Causal Inference for Time-Varying Instructional Treatments

Guanglei Hong and Stephen W. Raudenbush

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2008; 33; 333
originally published online Jan 22, 2008;
DOI: 10.3102/1076998607307355

Journal of Educational and Behavioral Statistics
September 2008, Vol. 33, No. 3, pp. 333–362
DOI: 10.3102/1076998607307355

Causal Inference for Time-Varying Instructional Treatments

Guanglei Hong
Ontario Institute for Studies in Education of the University of Toronto

Stephen W. Raudenbush
University of Chicago

The authors propose a strategy for studying the effects of time-varying instructional treatments on repeatedly observed student achievement. This approach responds to three challenges: (a) The yearly reallocation of students to classrooms and teachers creates a complex structure of dependence among responses; (b) a child's learning outcome under a certain treatment may depend on the treatment assignment of other children, the skill of the teacher, and the classmates and teachers encountered in the past years; and (c) time-varying confounding poses special problems of endogeneity. The authors address these challenges by modifying the stable unit treatment value assumption to identify potential outcomes and causal effects and by integrating inverse probability of treatment weighting into a four-way value-added hierarchical model with pseudolikelihood estimation. Using data from the Longitudinal Analysis of School Change and Performance, the authors apply these methods to study the impact of "intensive math instruction" in Grades 4 and 5.

Keywords: *potential outcomes; stable unit treatment value assumption; value-added model; inverse probability of treatment weighting; pseudolikelihood estimation*

1. Introduction

Instructional practice is the proximal cause of students' academic learning. Identifying effective instructional interventions has therefore been a central aim of educational research and one that has attracted especially widespread interest

The names of the authors are in alphabetical order. This research was supported by a grant from the Spencer Foundation titled *Improving Research on Instruction: Models, Designs, and Analytic Methods*. Additional support came from a National Academy of Education/Spencer Postdoctoral Fellowship awarded to the first author. This article is a substantial revision of an earlier paper supported by Westat subcontract 9844—Longitudinal Evaluation of School Change and Performance (LESCP): A Secondary Analysis. Richard Congdon deserves special thanks for the applications programming. We wish to thank Susan Murphy for her numerous and extremely helpful discussions about the issues we address in this article.

recently (Cohen, Raudenbush, & Ball, 2003). Causal-comparative studies of instruction tend to be bounded by a single academic year. Although understanding instructional effects during a single year is necessary, it is essential to understand how sequences of instruction extending over 2 or more years cumulatively affect learning. The effect of a multiyear sequence of instructional experiences cannot logically be equated to the sum of the effects of instruction occurring each year. For example, if the instruction in a later year builds effectively on students' instructional experiences in the earlier years, the benefit of the sequence may substantially exceed the sum of the yearly effects.

Our aim in this article is to adapt causal inference concepts and methods based on potential outcomes (Holland, 1986; Neyman, 1923/1990; Rubin, 1978; see also Haavelmo, 1943; Heckman, 2005) to the nonexperimental study of time-varying instructional treatments. For illustration, we estimate the causal effect of a sequence of intensive math instructional treatments on student learning in Grades 4 and 5. Data are from the Longitudinal Evaluation of School Change and Performance (LESCP). The novel treatment exposes children to instruction that emphasizes comparatively high-level content and devotes substantial classroom time to mathematics. We ask: (a) What is the effect of intensive math instruction in Grade 4 on Grade 4 outcome? (b) What is the effect of this treatment in Grade 4 alone on Grade 5 outcome? (c) What is the effect of intensive math instruction in Grade 5 alone on Grade 5 outcome? (d) Does experiencing intensive math instruction again in Grade 5 enhance the effect of intensive math instruction received in Grade 4?

The case study provides a typical example for evaluating multiyear instructional sequences, posing three characteristic methodological challenges.

1. Complex multilevel structure. Longitudinal studies of classroom treatments are complicated by the reassignment of students to teachers and classes at the beginning of each year. As a result, repeated assessments of students are cross-classified by teachers who are in turn typically nested within schools. Model specification and statistical adjustment pose tricky problems in the context of this complex data structure. We develop a four-way hierarchical linear model to be analyzed via pseudolikelihood estimation.

2. Violation of stable unit treatment value assumption. In causal comparative studies, it is common to assume that each participant has a single potential outcome under each treatment. This is the stable unit treatment value assumption (SUTVA), articulated by Rubin (1986). SUTVA requires that a participant's potential outcome under a given treatment should not depend on the treatment assignment of other participants and that the causal effect of a treatment should not depend on the assignment mechanism. In our application, a teacher delivers intensive math instruction to a class of students within a year. Hence, a child's learning outcome may depend on the teacher and classmates as well as the school context during that year. A child's learning outcome may also depend on the teachers and classmates that the child has encountered in the previous years, generating, in principle, a vast number of potential outcomes per student per treatment. To cope with this problem, we propose a weaker form of SUTVA that allows potential outcomes to depend on current and past school and class assignments. We offer a rationale for this approach based on theory and empirical evidence.

3. Time-invariant and time-varying confounding. The assignment of a class to receive intensive math instruction may depend on a host of factors including school and teacher characteristics, class composition of student demographic characteristics, their instructional histories, and past learning outcomes. Standard methods of adjustment, though sufficient for removing observed time-invariant confounding, can lead to bias in the presence of *time-varying confounders*, defined as covariates that are outcomes of prior treatments but also predictors of later treatment assignments. To cope with this problem, we adapt inverse probability of treatment weighting (IPTW) as developed by Robins (2000) to complex multilevel data.

In Section 2 we present our theoretical approach to defining the potential outcomes, causal effects, and causal estimands under a relaxed version of SUTVA. Section 3 embeds the potential outcomes and causal effects in a growth model for children who are moving across classrooms that are in turn nested within schools. Section 4 defines the model for the observed data. Section 5 provides a rationale for IPTW to remove observed confounding in the multilevel context and describes our pseudolikelihood approach to estimation. Section 6 applies the model and the adjustment method to the case study data. Section 7 revisits key assumptions made in case study and highlights some unsolved methodological issues.

2. Defining Potential Outcomes and Causal Effects of Instruction

Single-Year Treatments

We first consider the potential outcomes causal framework in its general form for a single-year study. Let $z = 1$ denote intensive math instruction and $z = 0$ for conventional instruction. Formally, with N units in the population, we have the 1 by N vector of possible treatment assignments, $\mathbf{z} = (z_1, z_2, \ldots, z_N)$. Unless we impose constraints, student i's potential outcome $Y_i\ (\mathbf{z})$ may depend on the treatment assignments for all the N units. Hence the general causal estimand takes the form $E[Y(\mathbf{z}) - Y(\mathbf{z}')]$, where $\mathbf{z}$ and $\mathbf{z}'$ are alternative treatment assignment vectors. Under this setup, student i's potential outcome under each treatment can be affected by a shift in the treatment assignment of any other student. Without further simplification, causal inference becomes intractable. So our first challenge is to place sensible constraints on the potential outcomes.

SUTVA has been invoked in the past to simplify the causal effect of interest by stipulating that $Y_i(\mathbf{z}) = Y_i(z_i)$. When applied to school settings, this strong assumption ignores the interactions between teacher and students in a class and therefore fails to reflect the nature of instruction and its effect on student learning. Recent work has extended Rubin's causal model by invoking weaker and relatively more plausible assumptions that support causal inference in multilevel settings in a single time period (Gitelman, 2005; Hong & Raudenbush, 2006; Sobel, 2006).

Our current study is focused on instructional treatments assigned to classes within schools. Following Hong and Raudenbush (2006), we adopt a weaker form of SUTVA appropriate to the school setting. We assume that generalization of causal inferences is restricted to current school assignments (i.e., intact schools) and that there is no interference between schools. In addition, when the class-level treatment is given, a child's learning outcome depends mainly on the teacher and classmates that the child directly encounters and is unlikely affected by teachers and students in other classes. Hence, it seems reasonable to assume no interference between classes within an intact school. Under these assumptions, we have the generic potential outcome for student i attending classroom j in school k as $Y_{ijk}(z_{jk})$. In words, a student's potential outcome value is assumed stable given the school assignment and class assignment and given the treatment assigned to the class. We are interested in causal effects having the form $Y_{ijk}(1) - Y_{ijk}(0)$, holding constant the class and school attended. We are not interested, for example, in causal effects of the form $Y_{ijk}(1) - Y_{ij'k'}(0)$ for $j \neq j'$ or $k \neq k'$.

Multiyear Sequences of Treatments

Let us now consider the causal effects of a 2-year sequence of instructional treatments. The logic can easily be extended to treatments over more than 2 years. A student's learning outcome depends not only on the current year's instructional treatment but also on the treatments received in the earlier years. In our current study, in addition to assuming intact schools and no interference between schools, we also assume that a student's potential outcome values associated with a treatment sequence can be affected only by the sequence of teachers and classmates that the student has directly encountered. In the discussion section, we consider conditions under which such an assumption may become unreasonable.

Instructional treatments over several years might be prescribed as a mandatory sequence regardless of students' intermediate status. This would happen, for example, if Year 1 and Year 2 teachers were to follow a standard curriculum. In an alternative scenario, teachers in a school may follow a common set of dynamic rules in assigning instructional treatments given students' current cognitive status. For example, one might assign all students scoring above a cutoff point on a test to receive intensive math instruction while those scoring below that cutoff point would receive conventional math instruction.

However, schooling in the United States has been characterized as a "loosely coupled system" (Weick, 1976) in which elementary school teachers in particular have considerable autonomy in determining the pace and difficulty level of instruction. Moreover, analysis of our data suggests considerable uncertainty in predicting which classes will receive intensive math instruction given past test scores, past treatments, and other background characteristics of students. In particular, given that future treatment assignment is not strongly predicted by past observable outcomes, we reason that instructional assignments reflected in our data are neither prescribed nor entirely dynamically adaptive and that all different combinations of Year 1 and Year 2 treatments are possible for most students.

Potential Outcomes and Causal Effects

We formalize the treatment assignment process in a 2-year study as follows.

Year 1. In the fall of Year 1 of the study, student i attending school k, having been assigned to teacher j_0k in the previous year, is now assigned to teacher j_1k who decides whether to adopt intensive math instruction ($z_{ij_1k}=1$) or nonintensive math instruction ($z_{ij_1k}=0$). The teacher's decision can be influenced by teacher, school, and child characteristics captured in the covariate vector $\mathbf{X}_{1j_1k}$ and by her students' past math achievement $\mathbf{Y}_{0j_1k}$. The conditional probability that the teacher will adopt the intensive instruction is a function $h(z_{1j_1k}=1|\mathbf{X}_{1j_1k},\ \mathbf{Y}_{0j_1k})=h_{1j_1k}$ of the aforementioned factors. This process generates for each student two potential outcomes $Y_{1ij_0j_1k}(z_{1j_1k})$ for $z_{1j_1k}\in\{0,1\}$. The difference between these is the child-specific causal effect $\Delta_{1ij_0j_1k}=Y_{1ij_0j_1k}(1)-Y_{1ij_0j_1k}(0)$. The causal estimand $\delta_1=E(\Delta_1)$ defines the average causal effect of intensive math instruction in Year 1 on Year 1 outcome, which answers our first causal question.

