Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Evaluating Kindergarten Retention Policy

Guanglei Hong and Stephen W. Raudenbush
Guanglei Hong is Assistant Professor, Ontario Institute for Studies in Education of the University of Toronto, Toronto, Ontario, Canada M5S 1V6 . Stephen W. Raudenbush is Lewis-Sebring Distinguished Service Professor, Department of Sociology, University of Chicago, Chicago, IL 60637 . This research was based on the first author's dissertation, supported by a grant from the American Educational Research Association, which received funds for its grants program from the National Center for Education Statistics and the Office of Educational Research and Improvement (U.S. Department of Education) and the National Science Foundation under grant REC-9980573. Additional support came from the Spencer Foundation in the form of a 2003-2004 Spencer Dissertation Fellowship for Research Related to Education, from the Consortium for Policy Research in Education, and from the Connaught Startup Fund granted by the University of Toronto. Opinions herein are those of the authors and do not necessarily reflect the views of the granting agencies. The authors benefited from the ideas of Susan Murphy, Ben Hansen, Natalya Verbitsky, and participants in the hierarchical linear modeling class at the University of Michigan in the Fall 2003 term. The authors appreciate the very helpful feedback from the editor, the associate editor, and two anonymous reviewers.

Available online: 01 Jan 2012

Evaluating Kindergarten Retention Policy: A Case Study of Causal Inference for Multilevel Observational Data

Guanglei HONG and Stephen W. RAUDENBUSH

This article considers the policy of retaining low-achieving children in kindergarten rather than promoting them to first grade. Under the stable unit treatment value assumption (SUTVA) as articulated by Rubin, each child at risk of retention has two potential outcomes: $Y(1)$ if retained and $Y(0)$ if promoted. But SUTVA is questionable, because a child's potential outcomes will plausibly depend on which school that child attends and also on treatment assignments of other children. We develop a causal model that allows school assignment and peer treatments to affect potential outcomes. We impose an identifying assumption that peer effects can be summarized through a scalar function of the vector of treatment assignments in a school. Using a large, nationally representative sample, we then estimate (1) the effect of being retained in kindergarten rather than being promoted to the first grade in schools having a low retention rate, (2) the retention effect in schools having a high retention rate, and (3) the effect of being promoted in a low-retention school as compared to being promoted in a high-retention school. This third effect is not definable under SUTVA. We use multilevel propensity score stratification to approximate a two-stage experiment. At the first stage, intact schools are blocked on covariates and then, within blocks, randomly assigned to a policy of retaining comparatively more or fewer children in kindergarten. At the second stage, "at-risk" students within schools are blocked on covariates and then assigned at random to be retained. We find evidence that retainees learned less on average than did similar children who were promoted, a result found in both high-retention and low-retention schools. We do not detect a peer treatment effect on low-risk students.

KEY WORDS: Grade retention; Multilevel design; Potential outcomes; Propensity score; Stable unit treatment value assumption.

1. INTRODUCTION

Schools typically allocate children to age-based grade levels; however, children who make inadequate progress may be retained in their current grade rather than being promoted to the next grade. For example, based on the National Household Education Survey, Zill, Loomis, and West (1997) reported that the kindergarten retention rate was about 6% in 1993 and 5% in 1995. Not all educators agree that grade retention is helpful in these cases; indeed, many school districts have adopted a policy of "social promotion" that enables children having academic difficulty to proceed to the next grade with their same-age peers. But "ending social promotion" has recently become a popular slogan to justify grade retention (Ellwein and Glass 1989; Hauser 1998; Roderick, Bryk, Jacobs, Easton, and Allensworth 1999). In principle, the retention policy may affect not only the children who are at risk of repetition, but also those who are at no risk.

In Section 1 we clarify key causal questions, introduce the sample and data, and describe major methodological challenges in studying the effects of grade retention. In Section 2 we propose a framework for causal inference that incorporates school and peer effects. We give empirical results in Section 3 and provide conclusions in Section 4.

