PSYCHOMETRIKA
https://doi.org/10.1007/s11336-018-9606-0

IDENTIFICATION AND SENSITIVITY ANALYSIS FOR AVERAGE CAUSAL MEDIATION EFFECTS WITH TIME-VARYING TREATMENTS AND MEDIATORS: INVESTIGATING THE UNDERLYING MECHANISMS OF KINDERGARTEN RETENTION POLICY

SOOJIN PARK

UNIVERSITY OF CALIFORNIA, RIVERSIDE

PETER M. STEINER AND DAVID KAPLAN

UNIVERSITY OF WISCONSIN-MADISON

Considering that causal mechanisms unfold over time, it is important to investigate the mechanisms over time, taking into account the time-varying features of treatments and mediators. However, identification of the average causal mediation effect in the presence of time-varying treatments and mediators is often complicated by time-varying confounding. This article aims to provide a novel approach to uncovering causal mechanisms in time-varying treatments and mediators in the presence of time-varying confounding.queryPlease check and confirm edit made in the article title. We provide different strategies for identification and sensitivity analysis under homogeneous and heterogeneous effects. Homogeneous effects are those in which each individual experiences the same effect, and heterogeneous effects are those in which the effects vary over individuals. Most importantly, we provide an alternative definition of average causal mediation effects that evaluates a partial mediation effect; the effect that is mediated by paths other than through an intermediate confounding variable. We argue that this alternative definition allows us to better assess at least a part of the mediated effect and provides meaningful and unique interpretations. A case study using ECLS-K data that evaluates kindergarten retention policy is offered to illustrate our proposed approach.

Key words: causal mediation analysis, time-varying treatment and mediator, time-varying confounding variable, homogeneous effects, heterogeneous effects, sensitivity analysis.

1. Introduction

Retention is a school policy that requires students to repeat a grade if they fail to make adequate progress. Recent studies using same-age comparisons have found negative effects of kindergarten retention on student math achievement among those students who are at risk for grade retention (Hong & Raudenbush, 2005, 2006; Vandecandelaere, Vansteelandt, De Fraine, & Van Damme, 2016). Vandecandelaere et al. (2016) demonstrated that this negative effect persists throughout primary education, although the effect attenuates over time. As a next step, it is important to investigate why kindergarten retention has a negative effect and how this adverse effect attenuates over time.

In this article we discuss a novel approach to causal mediation analysis with two time-varying treatments and mediators, in which we investigate causal mechanisms underlying the relationship

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11336-018-9606-0) contains supplementary material, which is available to authorized users.

Correspondence should be made to Soojin Park, University of California, Riverside, Riverside, USA. Email: soojinp@ucr.edu

Published online: 01 March 2018

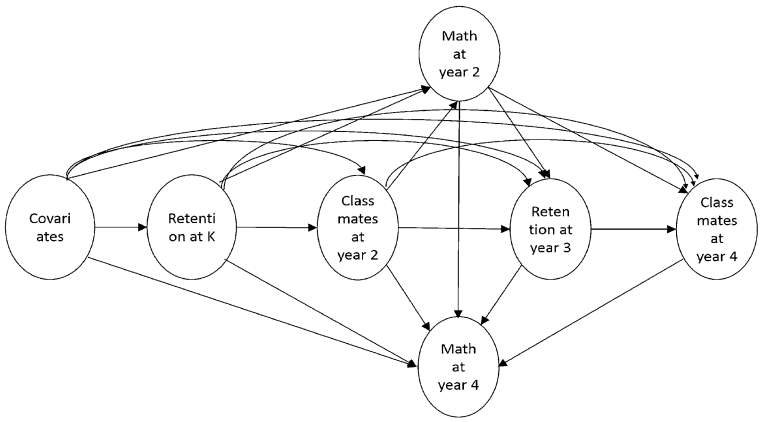

FIGURE 1.
The causal structural model used in the case study. *Note* Retention at K: Retention status at Kindergarten, Classmates: Classmates' math score, Year n: nth year of the study, and Math: math score.

between the kindergarten retention policy and student math achievement. Figure 1 is a graphical representation of the model. The time-varying treatments are the retention status at kindergarten and year 3, and the time-varying mediators are classmate math ability measured at year 2 and year 4. The outcome of interest is student math achievement measured at the end of year 4. We hypothesize that grouping retained students with classmates who are one year younger may provide fewer chances to interact with the more advanced, promoted peers of the same age, thereby slowing down the retained students' learning. We specifically focus on the emerging mediation effect via classmate math ability one year after kindergarten retention to determine how the adverse effect of kindergarten retention attenuates over time.

The substantive causal questions examined in this article are

1. Does classmate math ability mediate the negative effect of kindergarten retention on math achievement during the first year after kindergarten retention?
2. Are there significant emerging mediation effects via classmate math ability one year after kindergarten retention?
3. Do the emerging mediation effects differ by retention status and/or by intermediate math score?

The first question can be answered by conducting a single-time-period causal mediation analysis. The second and third causal questions, which involve more than a single time period, have three methodological challenges to answer. First, in order to investigate causal mechanisms with two time-varying treatments and mediators, it is important to account for time-varying confounding variables in addition to pre-treatment covariates. For example, the math score measured at the end of year 2 may be an important intermediate confounding variable. The issue is that this intermediate math score is a posttreatment variable, which is affected by kindergarten retention status (treatment), and also confounds the relationship between future classmate math ability (mediator) and final math score (outcome). In the presence of a posttreatment confounding variable, the average causal mediation effect (ACME) and the average natural direct effect (ANDE) are not identified *nonparametrically* unless one assumes the absence of interaction effects in the mediator–mediator or treatment–mediator relationships with respect to the outcome (see, e.g., Avin, Shpitser, & Pearl, 2005; VanderWeele & Vansteelandt, 2014; Imai & Yamamoto, 2013).

Here, the ACME is the average effect of kindergarten retention on math achievement mediated by classmate math ability, and the ANDE is the average direct effect of kindergarten retention on math achievement, which is not mediated by classmate math ability. The inability to nonparametrically identify the ACME and ANDE prevents investigation of causal mechanisms in longitudinal studies unless one is willing to make strong linearity and no-interaction effect assumptions. Under these modeling assumptions, the mediation effect would be identified and estimable using the product of coefficients approach as in the SEM framework (De Stavola, Daniel, Ploubidis, & Micali, 2014).

Second, the validity of results also depends on whether the ACME and ANDE are *homogeneous* or *heterogeneous* (see, e.g., Imai & Yamamoto 2013; VanderWeele & Vansteelandt, 2014) for heterogeneous effects; and Daniel, De Stavola, Cousens, and Vansteelandt (2015) for homogeneous effects). Homogeneous effects refer to a situation where every subject has the same constant effect, whereas heterogeneous effects allow for systematic or random variations of the effects. If effect homogeneity does not hold, we need to use a different strategy to identify ACME and ANDE. Therefore in order to estimate consistent and meaningful mediation effects from real data, it is crucial to understand differences in causal estimands (effects of interest) and the respective assumptions that are required to identify the causal estimands under homogeneous and heterogeneous effects.

Third, the effect homogeneity assumption is unrealistic in many cases; it is therefore essential to appropriately account for heterogeneous effects. Imai and Yamamoto (2013) developed a sensitivity analysis in the context of multiple mediators that can be used under heterogeneous effects. The sensitivity analysis is designed to check the robustness of findings to potential violations of the no-interaction effect (in the treatment and mediator relationship) assumption under heterogeneous effects. However, generalizing their sensitivity analysis does not yield much information with respect to ACME and ANDE when time-varying confounding variables are present. This is because the corresponding sensitivity analysis depends on too many unknown sensitivity parameters (due to the added complexities of time-varying confounding variables).

The goal of this article is, therefore, to provide a comprehensive approach for investigating causal mechanisms with time-varying treatments and mediators that allow for time-varying confounding variables under both the homogeneous and heterogeneous effects assumption. Importantly, we suggest alternative definitions of ACME and ANDE that require fewer assumptions than the original definitions of ACME and ANDE. These alternative definitions provide insights regarding how the adverse effect of kindergarten retention attenuates over time, by fixing the effect via the intermediate math score. Our alternative definitions of ACME and ANDE also enable us to examine whether and how mediation effects vary with the level of the intermediate math score. Given the complex nature of causal mediation analysis with time-varying treatments and mediators, we provide practical suggestions for drawing meaningful conclusions from mediation analyses with real-life data.

The remainder of this article is organized as follows. We begin by providing a brief introduction to the data and our example. This is followed by introducing notation and definitions. Sections 4 and 5 provide a partial identification strategy and corresponding analyses of sensitivity to the alternative definitions of ACME and ANDE under homogeneous effects and heterogeneous effects, respectively. Then, we present a case study in which our proposed method is applied to the kindergarten retention study. The article concludes with a discussion.

2. Data

The Early Childhood Longitudinal Study, Kindergarten (ECLS-K) data provide representative information about a cohort of American children who attended kindergarten in 1998–1999

TABLE 1.
Structure of ECLS-K data.

Years	1	2	3	4
Never retained	K	G1	G2	G3
Retained at K	K	K	G1	G2
Variables	Retention at K(T_1)	Classmate math ability(M_1), Math score(L)	Retention at year 3(T_2)	Classmate math ability(M_2), Math score(Y)
Time periods	1st	1st	2nd	2nd

(1) K = kindergarten; G1–G3 = grade 1 through grade 3.
(2) Descriptions for variables (T_1, M_1, L, T_2, M_2, and Y) are given in Sect. 3.

(Tourangeau, Nord, Lê, Sorongon, & Najarian, 2009). The cohort was followed from kindergarten to the 8th grade. For ease of explanation, the 1998–1999 to 2001–2002 school years are referred to as year 1 to year 4. Table 1 shows the structure of the data. For example, if a student has never been retained at the end of year 1, s/he will be promoted to the first, second, and third grades in years 2, 3, and 4, respectively; and if a student has retained at the end of year 1, s/he will repeat kindergarten at year 2 and be promoted to the first, and second grades in years 3 and 4, respectively. Our sample includes 342 kindergarten retainees and 11,248 students who were promoted.

As shown in Fig. 1, retention statuses measured at the end of year 1 and year 3 serve as our time-varying treatments. The level of classmate math ability serves as our time-varying mediator and is measured by aggregating the math scores at the class level in years 2 and 4, except him/herself. Our final outcome is student math achievement measured at the end of year 4. Student math scores are calibrated by item response theory (IRT) and are vertically equated over time, which enables us to compare outcomes between those who are retained and promoted. Note that we compare students' vertically equated scores. The consequences of using vertically equated scores that are invalid with respect to underlying student ability are discussed in Steiner, Park, and Kim (2016).