Year 2. In the fall of Year 2 of the study, student i attending school k is now assigned to teacher j_2k. The Year 2 teacher may observe not only past covariates $\mathbf{X}_{1j_2k}$ but also a vector of time-varying covariates $\mathbf{X}_{2j_2k}$ in addition to the past treatment experiences $\mathbf{Z}_{1j_2k}$ and past achievement records $\mathbf{Y}_{0j_2k}$ and $\mathbf{Y}_{1j_2k}$ of all her students. She therefore selects intensive math instruction for her class with probability $h(z_{2j_2k}=1|\mathbf{X}_{1j_2k},\mathbf{X}_{2j_2k},\mathbf{Y}_{0j_2k},\mathbf{Y}_{1j_2k},\mathbf{Z}_{1j_2k})=h_{2j_2k}$. This process generates for child i in her class four potential outcomes having the form $Y_{2ij_0j_1j_2k}\ (z_{1j_1k},z_{2j_2k})$ for $z_{1j_1k},z_{2j_2k}\in\{0,1\}$. Three child-specific causal effects of interest to us are $\Delta_{21ij_0j_1j_2k}=Y_{2ij_0j_1j_2k}(1,0)-Y_{2ij_0j_1j_2k}(0,0)$; $\Delta_{22ij_0j_1j_2k}=Y_{2ij_0j_1j_2k}(0,1)-Y_{2ij_0j_1j_2k}(0,0)$; and $\Delta^*_{ij_0j_1j_2k}=Y_{2ij_0j_1j_2k}(1,1)-Y_{2ij_0j_1j_2k}(0,0)-\Delta_{22ij_0j_1j_2k}-\ \Delta_{21ij_0j_1j_2k}$. The causal estimands $\delta_{21}=E(\Delta_{21})$, $\delta_{22}=E(\Delta_{22})$, and $\delta^*=E(\Delta^*)$, taking expectations over all the children and all their past and current teacher/class assignments in all schools, correspond to causal Questions 2 through 4, respectively. Specifically, δ_{21} is the average causal effect of intensive math instruction in Year 1 alone on Year 2 outcome; δ_{22} is the average causal effect of intensive math instruction in Year 2 alone on Year 2 outcome; and δ^* is the average amplifying effect of having a second year of intensive math instruction given that the child also received it in Year 1.

3. Embedding the Causal Effects in a Four-Way Hierarchical Linear Model

Value-Added Model

In the current study, repeated assessments of students are cross-classified by children and teachers who are in turn nested within schools. We therefore formulate a value-added model that reflects classroom contributions to student growth (Raudenbush & Bryk, 2002, chapter 12, example 2). Suppose that over the 3 study years, student i in school k encountering typical classmates and teachers would display a linear growth trajectory. Indeed, extensive exploratory analysis of our data yielded no evidence against this assumption. However, this person-specific straight-line trajectory can be deflected by experiences in classrooms j_0k during the pretreatment year, j_1k during the first treatment year, and j_2k during the second treatment year. We represent these deflections with additive random teacher/class effects v_{j_0k}, v_{j_1k}, and v_{j_2k} assumed to be cumulative (see McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004, for a discussion of the cumulative effects assumption). Thus, the value-added model in the absence of specific intervention effects is

$$Y_{tij_0\ldots j_tk} = \beta_{0ik} + \beta_{1ik}(t-1) + \sum_{m=0}^{t} v_{j_mk} + \varepsilon_{tik} \tag{1}$$

for $t = 0, 1, 2$ in the current study. Here β_{0ik} is the child's status at $t = 1$, β_{1ik} is the child's growth rate, and ε_{tik} is a random error assumed independently and identically distributed as $N(0, \sigma^2)$. The child-specific intercepts and growth rates may vary within and between schools as a function of school random effects u_{0k}, u_{1k} and child random effects r_{0ik}, r_{1ik}:

$$\begin{aligned} \beta_{0ik} &= \gamma_0 + u_{0k} + r_{0ik}, \\ \beta_{1ik} &= \gamma_1 + u_{1k} + r_{1ik}. \end{aligned} \tag{2}$$

Here $\mathbf{u}_k = (u_{0k}, u_{1k})^T \sim N(0, \boldsymbol{\omega})$ and $\mathbf{r}_{ik} = (r_{0ik}, r_{1ik})^T \sim N(0, \boldsymbol{\tau})$, where $\boldsymbol{\omega}$ and $\boldsymbol{\tau}$ are positive-definite 2 by 2 covariance matrices. We assume $\mathbf{u}_k$, $\mathbf{r}_{ik}$, and ε_{tik} to be mutually independent.

Causal Effects of Interventions in a Value-Added Model

We are interested in the impact of a 2-year sequence of instructional interventions. At the end of Year 0, no treatment has been implemented, so we write

$$Y_{0ij_0k} = \beta_{0ik} - \beta_{1ik} + v_{j_0k} + \varepsilon_{0ik}. \tag{3}$$

During the treatment years, the random effects associated with teachers and students may depend on treatment assignments. The model for the two potential outcomes in Year 1 is

$$Y_{1ij_0j_1k}(z_{1j_1k}) = \beta_{0ik} + \delta_1 z_{1j_1k} + v_{j_0k} + v_{j_1k}(z_{1j_1k}) + \varepsilon_{1ik}(z_{1j_1k}) \tag{4}$$

for $z_{1j_1k} = 0$, 1, and the model for the four Year 2 potential outcomes is

$$\begin{aligned} Y_{2ij_0j_1j_2k}(z_{1j_1k}, z_{2j_2k}) = {} & \beta_{0ik} + \beta_{1ik} + \delta_{21} z_{1j_1k} + \delta_{22} z_{2j_2k} + \delta^* z_{1j_1k} z_{2j_2k} \\ & + v_{j_0k} + v_{j_1k}(z_{1j_1k}) + v_{j_2k}(z_{2j_2k}) + \varepsilon_{2ik}(z_{1j_1k}, z_{2j_2k}) \end{aligned} \tag{5}$$

for $z_{1j_1k} = 0$, 1 and $z_{2j_2k} = 0$, 1. For simplicity, the teacher-specific and student-specific increments are assumed additive. The random effects $v_{j_0k}, v_{j_1k}(z_{1j_1k}), v_{j_2k}(z_{2j_2k}), \varepsilon_{1ik}(z_{1j_1k}), \varepsilon_{2ik}(z_{1j_1k}, z_{2j_2k})$ are assumed to have zero means. To facilitate statistical inference, we impose further distributional assumptions about these random effects in the next section.

4. Model for the Observed Data

Selection of the Observed Data From the Potential Outcomes

In defining the model for a specific child, we will omit subscripts i, j_0, j_1, j_2, and k. For example, we will use z_1 to represent z_{1j_1k} and use y_2 for $y_{2ij_0j_1j_2k}$. For each student, we can observe only one potential outcome in each year. Equation 3 defines the Year 0 outcome. In the following years, the observed outcome depends on treatment assignment. We write the observed Year 1 outcome as a function of the random variable Z_1 that can take on values $z_1 \in \{0, 1\}$:

$$Y_1 = (1 - Z_1 \quad Z_1)\begin{pmatrix} Y_1(0) \\ Y_1(1) \end{pmatrix} = \gamma_0 + u_0 + r_0 + \delta_1 Z_1 + v_0 + v_1 + \varepsilon_1 \tag{6}$$

where $v_1 = Z_1 v_1(1) + (1 - Z_1) v_1(0)$ and $\varepsilon_1 = Z_1 \varepsilon_1(1) + (1 - Z_1)\varepsilon_1(0)$. In Year 2, we observe:

$$\begin{aligned} Y_2 &= [(1 - Z_1)(1 - Z_2) \quad Z_1(1 - Z_2) \quad (1 - Z_1)Z_2 \quad Z_1 Z_2] \begin{pmatrix} Y_2(0,0) \\ Y_2(1,0) \\ Y_2(0,1) \\ Y_2(1,1) \end{pmatrix} \\ &= \gamma_0 + u_0 + r_0 + (\gamma_1 + u_1 + r_1) + \delta_{21} Z_1 + \delta_{22} Z_2 + \delta^* Z_1 Z_2 + v_0 + v_1 + v_2 + \varepsilon_2 \end{aligned} \tag{7}$$

where $v_2 = Z_2 v_2(1) + (1 - Z_2) v_2(0)$ and $\varepsilon_2 = (1 - Z_1)(1 - Z_2)\varepsilon_2(0,0) + Z_1(1 - Z_2)\varepsilon_2(1,0) + (1 - Z_1) Z_2 \varepsilon_2(0,1) + Z_1 Z_2 \varepsilon_2(1,1)$. Under randomization, Z_1 and Z_2 are independent of $Y_1(z_1), Y_2(z_1, z_2), v_1(z_1), v_2(z_2)$, $\varepsilon_1(z_1)$, and $\varepsilon_2(z_1, z_2)$. Therefore, we have that $0 = E[v_t(z_t)] = E[v_t(z_t)|Z_t] = E[v_t|Z_t]$ for $t = 1$, 2; and similarly, $E[\varepsilon_2|Z_1, Z_2] = E[\varepsilon_1|Z_1] = 0$.

Mixed Model Formulation

The observed outcomes specified by Equations 3, 6, and 7 can be collected in the form of a four-way hierarchical linear model:

$$\begin{pmatrix} Y_0 \\ Y_1 \\ Y_2 \end{pmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & Z_1 & 0 & 0 & 0 \\ 1 & 1 & 0 & Z_1 & Z_2 & Z_1 Z_2 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \delta_1 \\ \delta_{21} \\ \delta_{22} \\ \delta^* \end{bmatrix}$$

$$+ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ v_2 \end{bmatrix} + \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \end{bmatrix} + \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} r_0 \\ r_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}.$$

