderator at time 2 for subject i. The complete set of all variables in temporal order is $\{M_{i1}, A_{i1}, M_{i2}(a_1), A_{i2}, Y_i(a_1, a_2)\}$.

With two time points, there are two sets of moderated intermediate causal effects, one set for each time point. The first set of causal effects is defined as

$$u_1(M_{i1}, a_1) = E(Y_i(a_1, 0) - Y_i(0,0)|M_{i1} = m_1), \quad (13)$$

which gives the average causal effect of being exposed to treatment only at time 1 (versus never being exposed) among the subgroups defined by M_{i1}, and M_{i1} is said to be a moderator for the effect of treatment at time 1 if $u_1(M_{i1}, a_1)$ is nonconstant in M_{i1}. Note that the function $u_1(M_{i1}, a_1)$ equals zero when a_1 is equal to zero. The second set of causal effects is defined as

$$u_2(M_{i1}, a_1 M_{i2}(a_1), a_2) = E\big(Y_i(a_1, a_2) - Y_i(a_1, 0)|M_{i1}, M_{i2}(a_1)\big), \quad (14)$$

which gives the average causal effect of being exposed to treatment at time 2 (versus not being exposed) among subgroups defined by $M_{i1}$and $M_{i2}(a_1)$had subjects initially been exposed to treatment a_1. If $u_2(M_{i1}, a_1 M_{i2}(a_1), a_2)$ is nonconstant in $\big(M_{i1}, M_{i2}(a_1)\big)$, then these variables are said to be moderators for the effect of treatment at time 2.[2] As before, the function $u_2(M_{i1}, a_1 M_{i2}(a_1), a_2)$ equals zero when a_2 is equal to zero.

Moderated intermediate causal effects involve fairly complex counterfactual contrasts. To better understand these contrasts, it can helpful to consult the language and logic of sequential experiments. Consider a hypothetical experiment in which the researcher measures M_{i1} and then randomly assigns subjects to different treatments at time 1 but the same treatment at time 2. Comparing mean end-of-study outcomes for subjects assigned to different treatments at time 1, separately among the subgroups defined by M_{i1}, would be an experimental estimate of

[2] This causal function may also be nonconstant in a_1, indicating that there is an interaction effect between treatments received at time 1 and time 2.

$u_1(M_{i1}, a_1)$. Now consider another hypothetical experiment where the researcher measures M_{i1}, assigns all subjects to the same treatment at time 1, measures $M_{i1}(a_1)$, and then randomly assigns subjects to different treatments at time 2. Comparing mean end-of-study outcomes for subjects assigned to different treatments at time 2, separately among subgroups defined by M_{i1} and $M_{i1}(a_1)$, would be an experimental estimate of $u_2(M_{i1}, a_1 M_{i2}(a_1), a_2)$. Rather than conducting two separate experiments, these effects can also be estimated from a single sequentially randomized experiment in which subjects are assigned to different treatments at each time point and measures of the moderator are taken just prior to treatment assignment (Almirall, Compton, Gunlicks-Stoessel, Duan, and Murphy 2012).

The Structural Nested Mean Model

The structural nested mean model (SNMM) formally relates $u_1(M_{i1}, a_1)$ and $u_2(M_{i1}, a_1 M_{i2}(a_1), a_2)$ to the conditional mean of the potential outcomes (Robins 1994). Specifically, the moderated intermediate causal effects of interest can be linked to the conditional mean of $Y_i(a_1, a_2)$ through the following additive decomposition:

$$E\big(Y_i(a_1, a_2)|M_{i1}, M_{i2}(a_1)\big) = \beta_0 + \varepsilon_1(M_{i1}) + u_1(M_{i1}, a_1) + \varepsilon_2\big(M_{i1}, a_1, M_{i2}(a_1)\big) + u_2(M_{i1}, a_1 M_{i2}(a_1), a_2), \quad (15)$$

where the intercept $\beta_0 = E\big(Y_i(0,0)\big)$ is the mean under no treatment, $\varepsilon_1(M_{i1}) = E(Y_i(0,0)|M_{i1}) - E\big(Y_i(0,0)\big)$ is the association between M_{i1} and the outcome had no subjects been exposed to treatment, and $\varepsilon_2\big(M_{i1}, a_1, M_{i2}(a_1)\big) = E\big(Y_i(a_1, 0)\big|M_{i1}, M_{i2}(a_1)\big) - E(Y_i(a_1, 0)|M_{i1})$ is the association between $M_{i2}(a_1)$ and the outcome had subjects in the groups defined by M_{i1}initially been exposed to treatment a_1 and then no treatment at time 2. Because our primary interest is in the causal functions $u_1(M_{i1}, a_1)$ and $u_2(M_{i1}, a_1 M_{i2}(a_1), a_2)$, the

associational effects of the moderators, $\varepsilon_1(M_{i1})$ and $\varepsilon_2\big(M_{i1}, a_1, M_{i2}(a_1)\big)$, are called nuisance functions.

An important property of the nuisance functions is that, conditional on the past, they have mean zero. That is,

$$E\big(\varepsilon_2\big(M_{i1}, a_1, M_{i2}(a_1)\big)\big|M_{i1}\big) = E\big(E\big(Y_i(a_1,0)\big|M_{i1}, M_{i2}(a_1)\big) - E(Y_i(a_1,0)|M_{i1})\big|M_{i1}\big)$$

$$= E\big(E\big(Y_i(a_1,0)\big|M_{i1}, M_{i2}(a_1)\big)\big|M_{i1}\big) - E(Y_i(a_1,0)|M_{i1}) = 0, \text{ and} \quad (16)$$

$$E\big(\varepsilon_1(M_{i1})\big) = E\left(E(Y_i(0,0)|M_{i1}) - E\big(Y_i(0,0)\big)\right)$$

$$= E\big(E(Y_i(0,0)|M_{i1})\big) - E\big(Y_i(0,0)\big) = 0. \quad (17)$$

This property of the nuisance functions gives them an interpretation as error terms and will inform their parameterization below.

We consider linear parametric models for the causal and nuisance functions of the SNMM. Any parameterization of the causal function $u_1(M_{i1}, a_1)$ must satisfy the constraint that it equals zero when a_1 is equal to zero. With a binary treatment and moderator, a saturated model for $u_1(M_{i1}, a_1)$ is

$$u_1(M_{i1}, a_1; \beta_1) = a_1(\beta_{10} + \beta_{11}M_{i1}). \quad (18)$$

This model includes the familiar interaction term between treatment and the moderator at time 1, where $\beta_{10}$is the average causal effect of treatment at time 1 among subjects in group $M_{i1} = 0$ and β_{11} increments this effect for subjects in group $M_{i1} = 1$. If $\beta_{11} = 0$, then M_{i1} is not a moderator for treatment at time 1.

Similarly, a saturated model for $u_2(M_{i1}, a_1 M_{i2}(a_1), a_2; \beta_2)$ is

$$u_2(M_{i1}, a_1 M_{i2}(a_1), a_2; \beta_2) = a_2\big(\beta_{20} + \beta_{21}M_{i1} + \beta_{22}M_{i2}(a_1) + \beta_{23}M_{i1}M_{i2}(a_1) +$$

$$\beta_{24}a_1 + \beta_{25}M_{i1}a_1 + \beta_{26}a_1M_{i2}(a_1) + \beta_{27}M_{i1}a_1M_{i2}(a_1)\big). \quad (19)$$

This model includes all possible interactions between treatment at time 2, prior treatment, and prior moderators. Specific linear combinations of the beta parameters return the average causal effect of being exposed to treatment at time 2 (versus not being exposed) among the subgroups defined by $M_{i1}$and $M_{i2}(a_1)$ had subjects initially been exposed to treatment a_1. For example, among subjects in group $M_{i1} = 0$ at time 1, who had not been exposed to treatment at time 1, and who were in group $M_{i2}(0) = 0$ at time 2 under no prior treatment, β_{20} is the average causal effect being exposed to treatment at time 2. As another example, $\beta_{20} + \beta_{21}$ gives the same effect among subjects in group $M_{i1} = 1$ at time 1.

The key to parameterizing the nuisance functions is to ensure that the model satisfies their zero conditional mean property. With this constraint in mind, a saturated model for $\varepsilon_1(M_{i1})$ is

$$\varepsilon_1(M_{i1}; \lambda_1) = \lambda_{10}\delta(M_{i1}), \tag{20}$$

where $\delta(M_{i1}) = M_{i1} - E(M_{i1})$ and λ_{10} gives the associational effect of the moderator at time 1 on the outcome had subjects not been exposed to treatment at any time point. It satisfies the zero conditional mean property because $E\big(\delta(M_{i1})\big) = E\big(M_{i1} - E(M_{i1})\big) = 0$.

A saturated model for the second nuisance function is

$$\varepsilon_2(M_{i1}, a_1, M_{i2}(a_1); \lambda_2) = \delta\big(M_{i2}(a_1)\big)(\lambda_{20} + \lambda_{21}M_{i1} + \lambda_{22}a_1 + \lambda_{23}a_1M_{i1}), \tag{21}$$

where $\delta\big(M_{i2}(a_1)\big) = M_{i2}(a_1) - E(M_{i2}(a_1)|M_{i1})$ and different combinations of the lambda parameters give the associational effect of the moderator at time 2 on the outcome. It satisfies the zero conditional mean constraint because $E\big(\delta\big(M_{i2}(a_1)\big)\big|M_{i1}\big) =$ $E(M_{i2}(a_1) - E(M_{i2}(a_1)|M_{i1})|M_{i1}) = E(M_{i2}(a_1)|M_{i1}) - E(M_{i2}(a_1)|M_{i1}) = 0$. It is important to note that $\delta(M_{i1})$ and $\delta\big(M_{i2}(a_1)\big)$ are similar to residual terms from regressions of the moderators on past covariates.

Combining the models for the causal and nuisance functions yields the following saturated SNMM:

$$E\big(Y_i(a_1,a_2)|M_{i1},M_{i2}(a_1)\big)=\beta_0+\lambda_{10}\delta(M_{i1})+a_1(\beta_{10}+\beta_{11}M_{i1})+\delta\big(M_{i2}(a_1)\big)(\lambda_{20}+\lambda_{21}M_{i1}+\lambda_{22}a_1+\lambda_{23}a_1M_{i1})+a_2\big(\beta_{20}+\beta_{21}M_{i1}+\beta_{22}M_{i2}(a_1)+\beta_{23}M_{i1}M_{i2}(a_1)+\beta_{24}a_1+\beta_{25}M_{i1}a_1+\beta_{26}a_1M_{i2}(a_1)+\beta_{27}M_{i1}a_1M_{i2}(a_1)\big). \quad (22)$$

This equation is similar to the familiar linear model with all possible interaction terms between treatments and moderators except that in several places the moderators are replaced with terms that resemble residuals. The saturated SNMM has a total of 16 parameters, one for every possible combination of binary treatments and moderators. For now, we focus on the saturated model, but in a subsequent section, we discuss several simplifying assumptions that might be imposed on the functional form of the SNMM, particularly in situations where multivalued treatments and moderators make a saturated model intractable.