Pre-treatment covariates include individual demographic characteristics; previous learning experiences, cognitive, emotional, and social development at the student level; teacher years of experience; and teacher educational degree at the class level. At the school level, we include the proportion of minorities, teacher salary, public/private school, number of students in kindergarten, number of students retained, and percentage of students retained at school. We consider student math scores measured before kindergarten and year 2 as the time-varying confounding variable. Particularly, student math achievement measured at year 2 is the intermediate confounding variable that is affected by the kindergarten retention status and has an impact on future classmate ability and math score measurements.

Our data have a hierarchical structure where students are nested within classes and schools. Although it is important to consider the hierarchical structure in estimating the mediation effects, it is very challenging with longitudinal data because class and school memberships changes over time. The intraclass correlation computed from our final outcome, student math achievement measured at the end of year 4, is 0.29, which indicates that the standard errors of estimates might be underestimated due to this hierarchical structure of the data.

3. Notation and Definitions

To begin, let $T_{1i} \in \{0, 1\}$ and $T_{2i} \in \{0, 1\}$ be the first- and second-time measured retention statuses for individual i, where $T = 1$ if retained and $T = 0$ if promoted. Let M_{1i} and M_{2i} be the

TABLE 2.
Key potential outcomes and a short version of the outcomes.

Potential outcomes	Interpretations	Short version
$M_{1i}(\mathrm{t})$	A potential value of M_1 that would have been observed under $T_{1i} = t$	$M_{1i}(\mathrm{t})$
$M_{2i}(t, 0, M_{1i}(t))$	A potential value of M_2 that would have been observed under $T_{1i} = t$ and $T_{2i} = 0$	$M_{2i}(t, 0)$
$Y_i(t, 0, M_{1i}(t'), M_{2i}(t'', 0))$	A potential outcome under $T_{1i} = t$, $T_{2i} = 0$, and the potential values of the first and second mediators under $T_{1i} = t'$ and $T_{1i} = t''$, respectively.	$Y_i(t, 0)$ if $t = t' = t''$
$M_{2i}(t, 0, M_{1i}(t), l)$	A potential value of M_2 that would have been observed under $T_{1i} = t$, $T_{2i} = 0$ and $L_i = l$.	$M_{2i}(t, 0, l)$
$Y_i(t, 0, M_{1i}(t'), M_{2i}(t'', 0, l), l)$	A potential outcome under $T_{1i} = t$, $T_{2i} = 0$, $L_i = l$, and the potential values of the first and second mediators under $T_{1i} = t'$ and $T_{1i} = t''$, respectively	$Y_i(t, 0, l)$ if $t = t' = t''$

time-varying mediators measured at the first and second time periods, respectively. Let Y_i be the final math achievement score for individual i. Let V_i be a set of pre-treatment covariates, and let L_i be an intermediate math score. In general, a time-varying variable (treatment T, mediator M, or covariate L) may represent either a single variable that is measured at multiple time points, or different variables at different time points. For instance, T_1 and T_2 can be the dosage (on/off) of a single treatment at the two time points or they may represent two entirely different treatments: treatment A is given at time 1, and treatment B is given at time 2 (analogously for the mediators and covariates). Under the potential outcomes framework of Rubin (1974),[1] we can write the potential mediators as $\{M_{1i}(t), M_{2i}(t, t', M_{1i}(t))\} \in \mu$, where μ represents the two-dimensional support region of potential mediators; and $Y_i(t, 0, M_{1i}(t'), M_{2i}(t'', 0)) \in \omega$, where ω represents the support of the potential outcome of Y and t, t' and $t'' \in \{0, 1\}$. Table 2 presents key potential outcomes that are used throughout.

In this notation, we first discuss the originally defined ACME and ANDE with two time-varying treatments and mediators and then introduce our alternative definitions of ACME and ANDE. There are several ways to decompose the total effect of kindergarten retention, but we focus on the following definitions of ACME and ANDE that are relevant to our causal questions.

3.1. *Definitions of ACME and ANDE*

The ACMEs via M_1 and via M_2 under t ($\delta^{M_1}(t)$, $\delta^{M_2}(t)$) and the ANDE under t' ($\zeta(t')$) are respectively defined as

$$\begin{aligned}
\delta^{M_1}(t) &= E[Y_i(t, 0, M_{1i}(1), M_{2i}(t, 0)) - Y_i(t, 0, M_{1i}(0), M_{2i}(t, 0)], \\
\delta^{M_2}(t) &= E[Y_i(t, 0, M_{1i}(t'), M_{2i}(1, 0)) - Y_i(t, 0, M_{1i}(t'), M_{2i}(0, 0)], \text{ and} \\
\zeta(t') &= E[Y_i(1, 0, M_{1i}(t'), M_{2i}(t', 0)) - Y_i(0, 0, M_{1i}(t'), M_{2i}(t', 0))] \qquad (1)
\end{aligned}$$

[1]Following Rubin (1974) each individual is assumed to have only one potential treatment and one potential control outcome, instead of an entire distribution as in Steyer's Theory of Causal Effects or in Neyman's setup. For detailed information, refer to Mayer, Thoemmes, Rose, Steyer, and West (2014) and Steyer, Mayer, and Fiege (2014).

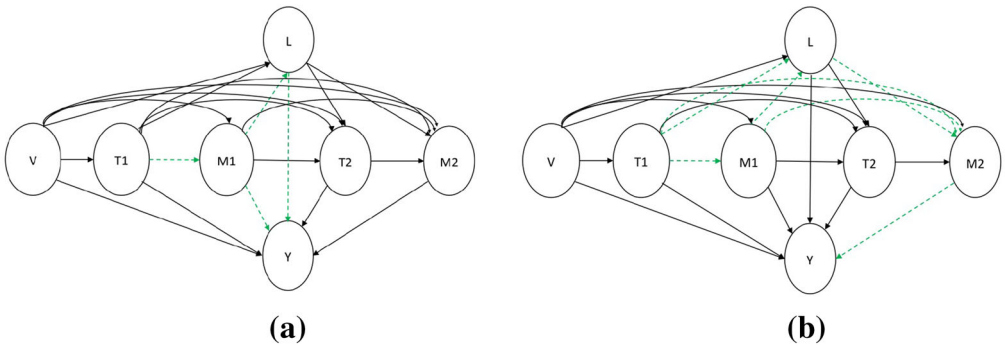

FIGURE 2.
Causal structural model highlighting the ACME. **a** ACME via M_1. **b** ACME via M_2

for $t \in \{0, 1\}$ and $t' = 1 - t$.[2] The path-specific effects of δ^{M_1} and δ^{M_2} are highlighted in Fig. 2. The ACME via M_1 ($\delta^{M_1}(t)$) is the effect of kindergarten retention on student math score transmitted along classmate math ability in year 2. The ACME via M_2 ($\delta^{M_2}(t)$) is the effect of kindergarten retention on student math score transmitted along classmate math ability in year 4, either directly or mediated by classmate math ability in year 2. Lastly, the ANDE $\zeta(t')$ states the direct effect of kindergarten retention on student math score through neither of the mediators.

The ACMEs via M_1 and via M_2 and ANDE are combined into the average kindergarten retention effect (τ), which is defined as the expected change in student math achievement in response to a change from not retained to retained at kindergarten if everyone was not retained at year 3. In formal expression, $\tau = E[Y_i(1, 0) - Y_i(0, 0)]$, which is the combined effects[3] of the $\delta^{M_1}(t)$, $\delta^{M_2}(t)$, and $\zeta(t')$.

These path-specific effects shown in Eq. (1) have been defined and discussed in the multiple mediators context (see, e.g., Daniel et al. 2015; Imai & Yamamoto, 2013; Steen, Loeys, Moerkerke, & Vansteelandt, 2017; VanderWeele & Vansteelandt, 2014); yet, there has been little discussion on these path-specific effects in the context of time-varying treatments and mediators. Identifying

[2]Following an inductive rule by Shpitser (2013), the definition can be rewritten as,

$$\begin{aligned}
\delta^{M_1}(t) &= E[Y_i(t, 0, M_{1i}(1), L_i(t, M_{1i}(1)), M_{2i}(t, 0, M_{1i}(t), L_i(t, M_{1i}(t))) \\
&\quad - Y_i(t, 0, M_{1i}(0), L_i(t, M_{1i}(0)), M_{2i}(t, 0, M_{1i}(t), L_i(t, M_{1i}(t)))], \\
\delta^{M_2}(t) &= E[Y_i(t, 0, M_{1i}(t'), L_i(t, M_{1i}(t')), M_{2i}(1, 0, M_{1i}(1), L_i(1, M_{1i}(1))) \\
&\quad - Y_i(t, 0, M_{1i}(t'), L_i(t, M_{1i}(t')), M_{2i}(0, 0, M_{1i}(0), L_i(0, M_{1i}(0)))], \text{ and} \\
\zeta(t') &= E[Y_i(1, 0, M_{1i}(t'), L_i(1, M_{1i}(t')), M_{2i}(t', 0, M_{1i}(t'), L_i(t', M_{1i}(t'))) \\
&\quad - Y_i(0, 0, M_{1i}(t'), L_i(0, M_{1i}(t')), M_{2i}(t', 0, M_{1i}(t'), L_i(t', M_{1i}(t')))]
\end{aligned} \tag{14}$$

for $t \in \{0, 1\}$ and $t' = 1 - t$.

[3]Suppose that $t = 1$; then, we have

$$\begin{aligned}
\tau &= \delta^{M_1}(1, 1) + \delta^{M_2}(1, 0) + \zeta(0) \\
&= E[Y_i(1, 0, M_{1i}(1), M_{2i}(1, 0)) - Y_i(1, 0, M_{1i}(0), M_{2i}(1, 0)] \\
&\quad + E[Y_i(1, 0, M_{1i}(0), M_{2i}(1, 0)) - Y_i(1, 0, M_{1i}(0), M_{2i}(0, 0)] \\
&\quad + E[Y_i(1, 0, M_{1i}(0), M_{2i}(0, 0) - Y_i(0, 0, M_{1i}(0), M_{2i}(0, 0))] \\
&= E[Y_i(1, 0, M_{1i}(1), M_{2i}(1, 0)) - Y_i(0, 0, M_{1i}(0), M_{2i}(0, 0))] \\
&= E[Y_i(1, 0) - Y_i(0, 0]
\end{aligned} \tag{15}$$

This is the same when $t = 0$.