Or, we express the model in matrix terms, bringing back subscripts for child i in school k,

$$\mathbf{Y}_{ik} = \mathbf{A}_{Fik}\boldsymbol{\theta}_F + \mathbf{A}_{vik}\mathbf{v}_k + \mathbf{A}_{uik}\mathbf{u}_k + \mathbf{A}_{rik}\mathbf{r}_{ik} + \boldsymbol{\varepsilon}_{ik}, \tag{8}$$

where $\mathbf{Y}_{ik}$is a 3 by 1 vector of observed outcomes, $\mathbf{A}_{Fik}$ is a 3 by 6 fixed effects design matrix, $\boldsymbol{\theta}_F$ is a 6 by 1 vector of fixed effects, $\mathbf{A}_{vik}$ is a 3 by 3 design matrix for teacher random effects, $\mathbf{v}_k$ is a 3 by 1 vector of teacher random effects, $\mathbf{A}_{uik}$ is a 3 by 2 design matrix for school random effects, $\mathbf{u}_k$ is a 2 by 1 vector of school random effects , $\mathbf{A}_{rik}$ is a 3 by 2 design matrix for child random effects, $\mathbf{r}_{ik}$ is a 2 by 1 vector of child random effects, and $\boldsymbol{\varepsilon}_{ik}$ is a 3 by 1 vector of time-specific random errors. Stacking person-specific models (Equation 8) formulates a school-level model

$$\begin{gathered} \mathbf{Y}_k = \mathbf{A}_{Fk}\boldsymbol{\theta}_F + \mathbf{A}_{vk}\mathbf{v}_k + \mathbf{A}_{uk}\mathbf{u}_k + \mathbf{A}_{rk}\mathbf{r}_k + \boldsymbol{\varepsilon}_k, \\ \mathbf{v}_k \sim N(0, \psi^2\mathbf{I}), \mathbf{u}_k \sim N(0, \boldsymbol{\omega}), \mathbf{r}_k \sim N(0, \mathbf{I} \otimes \boldsymbol{\tau}), \boldsymbol{\varepsilon}_k \sim N(0, \sigma^2\mathbf{I}) \end{gathered} \tag{9}$$

where $\mathbf{Y}_k = (\mathbf{Y}_{1k}^T, \mathbf{Y}_{2k}^T, \ldots, \mathbf{Y}_{n_k k}^T)^T$, $\mathbf{A}_{Fk} = (\mathbf{A}_{F1k}^T, \mathbf{A}_{F2k}^T, \ldots, \mathbf{A}_{Fn_k k}^T)^T$, $\mathbf{A}_{uk} = (\mathbf{A}_{u1k}^T,$ $\mathbf{A}_{u2k}^T, \ldots, \mathbf{A}_{un_k k}^T)^T$, $\mathbf{A}_{rk} = \bigoplus_{i=1}^{n_k} \mathbf{A}_{rik}$, $\mathbf{r}_k = (\mathbf{r}_{1k}^T, \mathbf{r}_{2k}^T, \ldots, \mathbf{r}_{n_k k}^T)^T$, and $\boldsymbol{\varepsilon}_k = (\boldsymbol{\varepsilon}_{1k}^T,\ \boldsymbol{\varepsilon}_{2k}^T,$ $\ldots, \boldsymbol{\varepsilon}_{n_k k}^T)^T$. The assumption that $\mathbf{v}_k$ and $\boldsymbol{\varepsilon}_k$ have zero means is valid when treatment assignments are independent of potential outcomes (see text following

Equation 7). The matrix $\mathbf{A}_{vk}$ has a special form that links students to the teachers they have encountered. Let T_{ik} be the number of time-series observations recorded for student i in school k. Let J_k denote the number of teachers in school k. The matrix $\mathbf{A}_{vk}$ is $\sum_{i=1}^{n_k} T_{ik}$ by J_k. Each of its elements is an indicator taking on a value of unity if student i has ever encountered teacher j by time t.

Remark. We maintain the assumption here that the teacher and student random effects have homogenous variance, keeping in mind that heterogeneity may be of interest substantively, especially as it reflects differential effects of treatment on teacher and student random effects, and that a failure of the homogeneity assumption may distort standard error estimates. To reduce this risk, we develop Huber-White robust standard errors (see Appendix B) that account for both heteroscadasticity and clustering at the school level.

Model 9 can in turn be regarded as a special case of the general mixed model

$$\mathbf{Y}_k = \mathbf{A}_{Fk}\boldsymbol{\theta}_F + \mathbf{A}_{Rk}\boldsymbol{\theta}_{Rk} + \boldsymbol{\varepsilon}_k, \quad \boldsymbol{\theta}_{Rk} \sim N(0, \boldsymbol{\Omega}), \quad \boldsymbol{\varepsilon}_k \sim N(0, \sigma^2\mathbf{I}), \tag{10}$$

where $\mathbf{A}_{Rk} = (\mathbf{A}_{vk}\,\mathbf{A}_{uk}\,\mathbf{A}_{rk})$, $\boldsymbol{\theta}_{Rk} = \left(\mathbf{v}_k^T\,\mathbf{u}_k^T\,\mathbf{r}_k^T\right)^T$ and

$$\boldsymbol{\Omega} = \begin{bmatrix} \psi^2\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\omega} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\tau} \end{bmatrix}. \tag{11}$$

This general form is useful in derivations and general proofs, as illustrated in the next section.

5. Endogeneity and IPTW for Multilevel Settings

Sequential Strong Ignorability

Estimation of Equation 10 by conventional means would yield unbiased estimates of the causal effects defined earlier if the sequences of treatments Z_1 and Z_2 were assigned at random to classrooms within schools. When the data are nonexperimental, causal inferences may nonetheless be possible if the assumption of sequential strong ignorability holds. Under this assumption, treatment assignment in each year is independent of all the future potential outcomes given past observables. In our example, for the Year 1 treatment,

$$Z_1 \perp Y_1(0), Y_1(1), Y_2(0,0), Y_2(1,0), Y_2(0,1), Y_2(1,1) | \mathbf{X}_1, \mathbf{Y}_0;$$
$$0 < \Pr(Z_1 = z_1 | \mathbf{X}_1, \mathbf{Y}_0) < 1.$$

For the Year 2 treatment, we have that

$$Z_2 \perp Y_2\,(0,0), Y_2(1,0), Y_2(0,1), Y_2(1,1)|\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}_0, \mathbf{Y}_1, \mathbf{Z}_1;$$
$$0 < \Pr(Z_2 = z_2|\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}_0, \mathbf{Y}_1, \mathbf{Z}_1) < 1.$$

The Problem of Statistical Adjustment

Under strong ignorability, the estimation of δ_1 does not pose a problem using standard methods of statistical adjustment. That is, pooling the association between Z_1 and Y_1 within levels of $h_1 = \Pr(Z_1 = 1|\mathbf{X}_1, \mathbf{Y}_0)$ would yield an unbiased estimate of δ_1. Unfortunately, conventional adjustment methods cannot be relied on for estimating δ_{21}, δ_{22}, and δ^* in the context of time-varying confounding without invoking much stronger and often implausible assumptions. Next we discuss, when using conventional methods, what one can estimate under sequential strong ignorability and what additional conditions are required for making inferences about the causal estimands of interest to us.

Average effect of Z$_1$ *on* Y$_2$. Using conventional methods, one might estimate the association between Z_1 and Y_2 within levels of h_1 without regard to Z_2. This would yield an estimate of $E\{E[Y_2(z_1 = 1)|Z_1 = 1, h_1] - E[Y_2(z_1 = 0)|Z_1 = 0, h_1]\}$, which under strong ignorability $Z_1 \perp Y_2(z_1, z_2)|h_1$, is equivalent to $E[Y_2(z_1 = 1) - Y_2(z_1 = 0)]$. Simple algebra reveals its association with our causal estimands δ_{21} and δ^*:

$$\begin{aligned} E[Y_2(z_1 = 1) - Y_2(z_1 = 0)] &= (\delta_{21} + \delta^*) \times \Pr(Z_2 = 1) + \delta_{21} \times \Pr(Z_2 = 0) \\ &= \delta_{21} + \delta^* \times \Pr(Z_2 = 1). \end{aligned}$$

Although this estimand may be of scientific interest, it does not correspond to our goal of separating the effect of receiving Year 1 treatment alone (that is, δ_{21}) from the amplifying effect of receiving the treatment both years (that is, δ^*). To achieve this separation, one might be tempted to estimate the association between Z_1 and Y_2 within levels of h_1 separately for $Z_2 = 0$ and $Z_2 = 1$. However, this approach of conditioning on Z_2, itself an intermediate outcome of Z_1, Y_1, and $\mathbf{X}_2$, would bias the estimate of the effect of Z_1 (alone or interaction with Z_2) on Y_2 (Rosenbaum, 1984).

Conditional effects of Z$_2$ *on* Y$_2$. We might use conventional methods to estimate the association between Z_2 and Y_2 pooled within levels of $h_2 = \Pr(Z_2 = 1|\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}_0, \mathbf{Y}_1, \mathbf{Z}_1)$ for units assigned to $Z_1 = 0$. Under strong ignorability $Z_2 \perp Y_2(z_1, z_2)|h_2$, this estimates

$$E\{E[Y_2(0,1)|Z_1 = 0, Z_2 = 1, h_2] - E[Y_2(0,0)|Z_1 = 0, Z_2 = 0, h_2]\} = E(\Delta_{22}|Z_1 = 0).$$

However, $E(\Delta_{22}|Z_1 = 0)$ would not in general be equivalent to δ_{22}. This is because we cannot assume that $Z_1 \perp Y_2(z_1, z_2)|h_2$ when h_2 is a function of $\mathbf{X}_2$ and

$\mathbf{Y}_1$ that are plausibly outcomes of Z_1. The previous result would equal δ_{22} only under the overly strong assumption of a constant treatment effect for all units. Following the same logic, we might estimate

$$E\{E[Y_2(1,1)|Z_1=1,Z_2=1,h_2]-E[Y_2(1,0)|Z_1=1,Z_2=0,h_2]\}=E(\Delta_{22}+\Delta^*|Z_1=1)$$

as the association between Z_2 and Y_2 pooled within levels of h_2 for units assigned to $Z_1=1$. However, this quantity would not in general be equivalent to $\delta_{22}+\delta^*$ except when the treatment effect is constant for all units.

One may notice that when $\delta^*=0$, conventional methods may generate an unbiased estimate of δ_{21}, which becomes equal to the average effect of Z_1 on Y_2. Also, one may obtain an unbiased estimate of δ_{22} as a weighted average of the conditional effects of Z_2 on Y_2 for units assigned to $Z_1=0$ and for those assigned to $Z_1=1$ weighted by the proportion of units in each z_1 group. However, to proceed with analyzing a main effects model would require empirical evidence indicating that the effect of having a second-year treatment does not depend on the previous year's treatment assignment.

Inverse Probability of Treatment Weighting

The IPTW method proposed by Robins and his colleagues (Robins, Greenland, & Hu, 1999; Robins, Hernán, & Brumback, 2000) provides a viable solution to the endogeneity problem in single-level nonexperimental settings. The weighted estimates are consistent for the marginal treatment effects of interest given sequential strongly ignorable treatment assignment with no need to assume constant treatment effects. Robins (2000) showed that in single-level settings, a weight that is inversely proportional to the probability of one's assigned treatment sequence creates a pseudosample that approximates data from a sequential randomized experiment. In essence, the expected value of the weighted score function for the nonexperimental data is equivalent to the unweighted score function in a randomized study. By solving the weighted score equation, we obtain consistent estimates of the causal effects of time-varying treatments on time-varying outcomes. Once the treatment groups have been equated through weighting, there is no need for direct conditioning on the time-varying covariates in the outcome models. In principle, this solves the dilemma left unresolved by conditional statistical adjustment through linear regression or propensity stratification.

To see why and how the IPTW method applies to multilevel educational data, consider now the case in which student i or the class student i attends in school k is assigned at random to treatments z_1 at Time 1, z_2 at Time $2,\ldots,$ and z_T at Time T. Treatment assignment at Time t is a random variable Z_{tik} taking on values $z_{tik}\in\{0,1\}$, $t=1,\ldots,T$. The entire vector of treatment assignments $\mathbf{Z}_{ik}=(Z_{1ik},Z_{2ik},\ldots,Z_{Tik})^T$ takes on values $\mathbf{z}_{ik}=(z_{1ik},z_{2ik},\ldots,z_{Tik})^T$. Sequential strong ignorability implies that treatment assignment at Time t is independent of all potential outcomes given the past observables:

$$h(\mathbf{z}|\mathbf{x},\mathbf{y}_\zeta)=\prod_{t=1}^{T}h(z_t|z_1,\ldots,z_{t-1},\mathbf{x}_1,\ldots,\mathbf{x}_t,y_1,\ldots,y_{t-1}),\tag{12}$$

where $\mathbf{y}_\zeta$ is the vector of potential outcomes over all time points; ζ is the support for Z; $h(\mathbf{z}|\mathbf{x},\mathbf{y}_\zeta)$is the joint probability of the entire sequence of treatment assignments given all covariates and potential outcomes.