Estimation

The moderated intermediate causal effects defined previously can be identified from observed data under the assumption of sequential ignorability of treatment assignment (Almirall, Ten Have, and Murphy 2010; Robins 1994). This assumption is formally expressed in two parts:

$$Y_i(a_1,a_2)\perp A_{i1}|M_{i1}\ \forall\ (a_1,a_2), \quad (23)$$

$$Y_i(a_1,a_2)\perp A_{i2}|M_{i1},A_{i1},M_{i2}\ \forall\ (a_1,a_2). \quad (24)$$

Substantively, it states that at each time point there are not any variables other than the prior moderators and treatments that directly affect selection into future treatment and the outcome. This assumption is met by design in sequentially randomized experimental studies. Figure 3 displays two DAGs that describe time-dependent causal systems in which this assumption is

satisfied. Panel A describes the scenario in which treatment is sequentially randomized, while Panel B depicts the situation in which selection into treatment is determined only on the basis of prior treatments and moderators. In both DAGS, the treatment and moderator at each time point directly affect the outcome, and treatment at time 1 has an indirect effect on the outcome that operates through the moderator at time 2. Unobserved characteristics of a subject, denoted by U_i, affect the moderator and the outcome, but not treatment.

Limitations of Conventional Regression Estimation

Recall that in the point-in-time setting, moderated causal effects can be consistently estimated with a conventional regression that conditions on the treatment, moderator, and a treatment-moderator interaction term when the ignorability assumption is satisfied. In this section, we explain why estimates from an analogous conventional regression in the longitudinal setting are biased, even when the sequential ignorability assumption is satisfied. Consider the following observed data regression model:

$$E(Y_i|M_{i1}, A_{i1}, M_{i2}, A_{i2}) = \beta_0^* + \lambda_{10}^* M_{i1} + A_{i1}(\beta_{10}^* + \beta_{11}^* M_{i1}) + M_{i2}(\lambda_{20}^* + \lambda_{21}^* M_{i1} + \lambda_{22}^* A_{i1} + \lambda_{23}^* A_{i1} M_{i1}) + A_{i2}(\beta_{20}^* + \beta_{21}^* M_{i1} + \beta_{22}^* M_{i2} + \beta_{23}^* M_{i1} M_{i2} + \beta_{24}^* A_{i1} + \beta_{25}^* M_{i1} A_{i1} + \beta_{26}^* A_{i1} M_{i2} + \beta_{27}^* M_{i1} A_{i1} M_{i2}). \quad (25)$$

Least squares estimates of the parameters in this equation are biased for the causal parameters in the SNMM for two reasons. First, because this model conditions naively on M_{i2}, which mediates the effect of prior treatment, A_{i1}, on the outcome, the parameters $\{\beta_{10}^*, \beta_{11}^*\}$ do not capture the indirect effect of prior treatment that operates through future levels of the moderator. This problem is known as over-control of intermediate pathways. It is depicted visually with the stylized graph in Panel A of Figure 4.

Despite the problem of over-control, it may be tempting to assume that these parameters still recover the moderated *direct* effects of treatment at time 1, holding the future moderator constant. However, the parameters $\{\beta_{10}^{*}, \beta_{11}^{*}\}$ cannot even be interpreted as direct effects of treatment because of a second problem associated with conditioning on an intermediate variable: collider-stratification bias. As depicted graphically in Panel B of Figure 4, conditioning on M_{i2} induces an association between prior treatment and the unobserved determinants of Y_i (i.e. the error term of the outcome), which leads to bias. The same problems prevent estimation of moderated causal effects in the longitudinal setting with propensity score stratification methods (e.g., Xie, Brand, and Jann 2012). Indeed, even with data from an optimal sequentially randomized experiment, conventional methods fail to recover the moderated causal effects of interest if the moderator is time-varying and affected by prior treatment.

Another way to conceive of these problems is as a specification error in the observed data regression, and specifically, as an error in the specification of the nuisance functions. In the SNMM, the nuisance functions are specified using a residual transformation of the time-varying moderators, but the observed data regression considered here includes the untransformed values of these variables in the model. As we show in the next section, unbiased estimation of moderated intermediate causal effects can be achieved with an observed data regression that correctly specifies the nuisance functions with appropriate residual terms.

Regression-with-residuals

Regression-with-residuals (RWR) estimation is very similar to the conventional regression approach discussed in the previous section, but it avoids the problems of over-control and collider stratification bias by aiming to correctly model the nuisance functions in the SNMM

(Almirall et al. 2013; Almirall et al. 2010). This method proceeds in two stages. In the first stage, the time-varying moderators are regressed on the observed past to obtain estimates of the residual terms $\delta(M_{i1})$ and $\delta\big(M_{i2}(a_1)\big)$. Specifically, parametric models of the form $E(M_{i1}) = \phi$ and $E(M_{i2}|M_{i1}, A_{i1}) = f(M_{i1}, A_{i1}; \psi)$ are estimated and then used to construct the following residual terms: $\hat{\delta}(M_{i1}) = M_{i1} - \hat{E}(M_{i1}) = M_{i1} - \hat{\phi}$ and $\hat{\delta}(M_{i2}) = M_{i2} - \hat{E}(M_{i2}|M_{i1}, A_{i1}) = M_{i2} - f\big(M_{i1}, A_{i1}; \hat{\psi}\big)$. When M_{i2} is binary, $E(M_{i2}|M_{i1}, A_{i1})$ can be estimated by least squares using a saturated linear probability model, such as $E(M_{i2}|M_{i1}, A_{i1}) = \psi_0 + \psi_1 M_{i1} + A_{i1}(\psi_2 + \psi_3 M_{i1})$, or it could be estimated by maximum likelihood using a more complex nonlinear model (e.g., logit or probit).

In the second stage, the SNMM is estimated via the following observed data regression that replaces the untransformed values of the moderators in the nuisance functions with the residualized values:

$$E(Y_i|M_{i1}, A_{i1}, M_{i2}, A_{i2}) = \beta_0^* + \lambda_{10}^* \hat{\delta}(M_{i1}) + A_{i1}(\beta_{10}^* + \beta_{11}^* M_{i1}) + \hat{\delta}(M_{i2})(\lambda_{20}^* + \lambda_{21}^* M_{i1} + \lambda_{22}^* A_{i1} + \lambda_{23}^* A_{i1} M_{i1}) + A_{i2}(\beta_{20}^* + \beta_{21}^* M_{i1} + \beta_{22}^* M_{i2} + \beta_{23}^* M_{i1} M_{i2} + \beta_{24}^* A_{i1} + \beta_{25}^* M_{i1} A_{i1} + \beta_{26}^* A_{i1} M_{i2} + \beta_{27}^* M_{i1} A_{i1} M_{i2}). \quad (26)$$

Least squares estimates of this equation are unbiased and consistent for the moderated causal effects of interest under the sequential ignorability assumption outlined previously (Almirall et al. 2013; Almirall et al. 2010). The only difference between this model and the conventional regression in Equation 25 is that it correctly specifies the nuisance functions of the SNMM using residualized, rather than untransformed, values of the moderators.

Figure 5 displays a stylized graph that provides some intuition as to how RWR estimation overcomes the problems of over-control and collider-stratification. It shows that residualizing M_{i2} based on the observed past "purges" this variable of its association with prior treatment

while leaving its association with other variables (e.g., the outcome) intact. Thus, conditioning on the residualized moderator in an observed data regression for Y_i does not remove the indirect effect of prior treatment that operates through the moderator and does not induce an association between prior treatment and unobserved determinants of the outcome.

Adjustment for Confounding

It is often the case in the social sciences that the sequential ignorability assumptions defined in Equations 23 and 24 do not hold because there are variables other than the moderators and prior treatments that affect selection into future treatments and the outcome. When these variables change over time, they are called time-varying confounders, and they lead to bias if not properly accounted for. Panel A of Figure 6 contains a DAG that graphically depicts the problem of bias due to time-varying confounders. It shows additional variables, C_1 and C_2, that affect both treatment and the outcome. Specifically, C_1 is a confounder for the effects of A_1on Y, and both C_1 and C_2 are confounders for the effect of A_2 on Y. Under a slightly modified version of the sequential ignorability assumptions defined previously, the moderated intermediate causal effects of interest can still be identified from observed data in the presence of time-varying confounders, but more complicated estimation methods are required.

To appreciate the need for more complicated estimation methods in this setting, note that if the analysis adjusts naively for C_1 and C_2 by including them as covariates in either a conventional regression or even as part of a RWR (where the moderators, M_1 and M_2, are residualized), it may also incur biases due to over-control and collider stratification. This could happen for the very same reasons that adjusting naively for M_1 and M_2 may lead to such biases, which were previously described in Figure 4.

The stylized graph in Panel B of Figure 6 depicts what would happen if RWR were used to adjust appropriately for the time-varying moderators, but the confounders were adjusted for naively by including them directly in the outcome model. An example of such a regression model is

$$E(Y_i|C_{i1}, M_{i1}, A_{i1}, C_{i2}, M_{i2}, A_{i2}) = \beta_0^* + \lambda_{10}^*\hat{\delta}(M_{i1}) + \lambda_{11}^* C_{i1} + A_{i1}(\beta_{10}^* + \beta_{11}^* M_{i1}) + \hat{\delta}(M_{i2})(\lambda_{20}^* + \lambda_{21}^* M_{i1} + \lambda_{22}^* A_{i1} + \lambda_{23}^* A_{i1} M_{i1}) + \lambda_{24}^* C_{i2} + A_{i2}(\beta_{20}^* + \beta_{21}^* M_{i1} + \beta_{22}^* M_{i2} + \beta_{23}^* M_{i1} M_{i2} + \beta_{24}^* A_{i1} + \beta_{25}^* M_{i1} A_{i1} + \beta_{26}^* A_{i1} M_{i2} + \beta_{27}^* M_{i1} A_{i1} M_{i2}). \quad (27)$$

Note that Equation 27 does not include interaction terms between C_{i1} and A_{i1} or between (C_{i1}, A_{i1}, C_{i2}) and A_{i2}. This is consistent with the intent to adjust for C_{i1} and C_{i2} because they are time-varying confounders, not because they are moderators of scientific interest. Similar to the problems outlined in the previous section, Panel B of Figure 6 shows that least squares estimates of $(\beta_{10}^*, \beta_{11}^*)$ are biased due to over-control and collider stratification, which results from naively conditioning on C_{i2} in this model. In the sections that follow, we present two approaches for estimating moderated intermediate causal effects in the presence of time-varying confounders that avoid these problems.