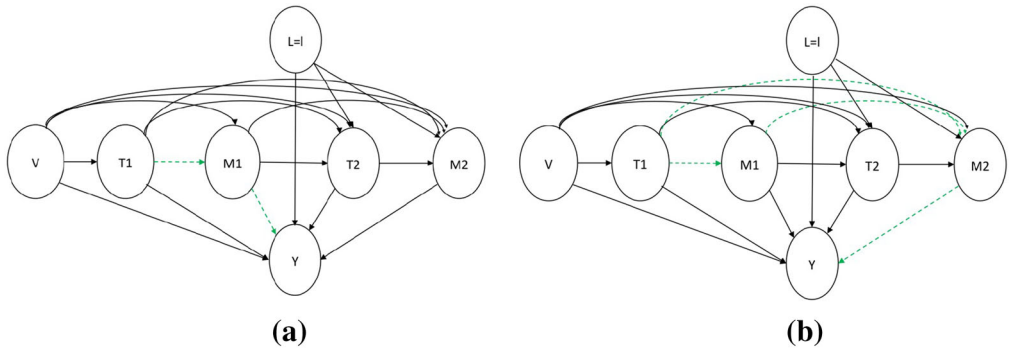

FIGURE 3.
Causal structural model highlighting the ACME-l. *Note* T_1: first treatment measurement, T_2: second treatment measurement, M_1: first mediator measurement, M_2: second mediator measurement, Y: outcome, V: pre-treatment covariates, and L: time-varying confounding variable. **a** ACME-l via M_1. **b** ACME-l via M_2

these path-specific effects in multiple mediators context shares some similar challenges such as a time-varying confounding or treatment-induced mediator and outcome confounding issue, which gives many implications to studies on time-varying treatments and mediators. Despite the similarity, a time-varying treatments and mediators case is often more complicated because of additional treatment variables and multiple time periods. Therefore, we propose alternative definitions that require less assumptions to be identified than these path-specific effects shown in Eq. (1).

3.2. *Alternative Definitions of ACME and ANDE*

We propose alternative definitions of the ACME and ANDE in which the intermediate math score is fixed to a specific value $L_i = l$ (ACME-l and ANDE-l) as below.

$$\begin{aligned}
\delta^{M_1}(t,l) &= E[Y_i(t,0,M_{1i}(1),M_{2i}(t,0,l),l) - Y_i(t,0,M_{1i}(0),M_{2i}(t,0,l),l)],\\
\delta^{M_2}(t,l) &= E[Y_i(t,0,M_{1i}(t'),M_{2i}(1,0,l),l) - Y_i(t,0,M_{1i}(t'),M_{2i}(0,0,l),l], \text{ and}\\
\zeta(t',l) &= E[Y_i(1,0,M_{1i}(t'),M_{2i}(t',0,l),l) - Y_i(0,0,M_{1i}(t'),M_{2i}(t',0,l),l)] \qquad (2)
\end{aligned}$$

for $t, t' \in \{0, 1\}$. The path-specific effects of $\delta^{M_1}(t,l)$, and $\delta^{M_2}(t,l)$ are highlighted in Fig. 3. The alternative definition of the ACME via M_1 ($\delta^{M_1}(t,l)$) is the effect of kindergarten retention on student math score transmitted by classmate math ability in year 2, after fixing the intermediate math score to $L_i = l$. The ACME-l via M_2 and ANDE-l are interpreted in the same manner as the counterparts shown in Sect. 3.1 while fixing the intermediate math score to $L_i = l$. These alternative definitions can be used both when L is continuous or categorical. With continuous L as in our example, we can fix L to a quantile, depending on researcher's judgment. With categorical L, we can fix L to each category.

Holding L fixed should be interpreted as hypothetical intervention, not as conditioning or adjusting on the variable (Pearl, 2009). This implies that the values of L are no longer determined by its parent variables (previously measured variables that affect L), and thus, no mediating effects are transmitted via L (see Fig. 3). As a result, the mediation effects defined in Eqs. (1) and (2) actually differ in terms of variables through which the effects are transmitted. For example, the ACME via M_2 ($\delta^{M_2}(t)$) includes the mediation effect of $T_1 \to L \to M_2 \to Y$ and $T_1 \to M_1 \to L \to M_2 \to Y$, whereas the ACME-$l$ via M_2 ($\delta^{M_2}(t,l)$) does not include these paths because,

after fixing L, its parent variables T_1 and M_1 no longer affect L (the corresponding arrows are gone).

The combined effect of ACME-l and ANDE-l can be interpreted as a controlled direct effect if T_1, L, and Y represent treatment, mediator, and outcome variables, respectively. A controlled direct effect represents how much math scores at year 4 would change if math scores at year 2 were fixed at level $L_i = l$ in the entire population, but the kindergarten retention status was changed from not retained to retained. In formal expression, $\tau(l) = E[Y_i(1, 0, l) - Y_i(0, 0, l)]$. This controlled direct effect for a fixed $L = l$ is further decomposed into the ACME-l and ANDE-l. The usefulness of this controlled direct effect becomes evident with the following example. Suppose that grouping with younger kids has a side effect of lowering self-esteem of retainees, and hence, intermediate confounding variables include self-esteem. The interest now is focused on whether kindergarten retention has an effect on the final math score, regardless of the indirect effects of kindergarten retention might have on the final math score by way of deteriorating self-esteem. It is because this controlled direct effect represents a stable kindergarten retention effect that is invariant to personal and social factors (such as gender, socioeconomic status, parent support, etc.) that may affect self-esteem.

Likewise, the controlled direct effect (and subsequently, the ACME-l and ANDE-l) with our example, in which the intermediate confounding variable is the math score, allows us to learn about the mediating process by removing the effect transmitted through the intermediate math score. The alternative mediation effect (ACME-l) is now focused on whether kindergarten retention has an effect on the final math score via interacting with classmates, regardless of what indirect effects kindergarten retention might have via the intermediate math score. This removed indirect effect is the average kindergarten retention effect (T_1) on the intermediate math score (L) either directly or mediated by classmate math ability (M_1), which can be obtained from a single-time-period causal mediation analysis, being prolonged to the next time period (e.g., $T_1 \rightarrow M_1 \rightarrow L \rightarrow Y$ for δ^{M_1} and $T_1 \rightarrow L \rightarrow M_2 \rightarrow Y$ for δ^{M_2}). Hence, the alternative mediation effect can be regarded as the mediation effect that has emerged one year after kindergarten retention, considering that this alternative mediation effect was never captured in the single-time-period analysis (either by means of computing average direct effects or average mediation effects). This alternative definition is particularly informative in our case by identifying the effects that have emerged one year after kindergarten retention.

4. Identification and Sensitivity Analysis Under Homogeneous Effects

This section presents identification results for the alternative definitions of ACME and ANDE and provides a sensitivity analysis when effects across individuals are assumed to be constant. As previously discussed, the existence of intermediate confounding variables prevents identification of the ACME and ANDE. A simple solution would be to assume the absence of intermediate confounding, but this assumption is frequently unrealistic in practice. Therefore, recent studies on a time-varying treatments and mediators model attempted to solve this intermediate confounding issue. For example, VanderWeele and Tchetgen Tchetgen (2016) proposed the use of randomized interventional analogues of natural direct and indirect effects that are identified even when intermediate confounding is present (see "Appendix A" for more discussion on the randomized interventional analogues). Another approach, which was proposed by Bind, Vanderweele, Coull, and Schwartz (2016), is to identify the ACME with exogenous treatments using generalized mixed-effects models. The model proposed by Bind and her colleagues requires the absence of treatment–mediator confounding; this absence can be achieved by using exogenous treatments or sequential randomization of treatments. Their proposed approach would not be applicable in our

example, where the assumption is not met. Therefore, we focus only on ACME-l and ANDE-l instead of ACME and ANDE, respectively.

For the identification of ACME-l and ANDE-l shown in Fig. 3, we apply the approach suggested by Daniel et al. (2015), which introduces a set of sensitivity parameters under homogeneous effects to facilitate identification. First, the following sequential ignorability assumption is required to identify ACME-l and ANDE-l, up to a sensitivity parameter.

$$\begin{aligned}
&A1.\ \{M_{1i}(t), M_{2i}(t,0,m_1,l), Y_i(t,0,m_1,m_2,l)\} \perp T_{1i}|V_i = v,\\
&A2.\ \{M_{2i}(t,0,m_1,l), Y_i(t,0,m_1,m_2,l)\} \perp T_{2i}|T_{1i} = t, M_{i1} = m_1, V_i = v, L_i = l,\\
&A3.\ \{M_{2i}(t,0,m_1,l), Y_i(t,0,m_1,m_2,l)\} \perp M_{1i}|T_{1i} = t, V_i = v, \text{and}\\
&\quad Y_i(t,0,m_1,m_2,l) \perp M_{2i}|T_{1i} = t, T_{2i} = 0, M_{1i} = m_1, V_i = v, L_i = l,
\end{aligned} \tag{3}$$

for any value of t, m_1, m_2, l, and v. In addition, we assume consistency and that L should include all time-varying confounding variables. Assumption A1 states that kindergarten retention status is ignorable given pre-treatment covariates. Assumption A2 states that student retention status in year 3 is ignorable given the observed first treatment status, first mediator value, pre-treatment covariates, and intermediate confounding variable.

The first line of assumption A3 states that there is no unmeasured confounding in the $M_1 - Y$ and $M_1 - M_2$ relationships. In our example, this implies that there is no confounding in the relationships between classmate math score in year 2 and final math score ($M_1 - Y$) and between classmate math score in years 2 and 3 ($M_1 - M_2$), given the first treatment and pre-treatment covariates. The second line of assumption A3 states that there is no unmeasured confounding between the $M_2 - Y$ relationship. This implies that there is no confounding in the relationship between classmate math score in year 3 and final math score given the treatment and mediator history, pre-treatment covariates, and intermediate confounding variable.

Under this sequential ignorability assumption, $E[Y(t,0,M_1(t'),M_2(t'',0,M_1(t''),l),l)|v]$ is identified up to a sensitivity parameter as follows.

$$\begin{aligned}
&E[Y(t,0,M_1(t'),M_2(t'',0,l),l)|v]\\
&= \sum_{m_2}\sum_{m_1}\sum_{m_1'} E(Y_i|t,0,m_1,m_2,l,v)\cdot P(M_{2i}=m_2|t'',0,m_1',l,v)\\
&\quad \cdot \boxed{P(M_{1i}(t')=m_1|M_{1i}(t'')=m_1',v)}\cdot P(M_{1i}=m_1|t'',v),
\end{aligned} \tag{4}$$

for every m_1, m_1', m_2, v, and l where $t, t', t'' \in \{0,1\}$. The proof is given in "Appendices B and C." The boxed quantity in Eq. (4) is not identified because the conditional correlation between $M_{1i}(t')$ and $M_{1i}(t'')$ given $V_i = v$, which is denoted as ρ_{M_1}, is not known. The issue here is that this correlation coefficient is not empirically determined because we never observe individual potential outcomes under different treatment statuses simultaneously. Therefore, this correlation is used as a sensitivity parameter that ranges from 0 (no correlation across two potential outcomes) to 1 (perfect correlation). Negative correlations are not considered because it is less likely that one potential mediator tends to decrease when the other potential mediator increases (or vice versa) after taking different treatment statuses and covariates into account.

