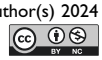
Original Article

Sociological Methods & Research
1–34



DOI: 10.1177/00491241241264562
journals.sagepub.com/home/smr


Causal Decomposition Analysis With Time-Varying Mediators: Designing Individualized Interventions to Reduce Social Disparities

Soojin Park[1], Namhwa Lee[2], and Rafael Quintana[3]

Abstract

Causal decomposition analysis aims to identify risk factors (referred to as "mediators") that contribute to social disparities in an outcome. Despite promising developments in causal decomposition analysis, current methods are limited to addressing a time-fixed mediator and outcome only, which has restricted our understanding of the causal mechanisms underlying social disparities. In particular, existing approaches largely overlook individual characteristics when designing (hypothetical) interventions to reduce disparities. To address this issue, we extend current longitudinal mediation approaches to the context of disparities research. Specifically, we develop a novel decomposition analysis method that addresses individual characteristics by (a) using optimal dynamic treatment regimes (DTRs) and (b) conditioning on a selective set of individual characteristics. Incorporating optimal DTRs into the design of interventions can be used to strike a balance between

[1]School of Education, University of California, Riverside, CA, USA
[2]Statistics, University of California, Riverside, CA, USA
[3]School of Education and Human Sciences, University of Kansas, Lawrence, KS, USA

Corresponding Author:
Soojin Park, School of Education, University of California, Riverside, CA, USA.
Email: soojinp@ucr.edu

equity (reducing disparities) and excellence (improving individuals' outcomes). We illustrate the proposed method using the High School Longitudinal Study data.

Keywords

longitudinal causal decomposition, time-varying mediators, individualized interventions, optimal dynamic treatment regimes, disparity reduction, disparity remaining

Introduction

Recently, there have been considerable methodological developments on approaches to decompose social disparities within the causal inference literature (e.g., VanderWeele and Robinson 2014; Jackson and VanderWeele 2018; Jackson 2021; Lundberg 2020; Park, Qin, and Lee 2022; Park et al. 2023). These developments in causal decomposition analysis extend traditional approaches (e.g., difference-in-coefficients and Blinder Oaxaca decomposition) to settings with nonlinear relationships, and have clarified the assumptions (e.g., no omitted confounding) required to permit causal interpretation of the results. Moreover, recently developed sensitivity analyses (Park, Qin, and Lee 2022; Park et al. 2023) enable researchers to assess the robustness of findings against a reasonable amount of omitted confounding. As a result, stronger causal interpretations of the estimated effects can be made.

A successful application of causal decomposition and sensitivity analysis can be found in Lee, Park, and Boylan (2021), which examines interventions to reduce cardiovascular health disparities across race/ethnicity and gender categories. That study begins by presenting a directed acyclic graph (DAG; Pearl 2012), which encodes the authors' understanding of the data-generating process. They used the DAG to determine which variables to control to eliminate confounding. The study concludes that approximately one-third of the cardiovascular health disparity between Black women and White men would be reduced if Black women's socioeconomic status (SES) was equal to that of White men. This reduction remains robust even in scenarios where a reasonable amount of omitted confounding (as large as the existing covariates, e.g., family history of cardiometabolic conditions) is assumed to exist. Causal decomposition analysis allows us to rigorously evaluate the effect of modifying risk factors or resources on reducing disparities, even when using observational data. The risk factors or resources are

referred to as "mediators," since they are believed to lie in the path between social groups (exposure) and the outcome.

Despite these promising developments in causal decomposition analysis, the methods currently available are restricted to scenarios involving only time-fixed mediators and outcomes. Consequently, this has limited our understanding of the causal mechanisms underlying the observed disparities. However, it is important to highlight the existence of the relevant literature on time-varying exposures and mediators in the causal inference framework. Specifically, Bind et al. (2016) proposed employing a generalized linear mixed model to identify natural direct and indirect effects (Pearl 2001) in settings with time-varying exposures, mediators, and outcomes. Natural direct and indirect effects are not nonparametrically identified when exposure-induced confounders exist (VanderWeele and Tchetgen Tchetgen 2017); hence, the analysis hinges on a strong assumption of no exposure-induced confounding. To circumvent this issue, VanderWeele and Tchetgen Tchetgen (2017) and Zheng and van der Laan (2017) used *interventional analogs of natural direct and indirect effects* (hereafter, interventional effects), which are identified even in the presence of exposure-induced confounding. In disparities research, such confounding is likely, since a myriad of life course factors influenced by the exposure (social groups) affect disparities. For example, in the United States, Blacks are more likely to be born into low-income families, which can affect their education (mediator) and math achievement (outcome). In this scenario, the causal decomposition framework based on interventional effects can be used to evaluate the effect of modifying mediators in reducing social disparities. However, the literature on causal decomposition analysis has not yet formally examined time-varying mediators.

An additional limitation of the current literature on causal decomposition analysis is that it largely overlooks individual characteristics when modifying risk factors or resources. Therefore, its capacity to inform the design of an actual intervention has been limited. In causal decomposition analysis, we estimate the effect of hypothetically intervening to set the mediator to a predetermined value or the mediator distribution equal to that of a reference group. Setting the mediator to a predetermined value would imply giving the same intervention to every individual, which is often unrealistic or unethical in practice. Equalizing the mediator distribution between groups may be more realistic, yet this approach still does not take into account individual characteristics. For example, providing the same math course-taking plan to all high school students or assigning Black students to follow the same course-taking pattern as White students,

regardless of their prior math achievement or motivation levels, may not be realistic or desirable.

The main goal of this paper is to (a) formally extend existing longitudinal mediation approaches to the context of disparities research and (b) propose a novel decomposition analysis that considers individual characteristics. The concept of individualized treatment is widely used in precision medicine, which aims to optimize treatments for each patient based on their unique genetic, environmental, and lifestyle factors, as opposed to a one-size-fits-all approach (Tsiatis et al. 2019). Dynamic treatment regimes (DTRs) are a set of decision rules that describe how treatments should be assigned in response to individual factors (Mahar et al. 2021). Optimal DTRs develop sequential decision rules that maximize an average outcome at the end of the time period (Murphy 2003). For example, optimal DTRs can be used to optimize the decision to take a series of math courses depending on each student's prior math achievement or motivation levels at each time interval, such that the final math score is maximized. Reducing social disparities to achieve equity is an important goal in and of itself. At the same time, we aim to maximize the outcome, assuming that larger outcomes (e.g., academic achievement) are preferable. Incorporating optimal DTRs into the design of interventions aimed at reducing disparities can serve as a means to strike a balance between equity (reducing disparities) and excellence (maximizing the outcome). In our study, we provide a novel contribution to the causal decomposition literature by incorporating individual characteristics and considering both excellence and equity. We distinguish between different quantities of interest and provide important considerations for deciding between them. These considerations involve the extent to which individual characteristics need to be considered, the feasibility of the intervention, and whether maximizing the outcome is a priority. The overarching goal is to examine the effect of interventions that are both realistic and appropriate given individuals' characteristics.

The article is organized as follows. We introduce a running example in the "Running Example" section, which is followed by extending existing longitudinal mediation approaches to disparities research in the "Extending Existing Longitudinal Mediation Approaches" section and a review of optimal DTRs in the "Review of Optimal DTRs" section. In the "Longitudinal Causal Decomposition Analysis (CDA) With Individualized Interventions" section, we propose a novel CDA that takes into account individual characteristics. In the "Recommendations for Empirical Researchers" section, we present key insights and recommendations for empirical researchers. Finally, we conclude with a discussion of the main contributions of the paper.

Running Example

Our motivating question is the following: "how much of the Black–White differential in math achievement would remain if we were to intervene on the courses they take?" Comparing other racial/ethnic groups might be of substantive interest, but we only focus on the Black–White differential in math scores in 11th grade for simplicity. The estimates we present here are based on data from the High School Longitudinal Study 2009 (HSLS:09).

Prior research suggests that the courses that students take in math have important consequences for a variety of educational and career outcomes (Attewell and Domina 2008; Kelly 2009). In particular, various studies have shown that taking advanced math courses affects students' math achievement and college enrollment (Byun, Irvin, and Bell 2015; Long, Conger, and Iatarola 2012; McEachin, Domina, and Penner 2020). In addition, researchers have argued that minority underrepresentation in advanced math courses can be a key driver of educational inequality (Attewell and Domina 2008; Riegle-Crumb and Grodsky 2010). Based on this concern, a variety of educational policies and efforts have been developed to reduce inequitable access to advanced math courses (Byun, Irvin, and Bell 2015; Long, Conger, and Iatarola 2012).

Considering these initiatives, an important question that arises is whether all students benefit from taking advanced math courses. Prior research suggests that taking courses without adequate preparation can actually have unintended negative consequences, such as a decline in students' motivation (Simzar, Domina, and Tran 2016a). A central question is, then, how to assign students to rigorous math courses (and thus increase access to educational opportunities), while making sure that students can succeed in these courses. This problem has motivated researchers to design effective course placement rules based on objective measures such as prior test scores (Dougherty et al. 2017).