1.1 Causal Questions

Many believe that some kindergarten children need more time to mature or more time to master basic skills before progressing to the first grade (Smith and Shepard 1988). Moreover, grouping these immature children with younger peers may increase their self-esteem, thus improving their learning (Plummer and Graziano 1987). In contrast, others argue that grade retention stifles children's cognitive and social development (Morrison, Griffith, and Alberts 1997) and stigmatizes those retained (Jackson 1975; Shepard 1989), slowing down their learning. Moreover, grade retention may simply reproduce an unsuccessful kindergarten experience (Karweit 1992; Leinhardt 1980; Peterson 1989; Reynolds 1992; Tanner and Galis 1997).

The effect of grade retention may not be confined to the individual children who are retained. When low-achieving students are retained in kindergarten, a more homogeneous classroom may ease the teacher's task in managing instructional activities (Shepard and Smith 1988), and also may allow the first-grade teacher to cover more advanced content. Therefore, it is natural to hypothesize that a child promoted to the first grade may fare better when a relatively larger proportion of children experiencing learning difficulties are retained in kindergarten. Similarly, the retention effect on a retainee may depend on the proportion of peers retained at the same time. These arguments rest on the belief that a student's learning outcome can be affected by the treatments assigned to other students.

In this article we ask three questions:

1. What is the effect of being retained in kindergarten versus being promoted to the first grade on the academic learning of retainees when few peers are retained?

Guanglei Hong is Assistant Professor, Ontario Institute for Studies in Education of the University of Toronto, Toronto, Ontario, Canada M5S 1V6 (E-mail: *ghong@oise.utoronto.ca*). Stephen W. Raudenbush is Lewis-Sebring Distinguished Service Professor, Department of Sociology, University of Chicago, Chicago, IL 60637 (E-mail: *sraudenb@uchicago.edu*). This research was based on the first author's dissertation, supported by a grant from the American Educational Research Association, which received funds for its grants program from the National Center for Education Statistics and the Office of Educational Research and Improvement (U.S. Department of Education) and the National Science Foundation under grant REC-9980573. Additional support came from the Spencer Foundation in the form of a 2003–2004 Spencer Dissertation Fellowship for Research Related to Education, from the Consortium for Policy Research in Education, and from the Connaught Startup Fund granted by the University of Toronto. Opinions herein are those of the authors and do not necessarily reflect the views of the granting agencies. The authors benefited from the ideas of Susan Murphy, Ben Hansen, Natalya Verbitsky, and participants in the hierarchical linear modeling class at the University of Michigan in the Fall 2003 term. The authors appreciate the very helpful feedback from the editor, the associate editor, and two anonymous reviewers.

Journal of the American Statistical Association
September 2006, Vol. 101, No. 475, Applications and Case Studies
DOI 10.1198/016214506000000447

2. What is the effect of being retained when comparatively many peers are retained?
3. What is the effect of a high retention rate versus a low retention rate on children at little or no risk of retention?

1.2 Sample and Data

We selected data from the Early Childhood Longitudinal Study Kindergarten cohort (ECLS–K), released by the National Center for Education Statistics, which contains repeated observations of a nationally representative sample of students, their families, teachers, and schools over two school years. Students were observed in the fall and the spring of the kindergarten year and then in the spring of the treatment year. Our analytic sample included 471 kindergarten retainees and 10,255 promoted students in 1,080 schools that allowed for kindergarten retention.

The outcome variables were reading and math scale scores calibrated by item response theory (IRT) (Hambleton, Swaminathan, and Rogers 1991). The test scores of each subject obtained from repeated assessments over the two study years were equated on the same scale. This enabled us to assess the reading and math growth of each student over time and also to compare the reading and math achievement of students from different grade levels. Table 1 presents the sample size, mean, and standard deviation of each of the three rounds of reading and math assessment scores.