We can estimate the ACME-l and ANDE-l for a fixed value of ρ_{M_1} using g-computation as below.

1. Fit regressions for mediators and outcome models.

2. Generate $M_{1i}(t)$ for $t \in \{0, 1\}$ using predicted values of the mediator model of M_1. The errors for $M_{1i}(1)$ and $M_{1i}(0)$ are obtained by randomly drawing from mean of zero and the covariance matrix with a fixed correlation value of ρ_{M_1}.
3. Generate $M_{2i}(t, 0, l)$ for $t \in \{0, 1\}$, incorporating the predicted values of the mediator model of M_1, which are obtained from step 2. Here, we set $L_i = l$ for every individual rather than incorporating the predicted values of L.
4. Generate $Y(t, 0, M_{1i}(t'), M_{2i}(t'', 0, l), l)$ for $t, t', t'' \in \{0, 1\}$, incorporating the results obtained from steps 2 and 3. Again, we put $L_i = l$ for every individual, rather than incorporating the predicted values of L.
5. Using the potential outcomes obtained from step 4, the ACME-l and ANDE-l are estimated by averaging over individuals. Calculate 95% confidence intervals using the bootstrap.

This g-computation approach is flexible in that the ACME-l and ANDE-l are estimable under both parametric and nonparametric settings. In a parametric setting, the sensitivity parameter, ρ_{M_1}, is only required when interaction effects exist between two mediators (M_1 and M_2) with respect to the outcome. To see this, suppose that we have the following simple data-generating model for two potential mediators under different treatment statuses: $M_1(t') = \alpha_0 + \alpha_1 t' + U_{1i}(t')$, and $M_2(t, 0, M_1(t), l) = \gamma_0 + \gamma_1 t + \gamma_2 M_1(t) + \gamma_3 l + U_{2i}(t)$, where α and γ are constant regression coefficients, and U_i is individual error. For notational simplicity, we did not include pre-treatment covariates and interaction effects between the treatment and the mediator in this data-generating model. However, the result does not change even after including pre-treatment covariates and interaction effects between the treatment and the mediator. Under this data-generating model, the interaction between the two potential mediators ($M_1(t') \times M_2(t, 0, M_1(t), l)$) with respect to the outcome requires a correlation between $U_{1i}(t')$ and $U_{1i}(t)$ in order to be uniquely identified. This implies that sensitivity analysis is unnecessary under a parametric setting if one can assume homogeneous effects and no-interaction effects between two mediators with respect to the outcome.

5. Bias Formula Under Heterogeneous Effects

In this section, we relax the assumption that every subject has the same constant effect and allow for systematic or random variations of the effects. Specifically, we introduce how to accommodate additional bias that emerges under heterogeneous effects by means of sensitivity analysis, which was used in Imai and Yamamoto (2013) in the context of multiple mediators.

We begin by presenting the parametric data-generating model with heterogeneous effects, since this sensitivity analysis accommodating additional bias is model-specific. This means that sensitivity analysis may no longer apply if the data-generating model was different than what was proposed here. The causal diagram in Fig. 1 can be expressed as shown below, using the following linear structural equations with varying coefficients.

$$\begin{aligned} M_1 &= \alpha_{0i} + \alpha_{1i} T_1 + \alpha_{2i} V + U_{1i} \\ M_2 &= \gamma_{0i} + \gamma_{1i} T_1 + \gamma_{2i} M_1 + \gamma_{3i} L + \gamma_{4i} T_2 + \gamma_{5i} T_1 M_1 + \gamma_{6i} V + U_{3i} \\ Y &= e_{0i} + e_{1i} T_1 + e_{2i} M_1 + e_{3i} L + e_{4i} T_2 + e_{5i} M_2 + e_{6i} T_1 M_1 + e_{7i} T_1 M_2 \\ &\quad + e_{8i} M_1 L + e_{9i} M_2 L + e_{10i} V + U_{4i} \end{aligned} \tag{5}$$

where α_i, γ_i, and e_i are structural coefficients, and U_i is an exogenous error for individual i. The structural coefficients' subscript i indicates that the effects are *heterogeneous* across individuals, and thus, each varying coefficient represents a random variable, rather than a constant

value. While this model is flexible enough to accommodate individual heterogeneous effects and various interaction effects (in the $T - M$ and $M - L$ relationships), we assume no interactions in the mediator–mediator relationship, as in Imai and Yamamoto (2013). Including the mediator–mediator interactions while addressing effect heterogeneity is not impossible, but it increases the complexity of the bias formulas.

Suppose now that we are interested in the ACME-l and ANDE-l and that we do not assume effect homogeneity. The bias is defined as the difference between the average effect estimate under the homogenous effects and the true effect under heterogeneous effects. In formal expression, $bias(\hat{\delta}^{M_1}(t,l)) = E[\hat{\delta}^{M_1}(t,l)] - \delta^{M_1}(t,l)$ where $E[\hat{\delta}^{M_1}(t,l)]$ is the expected estimate obtained as a result of assuming homogeneous effects, given assumptions A1–A3. Now, allowing for heterogeneous effects, the bias formulas for $\hat{\delta}_{M_1}(t,l)$, $\hat{\delta}_{M_2}(t,l)$, and $\hat{\zeta}(t',l)$ are given as follows.

$$\begin{aligned} bias(\hat{\delta}^{M_1}(t,l)) &= 0, \\ bias(\hat{\delta}^{M_2}(t,l)) &= \rho_{M_2}\sigma_{e_7}\sqrt{V(M_{2i}|T_{1i}=t', T_{2i}=0, L_i=l)}, \text{ and} \\ bias(\hat{\zeta}(t',l)) &= -\rho_{M_2}\sigma_{e_7}\sqrt{V(M_{2i}|T_{1i}=t', T_{2i}=0, L_i=l)}, \end{aligned} \tag{6}$$

for $t \in \{0,1\}$, $t = 1 - t'$ where $\sigma_{e_7} = \sqrt{V(e_{7i})}$, and ρ_{M_2} indicates the correlation between e_{7i} and $M_{2i}(t',0,l)$. A proof is shown in "Appendix D." This bias formula indicates that the ACME-l via M_1 is not affected by the effect heterogeneity. In contrast, the ACME-l via M_2 and ANDE-l will be biased if the effects are heterogeneous across individuals. The bias originates from these two random terms: 1) interaction effects between treatment and mediator across individuals (e_{7i}), and 2) the potential value of the second mediator ($M_{2i}(t',0,l)$). The standard deviation of the potential second mediator can be consistently estimated by the sample counterparts, while the standard deviation of the varying coefficient (σ_{e_7}) and the correlation between the two random terms (i.e., $M_{2i}(t',0,l)$ and e_{7i}) are not empirically known. Therefore, these two unknown terms will serve as sensitivity parameters. The correlation indicates the *direction of the bias*, which determines whether the bias is upward or downward; and the standard deviation of the varying coefficient represents the amount of *heterogeneity* across individuals. Using these sensitivity parameters, we can examine the sensitivity of effect estimates in response to the change in these two unknown sensitivity parameters. Specifically, we can obtain the upper and lower bounds of $\delta^{M_2}(t,l)$ or $\zeta(t',l)$ for every assumed value of σ_{e_7} because the correlation ρ_{M_2} is bounded to have values between -1 to 1. For example, the bias of $\hat{\delta}^{M_2}(t,l)$ falls between $0.5\sigma_{e_7}$ and $-0.5\sigma_{e_7}$ for a fixed value of σ_{e_7} if the standard deviation of the potential second mediator ($\sqrt{V(M_{2i}|T_{1i}=t', T_{2i}=0, L_i=l)}$) was 0.5, which can be obtained from a sample.

From Eq. (6), we realize that the bias would have been zero if either $\rho_{M_2} = 0$ (no correlation between e_{7i} and $M_{2i}(t',0,l)$) or $\sigma_{e_7} = 0$ (a constant effect of e_7). This indicates that the ACME-l and ANDE-l are identified without sensitivity parameters if we can assume no-interaction effects in the $T_1 - M_2$ relationship (in addition to assuming no interactions in the $M_1 - M_2$ relationship) with respect to the outcome under heterogeneous effects. If the no-interaction-effects assumption is satisfied, the ACME-l via M_2 is identified as $\delta^{M_2}(0,l) = \delta^{M_2}(1,l) = E[(e_{5i} + le_{9i}) \times (\gamma_{1i} + \alpha_{1i}(\gamma_{2i} + \gamma_{5i}))]$, using the product of coefficient approach.

One attractive feature of this alternative definition of ACME-l and ANDE-l is the generalizability to a model beyond two time periods. The bias formula for the ACME-l via M_k (where $k \in K$ denotes a number of time periods) remains the same even when the model is extended beyond two time periods if the model follows the same model specification as in Eq. (5). See "Appendix E" for extending this bias formulas to a model beyond two time periods.

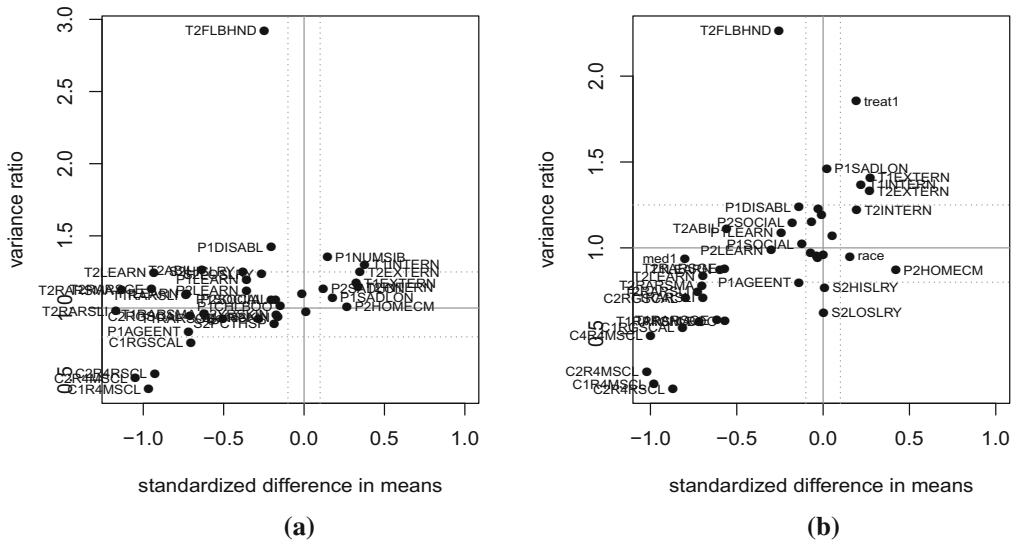

FIGURE 4.
Covariate imbalance by retention status before weighting. *Note* Standardized difference in means: standardized difference in covariate means, and variance ratio: covariates' variance ratio between treated and controlled units. **a** Year 1. **b** Year 3

6. Case Study

We illustrate the use of our proposed approach under homogeneous and heterogeneous effects with an empirical study about the mechanisms, underlying the relationship between kindergarten retention policy and final math score.