DAG. We define the social groups as Blacks ($R = 1$) and Whites ($R = 0$) and the outcome as math score in 11th grade (Y). We have two time-varying mediators: Algebra 1 (M_1) and advanced math courses (M_2). Using these variables, we draw a DAG as shown in Figure 1. We assume that the Black–White differential in math scores arises through multiple causal and noncausal paths: (P1) the path from race to math score through Algebra 1 or advanced courses (e.g., $R \rightarrow M_1 \rightarrow Y$ or $R \rightarrow M_2 \rightarrow Y$), (P2) the back-door path through history, childhood SES, and education (e.g., $R \leftarrow H \rightarrow X_{11} \rightarrow M_1 \rightarrow M_2 \rightarrow Y$), and (P3) the paths that do not operate through any mediators (e.g., $R \rightarrow X_{21} \rightarrow Y$ or $R \rightarrow Y$). Following Jackson and

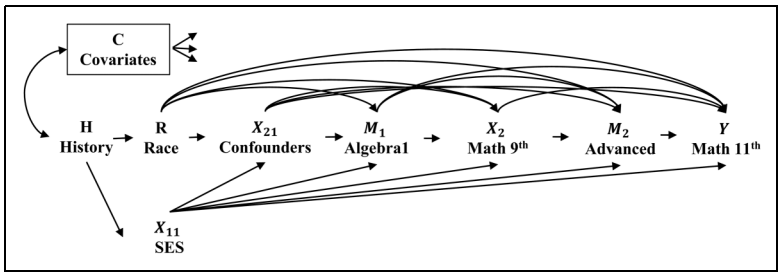


Figure 1. DAG showing the pathways to the racial disparity in math achievement in 11th grade. *Note.* DAG = directed acyclic graph; SES = socioeconomic status. (a) Baseline covariates (*C*) include gender and placing a box around *C* indicates conditioning on this variable. (b) The three arrows emanating from *C* indicate that they are confounders of all bivariate relations unless conditioned on, as visualized by the box around *C*. (c) Childhood SES (X_{11}) includes the individual's birth region, and father's and mother's years of education. (d) Confounders (X_{21}) include parental expectations, school climate, and students' readiness to take math courses.

VanderWeele (2018), Path P1 represents discrimination. For example, Blacks are more likely to be placed into less rigorous curriculum tracks (Kelly 2009) and are thus less likely to achieve high scores in math in 11th grade. Path P2 represents the effect of historical processes, including racism and segregation (Kaufman 2008). For example, Blacks are more likely to be born into families with low SES and live in neighborhoods with low-quality schools. Path P3 represents the direct effect of race on math achievement in 11th grade not via a math course-taking pattern.

The proposed DAG also includes baseline covariates and time-varying confounders. In the present example, the baseline covariates (C) include gender given that (a) the gender distribution varies in each racial group, and accounting for it addresses potential biases, and (b) gender is a source of outcome differences we do not consider in this study, making it an "allowable" covariate (Jackson 2021) when measuring Black–White differentials. In a different context, such as health outcomes, variables such as age or demographic status might fall under this category. In cases where these variables do not pertain to the study, one can specify "null" for the baseline covariates.

Time-varying confounders are the variables that are measured concurrently or after the group status and confound the mediator–outcome relationship. The first set of time-varying confounders includes childhood SES (X_{11}) and variables at different levels (individual, teacher, parent, peer, and other contextual factors) that are related to math readiness (X_{21}). For simplicity,

we use $X_1 = (X_{11}, X_{21})$. The second set (X_2) includes math achievement scores in nineth grade. We assume that X_1 is constant between nineth and 11th grades to avoid multicollinearity.

Sample and Measures. For subsequent analyses, we restricted our sample to those students who did not take Algebra 1 before nineth grade, as the time-varying confounders at Time 1 (X_{21}) were measured in nineth grade. Including students who took Algebra 1 in eighth grade would have raised analytic challenges, as the confounders (X_{21}) were measured after the mediator (M_1; taking Algebra 1 in eighth grade). This results in a sample size of 11,050 students, consisting of 8,919 White students and 2,127 Black students. We used the item response theory theta scores for math achievement in nineth (X_2) and 11th (Y) grades. Algebra 1 (M_1) was measured by identifying those students who took Algebra 1 in nineth grade, while advanced courses (M_2) were measured by identifying students who took courses beyond Algebra 2 (trigonometry, probability and statistics, precalculus, calculus, AP/IB calculus, and other AP/IB math) before graduating high school.

To ensure the validity of key identification assumptions, we carefully selected the set of intermediate confounders X_{21} based on the literature. Prior research suggests that achievement in previous courses is the main factor influencing students' course-taking patterns (Kalogrides and Loeb 2013; Long, Conger, and Iatarola 2012). Thus, we control for students' final grades in their most advanced eighth-grade math course. We also control for additional individual-level factors that can confound the mediator–outcome relationship: students' math course interest, math utility and identity beliefs, math self-efficacy, school engagement, and school belonging (see Ingels et al. 2013 for a detailed description of these variables).

Students' course-taking decisions can also be influenced by their teachers, parents, peers, and other contextual factors (Byun, Irvin, and Bell 2015; Kelly 2009; Long, Conger, and Iatarola 2012; Riegle-Crumb and Grodsky 2010). At the family level, we controlled for household SES, parental occupation, and parents' expectations and aspirations regarding their children. At the teacher level, we controlled for math teacher's sex and math teacher's emphasis on increasing students' interest in math. At the peer level, we controlled the academic disposition of the closest friend. Finally, at the school level, we controlled for the school's locale, school problems, and climate, the percentage of students in math courses that are unprepared, the science and math course requirements, and whether the school offers Science, Technology, Engineering, and Mathematics extracurricular activities.

It is worth noting, however, that this example is for illustrative purposes and should not be used to draw educational or policy conclusions.

We defined M_2 as having taken advanced courses by the time of high school graduation, as there was no item asking about the math courses students took by 11th grade. This could affect the validity of the results since some students may have taken advanced courses after 11th grade.

Extending Existing Longitudinal Mediation Approaches

Several approaches for addressing time-varying exposures and mediators within a causal inference framework have been proposed in the literature. In this section, we formally extend these existing methods in the context of disparities research.

To identify the disparity reduction and disparity remaining estimand (Jackson and VanderWeele 2018) with observational data, we assume the following:

- **A1. Conditional independence of M_1 and M_2**: $Y(m_1, m_2) \perp M_1 | R = r, X_1 = x_1, C = c$ and $Y(m_1, m_2) \perp M_2 | R = r, X_1 = x_1, M_1 = m_1, X_2 = x_2, C = c$ for all $x \in X, c \in C$ and $r, m_1, m_2 = 0, 1$ where $Y(m_1, m_2)$ is a potential outcome under $M_1 = m_1$ and $M_2 = m_2$. This assumption states that there is no omitted confounding in the mediator–outcome relationship given the corresponding conditioning sets.
- **A2. Positivity**: $0 < P(M_1 = m_1 | R = r, X_1 = x_1, C = c) < 1$ and $0 < P(M_1 = m_2 | R = r, X_1 = x_1, M_1 = m_1, X_2 = x_2, C = c) < 1$ for all $x \in X, c \in C$ and $r, m_1, m_2 = 0, 1$. This assumption implies that individuals of every group have a nonzero probability of receiving any level of the mediators (Algebra 1 and advanced math) given baseline covariates.
- **A3. Consistency**: The observed outcome (e.g., math score) of an individual who has a certain level of the mediators (e.g., Algebra 1 and advanced math) is the same as the potential outcome after intervening to set the mediators to that level.

All of these assumptions are strong, and their plausibility depends on the specific study. In the example provided, we have exercised caution in selecting baseline covariates and time-varying confounders based on the literature. However, the assumption of sequential ignorability (A1) may be compromised due to the presence of unobserved confounders. Furthermore, consistency may be violated if an individual student's math score is influenced by the course-taking patterns of their peers. Although these are strong assumptions, for the purposes of this article, we will assume that these assumptions hold.

Controlled Direct Effects

One potential strategy for addressing Black–White disparities in math achievement scores is to implement a standardized course-taking pattern for all students. To evaluate the potential impact of this intervention, one can estimate the controlled direct effect (CDE) using observational data. The CDE is defined as the effect of an exposure (race) on an outcome of interest (math score) after setting the mediators (math course-taking pattern) to a specific value.

The CDE with time-varying exposures and mediators can be estimated using the identification result and estimation method provided by VanderWeele and Tchetgen Tchetgen (2017). However, in the case where race is not a time-varying exposure, this method may not be directly applicable. To address this limitation, an extension can be made by defining the disparity remaining at each value of m_1 and m_2 as:

$$\zeta_c^{\mathrm{CDE}}(m_1, m_2) \equiv E[Y(m_1, m_2)|R = 1, C = c] - E[Y(m_1, m_2)|R = 0, C = c] \quad \text{for} \quad \mathrm{m}_1, \mathrm{m}_2 \in \{0, 1\} \tag{1}$$

where $Y(m_1, m_2)$ is a potential outcome under $M_1 = m_1$ and $M_2 = m_2$, and $R = 1$ denotes Black students and $R = 0$ denotes White students (the reference group). This definition of disparity remaining ($\zeta_c^{\mathrm{CDE}}(m_1, m_2)$) captures the degree to which the outcome disparity would persist if all individuals were to adopt the specified values of the mediators m_1 and m_2 given baseline covariates.

Under Assumptions A1 to A3, the disparity remaining at each value of m_1 and m_2 can be estimated by fitting the following weighted regression model, where baseline covariates are centered at $C = c$:

$$Y = \beta_1 + \beta_2 R + \beta_3 M_1 + \beta_4 M_2 + \beta_5 R \times M_1 + \beta_6 R \times M_2 + \beta_7 C + \epsilon_1, \tag{2}$$

given the weight of $W_{\mathrm{CDE}} = P(M_1|R, X_1, C)^{-1} P(M_2|R, X_1, M_1, X_2, C)^{-1}$. The disparity remaining at each value of m_1 and m_2 is then estimated as $\hat{\zeta}_c^{\mathrm{CDE}}(m_1, m_2) = \hat{\beta}_2 + \hat{\beta}_5 m_1 + \hat{\beta}_6 m_2$. To account for multiple stages of estimation, bootstrapping can be used to estimate standard errors.

The estimation of CDEs depends on W_{CDE}. When positivity (A2) is nearly violated, the weight can either be very large or very small, leading to unstable estimates. To overcome this issue, we can truncate the weight at the first and 99th percentiles of the weight distribution (Cole and Hernán 2008). Another common approach is to use stabilized weights (Robins, Hernan, and

Brumback 2000), where the weight is replaced by $P(M_1|R)P(M_2|R, M_1)$ $/[P(M_1|R, X_1, C)P(M_2|R, X_1, M_1, X_2, C)]$.

