Published in final edited form as:

Sociol Methodol. 2017 August ; 47(1): 212–245. doi:10.1177/0081175017701180.

Estimating Moderated Causal Effects with Time-varying Treatments and Time-varying Moderators: Structural Nested Mean Models and Regression with Residuals

Geoffrey T. Wodtke and

University of Toronto

Daniel Almirall

University of Michigan

Abstract

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

Keywords

effect heterogeneity; longitudinal; structural nested mean models

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 poor neighborhood is more severe for children in low-

income families than for children in high-income families (Jencks and Mayer 1990; Wilson 1987; Wodtke et al. 2016). 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; Wodtke et al. 2016).

Another common feature of treatments and the individuals, families, and households exposed to them is that they often 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, 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 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 nuanced assessment of the social theories that motivate research on effect moderation, 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 the risk of teen childbearing. With measures of exposure to neighborhood poverty over different time intervals, of family income over different time intervals, and of our primary 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 poor (versus non-poor) neighborhood during childhood and then a non-poor neighborhood during adolescence on the risk of subsequent childbearing among subjects in lower income families during childhood” and “what is the impact of living in a non-poor neighborhood during childhood and then a poor (versus non-poor) neighborhood during adolescence on the risk of subsequent childbearing among subjects in lower income families during adolescence?” These questions address the distal and proximal effects of exposures to neighborhood poverty conditional on the prior 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 (Robins et al. 2007; VanderWeele and Robins 2007a). At the conceptual level, seemingly reasonable questions, such as "what is the effect of living in a poor (versus non-poor) neighborhood throughout childhood and adolescence among subjects in lower income families throughout these periods," do not translate into well-defined causal contrasts. Causal contrasts compare the same subjects in two counterfactual states, but this question implies a comparison of different subjects because the children in families who would continue to have lower incomes had they continuously been exposed to poor neighborhoods and the children in families who would continue to have lower incomes had they been continuously exposed to non-poor neighborhoods are 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 (Wodtke et al. 2016).

At the methodological level, time-varying moderators affected by prior treatment complicate conventional estimation strategies. With cross-sectional data, effect moderation is typically examined using covariate by treatment interactions in conventional regression analyses or propensity score stratification methods (e.g., 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 may yield biased estimates of moderated causal effects.

To overcome these challenges, this article (1) introduces to sociology moderated intermediate causal effects and the structural nested mean model for examining effect moderation 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.

Briefly, the proposed method is a simple adaption of conventional regression methods for moderated effects familiar to all sociologists. Rather than fitting a conventional regression model for the outcome with "main effects" for time-varying measures of the treatment and moderator together with a set of moderator-by-treatment product terms, the regression-with-residuals procedure is implemented as follows: first, the moderators are regressed on prior treatments and then residualized, and second, a regression model for the outcome is fit that includes "main effects" for treatment and for *the residualized moderators* together with a set of moderator-by-treatment product terms. By residualizing the moderators in the first stage and then including these residuals in a second-stage model for the outcome, the method overcomes limitations of conventional regression in the longitudinal setting, where time-varying moderators may be affected by prior treatments.

Effect moderation refers to variation in the effects of a treatment across observed subgroups defined in terms of covariates measured prior to treatment exposure. Moderated effects are thus subtly distinct from several other types of effect heterogeneity, such as “effect interaction” and “principal stratification.” Effect interaction refers to how two or more treatments work with or against each other to cause an outcome (VanderWeele 2009, 2015); that is, effect moderation is defined in terms of variation in the effects of a treatment across pre-treatment random variables, while effect interaction is defined in terms of the joint effects of two or more separate treatments. It is also possible to examine the joint effects of two or more treatments conditional on pre-treatment covariates; in this case, interest is in both effect interaction and effect moderation. Principal stratification refers to variation in the effects of a treatment across neither observed subgroups defined in terms of pre-treatment covariates nor across levels of other treatments; rather, it refers to variation in the effects of a treatment across latent subgroups defined in terms of post-treatment covariates (Frangakis and Rubin 2004). Effect moderation is conceptually and empirically distinct from effect interaction and principal stratification. The methods we present in this article are only suitable for investigating effect moderation, and alternative methods are required for analyses of effect interaction and principle stratification.

EFFECT MODERATION IN THE LONGITUDINAL SETTING

This section formally defines moderated causal effects in the longitudinal setting using the potential outcomes framework (Holland 1986; Rubin 1974) and outlines several different models and estimation strategies for these effects. We define longitudinal settings as those in which both the treatment and moderator are time-varying. In these settings, a moderator is a variable which systematically modifies the form, direction, or strength of the effect of a future treatment on an outcome variable of interest. 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 outcome variable measured at a follow-up time period after the treatment exposures.

In the longitudinal setting, formulating coherent causal questions can be conceptually challenging. The possibility that future moderators may be affected by prior treatments precludes some intuitively appealing questions about how the effects of long-term treatment trajectories vary across subgroups defined in terms of long-term moderator trajectories. In this section, we resolve these conceptual difficulties by focusing on *moderated intermediate causal effects*, which isolate the average causal effect of one additional wave of treatment (versus no additional waves) conditional on prior treatments and moderators. 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.

The estimation procedures we propose in this section identify moderated intermediate causal effects under the assumption of sequential ignorability of treatment assignment, which we define in detail below. Briefly, this assumption requires that at each time point there not be any unobserved determinants of both treatment assignment and the outcome. Estimates from

the methods described in this section may not have a causal interpretation if this assumption is violated.

Effect moderation is metric dependent (Brumback and Berg 2008). Under the same data generating process, one measure of a treatment effect may differ across levels of a moderator while another measure does not. The most common effect metric is the mean difference in 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).