6.1. Assumptions and Models

As previously discussed, assumptions A1 and A2 in Eq. (3) are satisfied if the two retention statuses are sequentially randomized. Unfortunately, our data are non-experimental because researchers do not have control over retention decisions. Although we used an extensive set of pre-treatment covariates as shown in Sect. 2, the lack of control makes our study more sensitive to unobserved confounding. We carefully checked covariate balance and overlap between the treated and controlled units. As shown in Fig. 4, covariate distributions between those who were retained and promoted are substantially different. The left and right plots of Fig. 4 indicate covariate imbalance by retention statuses measured in year 1 and 3, respectively. The X axis represents the standardized difference in covariate means, and the Y axis represents the covariates' variance ratio between treated and controlled units. Covariate balance would be given if the standardized mean difference were zero or very close to zero, and if the variance ratio were close to one. Figure 5 assesses the overlap between the treatment and control group with respect to the distribution of the propensity score logit. Again, the left and right overlap plot refers to year 1 and year 3, respectively. The figures show that the treatment group is more likely to be retained than the control group and this leads to a lack of overlap on both tails of the distributions.

In fact, the proportion of students who are retained in year 1 and in year 3 is approximately 3%, which demonstrates that retention is a highly selective process (Hong & Raudenbush, 2006). The overlap between retained and promoted students is thus weak, as shown in Fig. 4. As a result, causal inference for areas of non-overlap would require extrapolation, which restricts the credibility of conclusions (Imbens & Rubin, 2015). To circumvent this issue, we trimmed our data

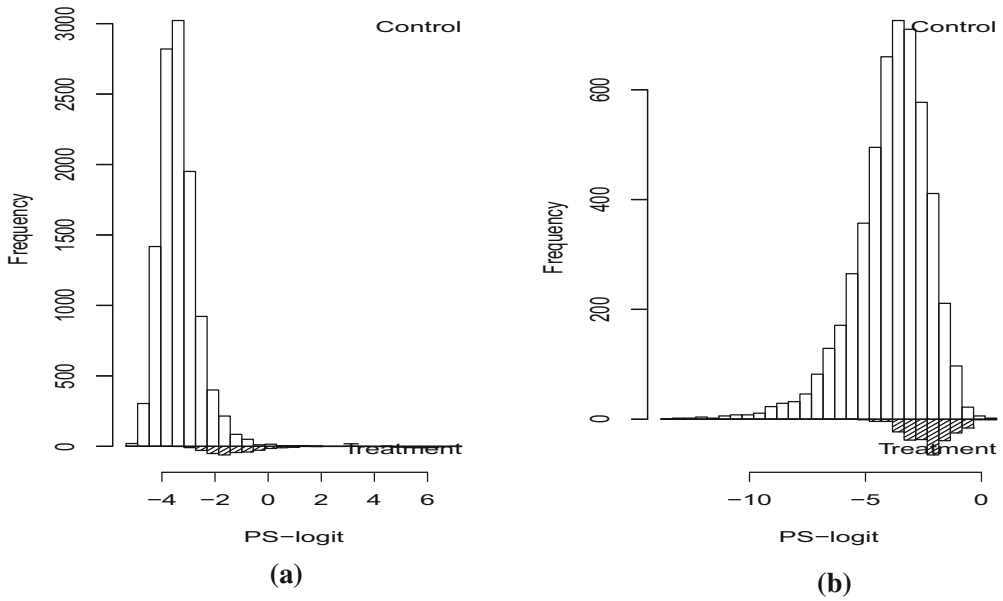

FIGURE 5.
Distributions of PS-Logit before trimming. *Note* PS-logit: The logit value of propensity scores. **a** Year 1, **b** Year 3

based on the logit of propensity scores to achieve overlap and better covariate balance between the two groups. We trimmed observations that were outside of the limits constructed based on the logit of propensity scores: $(\min - 0.05\sigma_{LPS})$ and $(\max + 0.05\sigma_{\text{LPS}})$, where min is the maximum of the treatment and control group's minima, max is the corresponding minimum of the two maxima, and σ_{LPS} is the standard deviation of the logit of propensity scores.

After trimming the data, we re-estimated the propensity scores (PSs) based on the remaining 286 kindergarten retainees and 4143 promoters. We obtained PS weights using the CBPS R package (Fong, Ratkovic, & Imai, 2014). By means of generalized method of moments, the PS is estimated such that it maximizes the covariate balance as well as the prediction of the treatment assignment (Imai & Ratkovic, 2015). After weighting, the covariate balance substantially improved, as shown in Fig. 6. In year 1, all standardized mean differences are below the threshold of 0.25 except for one variable (percentage of students retained at school). In year 3, eight out of 42 variables had a difference larger than the threshold of 0.25, but none are larger than the threshold of 0.5. Figure 7 also shows the overlap between the treatment and control groups based on the trimmed data. In fact, trimming data is not only helpful to achieve overlap between the two groups, but it also restricts the analytical sample to children who are more homogeneous in terms of their retention probabilities. After trimming, the retention probability at kindergarten and in year 3 ranged from 3 to 98% and from 1 to 61%, respectively.

Regarding assumption A3, we calculated Pearson's correlation between classmate ability and each covariate to examine whether there is a systematic relationship between them. From Figs. 8 and 9, it is clear that the correlations substantially reduce after weighting. In addition to A1–A3, we assume that L includes all exposure-induced confounding variables. In the context of our example this assumption implies that student math achievement at year 2 is the only time-varying confounding variable. This assumption is probably not very realistic and, unfortunately, not directly testable.

Given these assumptions, we use the same causal structural models as in Eq. (5) for mediators, time-varying confounder, and outcome models. Although the same regression models were used

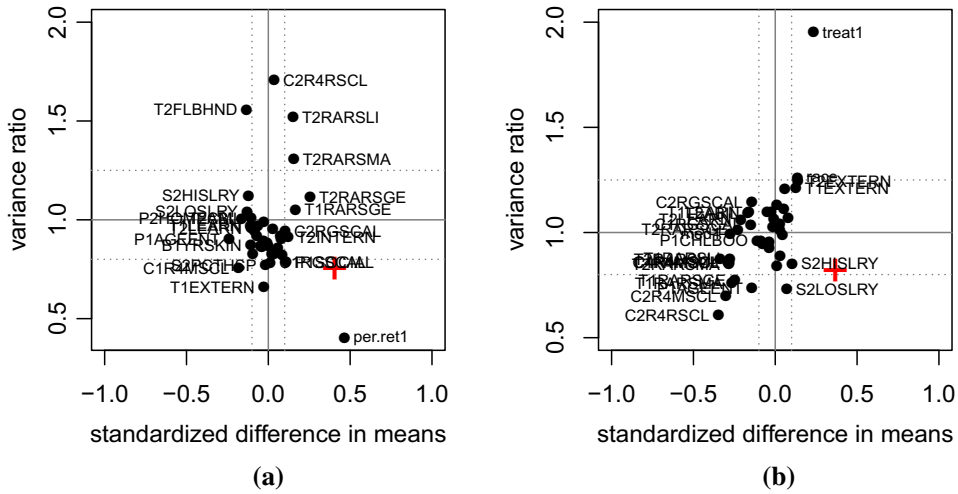

FIGURE 6.
Covariate balance by retention status after weighting. *Note* Standardized difference in means: standardized difference in covariate means, and variance ratio: covariates' variance ratio between treated and controlled units. **a** Year 1. **b** Year 3

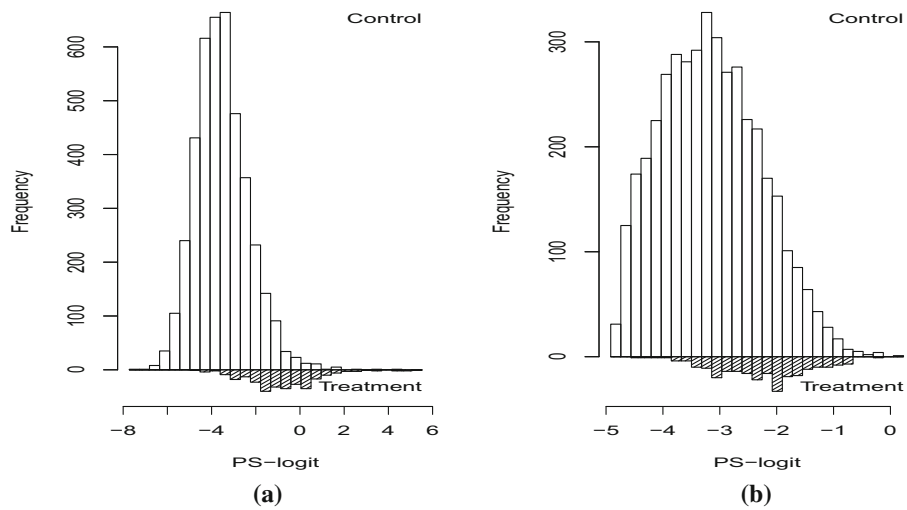

FIGURE 7.
Distributions of PS-Logit after trimming. *Note* PS-logit: The logit value of propensity scores. **a** Year 1. **b** Year 3

as shown in Eq. (5) in this article, researchers in practice need to make their own judgments about which interactions to include in their model. The impact of including interactions may differ depending on whether the effects are assumed to be heterogeneous or homogenous. If the effects are homogenous the approach of choosing a parsimonious model is reasonable (e.g., Daniel et al. (2015)). However, this approach of choosing a parsimonious model may not be adequate to address heterogeneity. This is because the interaction effects may vary across individuals (and thus not zero) even when the interaction effect is not significantly different from zero on average. Therefore, a decision to include interaction effects should be based on theoretical knowledge

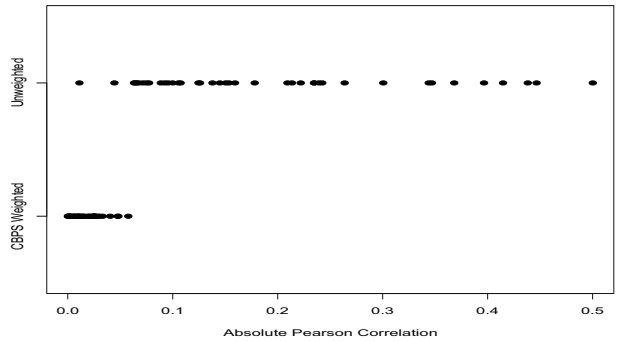

FIGURE 8.
Pearson correlations with the first mediator before and after weighting.