In equation (2), we assume a differential effect of mediators by race (i.e., $R \times M_1$ and $R \times M_2$). However, the equation can be modified to assume a constant effect of mediators by deleting the interaction terms. In the example, estimating the CDE using observational data can provide valuable insights into the potential impact of implementing a standardized course-taking pattern for all students on reducing disparities in math achievement given Assumptions A1 to A3.

Interventional Marginal Effects

An important limitation of the CDE approach is that it may not be feasible or beneficial to require all students to take the same math courses. An alternative approach to address racial disparities in math achievement could be to ensure that Black students are randomly placed in Algebra 1 and advanced math classes at the same rate as White students within the same gender status (baseline covariates). We could estimate the potential impact of this intervention using interventional analogs of natural direct and indirect effects (VanderWeele Vansteelandt, and Robins 2014; Jackson and VanderWeele, 2018: interventional effects,) with observational data. The interventional direct effect is defined as the effect of an exposure on an outcome after intervening to equalize the distribution of the mediators (e.g., math course-taking pattern) between groups given baseline covariates.

VanderWeele and Tchetgen Tchetgen (2017) proposed an identification method and estimation technique for mediational effects in situations where exposures and mediators are time-varying. This approach is based on the mediational g-formula (Pearl 2001), which is a method of estimating mediation effects by fitting parametric models. However, this method can be computationally intensive, as it involves numerous integrations and requires correctly specifying models for intermediate confounders. As an alternative, the authors suggest using marginal structural models (MSMs) and inverse probability of treatment weighting (Robins, Hernan, and Brumback 2000). Again, given that race is not a time-varying exposure, we extend their method by defining the disparity reduction and disparity remaining as:

$$\begin{aligned}&\delta_c^{\text{IME}}(1) \equiv E[Y|R=1, C=c] - E[Y(G_{M_1|R=0,C}, G_{M_2|R=0,C})|R=1, C=c], \text{ and}\\ &\zeta_c^{\text{IME}}(0) \equiv E[Y(G_{M_1|R=0,C}, G_{M_2|R=0,C})|R=1, c] - E[Y|R=0, c],\end{aligned} \tag{3}$$

for $c \in C$, where $G_{M_j|R=0,C}$ for $j \in \{1, 2\}$ is a random draw from the mediator distribution of the reference group given baseline covariates. Using potential outcomes notation, $Y(G_{M_1|R=0,C}, G_{M_2|R=0,C})$ is an outcome that is realized under a random draw from the mediator distributions of the reference group given baseline covariates. The disparity reduction ($\delta_c^{\text{IME}}(1)$) represents the change in the outcome among Black students given baseline covariates after intervening to set the mediator distribution equal to that of White students with the same level of baseline covariates; disparity remaining ($\zeta_c^{\text{IME}}(0)$) represents the difference in the outcome given baseline covariates that persists between Black and White students, even after the hypothetical intervention. In the example, disparity reduction and disparity remaining capture, respectively, the extent to which the outcome disparity would be reduced or persist if Black students were enrolled in Algebra 1 and advanced math at the same rate as White students within the same gender status.

Given Assumptions A1 to A3, the disparity reduction and disparity remaining can be estimated by fitting the outcome model as in equation (2) and the following mediator model:

$$\begin{aligned} P(M_1 = 1|R, C) &= \text{logit}^{-1}(\alpha_{11} + \alpha_{12}\text{R} + \alpha_{13}\text{C}), \quad \text{and} \\ P(M_2 = 1|R, C) &= \text{logit}^{-1}(\alpha_{21} + \alpha_{22}\text{R} + \alpha_{23}\text{C}). \end{aligned} \tag{4}$$

Then, $\hat{\delta}_c^{\text{IME}}(1)$ can be obtained as

$$\left\{\frac{\exp(\hat{\alpha}_{11} + \hat{\alpha}_{12} + \hat{\alpha}_{13}\hat{E}[C])}{1 + \exp(\hat{\alpha}_{11} + \hat{\alpha}_{12} + \hat{\alpha}_{13}\hat{E}[C])} - \frac{\exp(\hat{\alpha}_{11} + \hat{\alpha}_{13}\hat{E}[C])}{1 + \exp(\hat{\alpha}_{11} + \hat{\alpha}_{13}\hat{E}[C])}\right\} \times (\hat{\beta}_3 + \hat{\beta}_5)$$
$$+ \left\{\frac{\exp(\hat{\alpha}_{21} + \hat{\alpha}_{22} + \hat{\alpha}_{23}\hat{E}[C])}{1 + \exp(\hat{\alpha}_{21} + \hat{\alpha}_{22} + \hat{\alpha}_{23}\hat{E}[C])} - \frac{\exp(\hat{\alpha}_{21} + \hat{\alpha}_{23}\hat{E}[C])}{1 + \exp(\hat{\alpha}_{21} + \hat{\alpha}_{23}\hat{E}[C])}\right\} \times (\hat{\beta}_4 + \hat{\beta}_6);$$

$\hat{\zeta}_c^{\text{IME}}(0)$ can be obtained as

$$\hat{\beta}_2 + \hat{\beta}_5 \times \frac{\exp(\hat{\alpha}_{11} + \hat{\alpha}_{13}\hat{E}[C])}{1 + \exp(\hat{\alpha}_{11} + \hat{\alpha}_{13}\hat{E}[C])} + \hat{\beta}_6 \times \frac{\exp(\hat{\alpha}_{21} + \hat{\alpha}_{23}\hat{E}[C])}{1 + \exp(\hat{\alpha}_{21} + \hat{\alpha}_{23}\hat{E}[C])},$$

or alternatively as $\hat{\tau}_c - \hat{\delta}_c^{\text{IME}}(1)$. Here, $\hat{\beta}$s are obtained from the outcome model weighted in the same way as equation (2). We refer to the disparity reduction and remaining as *interventional marginal indirect and direct effects*, respectively. These effects are considered marginal because they involve equalizing the mediator distribution across intermediate confounders (X_1, X_2), given baseline covariates (C).

An Application to HSLS:09

We estimated the initial disparity, disparity remaining, and disparity reduction with CDEs and interventional marginal effects (IMEs). Table 1 shows the estimated quantities of interest. For illustration purposes, the baseline covariate (gender) is centered at the mean[1] . The results are based on truncated weights at the first and 99th percentiles of the weight distribution. The initial disparity in math achievement in 11th grade between Black and White students is negative (−0.413 SD) and significant at the 95% confidence level, meaning that Black students have significantly lower math scores than White students within the same gender status.

First, we estimated the disparity remaining using CDEs. In our study, we included a group status–mediator interaction effect in the MSM if it was found to be significant. In our data, we observed a lower return from taking advanced courses for Black students compared to White students (0.058 for Black students versus 0.231 for White students). In contrast, Black students benefit equally from taking Algebra 1 as White students (0.341). Thus, we included the interaction between group status and whether or not the students took advanced courses. Enforcing all students to enroll in Algebra 1 but not in advanced math courses results in a remaining disparity of $\zeta_c^{\mathrm{CDE}}(1,0) = -0.322$ SD, which represents a 22.0% reduction from the initial disparity. Enforcing all students to enroll in both Algebra 1 and advanced courses results in a remaining disparity of

Table 1. Estimates of the Initial Disparity, Disparity Reduction, and Disparity Remaining.

	Estimate (SE)		
	`CDE(1,0)`	`CDE(1,1)`	`IME`
Initial disparity	−0.413*** (0.022)	−0.413*** (0.022)	−0.413*** (0.022)
Disparity remaining	−0.322*** (0.038)	−0.498*** (0.056)	−0.398*** (0.022)
Disparity reduction			−0.015*** (0.004)
% reduction	22.0%	−20.6%	3.6%

Note. SE = standard error; `CDE` = controlled direct effect; `IME` = interventional marginal effect; `ICE` = interventional conditional effect. The asterisk followed by estimates indicates the level of statistical significance (*: significant at 0.05, **: at 0.01, ***: at 0.001). Gender is centered at the mean.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. High School Longitudinal Study of 2009 (HSLS:09) Base-Year Restricted-Use File (NCES 2011-333)

$\zeta_c^{\mathrm{CDE}}(1, 0) = -0.498$ SD, that is, it increases the initial disparity by 20.6%. This larger disparity in math achievement in 11th grade after the intervention is due to the lower return of taking advanced courses for Black students compared to White students.

Second, we used interventional marginal indirect and direct effects to estimate disparity reduction and remaining, respectively. Compared to enforcing a standard math course pattern, intervening to equalize the proportion of students taking Algebra 1 and advanced courses between groups has a relatively smaller effect. Equalizing between groups within the same gender group results in a disparity reduction of $\delta_c^{\mathrm{IME}}(1) = -0.015$ SD, which represents 3.6% of the initial disparity. This suggests that equalizing access to math courses alone may have a limited effect on reducing the math achievement gap in 11th grade, as there is only a 4% difference in course-taking patterns between Black and White students. Specifically, 83.0% of White students took Algebra 1 and 43.3% took advanced courses, while 79.3% of Black students took Algebra 1 and 39.6% took advanced courses.

Implications for Existing Causal Decomposition Analysis

The implementation of CDEs sheds light on the potential consequences of setting the mediators to a pre-specified value for all students in reducing disparities. However, this static intervention may not be practical or advantageous. Implementing interventional marginal direct effects provides insights into the potential impact of equalizing the mediator distribution between the groups in reducing disparities. This stochastic intervention may be more realistic than fixing the mediator to a single value for all individuals.

However, equalizing the mediator distribution between the two groups has important limitations. First, not all subjects in the reference group may have optimal mediator values. White students are commonly used as a reference group and their mediator distribution is compared with that of Black students. The logic behind this is that the mediator distribution of the reference group may result in greater rewards in terms of the outcome. However, as shown in our case study in the "An Application to HSLS:09' section', there is no substantial difference in course-taking patterns between Black and White students, so the effect of the hypothetical intervention would be limited.