Moderated Intermediate Causal Effects

Let a_t indicate exposure to a dichotomous treatment at time $t = 1, 2$. That is, $a_t = 1$ indicates exposure to treatment at time t , and $a_t = 0$ indicates no exposure to treatment at this time point. Next, let $Y_{i3}(a_1, a_2)$ be the outcome of interest at time $t = 3$ (a follow-up time period after treatment exposures) for subject i had she previously received the treatment sequence (a_1, a_2) . The set $\{Y_{i3}(0,0), Y_{i3}(1,0), Y_{i3}(0,1), Y_{i3}(1,1)\}$ gives all possible potential outcomes for subject i . Now, let M_{i1} be the binary moderator of interest for subject i at time $t = 1$, which is measured prior to treatment at time 1 but shares the same time subscript for notational simplicity. 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 . Note that a_t is not a random variable but rather a fixed value of treatment, while M_{i1} , $M_{i2}(a_1)$, and $Y_{i3}(a_1, a_2)$ are random variables. The distinction between fixed treatments and random variables is reflected in the use of lowercase versus uppercase letters, respectively.

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

$$\mu_1(M_{i1}, a_1) = E(Y_{i3}(a_1, 0) - Y_{i3}(0, 0) | M_{i1}), \quad (1)$$

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} . Formally, the variable M_{i1} is said to be a moderator for the effect of treatment at time 1 if $\mu_1(M_{i1}, a_1)$ is nonconstant in M_{i1} . Note that the function $\mu_1(M_{i1}, a_1)$ equals zero when a_1 is equal to zero. The second set of causal effects is defined as

$$\mu_2(M_{i1}, a_1, M_{i2}(a_1), a_2) = E(Y_{i3}(a_1, a_2) - Y_{i3}(a_1, 0) | M_{i1}, M_{i2}(a_1)), \quad (2)$$

which gives the average causal effect of being exposed to treatment at time 2 (versus not being exposed at time 2) among subgroups defined by M_{i1} and $M_{i2}(a_1)$ had subjects initially been exposed to treatment a_1 . If $\mu_2(M_{i1}, a_1, M_{i2}(a_1), a_2)$ is nonconstant in $(M_{i1}, M_{i2}(a_1))$,

then these variables are said to be moderators for the effect of treatment at time 2. As before, the function $\mu_2(M_{i1}, a_1, M_{i2}(a_1), a_2)$ equals zero when a_2 is equal to zero.

To better understand moderated intermediate causal effects, it can be helpful to consult the language and logic of sequential experiments (e.g., Wodtke et al. 2012). 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 $\mu_1(M_{i1}, a_1)$. Now consider another hypothetical experiment where the researcher measures M_{i1} , assigns all subjects to the same treatment, a_1 , at time 1, measures $M_{i2}(a_1)$, 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 $M_{i2}(a_1)$, would be an experimental estimate of $\mu_2(M_{i1}, a_1, M_{i2}(a_1), a_2)$. 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 prior to each subsequent treatment assignment (Almirall et al. 2012; Lavori and Dawson 2000; Murphy 2005).

The Structural Nested Mean Model

The structural nested mean model (SNMM) formally relates $\mu_1(M_{i1}, a_1)$ and $\mu_2(M_{i1}, a_1, M_{i2}(a_1), a_2)$ to the conditional mean of the potential outcomes (Robins 1994).¹ Specifically, the moderated intermediate causal effects of interest can be linked to the conditional mean of $Y_{i3}(a_1, a_2)$ through the following additive decomposition:

$$E(Y_{i3}(a_1, a_2) | M_{i1}, M_{i2}(a_1)) = \beta_0 + \varepsilon_1(M_{i1}) + \mu_1(M_{i1}, a_1) + \varepsilon_2(M_{i1}, a_1, M_{i2}(a_1)) + \mu_2(M_{i1}, a_1, M_{i2}(a_1), a_2), \quad (3)$$

where the intercept $\beta_0 = E(Y_{i3}(0,0))$ is the mean under no treatment, $\varepsilon_1(M_{i1}) = E(Y_{i3}(0,0) | M_{i1}) - E(Y_{i3}(0,0))$ is the association between M_{i1} and the outcome had no subjects been exposed to treatment, and $\varepsilon_2(M_{i1}, a_1, M_{i2}(a_1)) = E(Y_{i3}(a_1, 0) | M_{i1}, M_{i2}(a_1)) - E(Y_{i3}(a_1, 0) | M_{i1})$ 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 $\mu_1(M_{i1}, a_1)$ and $\mu_2(M_{i1}, a_1, M_{i2}(a_1), a_2)$, the associational effects of the moderators, $\varepsilon_1(M_{i1})$ and $\varepsilon_2(M_{i1}, a_1, M_{i2}(a_1))$, are called nuisance functions.

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

¹Robins (1994; Sections 3a–3c) uses the phrase “blip” (and the notation $\gamma^{[rz]}(k)$, where k denotes time) for what we call the “moderated intermediate causal effect” (or u_p). In addition, Robins (1994) refers to the SNMM as the parametric model used for the moderated intermediate causal effects, whereas we refer to the SNMM more broadly as the complete collection of causal and non-causal terms making up the conditional mean of the potential outcomes, $E(Y_{i3}(a_1, a_2) | M_{i1}, M_{i2}(a_1))$.

$$\begin{aligned}
E(\varepsilon_2(M_{i1}, a_1, M_{i2}(a_1)) | M_{i1}) &= \\
E\left(E\left(Y_{i3}(a_1, 0) | M_{i1}, M_{i2}(a_1)\right) - E(Y_{i3}(a_1, 0) | M_{i1}) \middle| M_{i1}\right) &= \\
E\left(E\left(Y_{i3}(a_1, 0) | M_{i1}, M_{i2}(a_1)\right) \middle| M_{i1}\right) - E(E(Y_{i3}(a_1, 0) | M_{i1}) | M_{i1}) &= E(Y_{i3}(a_1, 0) | M_{i1}) - \\
E(Y_{i3}(a_1, 0) | M_{i1}) &= 0,
\end{aligned}
\tag{4}$$

and

$$\begin{aligned}
E(\varepsilon_1(M_{i1})) &= E(E(Y_{i3}(0, 0) | M_{i1}) - E(Y_{i3}(0, 0))) = E(E(Y_{i3}(0, 0) | M_{i1})) - \\
E(E(Y_{i3}(0, 0))) &= E(Y_{i3}(0, 0)) - E(Y_{i3}(0, 0)) = 0.
\end{aligned}
\tag{5}$$