1.3 Methodological Challenges

Applying Rubin's (1978) causal model in a conventional way, we define the retention effect on an at-risk child as the difference between the outcome the child would display if retained and the outcome if promoted. This child-specific effect is defined under the stable unit treatment value assumption (SUTVA) that there is a single value of each potential outcome associated with each treatment for each experimental unit, regardless of how the treatments are assigned and of what treatments are received by other experimental units (Rubin 1986). Most researchers simply invoke SUTVA without further theoretical or empirical scrutiny. However, as Rubin (1990) cautioned, SUTVA becomes problematic when, for example, educational treatments are given to children who interact with one another. Due to the multilevel structure of modern schooling, almost every educational treatment is conducted in an organizational setting that bears an impact on a child's potential outcomes. The organizational effect has two major sources: treatment enactment variation and interference between individuals. For example, the potential outcome of a retained child may depend on whether the school allocates additional resources to the retainees, whereas the potential outcome of a child who is not at risk of retention and gets promoted may depend on how many of his or her low-achieving peers are retained. Hence, conceptually there is a distinct set of potential outcome values associated with each treatment for each individual. The multiplicity of potential outcomes poses a problem that is largely unaddressed in the causal inference literature.

Table 1. Descriptive Statistics of the Three Rounds of Assessment

IRT scale score	*n*	*Mean*	*SD*
Reading			
Fall, kindergarten year	10,108	23.52	8.80
Spring, kindergarten year	10,391	33.87	10.91
Spring, treatment year	10,680	56.66	13.48
Mathematics			
Fall, kindergarten year	10,431	20.35	7.23
Spring, kindergarten year	10,556	28.66	8.64
Spring, treatment year	10,678	44.04	8.90

2. EXTENDED FRAMEWORK FOR CAUSAL INFERENCE

2.1 Potential Outcomes and Causal Effects With Relaxation of SUTVA

For the current study, we propose a framework for causal inference in which peer treatment assignments can affect each child's potential outcomes through a scalar function. Assumptions underlying the framework include the following: (a) Generalization of causal inferences is restricted to current school assignments; (b) there is no interference between schools, and (c) treatment assignment is strongly ignorable; that is, one's own and one's peers' treatment assignments are independent of the ensemble of potential outcomes given observed covariates. For clarity of exposition, we introduce basic ideas in the case in which all students of interest attend the same school. We then extend the approach to the study at hand, in which students are nested within many schools.

The Case of a Single School. Given a binary treatment, $z_i = 1$ if child i is retained and $z_i = 0$ if promoted, and N units overall, we have the $1 \times N$ vector of possible treatment assignments $\mathbf{z} = (z_1, z_2, \ldots, z_N) = (z_i, \mathbf{z}_{-i})$, where $\mathbf{z}_{-i}$ is the $1 \times (N-1)$ vector of treatment assignments with z_i removed, for $z_i \in \{0, 1\}$, $i = 1, \ldots, N$. Under this setup, subject i has 2^N potential outcomes, $Y_i(\mathbf{z})$, corresponding to all possible treatment assignments of the N subjects. In the continuous outcome case, the potential outcome for a single subject is a function mapping $\{0, 1\}^N$ to $\mathbb{R}^{2N}$. A contrast between any two of the 2^N potential outcomes for a given subject is a causal effect. Clearly, SUTVA is a special case where $Y_i(\mathbf{z}) \equiv Y_i(z_i, \mathbf{z}_{-i}) = Y_i(z_i)$.

Without imposing further structure, the sheer number of causal effects per subject undermines any attempt to summarize evidence in a readily interpretable way. Moreover, a shift in the treatment assignment of any subject changes the potential outcome of any other subject, making it difficult to conceive of average causal effects or to frame interesting questions for policy. To impose structure that can lead to analytic progress, we follow the work of Verbitsky and Raudenbush (2004) in the context of spatial data. Specifically, we model the impact of $\mathbf{z}$ on a focal subject's potential outcome as operating through z_i and a scalar function $v(\mathbf{z})$. The N-dimensional space is thus reduced to a two-dimensional space. Hence we have

$$Y_i(\mathbf{z}) \equiv Y_i(z_i, \mathbf{z}_{-i}) = Y_i[z_i, v(\mathbf{z})]. \quad (1)$$