Covariate-adjusted Regression-with-residuals

Covariate-adjusted regression-with-residuals (CA-RWR) is nearly identical to the RWR method discussed previously, but it avoids the problems of over-control and collider stratification bias due to conditioning on untransformed values of the time-varying confounders by conditioning instead on a residual transformation of these variables. As with RWR, CA-RWR proceeds in two stages, but it involves four additional modeling considerations. First, two more estimated residuals are obtained in the first stage—one each for C_{i1} and C_{i2}. These are defined as $\hat{\delta}(C_{i1}) =$

$C_{i1} - \hat{E}(C_{i1})$ and $\hat{\delta}(C_{i2}) = C_{i2} - \hat{E}(C_{i2}|C_{i1}, M_{i1}, A_{i1})$, respectively. Second, with CA-RWR, $\hat{\delta}(M_{i2})$ may now depend on C_{i1}—that is, $\hat{\delta}(M_{i2})$ is defined as $\hat{\delta}(M_{i2}) = M_{i2} - \hat{E}(M_{i2}|M_{i1}, C_{i1}, A_{i1})$. Third, instead of conditioning on the untransformed values of the time-varying confounders, the SNMM is estimated via a second-stage regression that replaces them with the residualized values $\hat{\delta}(C_{i1})$ and $\hat{\delta}(C_{i2})$. Fourth, the second-stage regression may now include additional terms involving the confounders as part of the nuisance functions. For example, the associational effect of the moderators on the outcome may vary by levels of the confounders, which would necessitate including higher order interaction terms for different cross-products of C_{i1} and C_{i2} with $\hat{\delta}(M_{i1})$ and $\hat{\delta}(M_{i2})$.

CA-RWR requires a different set of identification assumptions compared with RWR. For the CA-RWR approach, the moderated intermediate causal effects $u_1(M_{i1}, a_1; \beta_1)$ and $u_2(M_{i1}, a_1 M_{i2}(a_1), a_2; \beta_2)$ can be identified from observed data under an expanded version of the sequential ignorability assumptions defined previously. Specifically, CA-RWR requires that

$$Y_i(a_1, a_2) \perp A_{i1}|M_{i1}, C_{i1} \; \forall \; (a_1, a_2) \text{ and} \tag{28}$$

$$Y_i(a_1, a_2) \perp A_{i2}|M_{i1}, C_{i1}, A_{i1}, M_{i2}, C_{i2} \; \forall \; (a_1, a_2). \tag{29}$$

Substantively, this assumption states that at each time point there are not any variables other than the prior moderators, time-varying confounders, and treatments that directly affect selection into future treatment and the outcome. These conditions subsume those defined in Equations 23 and 24, which implies that CA-RWR requires a weaker set of ignorability assumptions than RWR. However, compared to RWR, CA-RWR requires additional modeling assumptions. In particular, CA-RWR requires that the time-varying confounders are not also moderators for the effects of treatment on the outcome. This assumption is encoded in the models for the moderated intermediate causal effects, which depend only on M_{i1} and M_{i2}, and not on C_{i1} and C_{i2}.

Figure 7 displays a stylized graph that provides some intuition as to how CA-RWR estimation overcomes the problems of over-control and collider stratification bias that result from naively conditioning on the untransformed time-varying confounders. Note that these are essentially the same problems that were encountered previously as a result of conditioning naively on the untransformed time-varying moderators, except that here the problems apply to C_{i1} and C_{i2}. The figure shows that residualizing both M_{i2} and C_{i2} based on the observed past "purges" these variables of their association with prior treatment while leaving their association with other variables (e.g., the outcome and future treatment) intact. This approach avoids controlling away part of the treatment effect that operates through future time-varying confounders, and it avoids inducing a non-causal association between A_{i1} and unobserved variables, such as V, that are joint determinants of both C_{i2} and the outcome, Y.

IPT-weighted Regression-with-residuals

IPT-weighted regression-with-residuals (IPTW-RWR) aims to overcome two related limitations of the CA-RWR approach (Almirall et al. 2014). The first limitation involves the modeling assumption associated with CA-RWR which states that time-varying confounders are not also moderators of treatment effects on the outcome. In many cases, social scientists are specifically interested in how a particular time-varying covariate, say M_{i1} and M_{i2}, moderate the effects of a time-varying treatment, but they do not wish to rule out the possibility that other time-varying covariates, such as C_{i1} and C_{i2}, are also moderators. With CA-RWR, however, one must consider the explicit role that C_{i1} and C_{i2} play in moderating the effects of treatment because this approach directly adjusts for these variables in the outcome model.

The second limitation of CA-RWR is the potentially high dimensionality of C_{i1} and C_{i2}—that is, the possibility that there are a large number of observed time-varying confounders. In many social science applications, the number of covariates in C_{i1} and C_{i2} is large, while the analysis is focused on only a limited set of putative moderators. As the number of covariates in C_{i1} and C_{i2} grows larger, it becomes more difficult to ensure that the models for the nuisance functions are correctly specified. Recall that each time-varying confounder may require up to four additional modeling considerations—all of them having to do with the nuisance functions, which are not of scientific interest. Finally, these two limitations are related in that the CA-RWR assumption stating that the effects of treatment are not moderated by the covariates in C_{i1} or C_{i2} becomes more and more untenable as the number of these covariates grows large.

IPTW-RWR overcomes these limitations by adjusting for time-varying confounders via weighting rather than via a modeling approach. Specifically, with IPTW-RWR, the following IPT weights are computed for each subject *i*:

$$w_{i1} = \frac{P(A_{i1} = a_{i1}|M_{i1})}{P(A_{i1} = a_{i1}|M_{i1}, C_{i1})} \text{ and} \tag{30}$$

$$w_{i2} = \frac{P(A_{i2} = a_{i1}|M_{i1}, A_{i1}, M_{i2})}{P(A_{i2} = a_{i1}|M_{i1}, C_{i1}, A_{i1}, C_{i2}, M_{i2})}. \tag{31}$$

The numerator of the weights is the conditional probability of treatment given prior moderators and treatments, while the denominator is the conditional probability of treatment given prior moderators, treatments, and confounders. At each time point, weighting by the ratio of these conditional probabilities balances prior time-varying confounders, but not prior moderators, across levels of future treatment. As in the point-in-time setting, the true IPT weights are unknown and must be estimated from data. This is typically accomplished by estimating the numerator and denominator using logistic regression models, but alternative methods are also available (e.g., McCaffrey et al. 2004).

After estimates of the IPT weights are computed, IPTW-RWR proceeds just as RWR but with weighted, rather than unweighted, regressions at each stage. Specifically, in the first stage, the estimated residuals $\hat{\delta}(M_{i1}) = M_{i1} - \hat{E}(M_{i1})$ and $\hat{\delta}(M_{i2}) = M_{i2} - \hat{E}(M_{i2}|M_{i1}, A_{i1})$ are obtained from weighted regressions for $E(M_{i1})$ and $E(M_{i2}|M_{i1}, A_{i1})$ with weights equal to $\widehat{w}_{i1}$ and $\widehat{w}_{i1} \times \widehat{w}_{i2}$, respectively. In the second stage, the SNMM is estimated via the same observed data regression used in Equation 26, but in this instance estimates are computed using weighted least squares with weights equal to $\widehat{w}_{i1} \times \widehat{w}_{i2}$.

IPTW-RWR requires the same sequential ignorability assumptions as CA-RWR and the same set of modeling assumptions as RWR (namely, correct models for the nuisance functions involving M_{i1} and M_{i2} and for the moderated intermediate causal effects). This approach additionally requires correctly specified models for the denominator probabilities in the IPT weights. Note, however, that this approach effectively replaces the four additional modeling considerations for each time-varying confounder in the CA-RWR approach with a single set of modeling considerations for the propensity score at each time point.

Figure 8 displays a stylized graph that provides some intuition as to how IPTW-RWR overcomes (1) the problems of over-control and collider stratification bias due to conditioning naively on time-varying moderators and (2) the problem of confounding by time-varying covariates. First, by residualizing M_{i2} based on the observed past, IPTW-RWR "purges" M_{i2} of its association with prior treatment while leaving the association between prior treatment and other variables intact. Second, by re-weighting the data based on the IPT at each time point, this approach eliminates the association between treatment and prior confounders while leaving intact the indirect effects of treatment that operate through future levels of the confounders.

Extensions for Multivalued Treatments, Multivalued Moderators, and Many Time Points

Thus far, we have focused largely on a saturated SNMM with a binary treatment and a binary moderator at 2 time points. Even in this simplest of scenarios, the SNMM is quite complex. With multivalued treatments, multivalued moderators, or many time points (i.e., $t > 2$), a saturated SNMM becomes intractable, and researchers have to explore simplifying functional form assumptions that reduce the number of free parameters in the model. These assumptions can be conceived of as additional parametric constraints imposed on the SNMM.

For example, consider a hypothetical scenario with a 3-level ordinal treatment, an interval-level moderator that takes on 10 different values, and 2 time points. In this scenario, a saturated SNMM would have $3^2 \times 10^2 = 900$ parameters! One way to simplify this model would be to assume that (1) the effect of treatment at each time point only varies across levels the moderator immediately preceding it; (2) the effect of treatment at each time point is linear within levels of the prior moderator; (3) a unit increase in the level of the prior moderator increments the effect of treatment by a constant amount; and (4) the associational effect of the moderator at each time point is also linear and does not depend on prior variables. Translating these assumptions into parametric constraints gives an unsaturated SNMM of the form

$$E\big(Y_i(a_1, a_2)|M_{i1}, M_{i2}(a_1)\big) = \beta_0 + \lambda_{10}\delta(M_{i1}) + a_1(\beta_{10} + \beta_{11}M_{i1}) + \lambda_{20}\delta\big(M_{i2}(a_1)\big) + a_2(\beta_{20} + \beta_{21}M_{i2}), \quad (32)$$

which uses only 7 parameters to summarize all 900 possible values that the conditional expectation can take on. Of course, many different types of constraints are possible, and their suitability in any given context will depend on the true data-generating process.