Second, even if we identify a reference group with a better mediator distribution, equalizing that distribution may not be suitable or desirable for another group with different individual characteristics. For example, equalizing access to math courses may not benefit students who lack motivation or

are not ready to take the course. As noted before, previous literature has shown that placing average- and low-performing students in Algebra 1 in eighth grade could lower their motivation (Simzar, Domina, and Tran 2016b). How can we take into account individual characteristics, such as prior achievement scores or motivation, when designing interventions to reduce social disparities in an outcome?

Review of Optimal DTRs

In this section, we review optimal DTRs and discuss how they can be used in the context of our example.

Notation and Definition

Consider two decision points M_j for $j \in \{1, 2\}$, where we have two options for the jth decision point. Let $h_j \in H_j$ be a collection of variables, or *history*, available on an individual at the jth decision point. Given the DAG in Figure 1, $h_1 = (r, x_1, c)$ and $h_2 = (r, x_1, m_1, x_2, c)$. A two-interval DTR consists of two *decision rules* (d_1, d_2), with $d_j \equiv d_j(h_j) \in D_j$, where D_j is all possible treatment regimes. An example of a decision rule is $d_1(h_1) = I(\texttt{math efficacy} > -1)$, where $I(\cdot)$ is the indicator function. Under this rule, students whose math self-efficacy[2] is >-1 SD will be recommended to take Algebra 1. Otherwise, students will be recommended not to take the course at that time.

Among all decision rules, our interest centers on identifying an *optimal decision rule* d^{opt}. Assuming that larger outcomes are preferred, d^{opt} is the one that maximized the *value* $V(d_1, d_2)$, which is the expected potential outcome $E\{Y(d_1, d_2)\}$ under that optimal decision (Tsiatis et al. 2019). Formally,

$$d^{\text{opt}} = \arg \max_{d \in D} V(d_1, d_2) = \arg \max_{d \in D} E\{Y(d_1, d_2)\}. \tag{5}$$

In the example, the maximized average math score would be achieved if all students followed the optimal course-taking patterns.

Identification Assumptions

To identify optimal DTRs, we need to make the same assumptions as longitudinal causal decomposition analysis, which are Assumptions A1 to A3.

However, these assumptions are not sufficient for identifying optimal DTRs. The following additional assumption is needed:

- **A4. Conditional independence of M_1 with respect to X_2:** $X_2(m_1) \perp M_1 | R = r, X_1 = x_1, C = c$ for all $x_1 \in X_1, c \in C$ and $r, m_1 = 0, 1$. This assumption asserts the absence of omitted confounding for the relationship between M_1 (enrollment in Algebra I) and X_2 (nineth-grade math achievement) given the group status, intermediate confounders at Time 1, and baseline covariates.

This implies that the assumptions for identifying optimal DTRs are stronger than those for longitudinal mediation analysis introduced in the "Controlled Direct Effects" and "Interventional Marginal Effects" sections. For example, if there is omitted confounding between M_1 and X_2, the optimal DTRs will not be identified while CDEs or interventional marginal in/direct effects can still be identified (as long as there is no omitted confounding in the relationship between each mediator and the final outcome). In practice, Assumptions A1 and A4 can be simultaneously met if the mediators are sequentially randomized.

Although these assumptions are strong, for the sake of this study, we assume that these assumptions are met. Given Assumptions A1 to A4, the optimal DTRs can be expressed in terms of the observed data.

Estimation

We review two common approaches to obtain the optimal DTRs, which are Q-learning and weighting.

Q-Learning. Here, we use backward induction to define optimal DTRs. The estimation begins at the second interval by identifying the optimal value for the second mediator M_2 (advanced math). Then, the optimal value for the first mediator M_1 (algebra 1) is identified by estimating the impact of M_1 on the pseudo-outcome, which is a predicted outcome under an optimal value for M_2 (Moodie, Chakraborty, and Kramer 2012).

Q-learning is based on postulating the outcome regression model. The two Q-functions can be defined as

$$\begin{aligned} Q_2(h_2, m_2) &= E[Y | H_2 = h_2, M_2 = m_2], \\ Q_1(h_1, m_1) &= E[\max_{m_2} Q_2(h_2, m_2) | H_1 = h_1, M_1 = m_1], \end{aligned} \tag{6}$$

where H_j includes history variables up to decision point j and $\max_{m_2} Q_2(h_2, m_2)$ represents the expected value of the Q_2 function when it is assessed at the m_2 value that maximizes the Q_2 function. Based on the Q-functions, the optimal DTRs could be derived such that each Q function is maximized. Given that the true Q-functions are unknown, we model the Q-functions using linear models as

$$Q_j(H_j, M_j; \beta_j) = \beta_{j0} + \beta_{j1}^T H_j + (\beta_{j2} + \beta_{j3}^T H_{j1}) M_j. \tag{7}$$

Among these variables, H_j represents history variables (main effects); H_{j1} represents a subset of variables in H_j that have heterogeneous (interaction) effects on the outcome based on the mediator value. We can estimate β_j using ordinary least squares. After substituting β_j with the least square estimate, denoted as $\hat{\beta}_j$, we can determine optimized DTRs that maximize $Q_j(h_j, m_j; \hat{\beta}_j)$. That is, by leveraging the heterogeneous effects, we can construct optimized DTRs for the time interval j as follows:

$$d_j^{\text{opt}}(H_j) = I(\hat{\beta}_{j2} + \hat{\beta}_{j3}^T H_{j1} > 0). \tag{8}$$

The performance of Q-Learning depends on the correct specification of the Q-functions. While the implementation of Q-Learning is straightforward, the estimated optimal regimes tend to generate poor results given that the misspecification of the Q-functions for each stage is highly likely (Tsiatis et al., 2019).

Weighting. To define optimal DTRs through weighting, we adopt the backward induction approach proposed by Zhao et al. (2015). This estimation method begins at the second interval, where the optimal value for M_2 (advanced courses) is identified as if it were a single-point decision. Next, the optimal value for M_1 (algebra 1) is identified by maximizing the expected outcome, given the optimal value for M_2. The value function for each decision point that is to be maximized can be formalized as follows:

$$\begin{aligned} V_2(d_2(H_2)) &= E\left[\frac{I(M_2 = d_2(H_2))}{P(M_2|H_2)} Y\right], \\ V_1(d_1(H_1), \hat{d}_2{}^{\text{opt}}(H_2)) &= E\left[\frac{I(M_1 = d(H_1), M_2 = \hat{d}_2{}^{\text{opt}}(H_2))}{P(M_1|H_1)P(M_2|H_2)} Y\right]. \end{aligned} \tag{9}$$

At each stage, the value function is maximized. To obtain the maximized value, the weighting method proposed by Zhang et al. (2012) requires

modeling the following contrast function:

$$C(Y, M_j, H_j) = \frac{M_j Y}{P(M_j|H_j; \hat{\gamma}_j)} - \frac{(1 - M_j)Y}{1 - P(M_j|H_j; \hat{\gamma}_j)}, \tag{10}$$

where $P(M_j|H_j; \hat{\gamma}_j)$ is a propensity model for M_j, which is regressed on the history variables up to the jth decision point. For example, this contrast function for Time 2 estimates the difference in potential math scores in 11th grade between taking advanced courses or not. We then define $Z = I(C(Y, M_j, H_j) > 0)$, so that $Z = 1$ indicates subjects who will benefit from the mediator $M_j = 1$ than $M_j = 0$. The optimal value can be determined based on Z. However, the accuracy of the estimation relies on the extent to which the model for $C(Y, M_j, H_j)$ aligns with the true contrast function.

Recognizing the possibility of misspecifying the regression models, Zhang et al. (2012) considered an alternative approach of recasting the original problem of finding the optimal treatment regime as a weighted classification problem. Based on the estimated contrast function, the optimized value is obtained by minimizing a weighted classification error as

$$d_j^{\text{opt}}(H_j) = \arg \min_{d \in D} \sum |C(Y, M_j, H_j)| \left[\hat{Z} - d_j(H_j)\right]^2. \tag{11}$$

This optimization problem can be solved by existing classification techniques, such as classification and regression trees (Breiman 2017). This approach of minimizing the classification error separates the estimation of the contrast function and the maximization of the value function. Therefore, it offers increased flexibility and robustness compared to approaches that rely on a specific model for the outcome or contrast function only (Zhang et al. 2012). However, it is important to note that even with this approach, there may still be concerns related to the model specification of contrast functions.

An Application to HSLS:09

Our two decision points are whether to take Algebra 1 (M_1) and advanced courses (M_2). To estimate the optimal decision rules, we considered math self-efficacy, math course interest, and math achievement in nineth grade for M_2, and math self-efficacy, math course interest, and grades in their most advanced eighth grade math course for M_1. These variables, denoted as H_{j1}, were selected because, based on prior research, they can affect

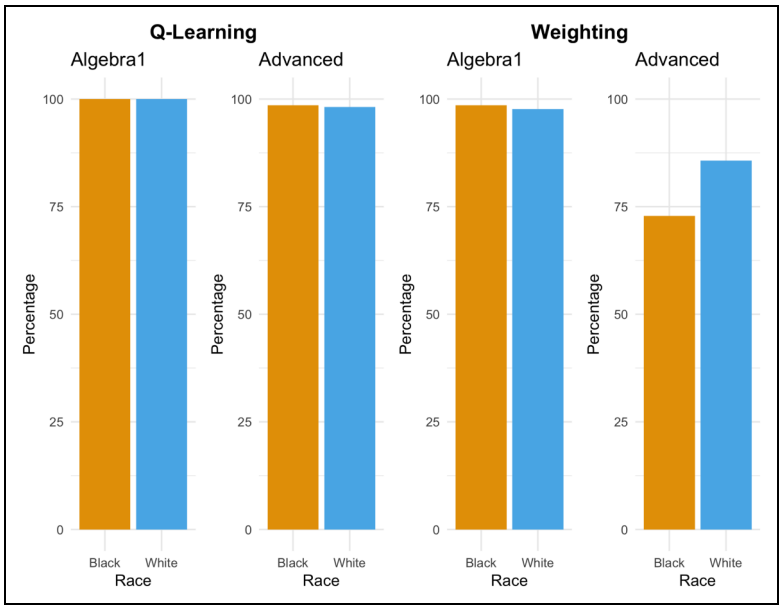


Figure 2. Percentage of Recommendation by Race. *Source:* U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. High School Longitudinal Study of 2009 (HSLS:09) Base-Year Restricted-Use File (NCES 2011-333).

whether the person benefits from taking advanced courses. The rest of the variables were used to estimate the main effects (Q-Learning) or propensity scores for each mediator (weighting).