Many functions can be formulated to represent theoretical conceptions of the influence of the ensemble of treatment assignments on the outcomes of any focal subject. A generic estimand is $E\{Y[z, v(\mathbf{z})] - Y[z', v(\mathbf{z}')]\}$, where z and z' are alternative treatment assignments for an individual and $\mathbf{z}$ and $\mathbf{z}'$ are alternative treatment assignments for all individuals. To pose questions of interest in the current study, let $v(\mathbf{z}) = 1$ if

a high proportion of kindergartners are retained and $v(\mathbf{z}) = 0$ if not. Thus $Y[z, v(\mathbf{z})]$ can take on four possible values—$Y(1, 1), Y(0, 1), Y(1, 0)$, and $Y(0, 0)$—generating estimands corresponding to the questions of interest in our study. $E[Y(1, 1) - Y(0, 1)]$ is the average causal effect of being retained if the school retains a high proportion of kindergartners; $E[Y(1, 0) - Y(0, 0)]$ is the average causal effect of being retained if the school retains a low proportion of kindergartners; and $E[Y(0, 1) - Y(0, 0)]$ is the average causal effect of attending a high- versus a low-retention school if a student is promoted.

The Case of Multiple Schools. Suppose now that students are level-1 units nested within schools at level 2. We introduce, along with the $1 \times N$ treatment assignment vector $\mathbf{z} = (z_1, z_2, \ldots, z_N)$ for $z_i \in \{0, 1\}$, the $1 \times N$ school assignment vector $\mathbf{s} = (s_1, s_2, \ldots, s_N)$, where the possible values of s_i are the school identification numbers $j = 1, \ldots, J$. The generic potential outcome for subject i is then $Y_i[z_i, v(\mathbf{z}), \mathbf{s}]$, modifying our causal estimands to have the form

$$E\big\{Y[z, v(\mathbf{z}), \mathbf{s}] - Y[z', v(\mathbf{z}'), \mathbf{s}']\big\}, \tag{2}$$

that is, the average causal effect of treatment assignments z, $v(\mathbf{z})$, and school assignments $\mathbf{s}$ compared with treatment assignments z', $v(\mathbf{z}')$, and school assignments $\mathbf{s}'$. Such a causal effect would be estimable if assignments to schools and treatments were ignorable (e.g., students were assigned at random to schools, retention rates were assigned at random to schools, and students within schools were assigned at random to be retained). However, in the real world, students are not assigned at random to schools, and thus estimand (2) is not of practical interest. Therefore, we modify our estimand as follows:

a. *Intact schools*. Given the social and geographic segregation of schools in many regions on the basis of student social background and ethnicity, we are typically interested in generalizing results to a set of existing schools rather than to a hypothetical world in which children are assigned at random to schools. In particular, in the current study we are interested in estimating the average treatment effects *given current membership* in those schools. Therefore, our estimands of interest will be conditional on current school assignments,

$$E\big\{Y[z, v(\mathbf{z}), \mathbf{s}^*] - Y[z', v(\mathbf{z}'), \mathbf{s}^*] | \mathbf{S} = \mathbf{s}^*\big\}, \tag{3}$$

where $\mathbf{s}^*$ is the vector of school assignments observed in the sample.

We impose two additional restrictions to make inferences tractable.

b. *No interference between schools*. We assume that the potential outcomes for student i depend on the identities and treatment assignments of that student's schoolmates, whereas the identities and treatment assignments of students attending other schools are assumed to be uninformative about student i's outcomes. Therefore, we sort $\mathbf{z}$ and $\mathbf{s}$ by school identification number, so that $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_J)$ and $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_J)$, where $\mathbf{z}_j = (z_{1j}, z_{2j}, \ldots, z_{n_j j})$ is the $1 \times n_j$ vector of treatment assignments of students assigned to school j and $\mathbf{s}_j = (j, j, \ldots, j)$ is the corresponding $1 \times n_j$ vector of school assignments. Our relaxed form of SUTVA allows no interference between intact schools; therefore, we write

$$Y_i[z_i, v(\mathbf{z}), \mathbf{s}^*] = Y_{ij}[z_{ij}, v(\mathbf{z}_j), \mathbf{s}_j^*] \equiv Y_{ij}[z_{ij}, v(\mathbf{z}_j)]. \tag{4}$$

Henceforth, to simplify our notation, we denote by $v(\mathbf{Z}) = V$ the random variable that takes on values $v(\mathbf{z}) = v = 0$ for low-retention schools or 1 for high-retention schools.