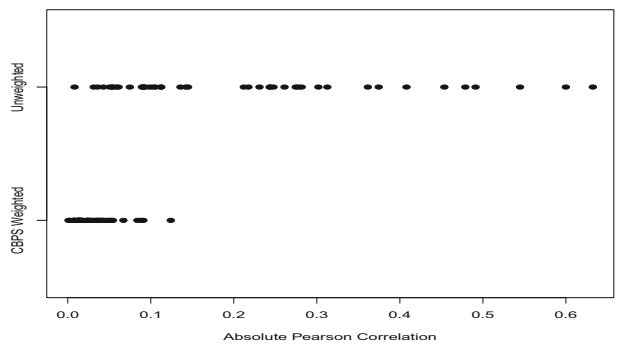

FIGURE 9.
Pearson correlations with the second mediator before and after weighting.

rather than relying only on the statistical significance since the more flexible model addresses the heterogeneity better.

The models for the first mediator, intermediate confounder, second mediator, and outcome are weighted by w_1, $w_1 \times w_2$, $w_1 \times w_2 \times w_3$, and $w_1 \times w_2 \times w_3 \times w_4$,[4] respectively, where

$$
\begin{aligned}
w_1 &= \frac{T_{1i} \times P(T_{1i} = t)}{P(T_{1i} = t|V_i)} + \frac{(1 - T_{1i}) \times (1 - P(T_{1i} = t))}{1 - P(T_{1i} = t|V_i)}, \\
w_2 &= \frac{f(M_{1i})}{f(M_{1i}|T_{1i}, V_i)}, \\
w_3 &= \frac{T_{2i} \times P(T_{2i} = t|T_{1i})}{P(T_{2i} = t|T_i, M_{1i}, L_i, V_i)} + \frac{(1 - T_{2i}) \times (1 - P(T_{2i} = t|T_{1i}))}{1 - P(T_{2i} = t|T_i, M_{1i}, L_i, V_i)}, \text{ and} \\
w_4 &= \frac{f(M_{2i})}{f(M_{2i}|T_{1i}, M_{1i}, L_i, T_{2i}, V_i)},
\end{aligned}
\tag{7}
$$

where $f(\cdot|\cdot)$ is the conditional density function of its arguments. After fitting the regressions incorporating weights, the ACME-l and ANDE-l are estimated by using the g-computation steps introduced in Sect. 4. We fixed the intermediate math scores to the 25, 50 and 75 percentiles, which represent low, medium, and high achievement conditions, respectively.

[4]We used the stabilized weight for W1 and W3 to avoid extreme weights as suggested by Robins, Hernan, and Brumback (2000).

TABLE 3.
Estimates for ACME and ANDE with the one-time period.

	Est.	95%	C.I.	P value
$\hat{\delta}(1)$	− 5.264	− 7.049	− 3.538	0.00
$\hat{\zeta}(1)$	− 2.013	− 3.570	− 0.397	0.01
$\hat{\delta}(0)$	− 5.736	− 6.585	− 4.967	0.00
$\hat{\zeta}(0)$	− 2.485	− 4.745	− 0.187	0.03
$\hat{\tau}$	− 7.749	− 9.302	− 5.935	0.00

Est. = estimate, and 95% C.I. = 95% confidence interval.

TABLE 4.
Estimates for ACME-l and ANDE-l with two time periods, by different level of L.

	L = low			L = med			L = high		
	Est.	95%	C.I.	Est.	95%	C.I.	Est.	95%	C.I.
$\hat{\delta}^{M_1}(1,l)$	2.087	− 0.694	4.749	1.681	− 1.012	4.542	1.288	− 1.389	4.087
$\hat{\delta}^{M_2}(1,l)$	− 2.371	− 4.944	− 0.217	− 2.467	− 5.188	− 0.161	− 2.548	− 5.250	0.154
$\hat{\zeta}(1,l)$	0.899	− 3.264	4.701	1.057	− 2.945	4.978	1.063	− 2.744	4.707
$\hat{\delta}^{M_1}(0,l)$	4.366	3.650	5.113	3.981	3.277	4.686	3.632	2.934	4.381
$\hat{\delta}^{M_2}(0,l)$	− 2.117	− 4.292	− 0.199	− 2.234	−4.476	− 0.139	− 2.327	− 4.557	0.125
$\hat{\zeta}(0,l)$	3.431	0.451	6.178	3.589	0.847	6.344	3.629	0.877	6.417
$\hat{\tau}(l)$	3.147	− 1.562	7.253	2.804	−2.013	7.155	2.368	− 2.167	6.764

Est. = estimate, and 95% C.I. = 95% confidence interval.

6.2. *Results*

In response to the first causal question given in the introduction, we begin by examining the results of the single-time-period causal mediation model using the intermediate math score as the outcome variable. Consistent with the previous studies, there is a negative effect of kindergarten retention on math score. Table 3 shows that students would have scored 7.7 points higher if they were promoted instead of retained, which is consistent with Hong and Raudenbush (2006). In addition, a large fraction of the negative effect is via the ability level of classmates during the first year after kindergarten retention. The retained students would have scored 5.3 points higher had they been interacting with same-age peers. Promoted students would have scored 5.7 points lower had they been interacting with one-year-younger peers.

Now our question centers on whether the emerging mediation effect via classmate math ability is still negative one year after the kindergarten retention (second and third causal questions). In order to answer this question, we computed effect estimates for ACME-l and ANDE-l, which are shown in Table 4. From the left hand side, we present estimates, and upper and lower 95% confidence intervals for L = low and the same quantities for L = medium and L = high. Interestingly, the results indicate that some emerging mediation effects via classmate math ability became positive after 1 year. The mediation effect via classmate math ability in year 1 ($\delta^{M_1}(0,l)$) for promoted low achievers is 4.37. This implies that the kindergarten retention effect on math score via classmate effects is strongly negative in the short term (1 year). However, a positive mediation effect appears in the long term (3 years), which was not captured in the single-time-period analysis. For both retained and promoted conditions, the positive effect would be stronger if students had a low intermediate math score as compared to a high score.

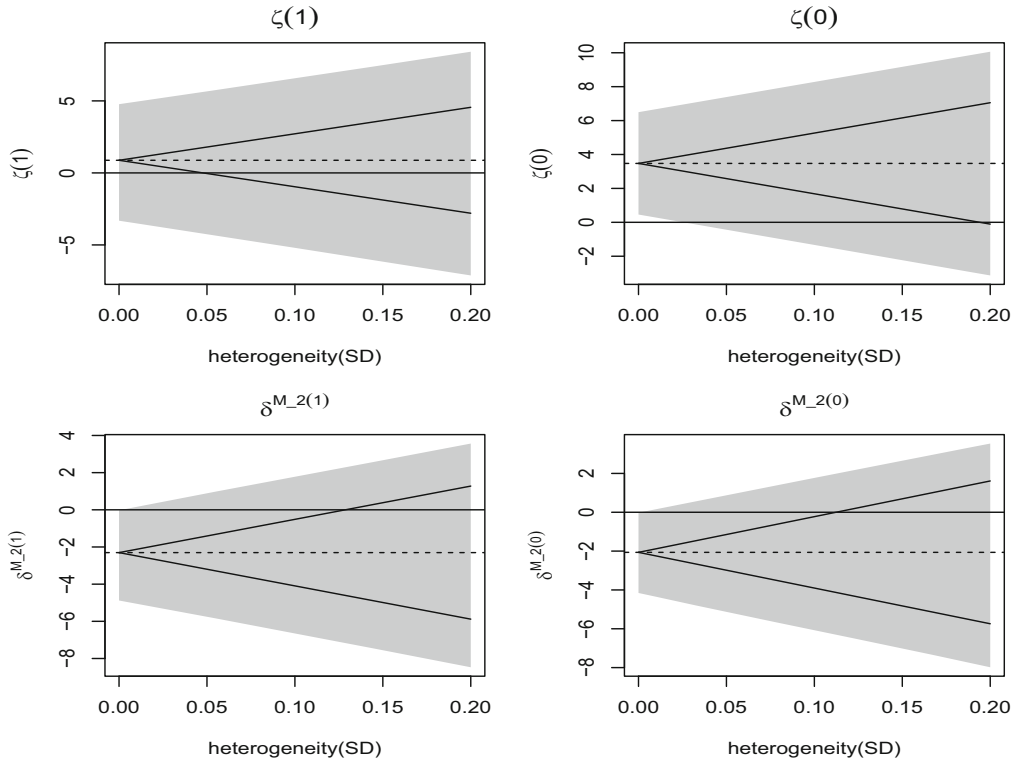

FIGURE 10.
Sensitivity analysis when L = low.

The results also indicate that a negative mediation effect of kindergarten retention through the ability level of classmates in year 4 ($\delta^{M_2}(0, l)$) persists even after removing the effect via the intermediate math score, although the magnitude of the effect is smaller. The negative mediation effects vary between -2.37 and -2.55 for the retained condition and between -2.12 and -2.33 for the promoted condition. For both retained and promoted conditions, the negative effect would be stronger if students had a high intermediate math score as compared to a low score.

Allowing for heterogeneous effects, our sensitivity analysis suggests that the results on the ANDE-l and the ACME-l via M_2 are fairly responsive to the amount of heterogeneity across individuals. Figures 10 and 11 present sensitivity plots when the intermediate math score is fixed to L = low and L = high, respectively. The X axis indicates the amount of heterogeneity (SD of the varying coefficient on the treatment–mediator interaction), and the Y axis indicates the upper and lower bounds of the estimates. A dotted line indicates the estimate when homogenous effects can be assumed. As the amount of heterogeneity increases, the bias becomes larger. The 95% confidence intervals are shown in gray. About 0.20 SD and 0.12 SD of effect heterogeneity can change the sign of the estimates of the ANDE-l and the ACME-l via M_2 for the promoted condition.