Definition. Consider the general model (Equation 10) for the observed outcomes. We regard $(\mathbf{Y},\boldsymbol{\theta}_R)$ to be the augmented data as contrasted with the observed data $\mathbf{Y}$. The "augmented data score" $\mathbf{S}_{ADtik}$ is the score for child i in school k at time t where the data include the observed data $\mathbf{Y}$ as well as the unobserved random effects $\boldsymbol{\theta}_R$:

$$\mathbf{S}_{ADtik}=\frac{d}{d\boldsymbol{\varphi}}\left[-\sigma^{-2}(Y_{tik}-\mathbf{A}_{Ftik}^T\boldsymbol{\theta}_F-\mathbf{A}_{Rtik}^T\boldsymbol{\theta}_{Rk})-\Omega^{-1}\boldsymbol{\theta}_{Rk}/T_{ik}\right],\tag{13}$$

where $\boldsymbol{\varphi}=\left(\boldsymbol{\theta}_F^T\ \boldsymbol{\theta}_{R1}^T\ \ldots\ \boldsymbol{\theta}_{RK}^T\right)^T$; $\mathbf{A}_{Ftik}^T$ and $\mathbf{A}_{Rtik}^T$ are, respectively, the itth rows of $\mathbf{A}_{Fk}$ and $\mathbf{A}_{Rk}$; T_{ik} is the number of time series observations for student i within school k (see Appendix A for details).

Theorem. In a nonrandomized study with sequential strongly ignorable treatment assignment (Equation 12), given variance components σ^2, Ω, solution to the weighted estimating equation

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{t=1}^{T_{ik}}w_{tik}\mathbf{S}_{ADtik}=0\tag{14}$$

jointly for $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_R$ will ensure consistent estimation of $\boldsymbol{\theta}_F$. Here n_k is the number of children in school k. We define the weight to be

$$\begin{aligned}w_{tik}=&\frac{h(z_{1ik})}{h(z_{1ik}|\mathbf{x}_{1ik})}\times\frac{h(z_{2ik}|z_{1ik})}{h(z_{2ik}|z_{1ik},\mathbf{x}_{1ik},\mathbf{x}_{2ik},y_{1ik})}\\&\times\cdots\times\frac{h(z_{tik}|z_{1ik},\ldots,z_{t-1,ik})}{h(z_{tik}|z_{1ik},\ldots,z_{t-1,ik},\mathbf{x}_{1ik},\ldots,\mathbf{x}_{tik},y_{1ik},\ldots,y_{t-1,ik})}.\end{aligned}\tag{15}$$

Our proof (see Appendix A) follows Robins (2000) for single-level data and extends the logic to multilevel models that require solving for the augmented-data score rather than simply solving for the observed data score. Under randomization, the conditional expectation of the augmented data score taken over the joint distribution of $\mathbf{z}$, $\mathbf{x}$, and $\mathbf{y}_\zeta$ is zero. We reveal the exact structure of the needed weight (Equation 15) for multilevel nonexperimental data when the sequential treatment assignments are strongly ignorable.

Estimation via Maximum Pseudolikelihood for Hierarchical Linear Models

Application of our theorem (Equations 14 and 15) assumes knowledge of covariance components σ^2 and Ω. The results will hold if we substitute consistent estimates of σ^2 and Ω into the solution of the estimating equations (Equation 14). To optimize efficiency, we adopt a maximum pseudolikelihood approach in the spirit of Pfefferman, Skinner, Homes, Goldstein, and Rasbash (1998). Using this approach, we maximize

$$L_w(\boldsymbol{\theta}_F, \sigma^2, \Omega; \mathbf{Y}) = \prod_{k=1}^{K} \int f_w(\mathbf{Y}_k | \boldsymbol{\theta}_F, \boldsymbol{\theta}_{Rk}, \sigma^2) p_w(\boldsymbol{\theta}_{Rk} | \Omega) d\boldsymbol{\theta}_{Rk} \tag{16}$$

where $L_w(\boldsymbol{\theta}_F, \sigma^2, \Omega; \mathbf{Y})$ is the weighted marginal likelihood of $\mathbf{Y}$ (i.e., integrating out the random effects), $f_w(\mathbf{Y}_k | \boldsymbol{\theta}_F, \boldsymbol{\theta}_{Rk}, \sigma^2)$ is the weighted conditional likelihood of $\mathbf{Y}_k$ given the random effects for school k, and $p_w(\boldsymbol{\theta}_{Rk} | \Omega)$ is the weighted marginal density of the random effects as defined in (10) for school k. The results, derived in Appendix B, yield point estimates of the fixed effects

$$\hat{\boldsymbol{\theta}}_F = \mathbf{B} \sum_{k=1}^{K} \mathbf{A}_{Fk}^T \mathbf{M}_k \mathbf{Y}_k \tag{17}$$

with model-based standard errors equal to the square roots of the diagonal elements of

$$Var_{mb}(\hat{\boldsymbol{\theta}}_F) = \mathbf{B} \sum_{k=1}^{K} \mathbf{A}_{Fk}^T \mathbf{M}_k \mathbf{V}_k \mathbf{M}_k \mathbf{A}_{Fk} \mathbf{B} \tag{18}$$

where $\mathbf{B} = \left(\sum_{k=1}^{K} \mathbf{A}_{Fk}^T \mathbf{M}_k \mathbf{A}_{Fk} \right)^{-1}$, $\mathbf{M}_k = \mathbf{W}_k - \mathbf{W}_k \mathbf{A}_{Fk}^T \mathbf{W}_k \mathbf{A}_{Rk} \mathbf{C}_k^{-1} \mathbf{A}_{Rk}^T \mathbf{W}_k \mathbf{A}_{Fk} \mathbf{W}_k$, $\mathbf{C}_k = \mathbf{A}_{Fk}^T \mathbf{W}_k \mathbf{A}_{Rk} + \sigma^2 \boldsymbol{\Omega}^{-1}$, $\mathbf{V}_k = \mathbf{A}_{Rk} \boldsymbol{\Omega} \mathbf{A}_{Rk}^T + \sigma^2 \mathbf{I}$, and $\mathbf{W}_k = diag\{w_{tik}\}$. The robust standard errors are the diagonal elements of

$$Var_{rob}(\hat{\boldsymbol{\theta}}_F) = \mathbf{B} \sum_{k=1}^{K} \mathbf{A}_{Fk}^T \mathbf{M}_k \hat{\mathbf{e}}_k \hat{\mathbf{e}}_k^T \mathbf{M}_k \mathbf{A}_{Fk} \mathbf{B}, \tag{19}$$

where $\hat{\mathbf{e}}_k = \mathbf{Y}_k - \mathbf{A}_{Fk} \hat{\boldsymbol{\theta}}_F$.

6. Case Study

Data

Data for this study were collected by the U.S. Department of Education's Planning and Evaluation Service for the Longitudinal Evaluation of School Change and Performance in 1997, 1998, and 1999 (Westat, 2001). LESCP drew its sample from 67 Title I schools located in 18 school districts in seven states. Our sample includes a longitudinal cohort of 4,216 students who progressed from Grade 3 to Grade 5 during the 3 study years. We use as a measure of students' math learning the Stanford Achievement Test 9 administered at the end of each year. Test scores in different years have been equated on the same scale so that we can assess the learning growth over years. We present in Table 1 the descriptive information of all the student, teacher, and school measures. Table 2 shows mean math achievement by missing data pattern over the three grade levels.

We construct a binary treatment variable (Z) for each grade level, with $Z = 1$ indicating a teacher's use of intensive math instruction characterized by emphasis on both instructional time and content difficulty and $Z = 0$ otherwise (see Raudenbush, Hong, & Rowan, 2002, for details). Among the 147 Grade 4 teachers, 36 of them provided intensive math instruction to their students. In Grade 5, 58 out of 147 teachers adopted the intensive math instruction in their classrooms. About 15% of the students received intensive math instruction in both Grade 4 and Grade 5, 8% of them had the treatment in Grade 4 only, 32% of them had it in Grade 5 only, and 45% of them had this treatment in neither of the 2 years.

Statistical Adjustment Procedure

Following Equation 15, we construct a weight w_{tik} for child i in school k in year t. Here $t = 0$, 1, 2 correspond to Grades 3, 4, and 5. Weights in Grade 3 are 1.0. To estimate the weights in Grade 4 and Grade 5, we compute for each classroom in each year the predicted probability of intensive math instruction given the past observed covariates, treatments, and outcomes. Predictors of the Grade 4 treatment include Grade 4 classroom-aggregated student background characteristics, Grade 3 instructional experiences, Grade 3 math test scores, and prior school and teacher characteristics. Predictors of the Grade 5 treatment include Grade 5 classroom aggregates of student background characteristics, the proportion of students who received intensive math instruction in Grade 4, average Grade 4 math test score, and teacher and school characteristics. These propensity models are estimated at the classroom level with logistic regression. We use missing indicators to represent the missing information in the predictors. Table 3 lists the results of Grade 4 and Grade 5 propensity analyses. Despite the large number of covariates entered in each of these two propensity models, neither of them shows strong explanatory power. The proportion of area under the ROC curve is .83 for the Grade 4 analysis and only .79 for the Grade 5 analysis. We assess the impact of potential unmeasured confounders through sensitivity analysis.