Figure 2 summarizes the percentages of students recommended to receive Algebra 1 and advanced courses by each estimator. Both the Q-learning and weighting methods recommended that almost all the students (Q-learn: 100%, weighting: 97.9%) take Algebra 1. The weighting method recommended Algebra 1 for all students except for those with math efficacy scores lower than −1.71 SD.

In contrast, the recommendation patterns for advanced courses between the methods were quite different. While Q-Learning recommended that more than 98% of students from both races receive advanced courses, the weighting method recommended advanced courses for 85.7% of White students and 72.8% of Black students. The weighting method did not recommend advanced courses for students with math achievement scores

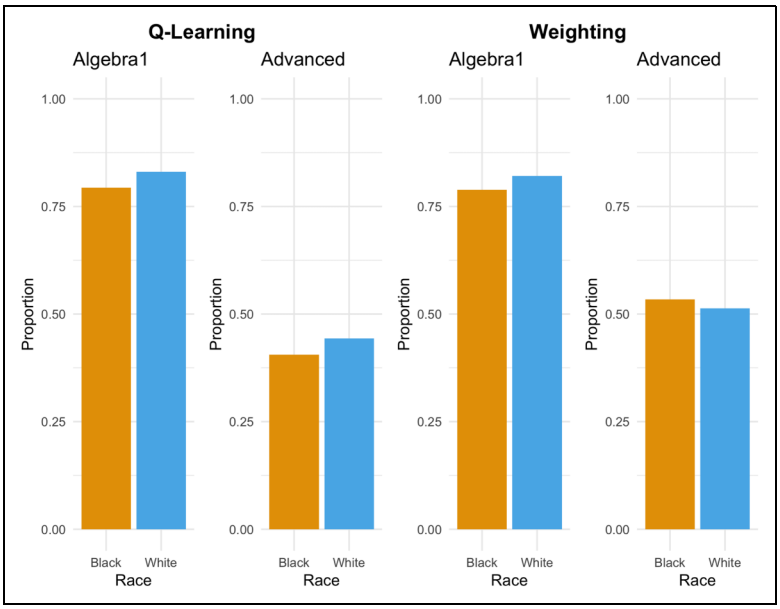


Figure 3. Proportion of Compliance by Race. *Source:* U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. High School Longitudinal Study of 2009 (HSLS:09) Base-Year Restricted-Use File (NCES 2011-333).

lower than −1.06 SD and Q-Learning did not recommend when $(0.098 \times \texttt{9th math score} + 0.04 \times \texttt{math efficacy} - 0.013 \times \texttt{math interest})$ is <-0.244.

We also examined the proportion of students who complied with the recommendations by race (see Figure 3). For Algebra 1, there was no significant difference in the compliance rates between Black and White students for both the Q-learn and weighting methods (Q-learning: 79.3% and 83.0%; weighting: 78.8% and 82.1%; for Blacks and Whites, respectively). However, for advanced courses, the proportion of students who complied with the weighting method recommendation was slightly higher for both White and Black students compared to the Q-learn recommendation (Q-learning: 40.5% and 44.4%; weighting: 53.4% and 51.3%; for Blacks and Whites, respectively).

Note that we use the term "compliance" to refer to instances where individuals' math course-taking patterns align with the optimal rules

(i.e., $I(M = d^{opt}) = 1$). It is noteworthy that our definition of compliance differs from the treatment noncompliance literature (Angrist, Imbens, and Rubin, 1996; Frangakis and Rubin 1999: e.g.,), which addresses the complications of individuals self-selecting into receiving the treatment. In our data, students make decisions regarding math courses without knowledge of the optimal rules, which differs from the context of the treatment noncompliance literature.

Longitudinal Causal Decomposition Analysis (CDA) With Individualized Interventions

In this section, we propose three strategies to reduce outcome disparities by tailoring interventions to individual characteristics. The first two strategies leverage optimal DTRs, which were introduced in the "Review of Optimal DTRs" section. The third strategy involves directly incorporating individual characteristics into the interventional effects.

Individualized Controlled Direct Effects

We propose a strategy for intervention wherein all students follow an optimal course-taking pattern based on their previous math achievement level and motivation. We use optimal DTRs as a reference to follow for each student and examine whether this would reduce the observed disparity in math achievement in 11th grade.

Using the optimal decision rules obtained from optimal DTRs, we define disparity remaining as

$$\zeta_c^{\mathrm{ICDE}}(d^{\mathrm{opt}}) \equiv E[Y(d_1^{\mathrm{opt}}, d_2^{\mathrm{opt}})|R = 1, C = c] - E[Y(d_1^{\mathrm{opt}}, d_2^{\mathrm{opt}})|R = 0, C = c], \tag{12}$$

for $c \in \mathcal{C}$, where d_j^{opt} is an optimal value for mediator M_j. This definition of disparity remaining ($\zeta_c^{\mathrm{ICDE}}(d^{\mathrm{opt}})$) states the difference in outcome between groups given baseline covariates after setting the mediator values to their optimal values, which are obtained from the optimal decision rules. We refer to this definition as individualized CDEs (ICDEs). In our example, ICDEs represent the degree to which the Black–White math disparity would remain if all students followed the optimal course-taking rule determined by their prior math achievement and motivation within the same gender status.

Given Assumption A1 to A4, $E[Y(d_1^{\text{opt}}, d_2^{\text{opt}})|R = r, C = c]$ is nonparametrically identified as

$$= E\Big[\frac{I(M_1 = d_1^{\text{opt}}, M_2 = d_2^{\text{opt}})}{P(M_1|R, X_1, C)P(M_2|R, X_1, M_1, X_2, C)} Y \Big| R = r, C = c\Big], \quad (13)$$

for $r = 0, 1$ and $c \in C$ where $I(\cdot)$ is an indicator function. A proof of this identification result is given in Appendix A of the Supplemental Materials.

One straightforward approach to estimate the disparity remaining after the individualized intervention is through an MSM. We fit the following model, where baseline covariates are centered at $C = c$:

$$Y = \gamma_1 + \gamma_2 R + \gamma_3 C + \epsilon_2, \quad (14)$$

given the weight of

$$W_{\texttt{ICDE}} = \frac{I(M_1 = d_1^{\text{opt}},\ M_2 = d_2^{\text{opt}})}{P(M_1|R, X_1, C)P(M_2|R, X_1, M_1, X_2, C)}.$$

Then, the disparity remaining is estimated as $\hat{\zeta}_c^{\texttt{ICDE}} = \hat{\gamma}_2$. Here, we assumed binary mediators and a continuous outcome. However, the MSM estimator can be easily modified for a binary outcome by fitting a logistic regression and computing the average difference in predicted probabilities. Alternatively, a weighting estimator can be used by directly applying equation (13) as:

$$\hat{\zeta}_c^{\texttt{ICDE}} = \hat{E}\big[\hat{W}_{\texttt{ICDE}} Y \big| R = 1, C = c\big] - \hat{E}\big[\hat{W}_{\texttt{ICDE}} Y \big| R = 0, C = c\big]. \quad (15)$$

The weighting estimator can be applied for both binary and continuous outcomes.

Individualized Interventional Effects

Optimal rules are not definitive and may not be suitable for every student. Hence, it is expected that not all students' course-taking patterns are consistent with the optimal rules. However, if one racial group has a tendency for their course-taking patterns to be more inconsistent with the optimal rules than other groups, it may be problematic in terms of reducing disparities. In this case, we could consider an intervention to equalize the rate of being consistent with optimal rules across groups. Ideally, the reference group's course-taking patterns should be more consistent with optimal rules. For illustrative purposes, we will use White students as the reference group although White students are not more compliant than Black students (see Figure 3).

Without loss of generality, let $K_j = G_{I(M_j=d_j^{\text{opt}})|R=0,C} \times d_j^{\text{opt}} + (1 - G_{I(M_j=d_j^{\text{opt}})|R=0,C}) \times (1 - d_j^{\text{opt}})$, where d_j^{opt} is an optimal value for mediator M_j, and $G_{I(M_j=d_j^{\text{opt}})|R=0,C}$ is a random draw from the compliance distribution for M_j of the reference group given baseline covariates. Then, the disparity reduction and disparity remaining after equalizing compliance rates across groups can be defined as:

$$\begin{aligned} &\delta_c^{\text{IIE}}(1) \equiv E[Y|R=1, c] - E[Y(K_1, K_2)|R=1, C=c], \quad \text{and} \\ &\zeta_c^{\text{IIE}}(0) \equiv E[Y(K_1, K_2)|R=1, c] - E[Y|R=0, C=c], \end{aligned} \tag{16}$$

for $c \in C$. Using potential outcomes notation, $Y(K_1, K_2)$ represents an outcome that is realized under mediators determined by a random draw from the compliance distributions for M_1 and M_2 of the reference group given baseline covariates. For instance, if a random draw indicates that the reference individual complied with the recommendation, then the mediators of an individual in the comparison group will likewise align with their assigned optimal values. Disparity reduction ($\delta_c^{\text{IIE}}(1)$) represents the change in the outcome among Black students given baseline covariates after intervening to set the compliance distribution equal to that of White students among those with the same level of baseline covariates; disparity remaining ($\zeta_c^{\text{IIE}}(0)$) represents the difference in the outcome that persists between Black and White students given baseline covariates, even after the hypothetical intervention. In the example, the disparity reduction (remaining) represents the degree to which the disparity in math achievement in 11th grade would be reduced (remain) if Black students complied with optimal rules at the same rate as White students among those with the same gender status. We refer to this definition as *individualized interventional in/direct effects*.