c. *Strongly ignorable treatment assignment*. In our case study, retention rates were not assigned at random to schools, and students within schools were not assigned at random to be retained. Let $\mathbf{X}$ be a vector of child-level covariates and let $\mathbf{W}$ be a vector of school-level covariates. Here $\mathbf{W}$ includes school-level aggregates of child-level covariates indicating, for example, the demographic composition of students attending the same school. Causal inferences are nevertheless possible if treatment assignments are strongly ignorable within levels of covariates (where $\mathbf{X} = \mathbf{x}$ and $\mathbf{W} = \mathbf{w}$) for intact schools such that $E[Y(z, v)|Z = z, V = v, \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}] = E[Y(z, v)|\mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}]$, in which case the conditional average causal effect, $E[Y(z, v)|\mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}] - E[Y(z', v')|\mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}]$, is equivalent to the observed data estimand, $E[Y(z, v)|Z = z, V = v, \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}] - E[Y(z', v')|Z = z', V = v', \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}]$.

2.2 Causal Estimands for Subpopulations and Multilevel Experimental Designs

In the current study, the potential outcomes for child i attending school j, denoted by $Y_{ij}(z_{ij}, v_j)$, could take up to four values: $Y_{ij}(1, 1)$, $Y_{ij}(0, 1)$, $Y_{ij}(1, 0)$, and $Y_{ij}(0, 0)$. However, not all of the four potential outcomes are defined for some children. A child who would likely be retained under a high retention rate may not be at such a risk under a low retention rate, whereas another child may never be retained even under a high retention rate. Let q_1 denote a child's probability of being retained under a high retention rate, and let q_0 denote the child's probability of repetition under a low retention rate. If monotonicity holds for every child (i.e., if one's probability of repetition under a high retention rate is always equal to or higher than that under a low retention rate), then we can identify three subpopulations of children:

(A) Children at risk of being retained under a low retention rate ($q_1 \geq q_0 > 0$)

(B) Children at risk of being retained only under a high retention rate ($q_1 > 0, q_0 = 0$)

(C) Children at no risk of being retained even under a high retention rate ($q_1 = q_0 = 0$).

Table 2 lists the potential outcomes and the causal effects of interest. For subpopulation (A), we are interested in estimating $E[Y(1, 0) - Y(0, 0)]$ and $E[Y(1, 1) - Y(0, 1)]$ as defined in Section 2.1. The difference between these two effects, $E[Y(1, 1) - Y(0, 1)] - E[Y(1, 0) - Y(0, 0)]$, provides information about whether the average retention effect depends on the school-level retention rate. Children in subpopulation (B) have three potential outcome values. The causal effect of particular interest is $E[Y(1, 1) - Y(0, 1)]$. The only estimand defined for the children in subpopulation (C) is $E[Y(0, 1) - Y(0, 0)]$.

A randomized experimental design is usually ideal for treatment effect estimation, because it ensures ignorable treatment assignment. For the causal questions at hand, we conceive a multilevel randomized experiment as follows.

Through a cluster randomized trial, intact schools that allow kindergarten retention are assigned at random to a high

Table 2. Potential Outcomes and Causal Effects for Subpopulations of Children

Subpopulations	*Probabilities of retention*	*Potential outcomes*	*Causal effects of interest*
(A)	$q_1 \geq q_0 > 0$	$Y(1,1),\ Y(0,1),\ Y(1,0),\ Y(0,0)$	$E[Y(1,0) - Y(0,0)]$, $E[Y(1,1) - Y(0,1)]$, $E[Y(1,1) - Y(0,1)] - E[Y(1,0) - Y(0,0)]$
(B)	$q_1 > 0$, $q_0 = 0$	$Y(1,1),\ Y(0,1),\ Y(0,0)$	$E[Y(1,1) - Y(0,1)]$
(C)	$q_1 = q_0 = 0$	$Y(0,1),\ Y(0,0)$	$E[Y(0,1) - Y(0,0)]$

retention rate with a probability Q. This is to be followed by a multisite randomized trial, in which at-risk students within each school are assigned at random to retention according to the fixed retention rate determined by the school-level treatment assignment such that $q_1 = P[Z = 1|V = 1]$ and $q_0 = P[Z = 1|V = 0]$. (This is similar to a standard split-plot design, in which schools are "whole plots" assigned to high or low retention rate policies and students are "subplots" to be retained or promoted.)