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 $\mu_1(M_{i1}, a_1)$ 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 $\mu_1(M_{i1}, a_1)$ is

$$\mu_1(M_{i1}, a_1; \beta_1) = a_1(\beta_{10} + \beta_{11} M_{i1}). \tag{6}$$

This model includes the familiar interaction term between treatment and the moderator at time 1, where β_{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 $\mu_2(M_{i1}, a_1, M_{i2}(a_1), a_2; \beta_2)$ is

$$\begin{aligned}
\mu_2(M_{i1}, a_1, M_{i2}(a_1), a_2; \beta_2) &= a_2(\beta_{20} + \beta_{21} M_{i1} + \beta_{22} M_{i2}(a_1) + \beta_{23} M_{i1} M_{i2}(a_1) + \\
&\quad \beta_{24} a_1 + \beta_{25} M_{i1} a_1 + \beta_{26} a_1 M_{i2}(a_1) + \beta_{27} M_{i1} a_1 M_{i2}(a_1)).
\end{aligned}
\tag{7}$$

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 of being exposed to treatment at time 2.

A key to parameterizing the nuisance functions is to ensure that the model satisfies their zero conditional mean property (i.e., Equations 4 and 5). With these constraints in mind, a saturated model for $\varepsilon_1(M_{i1})$ is

$$\varepsilon_1(M_{i1}; \lambda_1) = \lambda_{10} \delta(M_{i1}), \quad (8)$$

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 never been exposed to treatment. Note that $\delta(M_{i1})$ is just a mean-centered version of the time 1 moderator. It satisfies the zero conditional mean property because $E(\delta(M_{i1})) = E(M_{i1} - E(M_{i1})) = 0$. A saturated model for the second nuisance function is

$$\varepsilon_2(M_{i1}, a_1, M_{i2}(a_1); \lambda_2) = \delta(M_{i2}(a_1)) (\lambda_{20} + \lambda_{21} M_{i1} + \lambda_{22} a_1 + \lambda_{23} a_1 M_{i1}), \quad (9)$$

where $\delta(M_{i2}(a_1)) = M_{i2}(a_1) - E(M_{i2}(a_1)|M_{i1})$ 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 $E(\delta(M_{i2}(a_1))|M_{i1}) = E(M_{i2}(a_1) - E(M_{i2}(a_1)|M_{i1})|M_{i1}) = E(M_{i2}(a_1)|M_{i1}) - E(M_{i2}(a_1)|M_{i1}) = 0$. It is important to note that $\delta(M_{i1})$ and $\delta(M_{i2}(a_1))$ are similar to residual terms from regression models for the moderators given past covariates and treatments.

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

$$\begin{aligned} E(Y_{i3}(a_1, a_2)|M_{i1}, M_{i2}(a_1)) \\ = & \beta_0 + \lambda_{10} \delta(M_{i1}) \\ & + a_1 (\beta_{10} \\ & \quad + \beta_{11} M_{i1}) \\ & + \delta(M_{i2}(a_1)) (\lambda_{20} \\ & \quad + \lambda_{21} M_{i1} + \lambda_{22} a_1 + \lambda_{23} a_1 M_{i1}) \\ & + a_2 (\beta_{20} \\ & \quad + \beta_{21} M_{i1} + \beta_{22} M_{i2}(a_1) \\ & \quad + \beta_{23} M_{i1} M_{i2}(a_1) \\ & \quad + \beta_{24} a_1 + \beta_{25} M_{i1} a_1 \\ & \quad + \beta_{26} a_1 M_{i2}(a_1) \\ & \quad + \beta_{27} M_{i1} a_1 M_{i2}(a_1)). \end{aligned} \quad (10)$$

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 the 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 parsimonious model more appealing.

Estimation

In this section, we focus on estimating moderated intermediate causal effects from observed data. The set $\{M_{i1}, A_{i1}, M_{i2}, A_{i2}, Y_{i3}\}$ represents the observed data in temporal order. Note that the outcome and time 2 moderator are no longer indexed by prior treatments as potential outcomes. This notation reflects the fact that, for each subject, we only ever observe the one potential outcome that corresponds to the treatments actually received, and the other potential outcomes are counterfactuals. Also note that a subject's observed treatment is a random variable and is thus denoted by an uppercase A_{it} .

The moderated intermediate causal effects defined previously can be identified from the observed data under the assumptions of consistency and sequential ignorability of treatment assignment (Almirall et al. 2010; Robins 1994). The consistency assumption states that the observed outcome is consistent with one of the conceptualized potential outcomes; that is, $Y_{i3} = A_{i1}A_{i2}Y_{i3}(1,1) + A_{i1}(1 - A_{i2})Y_{i3}(1,0) + (1 - A_{i1})A_{i2}Y_{i3}(0,1) + (1 - A_{i1})(1 - A_{i2})Y_{i3}(0,0)$.² The sequential ignorability assumption is formally expressed in two parts:

$$Y_{i3}(a_1, a_2) \perp A_{i1} | M_{i1} \forall (a_1, a_2) \text{ and } Y_{i3}(a_1, a_2) \perp A_{i2} | M_{i1}, A_{i1}, M_{i2} \forall (a_1, a_2), \quad (11)$$

where $\perp$ denotes statistical independence. Informally, 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.³ Figure 1 displays two DAGs (Elwert 2013; Pearl 2009) 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.

These figures show hypothetical, but not unrealistic, causal systems where the moderators have their own causal effects on the outcome; however, we are not interested in estimating these effects. Rather, we have indexed our potential outcomes only by the treatment of interest. This simplified notation reflects that we have restricted our focus to moderated effects and that we do not consider other types of causal effects that may be of interest in certain situations, such as interaction effects. In other words, we use DAGs to illustrate a broader system of causal effects typical of social science applications, and we use potential

²The consistency assumption defined here should not be confused with asymptotic consistency (i.e., the property of an estimator that converges in probability to its estimand), which is an altogether different usage of the term "consistency."