Given Assumptions 1 to 4, $E[Y(K_1, K_2)|R=1, C=c]$ is nonparametrically identified as follows:

$$\begin{aligned} &= \sum_{\theta_1,\theta_2} P(I(M_1 = d_1^{\text{opt}}) = \theta_1 | R=0, C=c) P(I(M_2 = d_2^{\text{opt}}) = \theta_2 | R=0, C=c) \\ &\times E\left[\frac{I\big(M_1 = \theta_1 d_1^{\text{opt}} + (1-\theta_1)(1-d_1^{\text{opt}}), M_2 = \theta_2 d_2^{\text{opt}} + (1-\theta_2)(1-d_2^{\text{opt}})\big) Y}{P\big(M_1 | X_1, R, C\big) P\big(M_2 | M_1, X_1, X_2, R, C\big)} \,\middle|\, R=1, C=c \right], \end{aligned}$$

for $x_j \in X_j$, $m_j \in M_j$, $c \in C$, and $\theta_j \in \{0, 1\}$. A proof of this identification result is given in Appendix B of the Supplemental Materials.

For estimation, we assume the following MSM (where baseline covariates are centered at $C = c$):

$$Y = \lambda_1 + \lambda_2 R + \lambda_3 I(M_1 = d_1^{\text{opt}}) + \lambda_4 I(M_2 = d_2^{\text{opt}}) + \lambda_5 C + \epsilon_2, \tag{17}$$

given the weight of $W_{\text{CDE}} \equiv P(M_1|R = 1, X_1, C)^{-1} P(M_2|R = 1, X_1, M_1, X_2, C)^{-1}$. A more complex model could be specified to address nonlinear relationships, such as interactions between race and compliance status. Suppose further that the following compliance models hold.

$$\begin{aligned} &P(I(M_1 = d_1^{\text{opt}})|R, C) = \text{logit}^{-1}(\phi_{11} + \phi_{12}\text{R} + \phi_{13}\text{C}), \quad \text{and} \\ &P(I(M_2 = d_2^{\text{opt}})|R, C) = \text{logit}^{-1}(\phi_{21} + \phi_{22}\text{R} + \phi_{23}\text{C}). \end{aligned} \tag{18}$$

Then, $\hat{\delta}_c^{\text{IIE}}(1)$ can be obtained as

$$\begin{aligned} &\left\{ \frac{\exp(\hat{\phi}_{11} + \hat{\phi}_{12} + \hat{\phi}_{13}\hat{E}[C])}{1 + \exp(\hat{\phi}_{11} + \hat{\phi}_{12} + \hat{\phi}_{13}\hat{E}[C])} - \frac{\exp(\hat{\phi}_{11} + \hat{\phi}_{13}\hat{E}[C])}{1 + \exp(\hat{\phi}_{11} + \hat{\phi}_{13}\hat{E}[C])} \right\} \times \hat{\lambda}_3 \\ &+ \left\{ \frac{\exp(\hat{\phi}_{21} + \hat{\phi}_{22} + \hat{\phi}_{23}\hat{E}[C])}{1 + \exp(\hat{\phi}_{21} + \hat{\phi}_{22} + \hat{\phi}_{23}\hat{E}[C])} - \frac{\exp(\hat{\phi}_{21} + \hat{\phi}_{23}\hat{E}[C])}{1 + \exp(\hat{\phi}_{21} + \hat{\phi}_{23}\hat{E}[C])} \right\} \times \hat{\lambda}_4; \end{aligned}$$

$\hat{\zeta}_c^{\text{IIE}}(0)$ can be obtained as $\hat{\lambda}_2$, or alternatively as $\hat{\tau}_c - \hat{\delta}_c^{\text{IIE}}(1)$.

Alternatively, a weighting estimator for individualized interventional effect (IIE) can be employed using the following steps:

- **Step 1**: Compute the compliance rate among those with $R = 0$ and $C = c$, denoted as $\pi_{I_j=\theta_j|0c} \equiv P[I(M_j = d_j^{\text{opt}}) = \theta_j|R = 0, C = c]$.
- **Step 2**: Fit a logistic model, regressing each mediator on the history variables among those with $R = 1$ to estimate $P(M_1|R = 1, X_1, C)$ and $P(M_2|R = 1, X_1, M_1, X_2, C)$.
- **Step 3**: For each combination of θ_1 and θ_2, weights can be formed as

$$W_{\text{IIE}}^{\theta_1,\theta_2} \equiv \frac{I\left(M_1 = \theta_1 d_1^{\text{opt}} + (1-\theta_1)(1-d_1^{\text{opt}}), M_2 = \theta_2 d_2^{\text{opt}} + (1-\theta_2)(1-d_2^{\text{opt}})\right)}{P(M_1|R=1, X_1, C) \times P(M_2|R=1, X_1, M_1, X_2, C)}.$$

- **Step 4**: The disparity reduction is estimated as $\hat{\delta}_c^{\text{IIE}}(1) = \hat{E}[Y|R = 1, C = c] - \sum_{\theta_1,\theta_2} \hat{\pi}_{I_1=\theta_1|0c} \hat{\pi}_{I_2=\theta_2|0c} \hat{E}[\hat{W_{\text{IIE}}}^{\theta_1,\theta_2} Y|R = 1, C = c]$ and disparity remaining is estimated as $\hat{\zeta}_c^{\text{IIE}}(0) = \sum_{\theta_1,\theta_2} \hat{\pi}_{I_1=\theta_1|0c} \hat{\pi}_{I_2=\theta_2|0c} \hat{E}[\hat{W_{\text{IIE}}}^{\theta_1,\theta_2} Y|R = 1, C = c] - \hat{E}[Y|R = 0, C = c]$.

Individualized Conditional Effects

The last individualized intervention strategy we propose is to use interventional effects that incorporate individual characteristics. This strategy is similar to interventional conditional in/direct effects proposed by Zheng and van der Laan (2017). Interventional conditional direct effects estimate the potential effects of randomly assigning Black students to Algebra 1 and advanced math classes at a rate comparable to White students within the same level of all existing intermediate confounding variables (e.g., SES, parent career aspiration, math course availability at school, and prior math achievement). However, this approach has an important limitation. Taking into account SES or course availability at school when assigning students to mathematics courses carries the risk of reinforcing structural racism (McGee 2020; Jackson 2021), where Black students are more likely to be born into low SES families and attend schools with low quality that may not offer advanced mathematics courses.

In a sense, our proposed strategy resembles a time-varying version of Jackson (2021), as we assign math courses based on selective covariates driven by equity concerns. However, in addition to equity considerations, we suggest selecting covariates that modify the effect of taking math courses, allowing for a personalized intervention. The idea is to equalize the mediator distribution between the groups among those who would similarly benefit from the intervention, thus enabling tailored interventions based on their individual characteristics. Therefore, we propose here that we only condition on a subset of intermediate confounders, which were previously considered in obtaining optimal DTRs. As before, these variables are denoted as H_{j1} for $j = 1, 2$. We refer to these variables as *individualized* sources of difference, which must be considered when designing interventions to reduce disparities that are tailored to individual characteristics. Formally, we define the disparity reduction and disparity remaining as

$$\begin{aligned}\delta_c^{\mathrm{ICE}} &\equiv E[Y|R=1, C=c] - E[Y(G_{M_1|R=0,H_{11},C}, G_{M_2|R=0,H_{21},C})|R=1, C=c],\\ &\text{and}\\ \zeta_c^{\mathrm{ICE}} &\equiv E[Y(G_{M_1|R=0,H_{11},C}, G_{M_2|R=0,H_{21},C})|R=1, C=c] - E[Y|R=0, C=c],\end{aligned} \tag{19}$$

for $c \in C$, where $G_{M_j|R=0,H_{j1},C}$ is a random draw from the mediator distribution of the reference group given the individualized factors and

baseline covariates. This definition of disparity reduction represents the change in the outcome given baseline covariates after setting the mediator distributions equal between groups among those who have the same levels of the selected intermediate confounders and baseline covariates; disparity remaining represents the difference in the outcome that persists between groups given baseline covariates, even after the intervention. In the example, disparity reduction (remaining) reflects the degree to which the outcome disparity would be reduced (remain) after Black students were enrolled in Algebra 1 and advanced math at the same rate as White students who possess similar prior academic achievement and motivation levels (H_{11} and H_{21}) and gender status (C). We refer to this definition as *individualized conditional in/direct effects*.

Given Assumptions A1 to A4, $E[Y(G_{M_1|R=0,H_{11},C}, G_{M_2|R=0,H_{21},C})|R = 1, c]$ is identified as

$$= E\Big[\frac{P(M_1|R = 0, H_{11}, C)}{P(M_1|R, X_1, C)}\frac{P(M_2|R = 0, H_{21}, C)}{P(M_2|R, X_1, M_1, X_2, C)}Y\Big|R = 1, C = c\Big], \tag{20}$$

for $c \in C$. A proof of this identification result is given in Appendix C of the Supplemental Materials.