7. Conclusion

In this article we proposed some extensions to causal mediation analysis with time-varying treatments and mediators and discussed their application in investigating the mechanisms under-

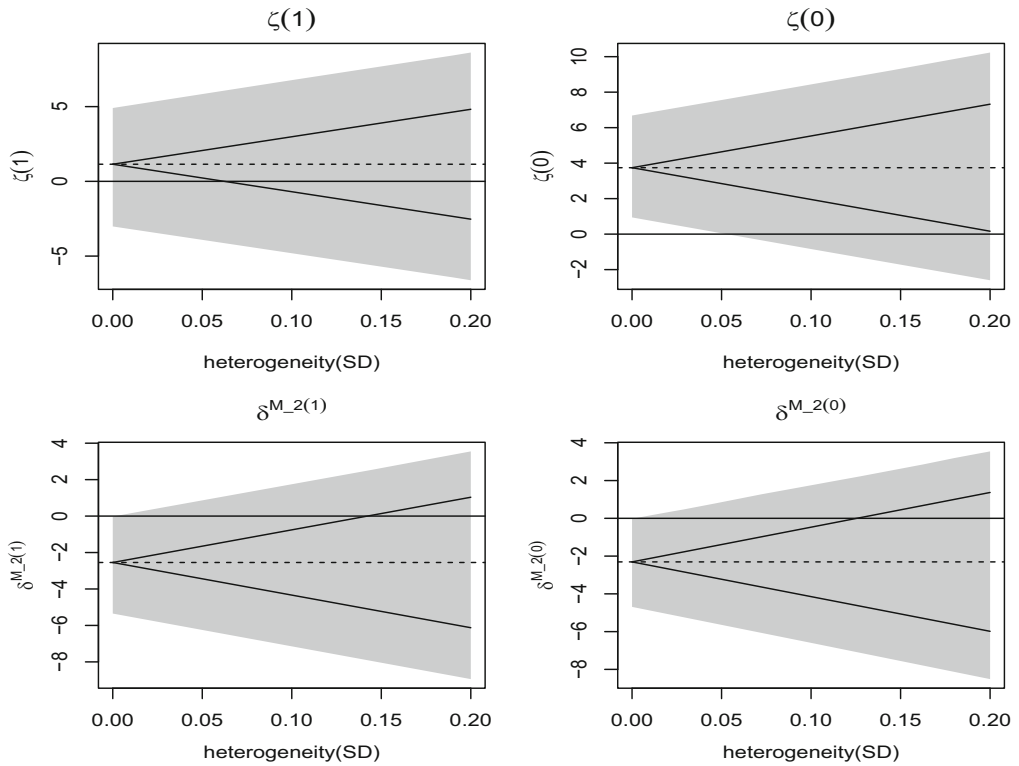

FIGURE 11.
Sensitivity analysis when L = high. *Note* X axis: standard deviation of varying coefficients on the $T - M_2$ interaction, Y axis: upper and lower bounds of effect estimates, and gray area: 95% confidence intervals of effect estimates.

lying the relationship between kindergarten retention and student math achievement scores. To overcome the limitations of existing causal mediation analyses with longitudinal data, we discussed a partial identification strategy (following Daniel et al. 2015) that identifies the mediated and natural direct effects up to some sensitivity parameters. In a sensitivity analysis the researcher can then assess the effects' robustness to a range of possible values of the unknown sensitivity parameters.

In addition, we discussed alternative definitions for ACME and ANDE for fixed values of the confounding intermediate variable. The alternative definitions fix the intermediate confounding variable at a certain value and thus investigate the mediation effect where the indirect effect via the intermediate confounding variable is blocked. That is, we only evaluate a partial mediation effect, the effect that is mediated by paths other than through the intermediate confounding variable. Evaluating the alternative ACME-l and ANDE-l at different values for the intermediate variable also allows for a useful investigation of effect heterogeneity. Though we cannot estimate the overall ACME and ANDE, we believe that the alternative versions of the effects are still valuable because they allow us to better assess at least a part of the mediated effect. As before, the alternative ACME-l and ANDE-l are only identified up to unknown sensitivity parameters. Corresponding sensitivity analyses can be used to assess the robustness of estimates under both homogeneous and heterogeneous effects.

A drawback of the suggested mediation analyses is that they require the specification of unknown sensitivity parameters. However, if the effect estimates are rather insensitive to a broad range of parameter settings, valid conclusions about the direction of the effects are still possible.

Another limitation of the proposed analyses is that they allow for a single intermediate confounder only (as is also the case for the analyses suggested by Daniel et al. 2015). For the alternative ACME and ANDE this limitation can be overcome by setting the entire vector of intermediate confounders to fixed values, but as more mediating pathways via intermediate confounding variables are blocked, partially identified effects become less meaningful. Thus, in practice researchers face a trade-off between restricting the meaningfulness of the alternative ACME and ANDE and making strong, probably implausible assumptions about the presence of intermediate confounders.

Acknowledgments

This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation under Grant #DRL-0941014. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies.

Appendix

A. *An Aside: Randomized Interventional Analogues*

VanderWeele and Tchetgen Tchetgen (2016) proposed the use of randomized interventional analogues of ACME and ANDE, which date back to Didelez, Dawid, and Geneletti (2012) and Geneletti (2007). Rather than fixing the mediators at values they would have had for each individual under a particular treatment status, the randomized analogues fix the mediators to values randomly drawn from the mediator distribution, given a particular treatment status. Their approach preserves the distribution of mediators, but at the individual level, the values are randomly assigned from the mediator distribution. As the authors argued in their work, these randomized analogues are useful and even of greater interest than the ACME and ANDE in some scenarios. However, one potential issue is that it is usually unknown how much of the true ACME is carried over to the randomized interventional analogues of the ACME due to its randomness. In contrast, our alternative definition does not include the indirect effect via the intermediate confounding variable (intermediate math score), but the remaining mediation effect provides a meaningful interpretation—that is, the mediation effects that have emerged over time.

B. *Implications of Having an Intermediate Confounding Variable on the Identification Assumptions*

We follow the same logic as Daniel et al. (2015). Suppose that our data are generated non-parametrically as,

$$\begin{aligned} M_1 &= g_{M_1}(V, T_1, U_{M_1}) \\ L &= g_L(V, T_1, M_1, U_L) \\ M_2 &= g_{M_2}(V, T_1, T_2, M_1, L, U_{M_2}) \\ Y &= g_Y(V, T_1, T_2, M_1, M_2, L, U_Y) \end{aligned} \tag{8}$$

where $g(\cdot)$ is a deterministic function and $\{U_{M_1}, U_{M_2}, U_L, U_Y\}$ are mutually independent.

Then, the potential outcomes derived from the above data-generating models are shown as below.

$$\begin{aligned}
M_1(t) &= g_{M_1}(V, t, U_{M_1}) \\
M_1(t') &= g_{M_1}(V, t', U_{M_1}) \\
L(t, m_1) &= g_L(V, t, m_1, U_L) \\
L(t', m_1') &= g_L(V, t', m_1', U_L) \\
M_2(t, 0, m_1) &= g_{M_2}(V, t, 0, m_1, g_L(V, t, m_1, U_L), U_{M_2}) \\
M_2(t', 0, m_1') &= g_{M_2}(V, t', 0, m_1', g_L(V, t', m_1', U_L), U_{M_2}) \\
Y(t, 0, m_1, m_2) &= g_Y(V, t, 0, m_1, m_2, g_L(V, t, m_1, U_L), U_Y) \\
Y(t, 0, m_1, m_2) &= g_Y(V, t, 0, m_1, m_2, g_L(V, t, m_1, U_L), U_Y)
\end{aligned} \tag{9}$$

Consider the following assumptions that are needed to identify ACME and ANDE.

$$\begin{aligned}
&M_{2i}(t, 0, m_1) \perp M_{1i}(t')|V = v \\
&Y_i(t, 0, m_1, m_2) \perp M_{1i}(t')|V = v \\
&Y_i(t, 0, m_1, m_2) \perp M_{2i}(t'0, m_1')|L = l, V = v
\end{aligned} \tag{10}$$

for every m_1, m_2 and v where $t, t' \in \{0, 1\}$. The first and second assumptions of (10) are satisfied as far as the first line of A3 (no unobserved confounding in $M_1 - M_2$ and $M_1 - Y$ relationships) is met. However, the third assumption of (10) is not satisfied because of L, even when the second line of A3 (no unmeasured confounding in the $M_2 - Y$ relationship) is met. It is because the intermediate confounding variable L is affected by T_1 and also has an impact to both M_2 and Y. As shown in the potential outcomes above, both $Y_i(t, 0, m_1, m_2)$ and $M_{2i}(t'0, m_1)$ involve U_L and thus violate the third assumption of (10). Due to this violation of this assumption, we need A4 instead in order to identify ACME and ANDE.

On the other hand, the third assumption is met if the intermediate confounding variable is fixed to $L = l$. Formally, $Y(t, 0, m_1, m_2, l) \perp M_2(t'0, m_1, l)|V = v$. It is because L is no longer affected by T_1 but is fixed to a certain value. Other assumptions in A1–A3 as well as the first two assumptions in (10) hold even after L is fixed to l.

C. Identification of the Alternative ACME-l and ANDE-l Under the Homogeneous Effects

First, we have

$$\begin{aligned}
&E[Y(t, 0, M_1(t'), M_2(t'', 0, M_1(t''), l), l)|v] \\
&= \sum_{m_2}\sum_{m_1}\sum_{m_1'} E(Y_i(t, 0, m_1, m_2, l)|M_2(t'', 0, m_1', l) = m_2, M_1(t') \\
&= m_1, M_1(t'') = m_1', v) \\
&\quad \cdot P(M_2(t'', 0, m_1', l) = m_2|M_1(t') = m_1, M_1(t'') = m_1', v) \cdot P(M_1(t') \\
&= m_1|M_1(t'') = m_1', v) \cdot P(M_1(t'') = m_1'|v) \\
&= \sum_{m_2}\sum_{m_1}\sum_{m_1'} E(Y_i(t, 0, m_1, m_2, l)|v) \cdot P(M_2(t'', 0, m_1', l) = m_2|v) \\
&\quad \cdot P(M_1(t') = m_1|M_1(t'') = m_1', v) P(M_1(t'') = m_1'|v) \\
&= \sum_{m_2}\sum_{m_1}\sum_{m_1'} E(Y_i|t, 0, m_1, m_2, l, v) \cdot P(M_{2i} = m_2|t'', 0, m_1', l, v)
\end{aligned}$$

$$\cdot \boxed{P(M_{1i}(t') = m_1 | M_{1i}(t'') = m_1', v)} \cdot P(M_{1i} = m_1 | t'', v) \tag{11}$$

for every m_1, m_1', m_2, v and l, where $t, t', t'' \in \{0, 1\}$. The first equality is due to the law of total probability. The second equality is because the intermediate confounding variable is fixed to $L_i = l$ with the implications of "Appendix B." The third equality holds because of A1–A3, which still hold after fixing $L = l$. The ACME-l and ANDE-l can be obtained after plugging this identification result to Eq. (2). This completes the proof.