³In fact, sequential experiments typically meet stronger ignorability assumptions—for example, $Y_{i3}(a_1, a_2) \perp A_{i1}, A_{i2} \forall (a_1, a_2)$.

outcomes to define a narrower set of moderated effects that address the types of research questions outlined in the introduction section. The use of DAGs for illustrative purposes and potential outcomes for definitional purposes is common in both the sociological and causal inference literatures (e.g., Robins et al. 2000; Sharkey and Elwert 2011; VanderWeele 2015).

Limitations of Conventional Regression Estimation—In this section, we explain why estimates from a conventional regression model are biased in the longitudinal setting, even when the sequential ignorability assumption is satisfied. Consider the following conventional regression model:

$$E(Y_{i3}|M_{i1}, A_{i1}, M_{i2}, A_{i2}) = \beta_0^* + \lambda_{10}^* M_{i1} + A_{i1}(\beta_{10}^* + \beta_{11}^* M_{i1}) + M_{i2}(\lambda_{20}^* + \lambda_{21}^* M_{i1} + \lambda_{22}^* A_{i1} + \lambda_{23}^* A_{i1} M_{i1}) + A_{i2}(\beta_{20}^* + \beta_{21}^* M_{i1} + \beta_{22}^* M_{i2} + \beta_{23}^* M_{i1} M_{i2} + \beta_{24}^* A_{i1} + \beta_{25}^* M_{i1} A_{i1} + \beta_{26}^* A_{i1} M_{i2} + \beta_{27}^* M_{i1} A_{i1} M_{i2}). \quad (12)$$

The asterisks on these parameters denote the distinction between the fundamentally unobservable mean differences between potential outcomes in Equation 10 and the observed differences between conditional means in Equation 12.

Least squares estimates of this equation may be biased for the causal parameters in the SNMM for two reasons. First, because this model conditions naively on M_{i2} , which may be affected by prior treatment, A_{i1} , 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 2.

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 (except under strong additional assumptions) because of a second problem associated with conditioning on an intermediate variable: collider-stratification bias. As depicted graphically in Panel B of Figure 2, conditioning on M_{i2} induces an association between prior treatment and unobserved determinants of Y_{i3} , which leads to bias. Even with data from a sequentially randomized experiment, conventional methods may 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 model 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 involve a residual transformation of the time-varying moderators, but the observed data regression considered here includes only 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 using 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 appropriately modeling 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 $\delta(M_{i2}|A_{i1})$. Specifically, parametric models of the form $E(M_{i1}) = \phi$ and $E(M_{i2}|M_{i1}, A_{i1}) = \ell(M_{i1}, A_{i1}; \psi)$ 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 $\hat{\delta}(M_{i2}) = M_{i2} - \hat{E}(M_{i2}|M_{i1}, A_{i1}) = M_{i2} - \ell(M_{i1}, A_{i1}; \hat{\psi})$. When M_{i2} is binary, $E(M_{i2}|M_{i1}, A_{i1})$ can be estimated by least squares using a 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:

$$E(Y_{i3}|M_{i1}, A_{i1}, M_{i2}, A_{i2}) = \beta_0^* + \lambda_{10}^* \hat{\delta}(M_{i1}) + A_{i1}(\beta_{10}^* + \beta_{11}^* M_{i1}) + \hat{\delta}(M_{i2})(\lambda_{20}^* + \lambda_{21}^* M_{i1} + \lambda_{22}^* A_{i1} + \lambda_{23}^* A_{i1} M_{i1}) + A_{i2}(\beta_{20}^* + \beta_{21}^* M_{i1} + \beta_{22}^* M_{i2} \beta_{23}^* M_{i1} M_{i2} + \beta_{24}^* A_{i1} + \beta_{25}^* M_{i1} A_{i1} + \beta_{26}^* A_{i1} M_{i2} + \beta_{27}^* M_{i1} A_{i1} M_{i2}). \quad (13)$$

Least squares estimates of this equation are unbiased for the moderated causal effects of interest under the assumptions outlined previously. The only difference between this model and the conventional regression in Equation 12 is that it correctly specifies the nuisance functions of the SNMM using residualized, rather than untransformed, values of the moderators.

Figure 3 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_{i3} 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.

To complement this graphical explanation, Part A of the Online Supplement provides a technical appendix that formally explains the logic underlying RWR. Specifically, we show that observed differences in conditional means can be equated with the fundamentally unobservable moderated intermediate causal effects under the assumptions outlined previously; that conventional regression implies a model for the moderated intermediate effects of treatment at time 1 that contains non-causal terms, whereas RWR ensures a clean separation of the causal and non-causal information contained in the conditional mean of the outcome; and finally, that the RWR estimator is unbiased under the assumption of a correctly specified SNMM.

Adjustment for Confounding

It is often the case in the social sciences that the sequential ignorability assumptions defined in Equations 11 and 12 do not hold because there are additional 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 they are not properly controlled. Panel A of Figure 4 contains a DAG that graphically depicts the problem of bias due to time-varying confounders. It shows additional variables, C_{i1} and C_{i2} , that affect both treatment and the outcome. Specifically, C_{i1} is a confounder for the effects of A_{i1} on Y_{i1} , and both C_{i1} and C_{i2} are confounders for the effect of A_{i2} on Y_{i2} . Under a slightly modified version of the sequential ignorability assumption, 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 methods in this setting, note that if the analysis adjusts naively for C_{i1} and C_{i2} by including them as covariates in either a conventional regression or even as part of a RWR (where the moderators are residualized but the confounders are not), it may also incur biases due to over-control and collider stratification, as depicted graphically in Panel B of Figure 4. This occurs for the same reasons that conditioning naively on M_{i1} and M_{i2} may lead to over-control and collider-stratification biases. In the sections that follow, we present two approaches for estimating moderated intermediate causal effects in the presence of observed 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 now 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, moderated intermediate causal effects can be identified from observed data under an expanded version of the sequential ignorability assumptions defined previously. Specifically, CA-RWR requires that

$$Y_{i3}(a_1, a_2) \perp A_{i1} | M_{i1}, C_{i1} \forall (a_1, a_2) \text{ and } Y_{i3}(a_1, a_2) \perp A_{i2} | M_{i1}, C_{i1}, A_{i1}, M_{i2}, C_{i2} \forall (a_1, a_2).$$

(14)

Substantively, this assumption states that at each time point there are not any variables other than the prior moderators, confounders, and treatments that directly affect selection into future treatment and the outcome. These conditions subsume those defined in Equation 11, 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 5 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. 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_{i2} that are joint determinants of both C_{i2} and the outcome, Y_{i3} .