Unbiased estimation of moderated intermediate causal effects requires a correctly specified SNMM. If the simplifying assumptions imposed via parametric constraints on the

functional form of the model are incorrect, then estimates of these effects will be biased. Thus, in practice, researchers should experiment with a variety of different functional forms, investigate the sensitivity of causal effect estimates to these different specifications, clearly delineate the assumptions underlying the favored specification, and justify these assumptions based on substantive knowledge of the underlying data-generating process.

Variance Estimation

For all of the estimation approaches described previously, the standard errors (SEs) reported from over-the-counter software packages, such as Stata or R, are inappropriate because they assume that the residuals terms and, in the case of IPTW-RWR, the weights are known rather than estimated. Consequently, hypothesis tests and confidence intervals for the moderated causal effects of interest that are based on these SEs will be invalid. Almirall et al. (2010) derives asymptotic SEs that additionally account for sampling error in the estimation of the residuals using standard Taylor series arguments. However, because the programming needed to compute these SEs is highly complex, we propose the use of bootstrap estimates, which are easier to calculate using over-the-counter software (Efron and Tibshirani 1993). Simulation studies reveal a close correspondence between the bootstrap and asymptotic SEs in large sample applications. With smaller samples, simulations suggest that bootstrap SEs perform better than the asymptotic SEs, as expected. To obtain bootstrap estimates of the SEs, any of the estimation methods described previously are first applied to b samples of size N chosen at random (with replacement) from the original data. For each sample, parameter estimates for the moderated causal effects of interest are stored, and then the SEs are estimated using the standard deviation of these estimates across the b samples. The larger the number of samples, the more accurate are

estimates of the SEs. In practice, $b = 200$ provides a sufficient degree of accuracy (Efron and Tibshirani 1993).

EMPIRICAL EXAMPLE: NEIGHBORHOOD EFFECTS ON TEEN CHILDBEARING

This section presents an example application of SNMMs and RWR estimation that investigates whether the impact of concentrated neighborhood poverty on the risk of teen childbearing is moderated by prior family income. Neighborhood exposures and family income levels both vary over time (Quillian 2003; Timberlake 2007), and several competing theories suggest that family income moderates the impact of neighborhood poverty on the risk of teen childbearing. For example, compound disadvantage theory contends that family poverty intensifies the effects of neighborhood poverty because children from poor families must rely more heavily on neighborhood networks and institutional resources than children from nonpoor families (Jencks and Mayer 1990; Wilson 1987). By contrast, relative deprivation theory posits that the effects of poor neighborhoods are less severe among children in poor families because these children lack the family resources needed to capitalize on the advantages available in affluent neighborhoods (Jencks and Mayer 1990).

We investigate the impact of different longitudinal patterns of exposure to neighborhood poverty among subgroups of children defined by their time-varying family incomes using data from the PSID (Michigan Survey Research Center 2013). The PSID is a longitudinal study that began in 1968 with a national sample of about 4,800 households. From 1968 to 1997, the PSID interviewed household members annually; after 1997, interviews were conducted biennially. Families are matched to census tracts using the restricted-use PSID geocode file, and data on the socioeconomic composition of census tracts come from the Geolytics Neighborhood Change

Database (GeoLytics 2003).[3] The analytic sample for this study includes the 7,816 subjects in the PSID who were age 3 at any time between 1968 and 1986. Using all available data for these subjects between ages 3 and 14, measurements of neighborhood poverty and family-level covariates are constructed separately by developmental period, where the time index t is used to distinguish between measurements taken during childhood $(t = 1)$ and adolescence $(t = 2)$.

The treatment of interest in this analysis is exposure to different levels of neighborhood poverty. We construct a three-level ordinal treatment variable coded 0, 1, or 2 to indicate that a child lived in a low-poverty (<10%), moderate-poverty (10-20%), or a high-poverty (>20%) neighborhood, respectively. The childhood measurement of neighborhood poverty is based on a subject's average tract poverty rate over the three survey waves from age 6 to 8. Neighborhood poverty during adolescence is based on the average tract poverty rate over the three survey waves from age 12 to 14. We also construct separate multi-wave averages of time-varying covariates during childhood and adolescence. Time-varying covariates during childhood are based on averages taken over the years in which a subject is age 3 to 5—the three survey waves immediately preceding measurement of childhood treatment. Similarly, time-varying covariates during adolescence are based on averages over the years in which a subject is age 9 to 11—the three survey waves preceding measurement of adolescent neighborhood poverty and following measurement of neighborhood poverty during childhood. The end-of-study outcome of interest is a binary variable indicating whether a subject experienced a childbirth event between ages 15

[3] The data used in this analysis are derived from Sensitive Data Files of the PSID, obtained under special contractual arrangements designed to protect the anonymity of respondents. These data are not available from the authors. Persons interested in obtaining PSID Sensitive Data Files should contact PSIDHelp@isr.umich.edu.

and 20.[4] This measurement strategy, which is depicted graphically in Figure 9, ensures appropriate temporal ordering of the treatment, moderator, confounders, and outcome.

The time-varying moderator of interest in this analysis is the family income-to-needs ratio. This variable is equal to a family's annual real income from all sources divided by the official poverty threshold, which is indexed to family size. For ease of interpretation, the income-to-needs ratio is centered at the poverty line (i.e., it is equal to 0 for families with poverty-level incomes, 1 for families with incomes equivalent to twice the poverty line, and so on). The time-varying confounders included in this analysis are the family head's marital status (married versus not married), employment status (employed versus not employed), most recent occupation (professional or managerial occupation versus others), and homeownership status (homeowner versus renter). In addition, we control for a number of time-invariant confounders, including gender, race, birth year, mother's age and marital status at the time of a subject's birth, and the family head's highest level of completed education.[5]

We focus on estimating SNMMs of the form:

$$E\big(Y_i(a_1,a_2)|M_{i1},M_{i2}(a_1)\big) = \beta_0 + \lambda_{10}\delta(M_{i1}) + a_1(\beta_{10}+\beta_{11}M_{i1}) + \delta\big(M_{i2}(a_1)\big)\lambda_{20} + a_2\big(\beta_{20}+\beta_{21}M_{i2}(a_1)\big), \quad (33)$$

where a_t denotes exposure to different levels of neighborhood poverty; M_{it} denotes the family income-to-needs ratio; and $Y_i(a_1,a_2)$ is the potential outcome of interest coded 1 if a subject would have experienced a childbirth event between age 15 and 20 had they been exposed to the trajectory of neighborhood conditions (a_1,a_2), and 0 otherwise. This equation is a linear

[4] This analysis includes childbirth events for both male and female subjects in the PSID. Childbearing data for males is likely of poorer quality than that for females because males are simply not as accurate as females in their fertility reporting. Nevertheless, analyses stratified by gender yield results similar to those based on the pooled sample.
[5] Missing values are simulated for all variables using multiple imputation with 10 replications (Rubin 1987). Results are based on combined estimates and standard errors.

probability SNMM. In previous sections, we focused largely on saturated SNMMs that did not require assumptions about functional form. This model, however, assumes that the effects of neighborhood poverty on the risk of teen childbearing are approximately linear and are moderated only by the income-to-needs ratio measured during the same developmental period. It also assumes that the associational effect of the family income-to-needs ratio during adolescence does not depend on prior treatment or prior family income. Experimentation with a variety of more complex SNMMs suggests that this simplified specification accurately captures the moderated effects of interest.

In this model, the parameter β_{10} gives the average causal effect of childhood exposure to different levels of neighborhood poverty, setting adolescent treatment to low-poverty neighborhoods, among subjects in families with poverty-level incomes during childhood; β_{11} increments this effect for subjects in families with incomes above or below the poverty line. The parameter β_{20} gives the average causal effect of adolescent exposure to different levels of neighborhood poverty, holding neighborhood conditions during childhood constant, among subjects in families that would have poverty-level incomes during adolescence under the fixed childhood exposure; β_{21} increments this effect for subjects in families that with incomes above or below the poverty line at this development stage. If $\beta_{11} = \beta_{21} = 0$, then the family income-to-needs ratio does not moderate the impact of neighborhood poverty on the risk of teen childbearing.

We estimate these effects using the different variants of RWR described previously. First, we use unadjusted RWR. This approach involves residualizing the family income-to-needs ratio based on prior treatment and income-to-needs, and then regressing the indicator for teen childbearing on treatment, treatment by moderator interactions, and the residualized moderators.

It assumes that exposure to neighborhood poverty at each time period is not confounded by variables other than the family income-to-needs ratio and that the functional form of the SNMM is correctly specified.

Second, we use CA-RWR. This approach involves residualizing the family income-to-needs ratio and all measured confounders based on the observed past, and then regressing the indicator for teen childbearing on treatment, treatment by moderator interactions, the residualized moderators, and the residualized confounders. It assumes that exposure to neighborhood poverty is confounded only by the income-to-needs ratio and other measured covariates described previously; that the functional form of the SNMM is correctly specified; and relatedly, that the covariates treated solely as confounders are not also effect moderators.

Third, we use IPTW-RWR. This approach involves estimating IPT weights via an ordinal logistic regression model for the conditional probability of exposure to different levels of neighborhood poverty at each time period given prior treatment, moderators, and confounders. Then, the family income-to-needs ratio is residualized based on prior treatment and prior measures of this moderator using an IPT-weighted regression. Finally, estimates of moderated causal effects are obtained by fitting an IPT-weighted regression of the indicator for teen childbearing on treatment, treatment by moderator interactions, and the residualized moderators. This approach assumes that exposure to neighborhood poverty is confounded only by the income-to-needs ratio and other observed covariates; that the ordinal logistic regression models for the conditional probability of treatment at each time period are correctly specified; and that the functional form of the SNMM is correctly specified.

Table 1 presents point estimates and bootstrap standard errors for the SNMM causal parameters. The first column of the table presents estimates based on unadjusted RWR. The

second and third columns present estimates from CA-RWR and IPTW-RWR, respectively. All three estimation strategies yield point estimates for the direct effect of childhood exposure to neighborhood poverty that are substantively small and do not reach conventional significance thresholds. In general, results suggest a negligible direct impact of exposure to neighborhood poverty during childhood on the risk of teen childbearing and provide no evidence of effect moderation by prior family income. These findings are not simply due to over-control or collider stratification biases, as the RWR approach avoids them.

Estimates for the effect of exposure to neighborhood poverty during adolescence, by contrast, indicate that living in moderate- or high-poverty neighborhoods during this developmental period has a strong and statistically significant positive effect on the risk of teen childbearing. Moreover, estimates also indicate that this effect is significantly moderated by prior family income levels. Consistent with compound disadvantage theory, exposure to neighborhood poverty during adolescence is estimated to have a larger inflationary effect on the risk of teen childbearing among individuals whose families are also poor during this developmental period.