The weighting estimation steps of disparity reduction and remaining based on individualized conditional effects (ICEs) are as follows:

- **Step 1**: Fit a numerator model, regressing each mediator on a subset of intermediate confounders and baseline covariates among those with $R = 0$. Using the fitted model, compute the predicted probability of each mediator using the confounders of those with $R = 1$ to estimate $P(M_j|R = 0, H_{j1}, C)$.
- **Step 2**: Fit a denominator model, regressing each mediator on the history variables among those with $R = 1$. Using the fitted model, compute the predicted probability of each mediator among those with $R = 1$ to estimate $P(M_1|R = 1, X_1, C)$ and $P(M_2|R = 1, X_1, M_1, X_2, C)$.
- **Step 3**: Weights can be formed as

$$W_{\texttt{ICE}} \equiv \frac{P(M_1|R = 0, H_{11}, C)}{P(M_1|R = 1, X_1, C)}\frac{P(M_2|R = 0, H_{21}, C)}{P(M_2|R = 1, X_1, M_1, X_2, C)}.$$

- **Step 4**: The disparity reduction is estimated as $\hat{\delta}_c^{\text{ICE}}(1) = \hat{E}[Y|R = 1, C = c] - \hat{E}[\hat{W}_{\text{ICE}}Y|R = 1, C = c]$ and disparity remaining is estimated as $\hat{\zeta}_c^{\text{ICE}}(0) = \hat{E}[\hat{W}_{\text{ICE}}Y|R = 1, C = c] - \hat{E}[Y|R = 0, C = c]$.

We do not present an MSM estimator for this ICE since it does not work when there are *M–X* interactions, as assumed in this study.

We conducted a simulation study to evaluate the performance of the proposed estimators for individualized effects when the outcome is continuous. The simulation settings and results are presented in Appendix D of the Supplemental Materials. Briefly, when the sample size is 2000 or more, most estimators demonstrate good performance in terms of bias and root-mean-squared errors. For sample sizes <2000, it is preferable to use the weighting estimators over the MSM estimators.

An Application to HSLS:09

In this section, we present findings on disparity reduction and disparity remaining using the proposed individualized effects (i.e., ICDEs, individualized interventional in/direct effects, and individualized conditional in/direct effects). We used optimal decision rules obtained by the weighting method as Q-Learning resulted in very few students who were not recommended to take Algebra 1 or advanced courses, which posed a modeling issue. Given the sample size of 11,050, we expect both MSM and weighting estimators to perform equally well. For simplicity, we present results obtained by the MSM estimators in Table 2.

First, we estimated the disparity remaining using ICDEs. Following the optimal course-taking rules results in a remaining disparity of $\zeta_c^{\text{ICDE}} = -0.482$ SD. This represents a 16.7% increase from the initial disparity. This increase may be due to either or both of the following reasons: (a) there are more White students who were recommended to take the courses but didn't, compared to Black students in the same situation, and (2) the returns of taking advanced courses for Black students are lower compared to White students. While this increase is substantial, it is lower than the increase observed after enforcing all students to enroll in both Algebra 1 and advanced courses (cf. $\zeta_c^{\text{CDE}}(1, 1) = -0.498$). In addition, following the optimal course-taking rules maximizes the average outcome for both groups of students.

Next, we used IIEs to estimate the disparity reduction and disparity remaining. These effects estimate the potential effect of equalizing the compliance rate between groups within the same gender status. This intervention

Table 2. Estimates of the Initial Disparity, Disparity Reduction, and Disparity Remaining.

	Estimate (SE)		
	ICDE	IIE	ICE
Initial disparity	−0.413***	−0.413***	−0.413***
(SE)	(0.022)	(0.022)	(0.022)
Disparity remaining	−0.482***	−0.407***	−0.424***
(SE)	(0.040)	(0.022)	(0.023)
Disparity reduction		−0.006	0.012
(SE)		(0.005)	(0.010)
% reduction	−16.7%	1.5%	−2.7%

Note: (a) ICDE = individualized controlled direct effect; IIE = individualized interventional effect; ICE = individualized conditional effect; SE = standard error. (b) The asterisk followed by estimates indicates the level of statistical significance (*: significant at 0.05, **: at 0.01, ***: at 0.001). (c) We did not take into account the uncertainty regarding obtaining the optimal values for the mediators. (4) Gender is centered at the mean.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. High School Longitudinal Study of 2009 (HSLS:09) Base-Year Restricted-Use File (NCES 2011-333).

does not yield a significant reduction in initial disparity. The remaining disparity is $\zeta_c^{\mathrm{IIE}} = -0.407$ SD, which is comparable to the initial disparity. This is because there is a similar compliance rate between Black and White students in the observed data (see the right panel of Figure 3: weighting).

Lastly, we used ICEs to estimate disparity reduction and disparity remaining. These effects determine the potential effect of equalizing course-taking patterns between groups among students with similar motivation and prior achievement levels within the same gender status. The ICE does not yield a significant reduction in disparity, with the remaining disparity of $\zeta_c^{\mathrm{ICE}} = -0.424$ SD.

Overall, our findings suggest that following optimal course-taking rules does not necessarily reduce the initial disparity. Additionally, simply equalizing compliance rates across groups or aligning course-taking patterns between groups among students with similar motivation and prior achievement levels may not effectively reduce the initial disparity.

Recommendations for Empirical Researchers

To reduce social disparities in relevant outcomes, policymakers and practitioners can explore the potential effect of hypothetical interventions using

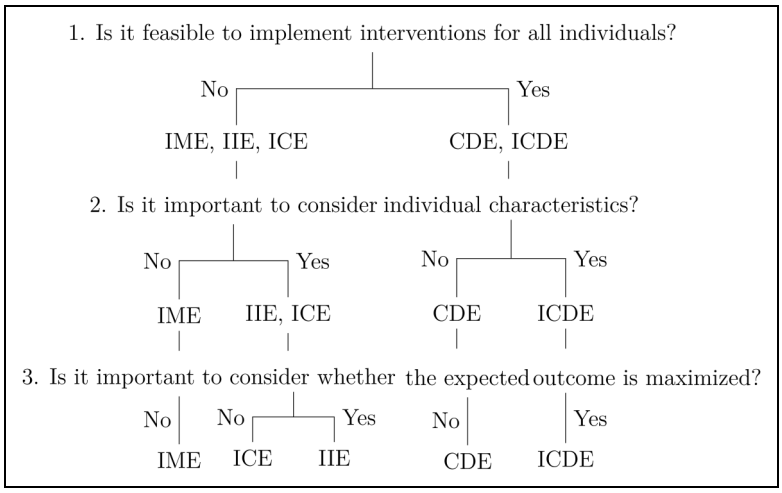


Figure 4. Tree diagram for optimal intervention decision-making.

approaches outlined in the "Extending Existing Longitudinal Mediation Approaches" and "Longitudinal CDA with Individualized Interventions" sections. To identify effective interventions in specific contexts, we propose three guiding questions for consideration. We illustrate the decision-making processes outlined in Figure 4.

The first question to consider is whether it is feasible or desirable to implement interventions for all individuals. For example, mandating that all students take Algebra 1 by the end of the nineth grade may be feasible and even desirable, as shown in Table 1. In this case, static interventions such as CDEs and ICDEs should be considered. If implementing interventions for all individuals is not feasible, stochastic interventions such as IMEs, IIEs, and ICEs become relevant options.

The second question is whether it is essential to consider individual characteristics when designing interventions. As demonstrated in our example, taking into account individual factors such as prior achievement and motivation is crucial in determining whether advanced courses are suitable for each student. In such cases, interventions that follow optimal treatment regimes (ICDE and IIE) or interventions to equalize the mediator distribution among those who have the same individual characteristics (ICE) should be chosen.

Lastly, one should ask whether maximizing each individual's outcome is a priority. In our example, both maximizing the final math score and reducing

racial disparities are important. In this case, interventions that follow optimal treatment regimes such as IIE and ICDE should be considered.

Another point of confusion arises when determining which variables to choose for different sets of confounders. It is crucial to distinguish between baseline covariates and intermediate confounders. Baseline covariates typically include demographic factors such as gender, home language status, or age. The remaining variables that confound the mediator–outcome relationships are intermediate confounders. When defining disparities in educational outcomes as well as risk factors for stochastic interventions, one should decide which variables to adjust based on whether the difference due to the variables is deemed allowable or fair (Jackson 2021). For example, it might be deemed unfair to remove differences attributable to SES when defining racial disparities in math achievement and math course-taking patterns.

In addition, in determining optimal regimes, it is necessary to select intermediate confounders H_{j1}, which represent a subset of H_j that has heterogeneous (interaction) effects based on the mediator level. Typically, H_{j1} variables are confounders in the mediator–outcome relationship. For instance, in the context of students' math course selection, prior math achievement and motivation might interact with taking advanced math courses while also being confounding variables. The variables that do not act as confounders but interact with the mediator can also be included in H_{1j} (Moodie, Chakraborty, and Kramer 2012).

We use the same set of variables, represented as H_{j1}, as conditioning variables for the ICE. However, it is important to highlight that conditioning on these H_{j1} variables is meaningful only if they serve as confounders in the mediator–outcome relationship. In the event that H_{j1} contains a variable that is not a confounder, the conditioning effect through that specific variable would be null.

Conclusion and Discussion

This paper contributes to the fast-growing literature on causal decomposition in three ways. First, we extend existing longitudinal mediation approaches to the context of disparities research. These approaches were previously developed for situations where exposures vary over time. However, in causal decomposition analysis, group status serves as the exposure, which rarely changes over time. We made a straightforward extension to the existing approaches by providing a formal definition, identification assumptions, and an illustrative example. Although the extension may seem methodologically trivial, it has practical implications by facilitating the investigation of contributing factors to social disparities that evolve over time.

Second, we combine the existing longitudinal mediation approaches with optimal DTRs that take into account individual characteristics. Simply applying the same intervention to all individuals or intervening to equalize the risk factors or resources between groups may be infeasible or not beneficial to individuals with different characteristics. Our new approach considers individual characteristics by using optimal DTRs, which were originally designed to maximize the average outcome. This new method represents a paradigm shift from existing causal decomposition methods that use a static or stochastic intervention by allowing interventions that are tailored to individual characteristics.