Inverse-probability-of-treatment-weighted Regression-with-residuals—Inverse-probability-of-treatment-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 one of the assumptions associated with CA-RWR, which states that time-varying confounders are not also moderators of treatment effects on the outcome. In many cases, researchers are specifically interested in how a particular time-varying covariate, say M_{it} , moderates 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_{it} , 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 dimension of C_{it} —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 large, 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} also becomes 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 weights are computed for each subject i :

$$w_{i1} = \frac{P(A_{i1}=a_{i1}|M_{i1})}{P(A_{i1}=a_{i1}|M_{i1}, C_{i1})} \text{ and } w_{i2} = \frac{P(A_{i2}=a_{i1}|M_{i1}, A_{i1}, M_{i2})}{P(A_{i2}=a_{i1}|M_{i1}, C_{i1}, A_{i1}, C_{i2}, M_{i2})}. \quad (15)$$

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. The true weights are unknown and must be estimated from observed 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 $\hat{w}_{i1} \times \hat{w}_{i2}$. In the second stage, the SNMM is estimated via the same observed data regression used in Equation 13, but estimates are now computed using weighted least squares with weights equal to $\hat{w}_{i1} \times \hat{w}_{i2}$.

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 6 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. Second, by re-weighting the data based on the inverse probability of treatment, 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.

Unsaturated Models

Thus far, we have focused on a saturated SNMM with a binary treatment and a binary moderator both measured at just 2 time points. With multivalued treatments, multivalued moderators, or many time points, a saturated SNMM becomes intractable. In this situation, researchers have to explore simplifying functional form assumptions that reduce the number of free parameters in the model.

For example, consider a hypothetical scenario with a 3-level ordinal treatment and an interval-level moderator that takes on 10 different values, each measured at 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

$$E(Y_{i3}(a_1, a_2) | M_{i1}, M_{i2}(a_1)) = \beta_0 + \lambda_{10} \delta(M_{i1}) + a_1(\beta_{10} + \beta_{11} M_{i1}) + \lambda_{20} \delta(M_{i2}(a_1)) + a_2(\beta_{20} + \beta_{21} M_{i2}), \quad (16)$$

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 is contingent upon the underlying 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.

Variance Estimation

For all of the estimation approaches described previously, the standard errors (SEs) reported from off-the-shelf software packages may be inappropriate because they assume that the residual 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 could be invalid. Almirall et al. (2010) derive asymptotic SEs that additionally account for sampling error in the estimation of the residuals using standard Taylor series arguments. However, because of the additional programming needed to compute these SEs, we propose the use of bootstrap estimates, which are easier to calculate using available software (Efron and Tibshirani 1993). Simulation studies reveal a close correspondence between the bootstrap and asymptotic SEs across a variety of large sample applications (Almirall et al. 2014).

NEIGHBORHOOD EFFECTS ON TEEN CHILDBEARING

This section presents a simplified application of SNMMs and RWR estimation that investigates whether the effects of neighborhood poverty on the risk of teen childbearing are moderated by family income. Theory and previous research suggest that neighborhood effects on adolescent outcomes are likely moderated by family-level characteristics (Sharkey and Faber 2014; Wodtke et al. 2016). Moreover, because both neighborhood exposures and characteristics of the family environment, like parental income, vary over time and are jointly endogenous (Wodtke 2013; Wodtke et al. 2011; Wodtke et al. 2016), research in this area requires new statistical tools capable of properly analyzing neighborhood effect moderation in the longitudinal setting.

We investigate moderated neighborhood effects using data from the PSID (Michigan Survey Research Center 2013). The PSID is a longitudinal study that began in 1968 with a sample of about 4,800 households. Respondents 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).⁴ 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 covariates are constructed separately by developmental period, where the time index t is used to distinguish between childhood ($t = 1$) and adolescence ($t = 2$).

The treatment of interest in this analysis is exposure to neighborhood poverty. We construct a binary treatment variable coded 1 if a child lives in a moderate- or high-poverty neighborhood ($>10\%$ tract poverty rate), and 0 if a child lives in a low-poverty neighborhood ($<10\%$ tract poverty rate). The childhood measure of neighborhood poverty is based on a subject's average tract poverty rate over the three survey waves from age 6 to 8, while the adolescent measure 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 outcome of interest is a binary variable indicating whether a subject experienced a childbirth between ages 15 and 20.⁵ This measurement strategy, which is depicted graphically in Figure 7, ensures appropriate temporal ordering of the treatment, moderator, confounders, and outcome.

⁴The 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.

⁵This analysis includes childbirth events for both male and female subjects in the PSID.

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. We construct a binary moderator variable equal to 1 if a child lives with a high-income family (income-to-needs ratio ≥ 3), and 0 if a child lives with a moderate- or low-income family (income-to-needs ratio < 3).

The time-varying confounders included in this analysis are the family head's current 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 (female versus male), race (black versus nonblack), mother's age (<20 versus ≥ 20 years) and marital status (married versus unmarried) at the time of a subject's birth, and the family head's highest level of completed education (<12 versus ≥ 12 years).⁶

We focus on estimating the following SNMM:

$$E(Y_{i3}(a_1, a_2) | M_{i1}, M_{i2}(a_1)) = \beta_0 + \lambda_{10} \delta(M_{i1}) + a_1(\beta_{10} + \beta_{11} M_{i1}) + \delta(M_{i2}(a_1)) \lambda_{20} + a_2(\beta_{20} + \beta_{21} M_{i2}(a_1)), \quad (17)$$

where a_t denotes exposure to a moderate- or high-poverty neighborhood; M_{it} denotes living with a high-income family; and $Y_{i3}(a_1, a_2)$ 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 an unsaturated linear probability SNMM. In previous sections, we focused mainly on saturated SNMMs that did not require strong assumptions about functional form. This model, however, assumes that the causal effects of neighborhood poverty depend only on family income measured during the same developmental period and that the associational effects of family income are additive.