In analyzing the causal effects of intensive math instruction, we include all the students regardless of their response pattern and construct nonresponse weights to adjust for various missing patterns (see Appendix C for details). We

TABLE 1
Descriptive Statistics

Variable	*M*	*SD*
Students ($N = 1, \ldots, 4{,}216$)		
Free lunch (1 = *yes*; 0 = *no*)	0.723	
African American (1 = *yes*; 0 = *no*)	0.440	
Hispanic (1 = *yes*; 0 = *no*)	0.102	
White (1 = *yes*; 0 = *no*)	0.429	
Other ethnic groups (1 = *yes*; 0 = *no*)	0.029	
Gender (1 = *male*; 0 = *female*)	0.492	
Title I (1 = *yes*; 0 = *no*)	0.672	
Individualized education program (1 = *yes*; 0 = *no*)	0.061	
Limited English proficiency (1 = *yes*; 0 = *no*)	0.052	
Migrant (1 = *yes*; 0 = *no*)	0.017	
English as a second language (1 = *yes*; 0 = *no*)	0.049	
Teachers/classrooms ($J = 386$)		
Grade (1 = *Grade 4*; 0 = *Grade 5*)	0.497	
Grade 4 teacher gender (1 = *male*; 0 = *female*)	0.171	
Grade 5 teacher gender (1 = *male*; 0 = *female*)	0.212	
Grade 4 teacher African American (1 = *yes*; 0 = *no*)	0.289	
Grade 4 teacher Hispanic (1 = *yes*; 0 = *no*)	0.000	
Grade 4 teacher White (1 = *yes*; 0 = *no*)	0.658	
Grade 4 teacher other ethnic groups (1 = *yes*; 0 = *no*)	0.053	
Grade 5 teacher African American (1 = *yes*; 0 = *no*)	0.259	
Grade 5 teacher Hispanic (1 = *yes*; 0 = *no*)	0.034	
Grade 5 teacher White (1 = *yes*; 0 = *no*)	0.660	
Grade 5 teacher other ethnic groups (1 = *yes*; 0 = *no*)	0.047	
Grade 4 teacher degree (1 = *master's or above*; 0 = *bachelor's or below*)	0.435	
Grade 5 teacher degree (1 = *master's or above*; 0 = *bachelor's or below*)	0.405	
Grade 4 teacher teaching experience	13.900	8.605
Grade 5 teacher teaching experience	13.652	8.437
Grade 3 math content difficulty	–0.190	0.840
Grade 3 math instructional time	0.022	0.802
Grade 3 proportion of low achievers in class	0.221	0.219
Grade 4 proportion of low achievers in class	0.240	0.294
Grade 5 proportion of low achievers in class	0.248	0.296
Grade 3 class size	17.214	7.539
Grade 4 class size	15.898	7.035
Grade 5 class size	15.877	7.094
Grade 4 intensive math (1 = *yes*; 0 = *no*)	0.245	
Grade 5 intensive math (1 = *yes*; 0 = *no*)	0.395	
Grade 4 class average pretest score	589.211	27.985
Grade 5 class average pretest score	610.148	27.238

(continued)

TABLE 1 *(continued)*

Variable	*M*	*SD*
Schools ($K = 67$)		
Year 1 school size	432.299	146.067
Year 1 percentage free lunch	0.731	0.191
Year 1 percentage African American	0.405	0.390
Year 1 percentage Hispanic	0.088	0.159
Year 1 schoolwide Title I (1 = *yes*; 0 = *no*)	0.806	0.398
Year 2 school size	440.239	142.641
Year 2 percentage free lunch	0.726	0.188
Year 2 percentage African American	0.421	0.392
Year 2 percentage Hispanic	0.092	0.159
Year 2 schoolwide Title I (1 = *yes*; 0 = *no*)	0.821	0.386

TABLE 2
Average Math Achievement and Proportion of Free Lunch by Grade and Response Pattern

		Average Math Achievement			Percentage
Response Pattern	*n*	Grade 3	Grade 4	Grade 5	Free Lunch
All 3 years	953	597.70	621.17	642.26	.67
Grades 3 and 4	730	593.46	616.51		.74
Grades 3 and 5	127	595.50		636.88	.69
Grades 4 and 5	363		611.80	635.96	.63
Grade 3 only	1,490	585.28			.76
Grade 4 only	435		605.05		.80
Grade 5 only	118			629.10	.72
Total	4,216	591.07	615.38	639.29	.72

estimate model (Equation 10) weighted by the product of the treatment weight and the nonresponse weight. The computation requires statistical software that allows the application of weights to a four-level model in which students are cross-classified by teachers who are nested within schools. We conduct the analysis using HLM6.4. The program and users' manual are available on request from the second author.

Causal Analysis Results

For the purpose of comparison, we present two sets of analytic results along with their robust standard errors. In both cases, growth modeling provides

TABLE 3
Propensity Model Results

	Grade 4 Treatment			Grade 5 Treatment		
Predictor	β	*SE* (β)	*p*	β	*SE* (β)	*p*
Average Grade 3 content difficulty	1.215	0.501	.015	−.654	0.421	.120
Average Grade 3 math time	−.249	0.520	.632	.558	0.466	.231
Percentage having Grade 4 intensive math	—	—	—	1.171	0.656	.075
Average math pretest score	.032	0.223	.885	.090	0.238	.706
Class size	−.032	0.049	.517	.137	0.048	.004
Percentage low achievers receiving services	−.379	0.993	.702	−1.440	0.968	.137
Teacher's educational degree	.046	0.619	.940	.118	0.495	.811
Teaching experience	−.046	0.031	.144	.007	0.024	.761
Teacher's gender	−1.566	0.780	.045	.355	0.580	.541
African American teacher	.941	0.710	.185	−.092	0.588	.876
Teacher of other non-White ethnicity	.761	1.123	.498	.356	0.969	.713
School size	−.006	0.003	.015	−.001	0.002	.701
Percentage free lunch students in school	−3.392	2.100	.106	.029	0.021	.168
Percentage Black students in school	.721	1.094	.510	−.569	0.994	.567
Percentage Hispanic students in school	4.607	1.989	.021	−2.166	1.890	.252
Schoolwide Title I program	.043	0.869	.960	.817	0.896	.362
Percentage missing Grade 3 instruction information	−2.969	1.211	.014	−.695	0.859	.419
Percentage missing Grade 4 treatment information	—	—	—	−1.322	0.739	.073
Missing at least one other covariate	−6.799	1.947	.000	3.528	1.792	.049

effective adjustment for the bias associated with students' pretest scores (Bryk & Weisberg, 1977). The weighted model makes a more comprehensive adjustment for all the observed confounding variables.

Growth modeling with no weighting. We generate the first set of results by analyzing value-added model (Equation 10) with no weights (see Table 4, panel 1). Comparing Grade 4 students attending treatment classes and those attending control classes shows a mean difference of 2.70 ($SE = 3.02$, $t = 0.89$) in Grade 4 outcome. When we compare students who had Grade 4 treatment but no Grade 5 treatment with those having treatment in neither year, we find a mean difference of 0.40 ($SE = 4.55$, $t = 0.09$) in Grade 5 outcome. Comparing Grade 5 students in treatment classes and those in control classes shows a mean difference of 7.79 ($SE = 3.07$, $t = 2.54$) in Grade 5 outcome, an effect that does not depend on Grade 4 treatment.

TABLE 4
Treatment Effect Estimation Results

	Unweighted Model			Weighted Model		
Fixed Effects	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*
Intercept, γ_0	609.83	1.96	310.78	609.88	2.96	205.97
Growth rate, γ_1	20.94	1.13	18.45	20.89	1.22	17.09
Grade 4 treatment on Grade 4 outcome, δ_1	2.70	3.02	0.89	4.80	4.09	1.17
Grade 4 treatment on Grade 5 outcome, δ_{21}	0.40	4.55	0.09	2.05	4.85	0.42
Grade 5 treatment on Grade 5 outcome, δ_{22}	7.79	3.07	2.54	6.46	3.35	1.93
Two-way interaction of Grade 4 and Grade 5 treatments on Grade 5 outcome, δ^*	0.59	6.39	0.09	−2.79	5.74	−0.49
Variance Components	Estimate			Estimate		
Within students						
σ^2	303.59			319.93		
Between students						
$\tau_{\pi 0}$	771.88			629.70		
$\tau_{\pi 1}$	21.65			1.37		
Correlation (π_o, π_1)	−0.15			−0.10		
Between schools						
$\omega_{\beta 00}$	171.62			136.22		
$\omega_{\beta 10}$	30.21			20.02		
Correlation (β_{00}, β_{10})	0.39			0.34		
Between classrooms						
$\Psi^2(v)$	172.20			171.41		

Growth modeling with IPTW. The second set of treatment effect estimates (Table 4, panel 2) is produced by applying inverse probability of treatment weights to the value-added model. Our point estimate for the causal effect of Grade 4 treatment on Grade 4 outcome is positive but not significantly different from zero, $\hat{\delta}_1 = 4.80$ ($SE = 4.09$, $t = 1.17$). The effect size is about a fifth of the yearly growth rate. If a student is assigned to the control condition in Grade 5, the carryover effect of Grade 4 treatment on Grade 5 outcome shows an even smaller magnitude, $\hat{\delta}_{21} = 2.05$ ($SE = 4.85$, $t = 0.42$). We find a positive and nearly significant effect of Grade 5 treatment on Grade 5 outcome, $\hat{\delta}_{22} = 6.46$ ($SE = 3.35$, $t = 1.93$), if a student has been in the control condition in Grade 4. The effect size is about a third of the yearly growth rate. There is no evidence

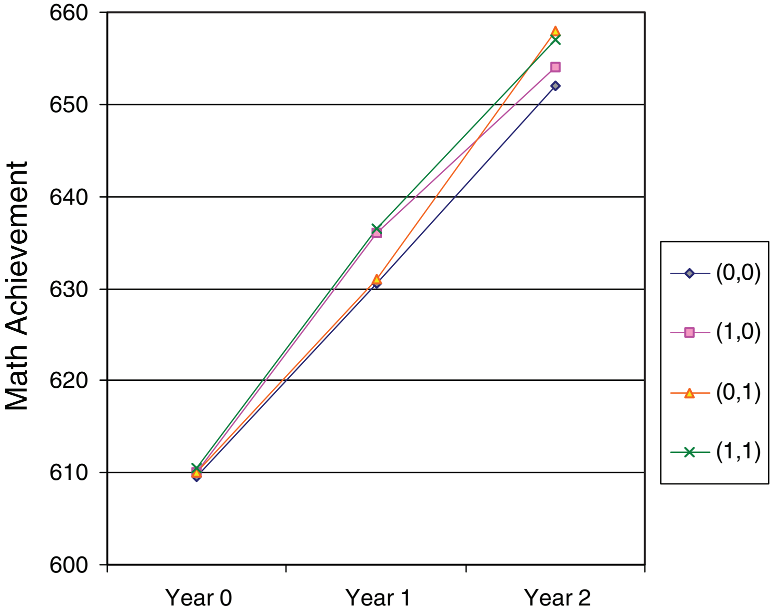

FIGURE 1. *Predicted treatment effects on linear growth trajectory.*

that the impact of the Grade 5 treatment depends on having had Grade 4 treatment, $\hat{\delta}^* = -2.79$ ($SE = 5.74$, $t = -0.49$). These results are displayed graphically in Figure 1.

Main Effects Model

Under IPTW, the point estimate of the amplifying effect δ^* is nonnegligible but not statistically significant. The large standard error estimate indicates that this effect has been estimated with poor precision, a result that is explainable by the comparatively small number of students experiencing intensive math instruction in Grade 4 but not Grade 5. Such imprecision might have disturbed other results. So we recompute the analysis with δ^* set to zero. The estimate of Grade 4 treatment effect on Grade 4 outcome is now significant, $\hat{\delta}_1 = 6.26$ ($SE = 3.00$, $t = 2.08$). The average effect of Grade 4 treatment on Grade 5 outcome, $E[Y(1, z_2) - Y(0, z_2)] = \delta_{21}$ when δ^* is zero, has an estimate of 3.75 ($SE = 2.82$, $t = 1.33$). The average effect of Grade 5 treatment on Grade 5 outcome, $E[Y(z_1, 1) - Y(z_1, 0)] = \delta_{22}$ when δ^* is zero, is estimated to be 9.65 ($SE = 3.70$, $t = 2.61$), almost a half of the yearly growth rate.