2.3 Estimable Causal Effects Under Strong Ignorability

Our current nonexperimental study can be approximated by a cluster-level randomized block design followed by an individual-level randomized block design within each cluster. First, intact schools are assigned at random to $V = 1$ within school-level blocks defined by $\mathbf{W}$ with a probability $Q = Q(\mathbf{W})$; thus

$$Q = P(V = 1|\mathbf{W}). \tag{5}$$

Next, given a school's assignment to high or low retention rate, children at risk of repetition within each school are assigned at random to $Z = 1$ or $Z = 0$ within blocks defined by $\mathbf{X}$ and $\mathbf{W}$. Hence $q_1 = q_1(\mathbf{X}, \mathbf{W})$ and $q_0 = q_0(\mathbf{X}, \mathbf{W})$, where

$$\begin{aligned} q_1 &= P(Z = 1|V = 1, \mathbf{X}, \mathbf{W}), \\ q_0 &= P(Z = 1|V = 0, \mathbf{X}, \mathbf{W}). \end{aligned} \tag{6}$$

In combination, a child's probability of receiving a certain treatment under a certain retention rate is $P(Z = z, V = v|\mathbf{X}, \mathbf{W}) = P(Z = z|V = v, \mathbf{X}, \mathbf{W})P(V = v|\mathbf{W})$.

With data from a cluster randomized block design followed by a multisite randomized block design, we expect to obtain an unbiased estimate of each of the following causal estimands. The first of these estimands is the average causal effect of retention relative to promotion, conditional on covariates, under a low retention rate for children in subpopulation (A),

$$E[Y_A(1,0) - Y_A(0,0)|\mathbf{X}, \mathbf{W}]. \tag{7}$$

The second is the average conditional retention effect under a high retention rate for children in subpopulation (A) and those in subpopulation (B). The union of these two subpopulations contains all of the children who would be ever at risk of repetition, denoted as subpopulation (AR),

$$E[Y_{AR}(1,1) - Y_{AR}(0,1)|\mathbf{X}, \mathbf{W}]. \tag{8}$$

For children in subpopulation (A), the difference between (7) and (8) indicates the extent to which the causal effect of the individual-level retention treatment depends on the school-level retention rate. For children in subpopulation (C), the third estimand, the average causal effect of high retention rate versus low retention rate, conditional on school-level covariates $\mathbf{W}$, is defined as

$$E[Y_C(0,1) - Y_C(0,0)|\mathbf{W}]. \tag{9}$$

In our nonexperimental data, kindergarten retention is a highly selective process that depends on various pretreatment conditions; so is a school's selection of the retention rate. In theory, a child's probability of being retained is likely associated with his or her demographic characteristics; cognitive, emotional, and social development; and previous learning experiences at home and in school. Meanwhile, a school's adoption of a high retention rate is associated mainly with school characteristics including compositions of students and teachers, instructional practices in kindergarten, principal characteristics, school resources and climate, and policy context.

Rosenbaum and Rubin (1983) showed a way of summarizing in a unidimensional propensity score all of the observed pretreatment information associated with the assignment to a particular treatment. For a binary treatment measure, the joint distribution of all of the observed pretreatment covariates is balanced between the treatment groups given the propensity score. In a nonrandomized experiment, if the treatment assignment is strongly ignorable given the observed pretreatment covariates, then statistical adjustment for the propensity score will be sufficient for producing unbiased estimates of the average treatment effect. Following (5) and (6), if the assignment of schools to a high or low retention rate is strongly ignorable given the observed school-level pretreatment covariates $\mathbf{W}$, then we can estimate school j's propensity of selecting a high retention rate, Q_j, as a function of $\mathbf{W}_j$. If the assignment of students to retention or promotion under high or low retention rate is strongly ignorable given the observed student-level pretreatment covariates $\mathbf{X}$ and the school-level covariates $\mathbf{W}$, then student i's propensities of being retained in school j, denoted by q_{1ij} if the school has adopted a high retention rate and by q_{0ij} otherwise, can be estimated as functions of $\mathbf{X}_i$ and $\mathbf{W}_j$. Because students are nested within schools, for strong ignorability to hold, the vector of school-level covariates $\mathbf{W}$ needs to capture the pretreatment commonalities of students from the same school that predict the student-level treatment assignment.