In this model, the parameter β_{10} gives the average causal effect of childhood exposure to moderate- or high-poverty (versus low-poverty) neighborhoods, setting adolescent treatment to low-poverty neighborhoods, among subjects in moderate- or low-income families during childhood; β_{11} increments this effect for subjects in high-income families at this developmental stage. The parameter β_{20} gives the average causal effect of adolescent exposure to moderate- or high-poverty (versus low-poverty) neighborhoods, holding neighborhood poverty during childhood constant, among subjects in families that would have moderate or low incomes during adolescence under the fixed childhood exposure; β_{21} increments this effect for subjects in high-income families at this development stage.

⁶Missing values are simulated for all variables using multiple imputation with 10 replications (Rubin 1987). Results are based on combined estimates and standard errors.

We estimate these effects using the different variants of RWR described previously. First, we use unadjusted RWR. This approach involves residualizing the moderators based on the observed past 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 family income and that the functional form of the SNMM is correctly specified.

Second, we use CA-RWR. This approach involves residualizing the moderators 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 family income and the other measured covariates; 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 weights via logistic regression models for the conditional probability of exposure to moderate- or high-poverty (versus low-poverty) neighborhoods at each time point given prior treatment, moderators, and confounders. Then, the moderators are residualized based on prior treatments and prior measures of the moderator using weighted regressions, and estimates of moderated causal effects are obtained by fitting a 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 family income and the other observed covariates; that the logistic regression models for the conditional probability of treatment 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, second, and third columns of the table present estimates from unadjusted RWR, CA-RWR, and IPTW-RWR, respectively. In general, results from all three different estimators suggest a negligible effect of exposure to neighborhood poverty during childhood on the risk of teen childbearing and provide no evidence of effect moderation by family income at this developmental stage. Estimates for the effect of exposure to neighborhood poverty during adolescence, by contrast, indicate that living in a moderate- or high-poverty neighborhood during this developmental period has a statistically significant and substantively large positive effect on the risk of teen childbearing. Moreover, estimates also indicate that this effect is significantly moderated by prior family income. Specifically, the inflationary impact of adolescent exposure to neighborhood poverty on the risk of subsequent childbearing is much stronger for individuals in moderate- or low-income families than for individuals in high-income families at this developmental stage.

For example, according to estimates based on IPTW-RWR, adolescent exposure to moderate- or high-poverty neighborhoods, rather than low-poverty neighborhoods, is estimated to increase the risk of a subsequent childbirth by about 10 percentage points among individuals in moderate- or low-income families during adolescence (i.e., $1(\hat{\beta}_{20} + \hat{\beta}_{21}(0)) = 0.100$). Among individuals in high-income families, estimates indicate that

exposure to moderate- or high-poverty neighborhoods, rather than low-poverty neighborhoods, during adolescence increases the risk of a subsequent childbirth event by just over 1 percentage point (i.e., $1(\hat{\beta}_{20} + \hat{\beta}_{21}(1)) = 0.100 - 0.089 = 0.011$).

All of the estimates in Table 1 are based on unsaturated linear probability SNMMs. As with any unsaturated model, researchers should consider the possibility of bias due to misspecification. A particular concern in our empirical application is that, because the outcome is bounded, models for the nuisance functions may constrain models for the moderated intermediate causal effects. We experimented with a variety of different linear specifications and found that estimates of moderated neighborhood effects are substantively similar in more complex models with less stringent constraints on functional form.⁷ With a binary outcome, researchers may also want to consider nonlinear models, such as structural nested logistic models (Clarke and Windmeijer 2010; Robins and Rotnitzky 2004; Vansteelandt and Goetghebeur 2003; Vansteelandt and Joffe 2014), which ensure that predicted probabilities do not fall outside the logical range. These models are also based on a different effect metric—the log odds ratio—that may be more appropriate than the risk difference in certain situations. As estimation procedures for structural nested logistic models become more accessible, we expect these methods to find wider application in the social sciences.

DISCUSSION

Treatment effect moderation 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 describes a new class of models and estimators for analyzing moderated causal effects in the longitudinal setting: SNMMs and RWR. It outlines how these methods avoid limitations associated with conventional regression when time-varying moderators are affected by prior treatments, and it adapts them to account for observed confounding.

To illustrate these methods, we present a simple empirical application with longitudinal data from the PSID. This analysis investigates whether the effects of childhood versus adolescent exposure to neighborhood poverty on the risk of teen childbearing are moderated by prior family income. Results indicate that exposure to neighborhood poverty during adolescence, but not during childhood, increases the risk of subsequent childbearing, particularly for individuals whose families have lower incomes during adolescence. This analysis 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 the moderated effects of time-varying treatments.

⁷Another specification test with an unsaturated linear probability SNMM is to check whether it yields nonsensical predicted values. Models estimated by unadjusted RWR and IPTW-RWR do not yield any predicted values outside the logical range, while the model estimated by CA-RWR yields predictions outside the logical range—between -0.04 and 0.00 —for about 1 percent of sample members. The suboptimal performance of CA-RWR likely reflects a failure to correctly model nuisance functions involving a large number of confounders, which is an additional complexity associated with this estimation strategy.

For expositional simplicity, this article focuses on SNMMs with a terminal outcome. These methods, however, can be adapted to investigate time-varying outcomes. Moreover, they can even be adapted to investigate whether treatment effects on a future time-varying outcome are moderated by past values of the 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 policymakers to develop adaptive interventions that tailor treatments across time to the evolving needs of different individuals, and it would allow researchers to better understand the dynamic etiology of labor income and academic achievement.

Although this study focuses 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). Despite the simplicity, convenience, and relative efficiency of RWR estimation, the G-estimator is not without its own advantages. In particular, the G-estimator provides unbiased estimates of moderated effects 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. RWR, on the other hand, requires correct specification of the nuisance functions. 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 (Almirall et al. 2010). In practice, researchers may want to consider implementing both G-estimation and RWR estimation to balance concerns about misspecification and precision.