Sensitivity Analysis

For the estimates of δ_1 and δ_{22} obtained from analyzing the main effects model, we test the sensitivity of our inferences to plausible departures from the strong ignorability assumption (Hong, 2004; Hong & Raudenbush, 2006; Lin, Psaty, & Kronmal, 1998; Rosenbaum, 1986, 2002). Suppose that an unobserved time-varying covariate, U_t, has confounded the estimation of the Year t treatment effect on Year t outcome. The association between U_t and the Year t treatment assignment Z_t and its association with the Year t outcome Y_t are assumed to be comparable in magnitude to the most important confounding variable observed. After adjusting for the potential hidden bias associated with U_t, we obtain a new estimate of the treatment effect and its 95% confidence interval. The original inference is considered to be insensitive to important omitted confounders if additional adjustment for U_t does not change the conclusion.

Among the observed covariates, Grade 4 class average pretest score shows the strongest confounding effect in estimating δ_1. We use its respective associations with the Grade 4 treatment and the Grade 4 outcome as reference values for the sensitivity parameters of U_4. To make additional adjustment for U_4, we subtract the product of the two sensitivity parameters from the original estimate and obtain 2.69 as a new estimate of δ_1. Adopting the same standard error as originally estimated, we find the 95% confidence interval of the new estimate of δ_1 to be $(-3.31, 8.69)$, which would lead to a decision of retaining the null hypothesis. Therefore, a conclusion about the positive effect of the Grade 4 treatment on Grade 4 outcome could be altered if an unmeasured confounder contains an additional positive bias as severe as that of the most important confounder observed.

In estimating δ_{22}, we find Grade 5 class size to be the strongest observed confounder. We imagine that our weighted estimate of δ_{22} could have been positively biased by U_5 in an amount comparable to the confounding effect of Grade 5 class size. Once we remove the hypothetical bias associated with U_5, our new estimate of δ_{22} is 8.96 and its 95% confidence interval is (1.56, 16.36). On the basis of this evidence, we conclude that our inference about the Grade 5 treatment effect on Grade 5 outcome is not highly sensitive to the omission of a confounding variable as important as the strongest observed confounder.

7. Conclusions

Our understanding of the impact of instruction depends ultimately on making inferences about the causal effects of sequences of instructional treatments. To study these effects, we have developed here an approach that copes with three characteristic methodological challenges. First, these effects unfold as students move across classrooms nested within schools, generating a special crossed-nested structure and requiring an appropriately complex mixed statistical model for analysis. Second, the potential outcomes causal framework now widely used in medical and social research is applicable to multiyear sequences of instruction only after careful modification of the stable unit treatment value assumption. Third, experiments that assign students to alternative multiyear sequences of instruction, though ideal for estimation of causal effects, are difficult to sustain so that inference will typically be based on nonexperimental studies that must overcome problems of endogenous treatment assignment. Our approach adapts the IPTW method to the multilevel context of schooling under a modified SUTVA. In illustrating this approach, we investigated the causal effects of intensive math instruction in Grades 4 and 5 using data from a recent evaluation of Title I. Our results suggest that intensive math instruction in Grade 5 leads to significant improvement in students' math learning in Grade 5. Here we revisit key assumptions invoked in the case study.

Intact schools and no interference between schools. Although these assumptions will often be plausible, they could be challenged if teachers from separate schools within districts collaborate closely or if children who are friends attend different schools. Moreover, if schools are closed and students are reassigned as a result of large-scale restructuring efforts in some school districts, such restructuring events will need to be modeled as either pretreatment covariates or concomitant treatments.

No indirect interference between classes within a school. We assumed that in general, a student's time-varying learning outcomes are not subject to the influence of teachers and students in other classes within the school. Therefore, a student's learning outcome in Year t is modeled as a function of the cumulative random effects of teachers/classes that the student has ever directly encountered up to and including Year t. The assumption would be violated in schools where teachers frequently exchange instructional information and share student work or where competition between classes for scarce resources limits students' learning opportunities. A possible solution is to explicitly model the fixed effects of between-class interference (Hong & Raudenbush, 2006).

Class assignment followed by treatment assignment. We reasoned that in upper elementary math instruction, after students have been assigned to classes at the beginning of a year, schools and teachers assign instructional treatments to intact classes. Our propensity modeling at the class level was justified under this assumption. Alternatively, at the end of each year, some schools may assign individual students to instructional treatments on the basis of their current performance. Students may then be assigned to classrooms in accordance with their treatment assignments, often under organizational constraints. This would require modeling propensity at the student level rather than the class level.

Sequential strong ignorability assumption. Using inverse probability of treatment weights to adjust for measured covariates will yield consistent estimates of average treatment effects under the assumption that the treatment assignment in each year is independent of the potential outcomes given prior observables and that there is a nonzero probability of receiving alternative treatments for each observed covariate pattern. Multiple pretreatment and time-varying measures of covariates in this study provide a reasonably promising basis for a quasi-experimental design. Meanwhile, we conceive of treatment effects at a given grade as deflections from a child-specific growth trajectory estimable from the repeated measures of math achievement. In this way, we were able to make adjustment for pretreatment growth rates under our modeling assumptions. In addition, we considered possible consequences of having omitted important confounding variables. A sensitivity analysis suggested that our conclusion about the positive effect of intensive math instruction in Grade 5 on Grade 5 outcome would be altered only by unmeasured confounders stronger than any of the observed covariates. However, the estimated positive effect of Grade 4 treatment on Grade 4 outcome was sensitive to unmeasured confounders as strong as the class average math pretest. If any of the cross-year treatment effects (i.e., δ_{21} and δ^*) were estimated to be statistically significant, a sensitivity analysis would require computing an inverse probability of treatment weight that incorporates the confounding effect of hypothetical unmeasured time-varying covariates.

In general, the theoretical and analytic approaches illustrated in the case study are applicable to multiyear studies of time-varying instructional treatments. Apparently, as the number of time points increases, the number of potential outcomes per student will increase exponentially. This can be dealt with by focusing on the causal effects of key scientific interest and by placing theoretical constraints on the analytic models. Examples of such constraints include a main effects model constraining interaction effects across years to be null or a cumulative treatment effects model assuming no decay of prior treatment effects on later outcomes.

Appendix A
Inverse Probability of Treatment Weights for Multilevel Nonexperimental Data

Our aim is to prove that in multilevel nonexperimental data, under the assumption of sequential strongly ignorable treatment assignment, solving the weighted augmented data score function (Equation 14) with respect to $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_R$ yields a consistent estimator of $\boldsymbol{\theta}_F$. The unweighted augmented data score under randomization has expectation zero over the joint distribution of $\mathbf{Z}$, $\mathbf{X}$, and $\mathbf{Y}_\zeta$. In the absence of randomization but under the assumption of strongly ignorable treatment assignment, the weighted augmented data score has expectation zero taken over the same joint distribution. Hence, equating this weighted score to zero and jointly solving for $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_R$ provides consistent estimation of the causal effects of interest. However, this solution requires a consistent estimate of the covariance components σ^2 and $\boldsymbol{\Omega}$, a problem we consider in Appendix B.

Expected Augmented Data Score Under Randomization

Definition 1. Consider the general model (Equation 10). Define the *joint density* of the observed data $\mathbf{Y}$ and the random effects $\boldsymbol{\theta}_R$ to be

$$g(\mathbf{Y},\boldsymbol{\theta}_R|\boldsymbol{\theta}_F,\sigma^2,\boldsymbol{\Omega}) = f(\mathbf{Y}|\boldsymbol{\theta}_F,\boldsymbol{\theta}_R,\sigma^2)p(\boldsymbol{\theta}_R|\boldsymbol{\Omega}), \tag{A1}$$

where $f(\mathbf{Y}|\boldsymbol{\theta}_F,\boldsymbol{\theta}_R,\sigma^2)$ is the density for the observed data $\mathbf{Y}\sim N(\mathbf{A}_F\boldsymbol{\theta}_F+\mathbf{A}_R\boldsymbol{\theta}_R, \sigma^2\mathbf{I})$ and $p(\boldsymbol{\theta}_R|\boldsymbol{\Omega})$ is the prior density for the random effect vector $\boldsymbol{\theta}_R\sim N(0,\boldsymbol{\Omega})$.

Definition 2. Let $N_k=\sum_{i=1}^{n_k}T_{ik}$ be the total number of observations in school k. The augmented-data score for $\boldsymbol{\varphi}$ is

$$\begin{aligned}\mathbf{S}_{AD}(\boldsymbol{\varphi}) &= \frac{d}{d\boldsymbol{\varphi}}\sum_{k=1}^{K}\left[Constant-\frac{N_k}{2}\ln(\sigma^2)-\frac{1}{2}\ln(|\boldsymbol{\Omega}|)-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}_k^T\boldsymbol{\varepsilon}_k-\frac{1}{2}\boldsymbol{\theta}_{Rk}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\theta}_{Rk}\right]\\ &= \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{t=1}^{T_{ik}}\mathbf{S}_{ADtik}(\boldsymbol{\varphi}),\end{aligned} \tag{A2}$$

where $\boldsymbol{\varepsilon}_k=\mathbf{Y}_k-\mathbf{A}_{Fk}\boldsymbol{\theta}_F-\mathbf{A}_{Rk}\boldsymbol{\theta}_{Rk}$. From Equation A2 we can derive Equation 13.

In a randomized experiment, conditional on σ^2 and $\boldsymbol{\Omega}$, the expected augmented data score is zero. The expectation is taken over the joint distribution denoted $f^*(\mathbf{G})$ where $\mathbf{G}=(\mathbf{z},\mathbf{y}_\zeta,\mathbf{x})$. In this section, we omit subscripts i and k for ease of presentation. Specifically, with random assignment of an individual unit to the treatment sequence $\mathbf{z}=(z_1,z_2,\ldots,z_T)$,

$$E\left(\sum_{t=1}^{T}\mathbf{S}_{ADt}\right)=\int\sum_{t=1}^{T}\mathbf{S}_{ADt}f^*(\mathbf{G})d\mathbf{G}=0. \tag{A3}$$

This is because

$$\begin{aligned}E\left(\sum_{t=1}^{T}\mathbf{S}_{ADt}\right) &= E\left(E\left\{\frac{d}{d\boldsymbol{\varphi}}\left[-\sum_{t=1}^{T}\sigma^{-2}\left(Y_t-\mathbf{A}_{Ft}^T\boldsymbol{\theta}_F-\mathbf{A}_{Rt}^T\boldsymbol{\theta}_R\right)-\Omega^{-1}\boldsymbol{\theta}_R|Z_t\right]\right\}\right)\\ &= \frac{d}{d\boldsymbol{\varphi}}E\left[-\sigma^{-2}\sum_{t=1}^{T}E\left(Y_t-\mathbf{A}_{Ft}^T\boldsymbol{\theta}_F-\mathbf{A}_{Rt}^T\boldsymbol{\theta}_R|Z_t\right)-E(\Omega^{-1}\boldsymbol{\theta}_R|Z_t)\right].\end{aligned} \tag{A4}$$

As shown in Equation 9 and the text following it, we have that $E(\boldsymbol{\theta}_R|Z_t)=E(\boldsymbol{\theta}_R)=0$ and $E\left(Y_t-\mathbf{A}_{Ft}^T\boldsymbol{\theta}_F-\mathbf{A}_{Rt}^T\boldsymbol{\theta}_R|Z_t\right)=E(\boldsymbol{\varepsilon}_t|Z_t)=E(\boldsymbol{\varepsilon}_t)=0$. The marginal expectation in Equation A4 is null as a result.