Third, we propose the individualized conditional in/direct effects by conditioning on a selected set of individual characteristics that modify the effect of risk factors. This approach allows researchers to equalize the distribution of risk factors between groups among those who similarly benefit from the interventions. Previously, the judgment on selecting conditioning variables for stochastic interventions has been solely based on equity considerations, that is, whether it is fair to remove the source of differences due to the conditioning variables (Jackson 2021). Our approach suggests another way of selecting conditioning variables based on individual characteristics that determine whether they benefit from the intervention or not.

It is important to acknowledge the limitations of our study. First, we did not consider the uncertainty involved in obtaining the optimal values for the mediators when combining causal decomposition analysis with optimal DTRs. This could potentially lead to smaller standard errors, resulting in an inflated Type I error rate. Second, the assumptions required to identify the effects of interest (disparity reduction and disparity remaining) are strong. Therefore, it is essential for future studies to develop sensitivity analysis techniques to examine the potential effect of violating the assumptions. Third, the proposed models are based on two time points and assume no nested structure of the data. Future studies can extend these models to more complex scenarios, such as cases with more than two time points or with multilevel models.

Data Availability

The HSLS:09 data utilized for the case study is accessible via NCES's restricted-use License Program. Details can be found at https://nces.ed.gov/surveys/hsls09.asp. Additionally, the R codes for both the simulation study and case study presented in this article are available on the first author's GitHub repository at https://github.com/soojinpark33/Individualized-Intervention.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Soojin Park gratefully acknowledges support from the National Science Foundation (NSF #2243119) and the American Educational Research Association Grants Program funded by the National Science Foundation (NSF-DRL #1749275).

ORCID iDs

Soojin Park https://orcid.org/0000-0003-0288-5589
Namhwa Lee https://orcid.org/0009-0009-3416-6512
Rafael Quintana https://orcid.org/0000-0003-4776-9362

Supplemental Material

The supplemental material for this article is available online.

Notes

1. Hence, the conditional estimates of disparity reduction and disparity remaining are equivalent to the marginal estimates, which are averaged over the gender distribution.
2. Math self-efficacy refers to individual's beliefs about their math abilities (Bandura et al. 2001).

References

Angrist, J.D., G.W. Imbens, and D.B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American statistical Association* 91(434):444–55.

Attewell, P. and T. Domina. 2008. "Raising the Bar: Curricular Intensity and Academic Performance." *Educational Evaluation and Policy Analysis* 30(1):51–71.

Bandura, A., C. Barbaranelli, G.V. Caprara, and C. Pastorelli. 2001. "Self-Efficacy Beliefs as Shapers of Children's Aspirations and Career Trajectories." *Child Development* 72(1):187–206.

Bind, M.-A., T. Vanderweele, B. Coull, and J. Schwartz. 2016. "Causal Mediation Analysis for Longitudinal Data With Exogenous Exposure." *Biostatistics (Oxford, England)* 17(1):122–34.

Breiman, L. 2017. *Classification and Regression Trees*. New York: Routledge.

Byun, S.-Y., M.J. Irvin, and B.A. Bell. 2015. "Advanced Math Course Taking: Effects on Math Achievement and College Enrollment." *The Journal of Experimental Education* 83(4):439–68.

Cole, S.R. and M.A. Hernàn. 2008. "Constructing Inverse Probability Weights for Marginal Structural Models." *American Journal of Epidemiology* 168(6):656–64.

Dougherty, S.M., J.S. Goodman, D.V. Hill, E.G. Litke, and L.C. Page. 2017. "Objective Course Placement and College Readiness: Evidence From Targeted Middle School Math Acceleration." *Economics of Education Review* 58:141–61.

Frangakis, C.E. and D.B. Rubin. 1999. "Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes." *Biometrika* 86(2):365–79.

Ingels, S.J., D.J. Pratt, D.R. Herget, J.A. Dever, L.B. Fritch, R. Ottem, and S. Leinwand. 2013. *High School Longitudinal Study of 2009 (HSLS: 09) Base Year to First Follow-Up Data File Documentation. Appendixes. NCES 2014-361*. National Center for Education Statistics.

Jackson, J.W. 2021. "Meaningful Causal Decompositions in Health Equity Research: Definition, Identification, and Estimation Through a Weighting Framework." *Epidemiology (Cambridge, Mass.)* 32(2):282–90.

Jackson, J.W. and T. VanderWeele. 2018. "Decomposition Analysis to Identify Intervention Targets for Reducing Disparities." *Epidemiology (Cambridge, Mass.)* 29(6):825–35.

Kalogrides, D. and S. Loeb. 2013. "Different Teachers, Different Peers: The Magnitude of Student Sorting Within Schools." *Educational Researcher* 42(6):304–16.

Kaufman, J.S. 2008. "Epidemiologic Analysis of Racial/Ethnic Disparities: Some Fundamental Issues and a Cautionary Example." *Social Science & Medicine* 66(8):1659–69.

Kelly, S. 2009. "The Black–White Gap in Mathematics Course Taking." *Sociology of Education* 82(1):47–69.

Lee, C., S. Park, and J.M. Boylan. 2021. "Cardiovascular Health at the Intersection of Race and Gender: Identifying Life-Course Processes to Reduce Health Disparities." *The Journals of Gerontology: Series B* 76(6):1127–39.

Long, M.C., D. Conger, and P. Iatarola. 2012. "Effects of High School Course-Taking on Secondary and Postsecondary Success." *American Educational Research Journal* 49(2):285–322.

Lundberg, I. 2020. "The Gap-Closing Estimand: A Causal Approach to Study Interventions That Close Disparities Across Social Categories." *Sociological Methods & Research* 53(2):507–70.

Mahar, R.K., M.B. McGuinness, B. Chakraborty, J.B. Carlin, M.J. IJzerman, and J.A. Simpson. 2021. "A Scoping Review of Studies Using Observational Data to Optimise Dynamic Treatment Regimens." *BMC Medical Research Methodology* 21:1–13.

McEachin, A., T. Domina, and A. Penner. 2020. "Heterogeneous Effects of Early Algebra Across California Middle Schools." *Journal of Policy Analysis and Management* 39(3):772–800.

McGee, E.O. 2020. "Interrogating Structural Racism in Stem Higher Education." *Educational Researcher* 49(9):633–44.

Moodie, E.E., B. Chakraborty, and M.S. Kramer. 2012. "Q-Learning for Estimating Optimal Dynamic Treatment Rules From Observational Data." *Canadian Journal of Statistics* 40(4):629–45.

Murphy, S.A. 2003. "Optimal Dynamic Treatment Regimes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65(2):331–55.

Park, S., S. Kang, C. Lee, and S Ma. 2023. "Sensitivity Analysis for Causal Decomposition Analysis: Assessing Robustness Toward Omitted Variable Bias." *Journal of Causal Inference* 11(1):20220031.

Park, S., X. Qin, and C. Lee. 2022. "Estimation and Sensitivity Analysis for Causal Decomposition in Health Disparity Research." *Sociological Methods & Research* 53(2):571–602.

Pearl, J. 2012. "The Causal Mediation Formula—A Guide to the Assessment of Pathways and Mechanisms." *Prevention Science* 13(4):426–36.

Pearl, J.2022 "Direct and Indirect Effects." In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 373–92. Association for Computing Machinery.

Riegle-Crumb, C. and E. Grodsky. 2010. "Racial-Ethnic Differences at the Intersection of Math Course-Taking and Achievement." *Sociology of Education* 83(3):248–70.

Robins, J.M., M.A. Hernan, and B. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology (Cambridge, Mass.)* 11(5):550–60.

Simzar, R., T. Domina, and C. Tran. 2016a. "Eighth-Grade Algebra Course Placement and Student Motivation for Mathematics." *AERA Open* 2(1):233285 8415625227.

Simzar, R., T. Domina, and C. Tran. 2016b. "Eighth-Grade Algebra Course Placement and Student Motivation for Mathematics." *AERA Open* 2(1):2332858415625227.

Tsiatis, A.A., M. Davidian, S.T. Holloway, and E.B. Laber. 2019. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. New York: Chapman and Hall/CRC.

VanderWeele, T. and W.R. Robinson. 2014. "On Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables." *Epidemiology (Cambridge, Mass.)* 25(4):473–83.

VanderWeele, T. and E.J. Tchetgen Tchetgen. 2017. "Mediation Analysis With Time Varying Exposures and Mediators." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79(3):917–38.

VanderWeele, T., S. Vansteelandt, and J.M. Robins. 2014. "Effect Decomposition in the Presence of an Exposure-Induced Mediator–Outcome Confounder." *Epidemiology (Cambridge, Mass.)* 25(2):300.

Zhang, B., A.A. Tsiatis, M. Davidian, M. Zhang, and E. Laber. 2012. "Estimating Optimal Treatment Regimes From a Classification Perspective." *Stat* 1(1):103–14.

Zhao, Y.-Q., D. Zeng, E.B. Laber, and M.R. Kosorok. 2015. "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes." *Journal of the American Statistical Association* 110(510):583–98.

Zheng, W. and M. van der Laan. 2017. "Longitudinal Mediation Analysis With Time-Varying Mediators and Exposures, With Application to Survival Outcomes." *Journal of Causal Inference* 5(2).

Author Biographies

Soojin Park is an assistant professor of quantitative methods in the School of Education at the University of California, Riverside. Her research focuses on developing and validating quantitative methods for causal inference, which she uses to investigate the factors contributing to racial and gendered disparities in educational and health outcomes.

Namhwa Lee is a PhD student in the Department of Statistics at the University of California, Riverside. His research interest revolves around causal inference, applying these methodologies to address real-world problems in areas such as mental health and educational inquiries.

Rafael Quintana is an assistant professor in the Department of Educational Psychology at the University of Kansas. His research focuses on skill development, educational inequality, and quantitative methods, with special emphasis on graphical models, causal inference, and longitudinal analysis.