For example, according to estimates based on IPTW-RWR, adolescent exposure to high-poverty rather than low-poverty neighborhoods is estimated to increase the risk of a subsequent childbirth event by about 9 percentage points among individuals in families with incomes at the poverty line during adolescence (i.e., $2\left(\hat{\beta}_{20} + \hat{\beta}_{21}(0)\right) = 2(0.046) = 0.092$). Among individuals in nonpoor families with incomes equivalent to three times the poverty line, estimates indicate that exposure to high-poverty rather than low-poverty neighborhoods during adolescence increases the risk of a subsequent childbirth event by just over 3 percentage points (i.e., $2\left(\hat{\beta}_{20} + \hat{\beta}_{21}(2)\right) = 2\big(0.046 - 0.014(2)\big) = 0.036$). In other words, the inflationary impact of adolescent exposure to neighborhood poverty on the risk of subsequent childbearing is

about twice as strong for individuals in poor families than for individuals in nonpoor families. Effect estimates based on CA-RWR versus IPTW-RWR are comparable to each other but smaller than estimates based on unadjusted RWR. This indicates that unadjusted RWR overstates the effects of neighborhood poverty because it ignores confounding by family-level characteristics, such as parental marital status and education.

In sum, results from this example application of SNMMs and RWR estimation indicate that exposure to poor neighborhoods, particularly during adolescence, has a strong positive effect on the risk of teen childbearing, and that this effect is much more pronounced for individuals in poor families. RWR estimation is premised on the strong assumptions of no unobserved confounding and correct model specification, but these assumptions are in fact much weaker than those required by other methods that might naively be used to investigate neighborhood effect heterogeneity in the longitudinal setting, such as conventional regression or propensity score stratification. To further investigate the sensitivity of results to potential violations of these key assumptions, a variety of robustness checks are available and have been implemented in other settings (e.g., Brumback et al. 2004; Sharkey and Elwert 2010; Wodtke et al. 2012).

DISCUSSION

Treatment effect heterogeneity is ubiquitous in the social sciences. In many situations, both the treatment and effect moderators of interest vary over time, and they may influence one another through a dynamic selection and feedback process. This article introduced to sociology a new class of models and estimators for analyzing moderated causal effects in the longitudinal setting: SNMMs and RWR. It outlined how these methods avoid the limitations associated with

conventional methods when time-varying moderators are affected by prior treatments, and it adapted them to account for observed confounding.

To illustrate these methods, we presented a simple empirical application with longitudinal data from the PSID. This analysis investigated whether the effects of exposure to neighborhood poverty during childhood versus adolescence on the risk of teen childbearing are moderated by prior family income levels. Results indicate that exposure to neighborhood poverty during adolescence (but not during childhood) increases the risk of subsequent childbearing, especially for individuals whose families are also poor during adolescence. This example application demonstrates the utility of these methods for neighborhood-effects research, and given the growing prevalence of longitudinal data in the social sciences, SNMMs and RWR estimation should be even more widely applicable, wherever there is interest in understanding heterogeneous effects of time-varying treatments.

For expositional simplicity, this article focused on SNMMs with a terminal end-of-study outcome. These methods, however, can be adapted to investigate time-varying outcomes. Moreover, they can even be adapted to investigate whether treatment effects on future values of a time-varying outcome are moderated by past values of this outcome. For example, these methods could be used to investigate whether the impact of continuing workforce education on subsequent earnings is moderated by a worker's prior earnings history. They could also be used to investigate whether the impact on academic achievement of an ongoing instructional intervention is moderated by a student's prior achievements. This type of information would allow social policymakers to develop adaptive interventions that tailor treatments across time to the evolving needs of different individuals in heterogeneous target populations, and it would

allow researchers to better understand the dynamic etiology of labor earnings and academic achievement.

Although this study focused on RWR estimation, another approach—termed G-estimation in the literature on causal inference—can also be used to consistently estimate moderated intermediate causal effects under a similar set of assumptions (Robins 1994). The G-estimator involves solving a complex system of estimating equations that do not require correctly specified models for the nuisance functions in the SNMM. RWR is our preferred estimation strategy because of its simplicity, transparency, and similarity to regression methods that are already familiar to sociologists; because of its greater relative efficiency (Almirall et al. 2010); and because it can be easily implemented with off-the-shelf software. Parts A and B of the Online Supplement provide code for the Stata and R statistical packages that executes the different RWR estimators and computes bootstrap standard errors with simulated data from a simple two time period example.[6]

Despite the simplicity, convenience, and greater efficiency of RWR estimation, the G-estimator is not without its own advantages. In particular, with an unsaturated SNMM, the G-estimator is unbiased for the causal parameters of interest under weaker assumptions than RWR. The G-estimator provides unbiased estimates of the causal functions in an unsaturated SNMM if *either* the models for the nuisance functions are correctly specified *or* models for the conditional probability of treatment are correctly specified. This double-robustness property of the G-estimator provides a degree of protection against bias due to model misspecification, but it comes at the price of higher variance. In practice, researchers may want to consider

[6] In addition, an R function that executes RWR and computes asymptotic SEs in a more general setting is available at http://methcenter.psu.edu.

implementing both G-estimation and RWR estimation in an attempt to balance concerns about misspecification and precision.

Empirical researchers interested in effect heterogeneity are most often concerned with what is widely thought to be the main challenge for drawing valid causal inferences in the social sciences: unobserved confounding of treatment. This concern is certainly not misplaced, and methods for assessing the robustness of findings to hypothetical patterns of unobserved confounding should be incorporated in empirical analyses much more frequently (Brumback et al. 2004). However, we also show that even in studies where there is no unobserved confounding of treatment, conventional methods for analyzing effect heterogeneity remain biased if they condition on time-varying moderators (or confounders) that are affected by past levels of a time-varying treatment. Although biases due to conditioning on an outcome of treatment are often overlooked or discounted in sociological research, their magnitude can be substantively large and in some cases even greater than confounding bias (Elwert and Winship Forthcoming; Greenland 2003). Thus, it is critically important to have flexible statistical tools, like SNMMs and RWR estimation, which are capable of accounting for the variety of different biases encountered in longitudinal research on effect heterogeneity.

REFERENCES

Almirall, Daniel, Beth A. Griffin, Daniel F. McCaffrey, Rajeev Ramchand, Robert A. Yuen, and Susan A. Murphy. 2014. "Time-varying Effect Moderation using the Structural Nested Mean Model: Estimation using Inverse-weighted Regression with Residuals." *Statistics in Medicine* 33:3466-87.

Almirall, Daniel, Scott N. Compton, Meredith Gunlicks-Stoessel, Naihua Duan, and Susan A. Murphy. 2012. "Preparing for a Sequential Multiple Assignment Randomized Trial for Developing an Adaptive Treatment Strategy: Designing a SMART Pilot Study." *Statistics in Medicine* 31:1887-902.

Almirall, Daniel, Daniel F. McCaffrey, Rajeev Ramchand, and Susan A. Murphy. 2013. "Subgroups Analysis when Treatment and Moderators are Time-varying." *Prevention Science* 14:169-78.

Almirall, Daniel, Thomas Ten Have, and Susan A. Murphy. 2010. "Structural Nested Mean Models for Assessing Time-Varying Effect Moderation." *Biometrics* 66:131-9.

Amato, Paul R. 2004. "The Consequences of Divorce for Adults and Children." *Journal of Marriage and Family* 62:1269-87.

Brand, Jennie E. and Yu Xie. 2010. "Who Benefits Most from College? Evidence for Negative Selection in Heterogenous Economic Returns to Higher Education." *American Sociological Review* 75:273-302.

Brooks-Gunn, Jeanne, Greg J. Duncan, Pamela K. Klebanov, and Naomi Sealand. 1993. "Do Neighborhoods Influence Child and Adolescent Development?" *American Journal of Sociology* 99:353-95.

Brumback, Babette A., Miguel A. Hernan, Sebastien J. P. A. Haneuse, and James M. Robins. 2004. "Sensitivity Analyses for Unmeasured Confounding Assuming a Marginal Structural Model for Repeated Measures." *Statistics in Medicine* 23:749-67.

Efron, Bradley and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Elwert, Felix. 2013. "Graphical Causal Models." Pp. 245-71 in *Handbook of Causal Analysis for Social Research,* edited by S. L. Morgan. New York: Springer.

Elwert, Felix and Christopher Winship. Forthcoming. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology*.

GeoLytics, Inc. 2003. *CensusCD Neighborhood Change Database, 1970-2000 Tract Data*. New Brunswick, NJ: GeoLytics.

Greenland, Sander. 2003. "Quantifying Biases in Causal Models: Classical Confounding vs Collider-stratification Bias." *Epidemiology* 14:300-6.

Heckman, James J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science* 312:1900-1.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Review of Economics and Statistics* 88:389-432.

Hirano, K. and G. Imbens. 2001. "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." *Health Services and Outcomes Research* 2:259-78.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945-60.

Jencks, Christopher and Susan E. Mayer. 1990. "The Social Consequences of Growing Up in a Poor Neighborhood." Pp. 111-86 in *Inner-City Poverty in the United States,* edited by L. E. Lynn and M. G. H. McGreary. Washington, D.C.: National Academy Press.

Manski, Charles. 2007. *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral. 2004. “Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies.” *Psychological Methods* 9:403-25.

Michigan Survey Research Center. 2013. "Panel Study of Income Dynamics, Restricted Use Data." Ann Arbor, MI: University of Michigan.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Parcel, Toby L. and Mikaela Dufur. 2001. "Capital at Home and at School: Effects on Student Achievement." *Social Forces* 79:881-911.

Quillian, Lincoln. 2003. "How Long Are Exposures to Poor Neighborhoods? The Long-Term Dynamics of Entry and Exit from Poor Neighborhoods." *Population Research and Policy Review* 22:221-49.

Robins, J. M., M. A. Hernan, and A. Rotnitzky. 2007. "Effect Modification by Time-Varying Covariates." *American Journal of Epidemiology* 166:994-1002.

Robins, James M. 1994. "Correcting for Noncompliance in Randomized Trials Using Structural Nested Mean Models." *Communications in Statistics-Theory and Methods* 23:2379-412.

Robins, James M., Miguel A. Hernan, and Babette A. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11:550-60.

Rosenbaum, P. and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41-55.

Rothman, Kenneth J., S. Greenland, and Alexander M. Walker. 1980. "Concepts of Interaction." *American Journal of Epidemiology* 112:467-70.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688-701.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons.

Sanders, William L., S. Paul Wright, and Sandra P. Horn. 1997. "Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation." *Journal of Personnel Evaluation in Education* 11:57-67.