Weighting the Augmented Data Score

Under randomization to treatment sequences, we have

$$f^*(\mathbf{G}) = f^*(\mathbf{y}_\zeta, \mathbf{x}, \mathbf{z}) = g(\mathbf{y}_\zeta|\mathbf{x})q(\mathbf{x})h^*(\mathbf{z}). \tag{A5}$$

Let us now consider the case of sequential strongly ignorable treatment assignments, that is, when randomization at Time t occurs within levels of past treatments, pretreatment covariates, and past observed outcomes. The joint density of $\mathbf{y}_\zeta$, $\mathbf{x}$, and $\mathbf{z}$ becomes

$$f(\mathbf{G}) = f(\mathbf{y}_\zeta, \mathbf{x}, \mathbf{z}) = g(\mathbf{y}_\zeta|\mathbf{x})q(\mathbf{x})h(\mathbf{z}|\mathbf{y}_\zeta, \mathbf{x}). \tag{A6}$$

Applying Equation 12, we have that

$$\begin{aligned} f(\mathbf{G}) = {} & g(\mathbf{y}_{\zeta 1}|\mathbf{x}_1)\cdots g(\mathbf{y}_{\zeta,T}|\mathbf{x}_1,\ldots\mathbf{x}_T,\mathbf{y}_{\zeta 1},\ldots,\mathbf{y}_{\zeta,T-1}) \times q(\mathbf{x}_1)\cdots q(\mathbf{x}_T|\mathbf{x}_1,\ldots,\mathbf{x}_{T-1}) \\ & \times h(z_1|\mathbf{x}_1)h(z_2|z_1,\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}_1)\cdots h(z_T|z_1,\ldots,z_{T-1},\mathbf{x}_1,\ldots,\mathbf{x}_T,\mathbf{y}_1,\ldots,\mathbf{y}_{T-1}). \end{aligned}$$

Following Robins (2000), we multiply and divide Equation A3 by this density:

$$E\left(\sum_{t=1}^{T}\mathbf{S}_{ADt}\right) = \int\sum_{t=1}^{T}\mathbf{S}_{ADt}\frac{f^*(\mathbf{G})}{f(\mathbf{G})}f(\mathbf{G})d\mathbf{G} = 0. \tag{A7}$$

We now make use of the following facts:

(i) $$\begin{aligned} \frac{f^*(\mathbf{G})}{f(\mathbf{G})} &= \frac{h^*(\mathbf{z})}{h(\mathbf{z}|\mathbf{y}_\zeta,\mathbf{x})} \\ &= \frac{h^*(z_1)h^*(z_2|z_1)\ldots h^*(z_T|z_1,\ldots,z_{T-1})}{h(z_1|\mathbf{x}_1)h(z_2|z_1,\mathbf{x}_1,\mathbf{x}_2,y_1)\ldots h(z_T|z_1,\ldots,z_{T-1},\mathbf{x}_1,\ldots,\mathbf{x}_T,y_1,\ldots,y_{T-1})}; \end{aligned}$$

(ii) $f(\mathbf{G}_t)$ is the joint density of the data up to time t,

$$\begin{aligned} f(\mathbf{G}_t) = {} & h(z_1|\mathbf{x}_1)h(z_2|z_1,\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}_1)\cdots h(z_t|z_1,\ldots,z_{t-1},\mathbf{x}_1,\ldots,\mathbf{x}_t,\mathbf{y}_1,\ldots,\mathbf{y}_{t-1}) \\ & \times g(\mathbf{y}_{\zeta 1}|\mathbf{x}_1)\cdots g(\mathbf{y}_{\zeta,t}|\mathbf{x}_1,\ldots\mathbf{x}_t,\mathbf{y}_{\zeta 1},\ldots,\mathbf{y}_{\zeta,t-1}) \times q(\mathbf{x}_1)\cdots q(\mathbf{x}_t|\mathbf{x}_1,\ldots,\mathbf{x}_{t-1}); \end{aligned}$$

(iii) $\iint\ldots\int f^*(\mathbf{G}_{>t})dz_{t+1}\cdots dz_T d\mathbf{x}_{t+1}\cdots d\mathbf{x}_T d\mathbf{y}_{\zeta,t+1}\cdots d\mathbf{y}_{\zeta,T} = 1$, where

$$\begin{aligned} f^*(\mathbf{G}_{>t}) = {} & h^*(z_{t+1}|z_1,\ldots,z_t)\cdots h^*(z_T|z_1,\ldots,z_{T-1}) \\ & \times g(\mathbf{y}_{\zeta,t+1}|\mathbf{x}_1,\ldots\mathbf{x}_{t+1},\mathbf{y}_{\zeta 1},\ldots,\mathbf{y}_{\zeta,t})\cdots g(\mathbf{y}_{\zeta,T}|\mathbf{x}_1,\ldots\mathbf{x}_T,\mathbf{y}_{\zeta 1},\ldots,\mathbf{y}_{\zeta,T-1}) \\ & \times q(\mathbf{x}_{t+1}|\mathbf{x}_1,\ldots,\mathbf{x}_t)\cdots q(\mathbf{x}_T|\mathbf{x}_1,\ldots,\mathbf{x}_{T-1}). \end{aligned}$$

Thus, we have

$$
\begin{aligned}
E\left(\sum_{t=1}^{T}\mathbf{S}_{ADt}\right) &= \int\sum_{t=1}^{T}\mathbf{S}_{ADt}\frac{f^{*}(\mathbf{G})}{f(\mathbf{G})}f(\mathbf{G})d\mathbf{G}\\
&= \sum_{t=1}^{T}\iint\cdots\int\mathbf{S}_{ADt}\frac{h^{*}(\mathbf{z})}{h(\mathbf{z}|\mathbf{y}_{\zeta},\mathbf{x})}\\
&\quad\times h(\mathbf{z}|\mathbf{y}_{\zeta},\mathbf{x})g(\mathbf{y}_{\zeta}|\mathbf{x})q(\mathbf{x})dz_{1}\cdots dz_{T}d\mathbf{x}_{1}\cdots d\mathbf{x}_{T}d\mathbf{y}_{\zeta 1}\cdots d\mathbf{y}_{\zeta,T}\\
&= \sum_{t=1}^{T}\iint\cdots\int\mathbf{S}_{ADt}\\
&\quad\frac{h^{*}(z_{1})h^{*}(z_{2}|z_{1})\cdots h^{*}(z_{t}|z_{1},\ldots,z_{t-1})}{h(z_{1}|\mathbf{x}_{1})h(z_{2}|z_{1},\mathbf{x}_{1},\mathbf{x}_{2},\mathbf{y}_{1})\cdots h(z_{t}|z_{1},\ldots,z_{t-1},\mathbf{x}_{1},\ldots,\mathbf{x}_{t},\mathbf{y}_{1},\ldots,\mathbf{y}_{t-1})}\\
&\quad\times h(z_{1}|\mathbf{x}_{1})h(z_{2}|z_{1},\mathbf{x}_{1},\mathbf{x}_{2},\mathbf{y}_{1})\cdots h(z_{t}|z_{1},\ldots,z_{t-1},\mathbf{x}_{1},\ldots,\mathbf{x}_{t},\mathbf{y}_{1},\ldots,\mathbf{y}_{t-1})\\
&\quad\times h^{*}(z_{t+1}|z_{1},\ldots,z_{t})\cdots h^{*}(z_{T}|z_{1},\ldots,z_{T-1})\\
&\quad\times g(\mathbf{y}_{\zeta 1}|\mathbf{x}_{1})\cdots g(\mathbf{y}_{\zeta,t}|\mathbf{x}_{1},\ldots\mathbf{x}_{t},\mathbf{y}_{\zeta 1},\ldots,\mathbf{y}_{\zeta,t-1})\\
&\quad\;\; g(\mathbf{y}_{\zeta,t+1}|\mathbf{x}_{1},\ldots\mathbf{x}_{t+1},\mathbf{y}_{\zeta 1},\ldots,\mathbf{y}_{\zeta,t})\cdots g(\mathbf{y}_{\zeta,T}|\mathbf{x}_{1},\ldots\mathbf{x}_{T},\mathbf{y}_{\zeta 1},\ldots,\mathbf{y}_{\zeta,T-1})\\
&\quad\times q(\mathbf{x}_{1})\cdots q(\mathbf{x}_{t}|\mathbf{x}_{1},\ldots,\mathbf{x}_{t-1})q(\mathbf{x}_{t+1}|\mathbf{x}_{1},\ldots,\mathbf{x}_{t})\cdots q(\mathbf{x}_{T}|\mathbf{x}_{1},\ldots,\mathbf{x}_{T-1})\\
&\quad\times dz_{1}\cdots dz_{t}dz_{t+1}\cdots dz_{T}d\mathbf{x}_{1}\cdots d\mathbf{x}_{t}d\mathbf{x}_{t+1}\cdots d\mathbf{x}_{T}d\mathbf{y}_{\zeta 1}\cdots\\
&\quad\;\; d\mathbf{y}_{\zeta,t}d\mathbf{y}_{\zeta,t+1}\cdots d\mathbf{y}_{\zeta,T}\\
&= \sum_{t=1}^{T}\iint\ldots\int\mathbf{S}_{ADt}w_{t}f(\mathbf{G}_{t})\\
&\quad\;\;\left(\iint\ldots\int f(\mathbf{G}_{>t})dz_{t+1}\cdots dz_{T}d\mathbf{x}_{t+1}\cdots d\mathbf{x}_{T}d\mathbf{y}_{\zeta,t+1}\cdots d\mathbf{y}_{\zeta,T}\right)\\
&\quad\;\; dz_{1}\cdots dz_{t}d\mathbf{x}_{1}\cdots d\mathbf{x}_{t}d\mathbf{y}_{\zeta 1}\cdots d\mathbf{y}_{\zeta,t}\\
&= \sum_{t=1}^{T}\iint\ldots\int\mathbf{S}_{ADt}w_{t}f(\mathbf{G}_{t})dz_{1}\cdots dz_{t}d\mathbf{x}_{1}\cdots d\mathbf{x}_{t}d\mathbf{y}_{\zeta 1}\cdots d\mathbf{y}_{\zeta,t}\\
&= 0, \qquad\qquad (A8)
\end{aligned}
$$

where w_t is the weight for an individual unit at time t as in Equation 15.

Assuming that $\hat{w}_t$ converges in probability to the true w_t as $n \to \infty$, the estimated sum of the weighted augmented data score in Equation A7 will converge in expectation to 0, ensuring consistent estimation of $\boldsymbol{\theta}_F$ in the case of sequential strongly ignorable treatment assignment.