Because of its similarity to conventional regression, RWR can be easily implemented with off-the-shelf software. Parts B and C of the Online Supplement provide code for the Stata and R statistical software packages that executes the different RWR estimators discussed previously and computes bootstrap standard errors using simulated data. These methods can be easily implemented with available software because they merely involve regressing the moderators on prior treatments and covariates, computing residuals from these regressions, and then regressing the outcome on these residuals, the treatments, and the appropriate cross-product terms. Although it may at first appear technically complex, RWR estimation of an SNMM involves only familiar computations that can be carried out by virtually any statistical software package.

Empirical researchers interested in effect moderation 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 not misplaced, and methods for assessing the robustness of findings to hypothetical patterns of unobserved confounding should be incorporated into empirical analyses more frequently (Brumback et al. 2004). However, we show that even in studies where there is no unobserved confounding of treatment, conventional methods for analyzing effect moderation are 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, their magnitude can be substantively large and in some cases even

greater than confounding bias (Greenland 2003). Thus, it is important to have flexible statistical tools, like SNMMs and RWR, that are capable of accounting for the variety of different biases encountered in longitudinal research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported 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 Nos. R03 MH097954 and RC4 MH092722 to the Survey Research Center at the University of Michigan.

References

- Almirall, Daniel, Griffin, Beth A., McCaffrey, Daniel F., Ramchand, Rajeev, Yuen, Robert A., Murphy, Susan A. Time-varying Effect Moderation using the Structural Nested Mean Model: Estimation using Inverse-weighted Regression with Residuals. *Statistics in Medicine*. 2014; 33:3466–87. [PubMed: 23873437]
- Almirall, Daniel, Compton, Scott N., Gunlicks-Stoessel, Meredith, Duan, Naihua, Murphy, Susan A. Preparing for a Sequential Multiple Assignment Randomized Trial for Developing an Adaptive Treatment Strategy: Designing a SMART Pilot Study. *Statistics in Medicine*. 2012; 31:1887–902. [PubMed: 22438190]
- Almirall, Daniel, McCaffrey, Daniel F., Ramchand, Rajeev, Murphy, Susan A. Subgroups Analysis when Treatment and Moderators are Time-varying. *Prevention Science*. 2013; 14:169–78. [PubMed: 21424793]
- Almirall, Daniel, Have, Thomas Ten, Murphy, Susan A. Structural Nested Mean Models for Assessing Time-Varying Effect Moderation. *Biometrics*. 2010; 66:131–9. [PubMed: 19397586]
- Amato, Paul R. The Consequences of Divorce for Adults and Children. *Journal of Marriage and Family*. 2004; 62:1269–87.
- Brand, Jennie E., Xie, Yu. Who Benefits Most from College? Evidence for Negative Selection in Heterogenous Economic Returns to Higher Education. *American Sociological Review*. 2010; 75:273–302. [PubMed: 20454549]
- Brumback, Babette A., Berg, Arthur. On Effect-measure Modification: Relationships Among Changes in the Relative Risk, Odds Ratio, and Risk Difference. *Statistics in Medicine*. 2008; 27:3453–65. [PubMed: 18335568]
- Brumback, Babette A., Hernan, Miguel A., Haneuse, Sebastien JPA., Robins, James M. Sensitivity Analyses for Unmeasured Confounding Assuming a Marginal Structural Model for Repeated Measures. *Statistics in Medicine*. 2004; 23:749–67. [PubMed: 14981673]
- Casella, George, Berger, Roger L. *Statistical Inference*. 2. Pacific Grove, CA: Duxbury; 2002.
- Clarke, Paul S., Windmeijer, Frank. Identification of Causal Effects on Binary Outcomes using Structural Mean Models. *Biostatistics*. 2010; 11:756–70. [PubMed: 20522728]
- Efron, Bradley, Tibshirani, Robert J. *An Introduction to the Bootstrap*. New York: Chapman and Hall; 1993.
- Elwert, Felix. Graphical Causal Models. In: Morgan, SL., editor. *Handbook of Causal Analysis for Social Research*. New York: Springer; 2013. p. 245–71.
- Frangakis, Constantine E., Rubin, Donald B. Principal Stratification in Causal Inference. *Biometrics*. 2004; 58:21–9.
- GeoLytics, Inc. *CensusCD Neighborhood Change Database, 1970–2000 Tract Data*. New Brunswick, NJ: GeoLytics; 2003.