Sharkey, P. and F. Elwert. 2010. "The Legacy of Disadvantage: Multigenerational Neighborhood Effects on Cognitive Ability." *CDE WP 2010-06. Center for Demography and Ecology, University of Wisconsin-Madison*.

South, Scott J. and Kyle D. Crowder. 1997. "Escaping Distressed Neighborhoods: Individual, Community, and Metropolitan Influences." *American Journal of Sociology* 102:1040-84.

Timberlake, Jeffrey M. 2007. "Racial and Ethnic Inequality in the Duration of Children's Exposure to Neighborhood Poverty and Affluence." *Social Problems* 54:319-42.

VanderWeele, Tyler J. and James M. Robins. 2007a. "Directed Acyclic Graphs, Sufficient Causes and the Properties of Conditioning on a Common Effect." *American Journal of Epidemiology* 166:1096-104.

VanderWeele, Tyler J. and James M. Robins. 2007b. "Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs." *Epidemiology* 18:561-68.

Wallerstein, Judith S. 1991. "The Long-term Effects of Divorce on Children: A Review." *Journal of the American Academy of Child and Adolescent Psychiatry* 30:349-60.

Wilson, William J. 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Chicago: University of Chicago Press.

Wilson, William J. 1996. *When Work Disappears: The World of the New Urban Poor*. New York: Vintage Books.

Wodtke, Geoffrey T., Felix Elwert, and David J. Harding. 2012. "Poor Families, Poor Neighborhoods: How Family Poverty Intensifies the Impact of Concentrated Disadvantage on High School Graduation." *PSC Research Report No. 12-776.*

Wodtke, Geoffrey T., David J. Harding, and Felix Elwert. 2011. "Neighborhood Effects in Temporal Perspective: The Impact of Long-term Exposure to Concentrated Disadvantage on High School Graduation." *American Sociological Review* 76:713-36.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Xie, Yu. 2007. "Otis Dudley Duncan's Legacy: The Demographic Approach to Quantitative Reasoning in Social Science." *Research in Social Stratification and Mobility* 25:141-56.

Xie, Yu, Jennie E. Brand, and B. Jann. 2012. "Estimating Heterogeneous Treatment Effects with Observational Data." *Sociological Methodology* 42:314-47.

TABLES

Table 1. Moderated effects of neighborhood poverty on the risk of adolescent parenthood by family income

Specification	Unadjusted RWR			Adjusted RWR			IPT-weighted RWR		
	est	se		est	se		est	se	
Intercept	.109	(.008)	***	.138	(.010)	***	.139	(.010)	***
Childhood									
NH pov	.014	(.012)		–.002	(.012)		–.003	(.013)	
Inc x NH pov	–.005	(.005)		–.001	(.006)		.003	(.006)	
Adolescence									
NH pov	.063	(.013)	***	.047	(.013)	***	.046	(.014)	**
Inc x NH pov	–.016	(.005)	**	–.013	(.006)	*	–.014	(.005)	**

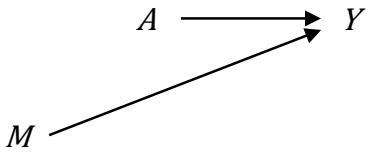
Notes: Sample includes children present in a PSID family at age 3 during the 1968-1986 waves. Results are combined estimates from 10 multiple imputation datasets. Standard errors are computed from 200 bootstrap samples.

*p < 0.05, **p < 0.01, and ***p < 0.001 for two-sided tests of no effect.

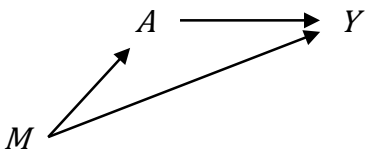
FIGURES

Figure 1. Point-in-time causal relationships between treatment, moderator, and outcome

A. Random assignment



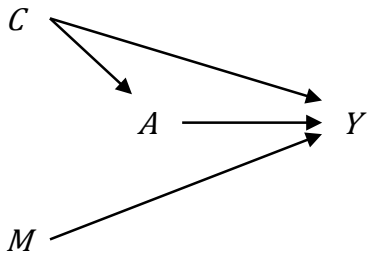
B. Confounding only by M



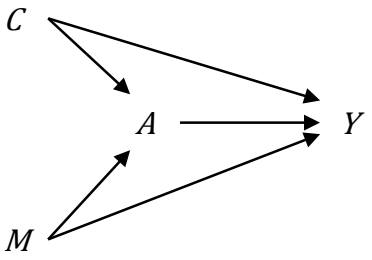
Notes: A = treatment, M = pre-treatment moderator, and Y = end-of-study outcome.

Figure 2. Point-in-time causal relationships between treatment, moderator, confounder, and outcome

A. Confounding only by C



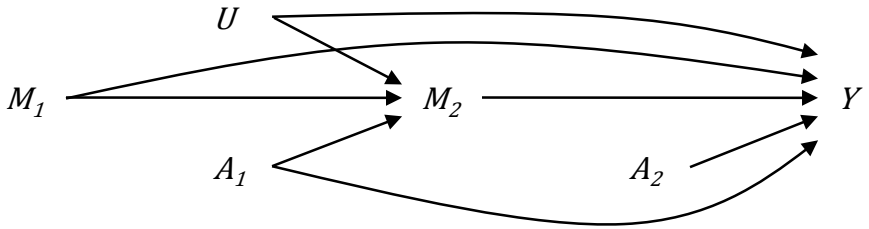
B. Confounding by M and C



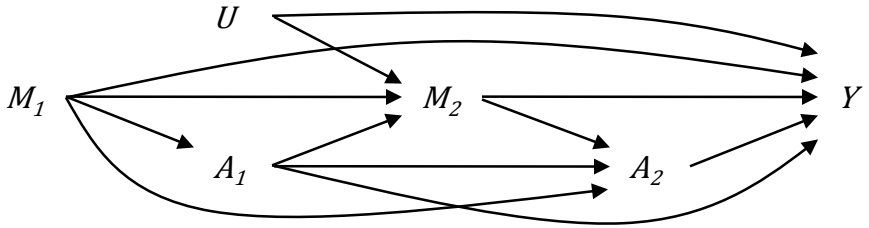
Notes: A = treatment, M = pre-treatment moderator, C = pre-treatment confounder, and Y = end-of-study outcome.

Figure 3. Causal relationships between time-varying treatments, moderators, and outcome

A. Sequential randomization



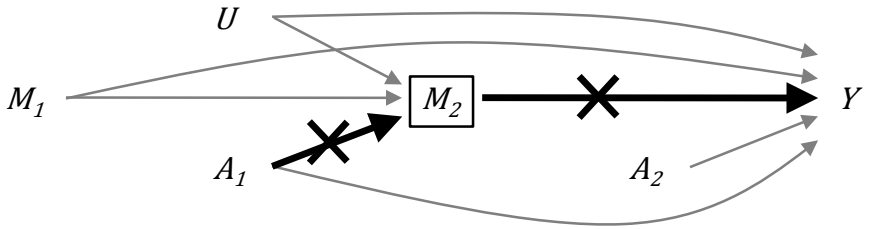
B. Selection on prior moderators



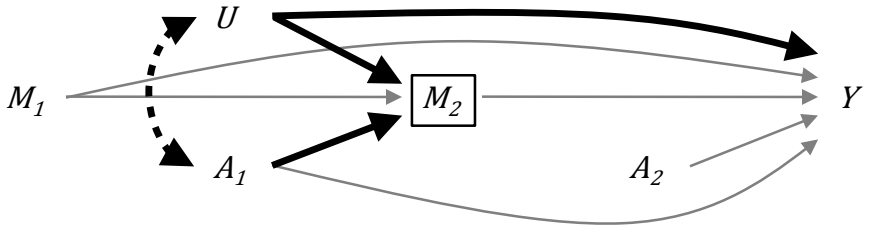
Notes: A_t = treatment, M_t = pre-treatment moderator, Y = end-of-study outcome, and U = unobserved factors.

Figure 4. Over-control of intermediate pathways and collider-stratification biases

A. Over-control of intermediate pathways



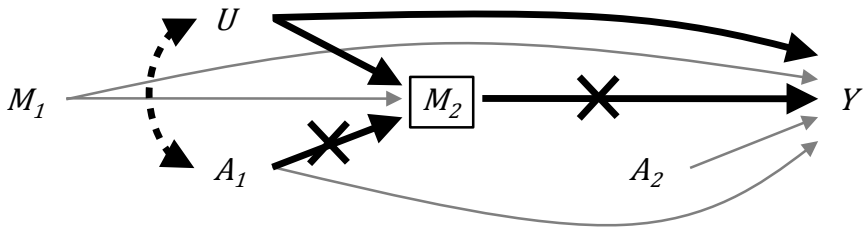
B. Collider-stratification



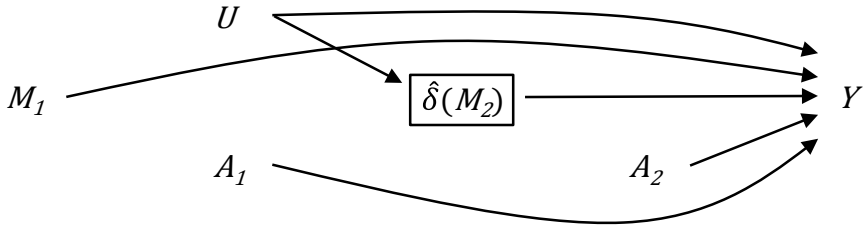
Notes: A_t = treatment, M_t = pre-treatment moderator, Y = end-of-study outcome, and U = unobserved factors. A box around a variable denotes conditioning.