Solutions to the Augmented Data Score Equations

The weighted augmented data score is

$$
\mathbf{WS}_{AD} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{t=1}^{T_{ik}} w_{tik}\mathbf{S}_{ADtik} \qquad (A9)
$$

where $\mathbf{S}_{ADtik}$ is given by Equation A2, yielding the augmented data score equation

$$\mathbf{WS}_{AD} = \frac{d}{d\boldsymbol{\varphi}}\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{t=1}^{T_{ik}} w_{tik}\varepsilon_{tik}^2 - \sum_{k=1}^{K}\bar{w}_k\boldsymbol{\theta}_{Rk}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\theta}_{Rk}\right] = 0, \qquad \text{(A10)}$$

where $\bar{w}_k = \sum_{i=1}^{n_k}\sum_{t=1}^{T_{ik}} w_{tik}/N_k$. Solving Equation A10 for elements of $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_{Rk}$ yields the equations

$$\hat{\boldsymbol{\theta}}_F = \left(\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{t=1}^{T_{ik}} w_{tik}\mathbf{A}_{Ftik}\mathbf{A}_{Ftik}^T\right)^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{t=1}^{T_{ik}} w_{tik}\mathbf{A}_{Ftik}(Y_{tik} - \mathbf{A}_{Rtik}^T\boldsymbol{\theta}_{Rk}),$$

$$\hat{\boldsymbol{\theta}}_{Rk} = \left(\sum_{i=1}^{n_k}\sum_{t=1}^{T_{ik}} w_{ti|k}\mathbf{A}_{Rtik}\mathbf{A}_{Rtik}^T + \sigma^2\Omega^{-1}\right)^{-1}\sum_{i=1}^{n_k}\sum_{t=1}^{T_{ik}} w_{ti|k}\mathbf{A}_{Rtik}(Y_{tik} - \mathbf{A}_{Ftik}^T\boldsymbol{\theta}_F), \qquad \text{(A11)}$$

where $w_{ti|k} = w_{tik}/\bar{w}_k$. Substituting $\hat{\boldsymbol{\theta}}_{Rk}$ for $\boldsymbol{\theta}_{Rk}$ yields the useful Equation 18 for $\boldsymbol{\theta}_F$.

Appendix B
Pseudolikelihood Estimation

Appendix A showed that given the covariance components σ^2 and $\boldsymbol{\Omega}$, solving the weighted augmented data score for the elements of $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_{Rk}$, $k = 1, \ldots, K$ will provide consistent estimation of the causal effects of time-varying instructional treatments (Equation A11) under the assumption of sequential strongly ignorable treatment assignment. This demonstration, although essential, is not sufficient to clarify how to compute the estimates for two reasons. First, Equation A11 is expressed in terms of the general model (Equation 10) and must be translated back into the terms of the problem at hand, which involves the four-way model given by Equation 9. Second, Equation A11 requires knowledge of the variance components. In this appendix we make the needed translation and show how to apply pseudomaximum likelihood estimation to produce consistent estimates of the variance components and therefore of the causal effects of interest.

Translating Back to the Four-Way Model

Based on Equation 9, the joint density of the observed data and the random effects in the four-way model is $\prod_{k=1}^{K} f(\mathbf{Y}_k|\boldsymbol{\theta}_F, \mathbf{v}_k, \mathbf{u}_k, \mathbf{r}_k, \sigma^2)p(\mathbf{v}_k, \mathbf{u}_k, \mathbf{r}_k|\psi^2, \boldsymbol{\omega}, \boldsymbol{\tau})$, yielding the logarithm of the joint density $L = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{t=1}^{T_{ik}} L_{tik}$, where

$$
\begin{aligned}
L_{tik} = {} & \text{constant} - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\varepsilon^2_{tik}/\sigma^2 - \frac{J_k}{2N_k}\log(\psi^2) - \frac{1}{2N_k}\log(|\boldsymbol{\omega}|) - \frac{n_k}{2N_k}\log(|\boldsymbol{\tau}|) \\
& - \frac{1}{2N_k}\sum_{j=1}^{J_k} v^2_{jk}/\psi^2 - \frac{1}{2N_k}\mathbf{u}^T_k\boldsymbol{\omega}^{-1}\mathbf{u}_k - \frac{1}{2T_{ik}}\mathbf{r}^T_{ik}\boldsymbol{\tau}^{-1}\mathbf{r}_{ik}.
\end{aligned}
\tag{B1}
$$

Applying weights yields the weighted log-joint density

$$
\begin{aligned}
WL = {} & \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{t=1}^{T_{ik}} w_{tik}L_{tik} \\
= {} & \text{constant} - \frac{N}{2}\log(\sigma^2) - \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{t=1}^{T_{ik}} w_{tik}\varepsilon^2_{tik}/\sigma^2 - \frac{J}{2}\log(\psi^2) - \frac{K}{2}\log(|\boldsymbol{\omega}|) \\
& - \frac{n}{2}\log(|\boldsymbol{\tau}|) - \frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\bar{w}_k v^2_{qk}/\psi^2 - \frac{1}{2}\sum_{k=1}^{K}\bar{w}_k\mathbf{u}^T_k\boldsymbol{\omega}^{-1}\mathbf{u}_k - \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\bar{w}_{ik}\mathbf{r}^T_{ik}\boldsymbol{\tau}^{-1}\mathbf{r}_{ik},
\end{aligned}
\tag{B2}
$$

where $J = \sum_{k=1}^{K} J_k$, the total number of teachers, and $\bar{w}_{ik} = \sum_{t=1}^{T_{ik}} w_{tik}/T_k$. We have normalized the weights as follows. First, we normalize w_{tik} to sum to N, the total number of observations. Next, we compute the mean of w_{tik} for each student and normalize these means $\bar{w}_{ik}$ to sum to n, the total number of students. We then normalize the school means $\bar{w}_k$ to sum to K, the total number of schools.

Our approach to pseudolikelihood estimation works by recognizing Equation B2 as the log density of a four-way model having Level-1 variance σ^2/w_{tik}, teacher random effects variance $\psi^2/\bar{w}_k$, school-level variance-covariance matrix $\boldsymbol{\omega}/\bar{w}_k$, and child-level variance–covariance matrix $\boldsymbol{\tau}/\bar{w}_{ik}$. This would be equivalent to a model having the form

$$
\begin{aligned}
& Y_{tik} = \mathbf{A}^T_{Ftik}\boldsymbol{\theta}_F + \left(\frac{1}{\sqrt{\bar{w}_k}}\right)\mathbf{A}^T_{vtik}\mathbf{v}_k + \left(\frac{1}{\sqrt{\bar{w}_k}}\right)\mathbf{A}^T_{utik}\mathbf{u}_k + \left(\frac{1}{\sqrt{\bar{w}_{ik}}}\right)\mathbf{A}^T_{rtik}\mathbf{r}_k + \left(\frac{1}{\sqrt{w_{tik}}}\right)\boldsymbol{\varepsilon}_k \\
& \mathbf{v}_k \sim N(0, \psi^2\mathbf{I}), \mathbf{u}_k \sim N(0, \boldsymbol{\omega}), \mathbf{r}_k \sim N(0, \boldsymbol{\tau}), \boldsymbol{\varepsilon}_k \sim N(0, \sigma^2\mathbf{I}).
\end{aligned}
\tag{B3}
$$

To maximize the weighted log density, we rescale the design matrices so that $\mathbf{A}^*_{vtik} = (\frac{1}{\sqrt{\bar{w}_k}})\mathbf{A}_{vtik}$, $\mathbf{A}^*_{utik} = (\frac{1}{\sqrt{\bar{w}_k}})\mathbf{A}_{utik}$, $\mathbf{A}^*_{rtik} = (\frac{1}{\sqrt{\bar{w}_{ik}}})\mathbf{A}_{rtik}$, and we assume the Level-1 variance to be σ^2/w_{tik}. We then apply the expectation-maximization algorithm as described in Raudenbush and Bryk (2002, chapter 14). The resulting point estimates are consistent based on the results of Appendix A. However, the conventional standard error estimates are inappropriate. We instead apply Equation 18 for model-based standard errors or Equation 19 for Huber-White robust standard errors.

Appendix C
Inverse Probability of Nonresponse Weights

We create a nonresponse weight to adjust for various missing patterns in the time-varying outcomes. Let $a_{tik} = 1$ if child i in school k has response data in Year t and 0 otherwise. The nonresponse weight is inversely proportional to his or her estimated probability of having the observed response pattern given the observed covariate history. We define $f(a)$ and $f(a|.)$ to be marginal and conditional distributions of a. The nonresponse weight ϕ_{tik} in its general form is

$$\phi_{tik} = \frac{f(a_{0ik})\cdots f(a_{tik}|a_{0ik},\cdots,a_{t-1,ik})}{f(a_{0ik}|\mathbf{x}_{0ik})\cdots f(a_{tik}|a_{0ik},\cdots,a_{t-1,ik},\mathbf{x}_{0ik},\cdots,\mathbf{x}_{tik},y_{0ik},\cdots,y_{t-1,ik},z_{1ik},\ldots,z_{t-1})}. \tag{C1}$$

Using nonresponse weights yields consistent estimation of Equation 10 under the assumption that the probability of missingness is unrelated to unobserved covariates given the covariates included in Equation C1.

References

Bryk, A., & Weisberg, H. (1977). Use of the nonequivalent control group design when subjects are growing. *Psychological Bulletin*, *84*, 950–962.

Cohen, D., Raudenbush, S., & Ball, D. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*, 119–142.

Gitelman, A. I. (2005). Estimating causal effects from multilevel group-allocation data. *Journal of Educational and Behavioral Statistics*, *30*, 397–412.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, *11*, 1–12.

Heckman, J. (2005). The scientific model of causality. *Sociological Methodology*, *35*, 1–97.

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

Hong, G. (2004). *Causal inference for multilevel observational data with application to kindergarten retention*. Unpublished doctoral dissertation, University of Michigan, School of Education.

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

Lin, D. Y., Psaty, B. M., & Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, *54*, 948–963.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67–101.

Neyman, J. S. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (D. M. Dabrowska & T. P. Speed, Eds., Trans.). *Statistical Science*, *5*, 465–472. (Original work published 1923)

Pfefferman, D., Skinner, C. J., Homes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection models in multilevel models. *Journal of the Royal Statistical Society, Series B*, *60*, 23–40.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Hong, G., & Rowan, B. (2002, March). *Studying the causal effects of instruction with application to primary-school mathematics*. Paper presented at the Research Seminar II: Instructional and Performance Consequences of High-Poverty Schooling, National Center for Educational Statistics, Washington, DC.

Robins, J. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp. 95–134). New York: Springer.

Robins, J., Greenland, S., & Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of repeated binary outcome. *Journal of the American Statistical Association*, *94*, 687–700.

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

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A*, *147*, 656–666.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics*, *11*, 207–224.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, *81*, 961–962.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association*, *101*, 1398–1407.

Weick, K. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, *21*, 1–19.

Westat. (2001). *The Longitudinal Evaluation of School Change and Performance (LESCP) in Title I schools*. Washington, DC: US Department of Education, Planning and Evaluation Services, DOC #2001–20.

Authors

GUANGLEI HONG is assistant professor, Department of Human Development and Applied Psychology, Ontario Institute for Studies in Education of the University of Toronto, 252 Bloor Street West, 9-184, Toronto, Ontario, Canada M5S 1V6; ghong@oise.utoronto.ca. Her areas of specialization include causal inference theories and methods, multilevel modeling, longitudinal data analysis, policy and program evaluation, and instructional effectiveness.

STEPHEN W. RAUDENBUSH is the Lewis-Sebring Distinguished Service Professor of Sociology and Chair, Committee on Education, University of Chicago, Chicago, IL

60637; sraudenb@uchicago.edu. His areas of specialization are hierarchical linear models, experimental design, and measurement of classrooms, schools, and neighborhoods.

Manuscript received June 27, 2006
Revision received January 9, 2007
Accepted April 4, 2007