The two propensity scores for every child, q_1 and q_0, could both be estimated if schools were assigned at random to a high or low retention rate even though the within-school assignments of students to retention or promotion might not be random. This is because, under randomization of the school-level retention rate, the joint distribution of all pretreatment covariates is balanced between high-retention schools and low-retention schools. Hence the propensity score function for q_1 estimated from the observed data of high-retention school students under strong ignorability is applicable to low-retention

school students as well. Similarly, the propensity score function for q_0 applies to both low-retention and high-retention school students.

But without randomization of the school-level retention rate, as is the case in our nonexperimental study, only one of these two propensity scores for every child can be estimated from the observed data. This fact constrains our ability to estimate the causal effects defined in (7) and (8). Equation (7) can be decomposed as $E[Y_A(1,0) - Y_A(0,0)|V = 1, \mathbf{X}, \mathbf{W}]P(V = 1|\mathbf{W}) + E[Y_A(1,0) - Y_A(0,0)|V = 0, \mathbf{X}, \mathbf{W}]P(V = 0|\mathbf{W})$. The problem is that the first term, $E[Y_A(1,0) - Y_A(0,0)|V = 1, \mathbf{X}, \mathbf{W}]$, cannot be directly estimated, because the propensity of retention under a low retention rate is not estimable for children attending high-retention schools. We can only estimate its second term,

$$\delta_{Z0} = E[Y_A(1,0) - Y_A(0,0)|V = 0, \mathbf{X}, \mathbf{W}], \qquad (10)$$

that is, the conditional effect of kindergarten retention under a low retention rate for children actually attending low-retention schools. For a similar reason, the low-retention school children's propensity of being retained under a high retention rate remains unknown. The interaction effect between retention and retention rate for those in subpopulation (A) is not estimable. Similarly, if we decompose (8) as $E[Y_{AR}(1,1) - Y_{AR}(0,1)|V = 1, \mathbf{X}, \mathbf{W}]P(V = 1|\mathbf{W}) + E[Y_{AR}(1,1) - Y_{AR}(0,1)|V = 0, \mathbf{X}, \mathbf{W}]P(V = 0|\mathbf{W})$, then we note that the second term, $E[Y_{AR}(1,1) - Y_{AR}(0,1)|V = 0, \mathbf{X}, \mathbf{W}]$, cannot be estimated directly from nonexperimental data, although we can estimate its first term,

$$\delta_{Z1} = E[Y_{AR}(1,1) - Y_{AR}(0,1)|V = 1, \mathbf{X}, \mathbf{W}], \qquad (11)$$

that is, the conditional effect of kindergarten retention under the high retention rate for children actually attending high-retention schools. For those in subpopulation (C), the estimand can be estimated as defined in (9),

$$\delta_{V0} = E[Y_C(0,1) - Y_C(0,0)|\mathbf{W}]. \qquad (12)$$

In summary, δ_{Z0}, δ_{Z1}, and δ_{V0} are among the causal estimands directly estimable under strong ignorability. These three causal estimands are of particular interest to the current study. They also exemplify possible ways of defining causal effects in multilevel nonexperimental data more generally when SUTVA is relaxed.

3. PROPENSITY SCORE–BASED CAUSAL INFERENCES

In this section we apply propensity score stratification to estimate causal effects defined by (10)–(12). We present in detail the estimation of the causal effects on the reading outcome and then briefly summarize the results for math. We used HLM6.0 (Raudenbush, Bryk, Cheong, Congdon, and du Toit 2005) for most of the analyses.