Figure 5. Consequences of residualizing time-varying moderators based on past treatment and covariates

A. Condition on M_2: over-control and collider stratification



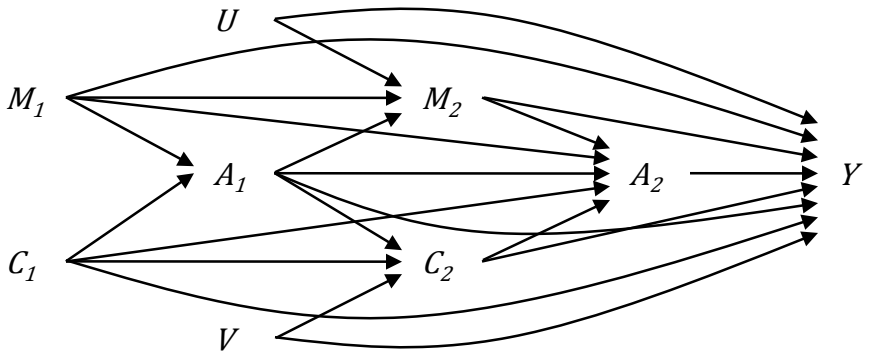
B. Condition on residualized M_2: transformed moderator independent of prior treatment and no bias



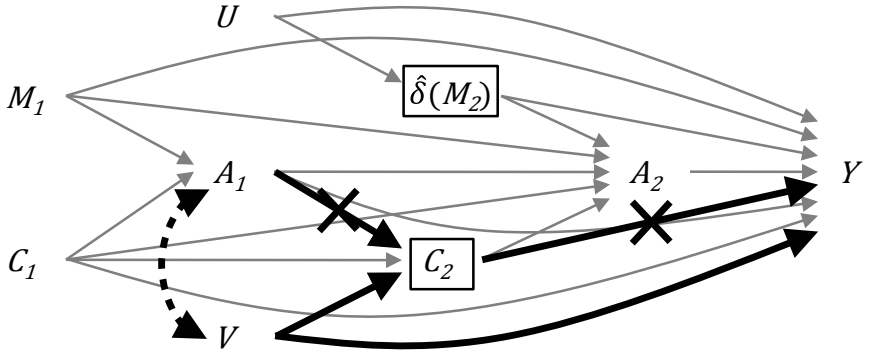
Notes: A_t = treatment, M_t = pre-treatment moderator, Y = end-of-study outcome, and U = unobserved factors. A box around a variable denotes conditioning. $\hat{\delta}(M_2)$ is equal to $M_2 - \hat{E}(M_2|M_1, A_1)$.

Figure 6. Causal relationships between time-varying treatments, moderators, confounders, and outcome

A. Selection on prior moderators and confounders



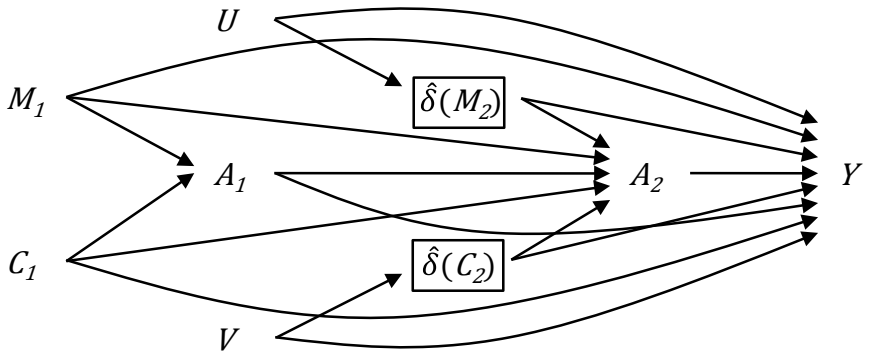
B. Over-control and collider stratification from conditioning on time-varying confounders



Notes: A_t = treatment, M_t = pre-treatment moderator, C_t = pre-treatment confounder, Y = end-of-study outcome, and U and V both represent unobserved factors. A box around a variable denotes conditioning. $\hat{\delta}(M_2)$ is equal to $M_2 - \hat{E}(M_2|M_1, A_1)$.

Figure 7. Consequences of residualizing both time-varying confounders and moderators based on past treatment and covariates

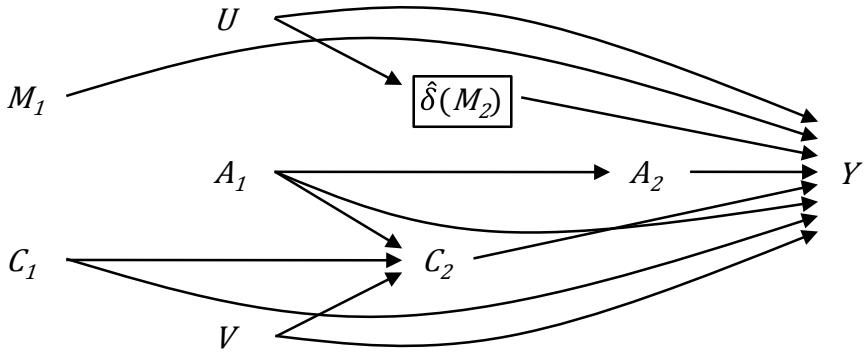
A. Condition on residualized C_2 and M_2: transformed confounder and moderator independent of prior treatment and no bias



Notes: A_t = treatment, M_t = pre-treatment moderator, C_t = pre-treatment confounder, Y = end-of-study outcome, and U and V both represent unobserved factors. A box around a variable denotes conditioning. $\hat{\delta}(M_2)$ is equal to $M_2 - \hat{E}(M_2|M_1, C_1, A_1)$ and $\hat{\delta}(C_2)$ is equal to $C_2 - \hat{E}(C_2|C_1, M_1, A_1)$.

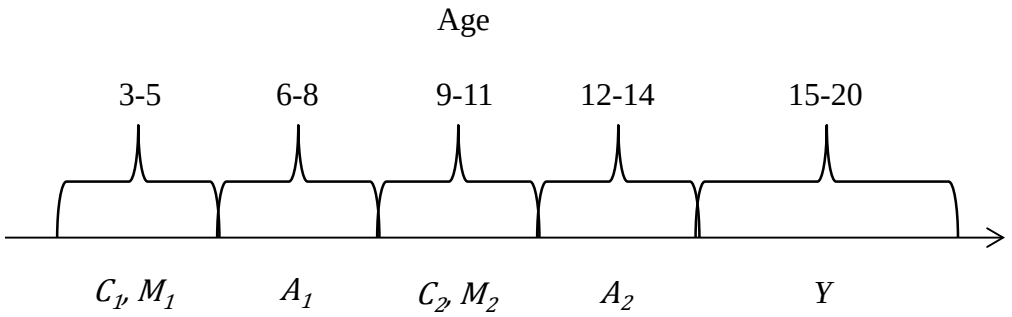
Figure 8. Consequences of weighting by the inverse probability of treatment (IPT) and residualizing time-varying moderators in the weighted pseudo-population

A. Weight by IPT and condition on residualized M_2: future treatment independent of past moderators and confounders, transformed moderator independent of prior treatment in weighted pseudo-population, and no bias



Notes: A_t = treatment, M_t = pre-treatment moderator, C_t = pre-treatment confounder, Y = end-of-study outcome, and U and V both represent unobserved factors. A box around a variable denotes conditioning. $\hat{\delta}(M_2)$ is equal to $M_2 - \hat{E}(M_2|M_1, C_1, A_1)$.

Figure 9. Longitudinal measurement strategy in the PSID



Notes: A_t = neighborhood poverty, M_t = family income-to-needs ratio, C_t = vector of observed confounders, and Y = childbirth event.

ONLINE SUPPLEMENT

Part A: Example STATA Code for Estimating a SNMM using RWR

```
#delimit ;
/***SIMULATE EXAMPLE DATA W/ TWO TIME PERIODS***/
/*NOTE: u and v are binary unobserved variables*/
/*NOTE: c1 and c2 are binary time-varying confounders, and c2 is affected by prior treatment*/
/*NOTE: m1 and m2 are binary time-varying moderators, and m2 is affected by prior treatment*/
/*NOTE: a1 and a2 are binary time-varying treatments*/
/*NOTE: y is a normally distributed end-of-study outcome*/
/*NOTE: effect of a1 is moderated only by m1, and effect of a2 is moderated only by m2*/
set obs 50000 ;
gen u=0.5>=uniform() ;
gen v=0.5>=uniform() ;
gen c1=0.5>=uniform() ;
gen m1=0.5>=uniform() ;
gen a1=0.3+0.2*c1+0.2*m1>=uniform() ;
gen c2=0.2+0.2*c1+0.2*a1+0.2*v>=uniform() ;
gen m2=0.2+0.2*m1+0.2*a1+0.2*u>=uniform() ;
gen a2=0.2+0.2*a1+0.2*c2+0.2*m2>=uniform() ;
gen y=(0+1*u+1*v)+1*(c1-0.5)+1.0*(m1-0.5)+a1*(1+1*m1)+
1*(c2-(0.2+0.2*c1+0.2*a1))+1*(m2-(0.2+0.2*m1+0.2*a1))+a2*(1+1*m2)+
3*invnorm(uniform()) ;

/***ESTIMATE (WITH OVER-CONTROL AND COLLIDER-STRATIFICATION BIAS) SNMM VIA CONVENTIONAL
REGRESSION***/
gen m1a1=m1*a1 ;
gen m2a2=m2*a2 ;
reg y c1 m1 a1 m1a1 c2 m2 a2 m2a2 ;
drop m1a1 m2a2 ;
```

```
/***ESTIMATE (WITH CONFOUNDING BIAS) SNMM VIA UNADJUSTED REGRESSION-WITH-RESIDUALS***/
/*FIRST STAGE REGRESSIONS*/
reg m1 ;
predict m1r, resid ;
reg m2 m1 a1 ;
predict m2r, resid ;
/*SECOND STAGE REGRESSION*/
gen m1a1=m1*a1 ;
gen m2a2=m2*a2 ;
reg y m1r a1 m1a1 m2r a2 m2a2 ;
drop m1a1 m2a2 m1r m2r ;

/***ESTIMATE (WITHOUT BIAS) SNMM VIA COVARIATE-ADJUSTED REGRESSION-WITH-RESIDUALS***/
/*FIRST STAGE REGRESSIONS*/
reg c1 ;
predict c1r, resid ;
reg m1 ;
predict m1r, resid ;
reg c2 c1 a1 ;
predict c2r, resid ;
reg m2 m1 a1 ;
predict m2r, resid ;
/*SECOND STAGE REGRESSION*/
gen m1a1=m1*a1 ;
gen m2a2=m2*a2 ;
reg y c1r m1r a1 m1a1 c2r m2r a2 m2a2 ;
drop m1a1 m2a2 c1r c2r m1r m2r ;

/***ESTIMATE (WITHOUT BIAS) SNMM VIA IPT-WEIGHTED REGRESSION-WITH-RESIDUALS***/
/*STABILIZED IPTWs*/
reg a1 ;
predict p1, xb ;
```

```
replace p1=a1*p1+(1-a1)*(1-p1) ;
reg a1 c1 m1 ;
predict p1x, xb ;
replace p1x=a1*p1x+(1-a1)*(1-p1x) ;
reg a2 ;
predict p2, xb ;
replace p2=a2*p2+(1-a2)*(1-p2) ;
reg a2 a1 c2 m2 ;
predict p2x, xb ;
replace p2x=a2*p2x+(1-a2)*(1-p2x) ;
gen iptw=(p1/p1x)*(p2/p2x) ;
/*WEIGHTED FIRST STAGE REGRESSIONS*/
reg m1 [pw=iptw] ;
predict m1r, resid ;
reg m2 m1 a1 [pw=iptw] ;
predict m2r, resid ;
/*WEIGHTED SECOND STAGE REGRESSION*/
gen m1a1=m1*a1 ;
gen m2a2=m2*a2 ;
reg y m1r a1 m1a1 m2r a2 m2a2 [pw=iptw] ;
drop iptw p1 p1x p2 p2x m1a1 m2a2 m1r m2r ;

/***COMPUTE BOOTSTRAP STANDARD ERRORS FOR CAUSAL PARAMETERS***/
program define covadj_rwr, rclass ;
     reg c1 ;
     predict c1r, resid ;
     reg m1 ;
     predict m1r, resid ;
     reg c2 c1 a1 ;
     predict c2r, resid ;
     reg m2 m1 a1 ;
     predict m2r, resid ;
```

```
      gen m1a1=m1*a1 ;
      gen m2a2=m2*a2 ;
      reg y c1r m1r a1 m1a1 c2r m2r a2 m2a2 ;
      return scalar b0=_b[_cons] ;
      return scalar b1=_b[a1] ;
      return scalar b2=_b[m1a1] ;
      return scalar b3=_b[a2] ;
      return scalar b4=_b[m2a2] ;
      drop m1a1 m2a2 c1r c2r m1r m2r ;
end ;
program define iptw_rwr, rclass ;
      reg a1 ;
      predict p1, xb ;
      replace p1=a1*p1+(1-a1)*(1-p1) ;
      reg a1 c1 m1 ;
      predict p1x, xb ;
      replace p1x=a1*p1x+(1-a1)*(1-p1x) ;
      reg a2 ;
      predict p2, xb ;
      replace p2=a2*p2+(1-a2)*(1-p2) ;
      reg a2 a1 c2 m2 ;
      predict p2x, xb ;
      replace p2x=a2*p2x+(1-a2)*(1-p2x) ;
      gen iptw=(p1/p1x)*(p2/p2x) ;
      reg m1 [pw=iptw] ;
      predict m1r, resid ;
      reg m2 m1 a1 [pw=iptw] ;
      predict m2r, resid ;
      gen m1a1=m1*a1 ;
      gen m2a2=m2*a2 ;
      reg y m1r a1 m1a1 m2r a2 m2a2 [pw=iptw] ;
      return scalar b0=_b[_cons] ;
```

```
      return scalar b1=_b[a1] ;
      return scalar b2=_b[m1a1] ;
      return scalar b3=_b[a2] ;
      return scalar b4=_b[m2a2] ;
      drop iptw p1 p1x p2 p2x m1a1 m2a2 m1r m2r ;
end ;
bootstrap beta0=r(b0) beta1=r(b1) beta2=r(b2) beta3=r(b3) beta4=r(b4), reps(100): covadj_rwr ;
bootstrap beta0=r(b0) beta1=r(b1) beta2=r(b2) beta3=r(b3) beta4=r(b4), reps(100): iptw_rwr ;
```