- Greenland, Sander. Quantifying Biases in Causal Models: Classical Confounding vs Collider-stratification Bias. *Epidemiology*. 2003; 14:300–6. [PubMed: 12859030]
- Heckman, James J. Skill Formation and the Economics of Investing in Disadvantaged Children. *Science*. 2006; 312:1900–1. [PubMed: 16809525]
- Heckman, James J., Urzua, Sergio, Vytlačil, Edward. Understanding Instrumental Variables in Models with Essential Heterogeneity. *Review of Economics and Statistics*. 2006; 88:389–432.
- Holland, Paul W. Statistics and Causal Inference. *Journal of the American Statistical Association*. 1986; 81:945–60.
- Jencks, Christopher, Mayer, Susan E. The Social Consequences of Growing Up in a Poor Neighborhood. In: Lynn, LE., McGreary, MGH., editors. *Inner-City Poverty in the United States*. Washington, D.C: National Academy Press; 1990. p. 111–86.
- Lavori, Philip W., Dawson, Ree. A Design for Testing Clinical Strategies: Biased Individually Tailored Within-subject Randomization. *Journal of the Royal Statistical Society*. 2000; 163:29–38.
- McCaffrey, Daniel F., Ridgeway, Greg, Morral, Andrew R. Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods*. 2004; 9:403–25. [PubMed: 15598095]
- Michigan Survey Research Center. Panel Study of Income Dynamics, Restricted Use Data. Ann Arbor, MI: University of Michigan; 2013.
- Murphy, Susan A. An Experimental Design for the Development of Adaptive Treatment Strategies. *Statistics in Medicine*. 2005; 24:1455–81. [PubMed: 15586395]
- Pearl, Judea. *Causality*. 2. New York: Cambridge University Press; 2009.
- Robins, James M. Correcting for Noncompliance in Randomized Trials Using Structural Nested Mean Models. *Communications in Statistics-Theory and Methods*. 1994; 23:2379–412.
- Robins, James M., Hernan, Miguel A., Brumback, Babette. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*. 2000; 11:550–60. [PubMed: 10955408]
- Robins, James M., Hernan, Miguel A., Rotnitzky, Andrea. Effect Modification by Time-Varying Covariates. *American Journal of Epidemiology*. 2007; 166:994–1002. [PubMed: 17875581]
- Robins, James M., Rotnitzky, Andrea. Estimation of Treatment Effects in Randomised Trials With Non-Compliance and a Dichotomous Outcome Using Structural Mean Models. *Biometrika*. 2004; 91:763–83.
- Rothman, Kenneth J., Greenland, S., Walker, Alexander M. Concepts of Interaction. *American Journal of Epidemiology*. 1980; 112:467–70. [PubMed: 7424895]
- Rubin, Donald B. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*. 1974; 66:688–701.
- Rubin, Donald B. *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons; 1987.
- Sanders, William L., Wright, S Paul, Horn, Sandra P. Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation. *Journal of Personnel Evaluation in Education*. 1997; 11:57–67.
- Sharkey, Patrick, Elwert, Felix. The Legacy of Disadvantage: Multigenerational Neighborhood Effects on Cognitive Ability. *American Journal of Sociology*. 2011; 116:1934–81.
- Sharkey, Patrick, Faber, Jacob W. Where, When, Why, and For Whom Do Residential Contexts Matter? Moving Away from the Dichotomous Understanding of Neighborhood Effects. *Annual Review of Sociology*. 2014; 40:559–79.
- VanderWeele, Tyler J. On the Distinction between Interaction and Effect Modification. *Epidemiology*. 2009; 20:863–71. [PubMed: 19806059]
- VanderWeele, Tyler J. *Explanation in Causal Inference*. New York: Oxford University Press; 2015.
- VanderWeele, Tyler J., Robins, James M. Directed Acyclic Graphs, Sufficient Causes and the Properties of Conditioning on a Common Effect. *American Journal of Epidemiology*. 2007a; 166:1096–104. [PubMed: 17702973]
- VanderWeele, Tyler J., Robins, James M. Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs. *Epidemiology*. 2007b; 18:561–68. [PubMed: 17700242]
- Vandsteelandt, Stijn, Goetghebeur, E. Causal Inference with Generalized Structural Mean Models. *Journal of the Royal Statistical Society: Series B*. 2003; 65:817–35.

- Vandsteelandt, Stijn, Joffe, Marshall. Structural Nested Models and G-estimation: The Partially Realized Promise. *Statistical Science*. 2014; 29:707–31.
- Wallerstein, Judith S. The Long-term Effects of Divorce on Children: A Review. *Journal of the American Academy of Child and Adolescent Psychiatry*. 1991; 30:349–60. [PubMed: 1810276]
- Wilson, William J. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Chicago: University of Chicago Press; 1987.
- Wodtke, Geoffrey T. Duration and Timing of Exposure to Poor Neighborhoods and the Risk of Adolescent Parenthood. *Demography*. 2013; 50:1765–88. [PubMed: 23720166]
- Wodtke, Geoffrey T., Elwert, Felix, Harding, David J. Neighborhood Effect Heterogeneity by Family Income and Developmental Period. *American Journal of Sociology*. 2016; 121:1168–222.
- Wodtke, Geoffrey T., Harding, David J., Elwert, Felix. Neighborhood Effects in Temporal Perspective: The Impact of Long-term Exposure to Concentrated Disadvantage on High School Graduation. *American Sociological Review*. 2011; 76:713–36. [PubMed: 22879678]
- Xie, Yu. Otis Dudley Duncan's Legacy: The Demographic Approach to Quantitative Reasoning in Social Science. *Research in Social Stratification and Mobility*. 2007; 25:141–56.
- Xie, Yu, Brand, Jennie E., Jann, B. Estimating Heterogeneous Treatment Effects with Observational Data. *Sociological Methodology*. 2012; 42:314–47. [PubMed: 23482633]

A. Sequential randomization

B. Selection on prior moderators

Notes: A = treatment, M = moderator, Y = outcome, and U = unobserved factors.

Figure 1.

Causal relationships between time-varying treatments, time-varying moderators, and outcome

A. Over-control of intermediate pathways

B. Collider-stratification

Notes: A = treatment, M = moderator, Y = outcome, and U = unobserved factors. A box around a variable denotes conditioning.

Figure 2.

Over-control of intermediate pathways and collider-stratification biases

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 = treatment, M = moderator, Y = 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 3.

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

A. Selection on prior moderators and confounders

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

Notes: A = treatment, M = moderator, C = confounder, Y = 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 4.

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

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

Notes: A = treatment, M = moderator, C = confounder, Y = 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 5.

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

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

Notes: A = treatment, M = moderator, C = confounder, Y = 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 6.

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

Notes: A = neighborhood poverty, M = family income-to-needs ratio, C = vector of observed confounders, and Y = childbirth event.

Figure 7.
Longitudinal measurement strategy in the PSID

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

Table 1

Specification	Unadjusted RWR		Adjusted RWR		IPT-weighted RWR				
	est	se	est	se	est	se			
Intercept	.080	(.007)	***	.129	(.009)	***	.110	(.013)	***
<i>Childhood</i>									
NH pov	.042	(.016)	**	-.003	(.016)		-.001	(.018)	
Fam inc × NH pov	-.032	(.022)		-.007	(.023)		.001	(.025)	
<i>Adolescence</i>									
NH pov	.111	(.019)	***	.078	(.018)	***	.100	(.020)	***
Fam inc × NH pov	-.077	(.020)	***	-.056	(.020)	**	-.089	(.024)	***

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.