D. Bias of $\hat{\delta}^{M_2}(t, 0, l)$ Under Heterogeneous Effects

Note that this proof is an extension of the proof shown in Imai and Yamamoto (2013). Based on the models shown in Eq. (5), $\delta^{M_1}(t, l)$ is identified as

$$\begin{aligned} &= E\{(e_{2i} + t \cdot e_{6i} + l \cdot e_{8i})(M_{1i}(1) - M_{1i}(0))\} \\ &= (e_2 + t \cdot e_6 + l \cdot e_8)\{E(M_{1i}|T_{1i} = 1) - E(M_{1i}|T_{1i} = 0)\} \end{aligned} \tag{12}$$

where $t \in \{0, 1\}$. The second equality follows because $M_{1i}(t')$ is independent to $Y_i(t, 0, m_1, m_2, l)$ for every t, t', m_1, m_2, l as implied in "Appendix A"ss; and $(e_{2i} + t \cdot e_{6i} + l \cdot e_{8i})$ is in fact $Y_i(t, 0, m_1, m_2, l) - Y(t, 0, m_1', m_2, l)$ for every t, m_1, m_1', m_2, and l. Therefore, the $\delta^{M_1}(t, l)$ is identified without further complications. In contrast, the identification of $\delta^{M_2}(t, l)$ requires two sensitivity parameters and is shown below.

$$\begin{aligned} \delta^{M_2}(t, l) &= E\{(e_{5i} + t \cdot e_{7i} + l \cdot e_{9i})(M_{2i}(1, 0, l) - M_{2i}(0, 0, l))\} \\ &= E\{(e_{5i} + e_{7i} + l \cdot e_{9i})M_{2i}(1, 0, l)|T_{1i} = 1\} - E\{(e_{5i} + l \cdot e_{9i})M_{2i}(0, 0, l)|T_{1i} = 0)\} \\ &\quad - E\{e_{7i} M_{2i}(t', 0, l)\} \\ &= E(e_{5i} + e_{7i} + l \cdot e_{9i})E(M_{2i}(1, 0, l)|T_{1i} = 1\}) \\ &\quad - E(e_{5i} + l \cdot e_{9i})E(M_{2i}(0, 0, l)|T_{1i} = 0) \\ &\quad - E\{e_{7i} M_{2i}(t', 0, l)\} \\ &= (e_5 + e_7 + l \cdot e_9)E(M_{2i}|T_{1i} = 1, T_{2i} = 0, L_i = l) \\ &\quad - (e_5 + l \cdot e_9)E(M_{2i}|T_{1i} = 0, T_{2i} = 0, L_i = l) \\ &\quad - e_7 E\{M_{2i}|T_{1i} = t', T_{2i} = 0, L_i = l)\} \\ &\quad - \rho\sigma_{e_7}\sqrt{V(M_{2i}|T_{1i} = t', T_{2i} = 0, L_i = l)} \end{aligned} \tag{13}$$

for every t, t', and l and $\sigma_{e_7} = \sqrt{V(e_{7i})}$; and ρ is the correlation between e_{7i} and $M_{2i}(t', l)$. The second equality follows from assumption A1 (the first treatment is ignorable given covariates), which still holds after fixing $L = l$. The third equality is because $e_{5i} + t \cdot e_{7i} + l \cdot e_{9i}$ is conditionally independent to $M_{2i}(t, 0, l)$ under assumption A3, which still holds after fixing $L = l$. It is because $e_{2i} + e_{6i} + l \cdot e_{8i}$ is same as $Y_i(t, 0, m_1, m_2, l) - Y_i(t, 0, m_1, m_2', l)$ for every m_1, m_2, m_2' and l. The last equality is because of the fact that two random variables, e_{7i} and $M_{2i}(t', 0, l)$, are not independent. This term, $\rho_1\sigma_{e_7}\sqrt{V(M_{2i}|T_{1i} = t', T_{2i} = 0, L_i = l)}$, indicates the covariance between the two random variables.

Assuming homogeneous effects, one will simply obtain $\hat{\delta}^{M_2}(t, l) = (e_5 + e_7 + l \cdot e_9)E(M_{2i}|T_{1i} = 1, T_{2i} = 0, L_i = l) - (e_5 + l \cdot e_9)E(M_{2i}|T_{1i} = 0, T_{2i} = 0, L_i = l)$ since e_7 is no longer a random variable. The bias given in Eq. (6) is obtained by calculating the difference between $\hat{\delta}^{M_2}(t, 0, l)$ and $\delta^{M_2}(t, 0, l)$. Bias for $\zeta(t', 0, l)$ is computed using the fact that the bias of combined effect of $\delta^{M_1}(t)$, $\delta^{M_2}(t)$ and $\zeta(t')$ is zero. This completes the proof.

E. Generalizability to an Extended Setting

One attractive feature of this alternative definition of ACME-l and ANDE-l is the generalizability to a model beyond two time periods. The bias formula for the ACME-l via M_k (where $k \in K$ denotes a number of time periods) remains the same even when the model is extended beyond two time periods if the model follows the same model specification as in Eq. (5). This implies that the model should not include the interaction effects among mediators (e.g., $M_1 \cdot M_2$) or higher order terms of mediators (e.g., M_2^2). Given this model specification, the bias formula for the ACME-l via M_3, for instance, is expressed as $bias(\hat{\delta}^{M_3}(t,l)) = \rho_{M_3}\sigma_{e_{T_1M_3}}\sqrt{V(M_{3i}|T_{1i}=t', T_{2i}=0, L_1=l, L_2=l)}$ where $\sigma_{e_{T_1M_3}}$ is $\sqrt{V(e_{T_1M_3i})}$ in which $e_{T_1M_3i}$ is the interaction effect between T_1 and M_3, and ρ_{M_3} is the correlation between $e_{T_1M_3i}$ and $M_{3i}(t',0,l,l)$. This bias formula has the same form as Eq. (6), but the only difference is that the bias is due the interaction effect in the $T_1 - M_3$ relationship as compared to the interaction in the $T_1 - M_2$ relationship. This bias formula holds even when L_k is affected by a previously measured time-varying confounding variable or L_k itself affects the following mediators.

However, this simple generalization to a model beyond two time periods only holds when the time-varying treatments measured after the first time period are all fixed as in our example (e.g., $T_2 = T_3 = T_4 = \cdots = 0$). If this does not hold, the bias formulas shown in Eq. (6) depend on more than two sensitivity parameters. To see this, suppose that T_2 is not fixed to zero. Then, $bias(\hat{\delta}(M_2)(t,t,l) = \rho_{M_2}\sigma_{e_7}\sqrt{V(M_{2i}|T_{1i}=t', T_{2i}=t', L_i=l)} + \rho_{M_2T_2}\sigma_{e_{T_2M_2}}\sqrt{V(M_{2i}|T_{1i}=t', T_{2i}=t', L_i=l)}$, where $e_{T_2M_2i}$ is the interaction effect between T_2 and M_2, and $\rho_{M_2T_2}$ is the correlation between $e_{T_2M_2}$ and $M_2(t',t',l)$. This bias formula shows that interaction effects between T_2 and M_2 contribute additionally to the bias formula shown in Eq. (6). As discussed above, extending the proposed bias formulas to a more general time-varying treatments and mediators setting is not impossible. However, sensitivity analysis based on this extended bias formula becomes dependent on more than two parameters, which is undesirable when examining the sensitivity of effect estimates.

References

Avin, C., Shpitser, I., & Pearl, J. (2005). *Identifiability of path-specific effects*. California: Department of Statistics, UCLA.

Bind, M.-A., Vanderweele, T., Coull, B., & Schwartz, J. (2016). Causal mediation analysis for longitudinal data with exogenous exposure. *Biostatistics*, *17*(1), 122–134.

Daniel, R., De Stavola, B., Cousens, S., & Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, *71*(1), 1–14.

De Stavola, B. L., Daniel, R. M., Ploubidis, G. B., & Micali, N. (2014). Mediation analysis with intermediate confounding: structural equation modeling viewed through the causal inference lens. *American Journal of Epidemiology*, *181*(1), 64–80.

Didelez, V., Dawid, P. & Geneletti, S. (2012). Direct and indirect effects of sequential treatments. arXiv:1206.6840.

Fong, C., Ratkovic, M. & Imai, K. (2014). Cbps: R package for covariate balancing propensity score. *Comprehensive R Archive Network (CRAN)*.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 199–215.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, *27*(3), 205–224.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*(475), 901–910.

Imai, K., & Ratkovic, M. (2015). Robust estimation of inverse probability weights for marginal structural models. *Journal of the American Statistical Association*, *110*(511), 1013–1023.

Imai, K., & Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, *21*, 141–171.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press.

Mayer, A., Thoemmes, F., Rose, N., Steyer, R., & West, S. G. (2014). Theory and analysis of total, direct, and indirect causal effects. *Multivariate Behavioral Research*, *49*(5), 425–442.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge: Cambridge University Press.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.
Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science*, *37*(6), 1011–1035.
Steen, J., Loeys, T., Moerkerke, B., & Vansteelandt, S. (2017). Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology*, *186*, 184–193.
Steiner, P. M., Park, S. & Kim, Y. (2016). Identifying causal estimands for time-varying treatments measured with time-varying (age or grade-based) instruments. *Multivariate Behavioral Research*, *51*, 1–6.
Steyer, R., Mayer, A. & Fiege, C. (2014). Causal inference on total, direct, and indirect effects. *Encyclopedia of Quality of Life Research*, 606–631.
Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G. & Najarian, M. (2009). Early childhood longitudinal study, kindergarten class of 1998–99 (ecls-k): Combined user's manual for the ecls-k eighth-grade and k-8 full sample data files and electronic codebooks. nces 2009-004. National Center for Education Statistics.
Vandecandelaere, M., Vansteelandt, S., De Fraine, B. & Van Damme, J. (2016). Time-varying treatments in observational studies: Marginal structural models of the effects of early grade retention on math achievement. *Multivariate Behavioral Research*, 1–22.
VanderWeele, T., & Tchetgen Tchetgen, E. (2016). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*, 917–938.
VanderWeele, T., & Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods*, *2*(1), 95–115.

Manuscript Received: 17 JUL 2015
Final Version Received: 23 JAN 2018