Part B: Example R Code for Estimating a SNMM using RWR

```
### SIMULATE EXAMPLE DATA W/ TWO TIME PERIODS
# NOTE: u and v are binary unobserved variables
# NOTE: c1 and c2 are binary time-varying confounders, and c2 is affected by prior treatment
# NOTE: m1 and m2 are binary time-varying moderators, and m2 is affected by prior treatment
# NOTE: a1 and a2 are binary time-varying treatments
# NOTE: y is a normally distributed end-of-study outcome
# NOTE: effect of a1 is moderated only by m1, and effect of a2 is moderated only by m2
u<-rbinom(50000,1,0.5)
v<-rbinom(50000,1,0.5)
c1<-rbinom(50000,1,0.5)
m1<-rbinom(50000,1,0.5)
a1<-rbinom(50000,1,0.3+0.2*c1+0.2*m1)
c2<-rbinom(50000,1,0.2+0.2*c1+0.2*a1+0.2*v)
m2<-rbinom(50000,1,0.2+0.2*m1+0.2*a1+0.2*u)
a2<-rbinom(50000,1,0.2+0.2*a1+0.2*c2+0.2*m2)
y<-rnorm(50000,(0+1*u+1*v)+1*(c1-0.5)+1*(m1-0.5)+a1*(1+1*m1)+1*(c2-(0.2+0.2*c1+0.2*a1))+1*(m2-
(0.2+0.2*m1+0.2*a1))+a2*(1+1*m2),3)

### ESTIMATE (WITH OVER-CONTROL AND COLLIDER-STRATIFICATION BIAS) SNMM VIA CONVENTIONAL REGRESSION
m1a1<-m1*a1
m2a2<-m2*a2
model1<-lm(y~c1+m1+a1+m1a1+c2+m2+a2+m2a2)
summary(model1)
rm(list=c('m1a1','m2a2','model1'))

### ESTIMATE (WITH CONFOUNDING BIAS) SNMM VIA UNADJUSTED REGRESSION-WITH-RESIDUALS
# FIRST STAGE REGRESSIONS
model1<-lm(m1~1)
m1r<-model1$residuals
model2<-lm(m2~m1+a1)
```

```
m2r<-model2$residuals
# SECOND STAGE REGRESSION
m1a1<-m1*a1
m2a2<-m2*a2
model3<-lm(y~m1r+a1+m1a1+m2r+a2+m2a2)
summary(model3)
rm(list=c('m1a1','m2a2','m1r','m2r','model1','model2','model3'))

### ESTIMATE (WITHOUT BIAS) SNMM VIA COVARIATE-ADJUSTED REGRESSION-WITH-RESIDUALS
# FIRST STAGE REGRESSIONS
model1<-lm(c1~1)
c1r<-model1$residuals
model2<-lm(m1~1)
m1r<-model2$residuals
model3<-lm(c2~c1+a1)
c2r<-model3$residuals
model4<-lm(m2~m1+a1)
m2r<-model4$residuals
# SECOND STAGE REGRESSION
m1a1<-m1*a1
m2a2<-m2*a2
model5<-lm(y~c1r+m1r+a1+m1a1+c2r+m2r+a2+m2a2)
summary(model5)
rm(list=c('m1a1','m2a2','m1r','m2r','c1r','c2r','model1','model2','model3','model4','model5'))

### ESTIMATE (WITHOUT BIAS) SNMM VIA IPT-WEIGHTED REGRESSION-WITH-RESIDUALS
# STABILIZED IPTWs
model1<-lm(a1~1)
p1<-model1$fitted.values
p1<-a1*p1+(1-a1)*(1-p1)
model2<-lm(a1~c1+m1)
p1x<-model2$fitted.values
```

```
p1x<-a1*p1x+(1-a1)*(1-p1x)
model3<-lm(a2~1)
p2<-model3$fitted.values
p2<-a2*p2+(1-a2)*(1-p2)
model4<-lm(a2~a1+c2+m2)
p2x<-model4$fitted.values
p2x<-a2*p2x+(1-a2)*(1-p2x)
iptw<-(p1/p1x)*(p2/p2x)
# WEIGHTED FIRST STAGE REGRESSIONS
model5<-lm(m1~1,weights=iptw)
m1r<-model5$residuals
model6<-lm(m2~m1+a1,weights=iptw)
m2r<-model6$residuals
# WEIGHTED SECOND STAGE REGRESSION
m1a1<-m1*a1
m2a2<-m2*a2
model7<-lm(y~m1r+a1+m1a1+m2r+a2+m2a2,weights=iptw)
summary(model7)
rm(list=c('m1a1','m2a2','m1r','m2r','p1','p1x','p2','p2x','iptw','model1','model2','model3','model
4','model5','model6','model7'))

### COMPUTE BOOTSTRAP STANDARD ERRORS FOR CAUSAL PARAMETERS
dataset<-data.frame(cbind(u,v,c1,m1,a1,c2,m2,a2,y))
covadj_rwr<-matrix(data=NA,nrow=100,ncol=5)
iptw_rwr<-matrix(data=NA,nrow=100,ncol=5)
for (j in 1:100)
      {
      bid<-sample(nrow(dataset),replace=T)
      bsamp<-dataset[bid,]
      model1<-lm(c1~1,data=bsamp)
      bsamp$c1r<-model1$residuals
      model2<-lm(m1~1,data=bsamp)
```

```
bsamp$m1r<-model2$residuals
model3<-lm(c2~c1+a1,data=bsamp)
bsamp$c2r<-model3$residuals
model4<-lm(m2~m1+a1,data=bsamp)
bsamp$m2r<-model4$residuals
bsamp$m1a1<-bsamp$m1*bsamp$a1
bsamp$m2a2<-bsamp$m2*bsamp$a2
model5<-lm(y~a1+m1a1+a2+m2a2+c1r+m1r+c2r+m2r,data=bsamp)
for (i in 1:5)
     {
     covadj_rwr[j,i]<-model5$coef[i]
     }
model1<-lm(a1~1,data=bsamp)
bsamp$p1<-model1$fitted.values
bsamp$p1<-bsamp$a1*bsamp$p1+(1-bsamp$a1)*(1-bsamp$p1)
model2<-lm(a1~c1+m1,data=bsamp)
bsamp$p1x<-model2$fitted.values
bsamp$p1x<-bsamp$a1*bsamp$p1x+(1-bsamp$a1)*(1-bsamp$p1x)
model3<-lm(a2~1,data=bsamp)
bsamp$p2<-model3$fitted.values
bsamp$p2<-bsamp$a2*bsamp$p2+(1-bsamp$a2)*(1-bsamp$p2)
model4<-lm(a2~a1+c2+m2,data=bsamp)
bsamp$p2x<-model4$fitted.values
bsamp$p2x<-bsamp$a2*bsamp$p2x+(1-bsamp$a2)*(1-bsamp$p2x)
bsamp$iptw<-(bsamp$p1/bsamp$p1x)*(bsamp$p2/bsamp$p2x)
model5<-lm(m1~1,weights=bsamp$iptw,data=bsamp)
bsamp$m1r<-model5$residuals
model6<-lm(m2~m1+a1,weights=bsamp$iptw,data=bsamp)
m2r<-model6$residuals
bsamp$m1a1<-bsamp$m1*bsamp$a1
bsamp$m2a2<-bsamp$m2*bsamp$a2
model7<-lm(y~a1+m1a1+a2+m2a2+m1r+m2r,weights=bsamp$iptw,
```

```
      data=bsamp)
      for (i in 1:5)
            {
            iptw_rwr[j,i]<-model7$coef[i]
            }
      }
sd(covadj_rwr)
sd(iptw_rwr)
```