3.1 Propensity Score Estimation and Stratification

Empirical Identification of High-Retention Schools. The ratio of the sampled number of kindergarten retainees during the treatment year to the sampled number of kindergartners during the pretreatment year in each school provides an unbiased estimate of that school's kindergarten retention rate. Using a two-level logistic regression model, we obtained an empirical Bayes estimate of each school's retention rate. We found considerable between-school variation in retention rates. With an average school-level log-odds of retention of -3.84 and a variance of .37, school-specific retention rates as low as 0 and as high as .27 were not implausible. For each school, we calculated a 95% posterior confidence interval for its retention rate and classified as "high-retention schools" those with a lower confidence limit exceeding the national average. Among the 790 sampled schools that had information on kindergarten retention rate, 57 enrolling 560 sampled children were thus labeled high-retention schools, and the remaining 733 schools serving 9,233 sampled children were labeled low-retention schools. The number of sampled retainees ranged from 2 to 10 in high-retention schools and from 0 to 2 in low-retention schools.

School Propensity of Adopting a High Retention Rate, Q. In the ECLS–K dataset, we found measures of many theoretically important predictors of school retention rate. The pretreatment information included school type (public or private), urbanicity, geographic region, principal characteristics (e.g., demographic background, credentials, work experience), school climate (e.g., goals, academic emphasis, security, outreach to parents), and school resources (e.g., funding, staffing, services, facilities, curriculum materials). We also aggregated student information to the school level. (A list of pretreatment covariates and information about the specification of each propensity model are available at the first author's website, *http://home.oise.utoronto.ca/~ghong*.) We tentatively assumed that, given all of these observed covariates, the assignment of school-level retention rate would be independent of other unmeasured covariates. Hence, following (5), we used logistic regression to compute the estimate, $\hat{Q}$, of each school's conditional probability of adopting a high retention rate given the observed school-level covariates. High-retention schools were notable, for example, in tending to enroll students having below-average mean teacher ratings on science and social science skills and also in having comparatively large numbers of students with disabilities. Such schools tended to provide comparatively more years of bilingual services, to base kindergarten assignment on children's preschool experience, and to enforce a sign-in policy. On the basis of the logit of $\hat{Q}$, we divided the sample of schools into seven strata (Table 3). The distribution of the logit of $\hat{Q}$ was balanced in all of the strata containing both low- and high-retention schools, and balance was achieved in >99% of the 161 school-level pretreatment covariates.

Table 3. Distribution of the Logit of Propensity for High Retention Rate

	Low retention rate			High retention rate		
Stratum	*n*	*Mean*	*SD*	*n*	*Mean*	*SD*
$Q = 0$	80	−6.79	.59	1	−6.70	
$Q = 1$	336	−4.83	.60	0		
$Q = 2$	202	−3.14	.40	13	−3.12	.36
$Q = 3$	61	−1.99	.24	8	−1.93	.20
$Q = 4$	43	−1.01	.33	9	−.98	.39
$Q = 5$	8	−.03	.21	18	−.02	.23
$Q = 6$	3	1.36	.38	8	1.31	.49

Child Propensity of Repetition in High-Retention Schools, q_1. The ECLS–K dataset provided a comparatively comprehensive set of pretreatment predictors of kindergarten repetition, including child demographic characteristics; academic performance at the beginning and the end of the first kindergarten year; physical and mental health; tardiness and absenteeism, day care, preschool, and Head Start experiences; learning experiences during the first kindergarten year; home literacy; parenting style; parent involvement in school; kindergarten teacher background; teacher beliefs and teaching practices; classroom composition; and school characteristics as listed before. We estimated q_1 for students attending high-retention schools as a function of the observed school-level covariates, student-level covariates, and a school-specific random effect estimated through an empirical Bayes procedure. Among high-retention school children, retainees differed from promoted children in a number of ways; for example, retainees were likely to be boys, to be young at entry to kindergarten, and to score low on pretests of reading and general knowledge. The schools that the retained children attended had comparatively high average scores in pretests.

The kindergarten retainees attending high-retention schools had a minimum value of -3.89 in the logit of $\hat{q}_1$. Choosing this minimum as the cutoff point, we identified 89 promoted children in high-retention schools whose logit of $\hat{q}_1$ was below this value. Because these children did not have counterparts in the retained group, they were identified as at extremely low risk of repetition under a high retention rate and hence likely would belong to subpopulation (C). We classified those children whose logit of $\hat{q}_1$ was above the cutoff value as members of the subpopulation (AR).

Then we divided the sample of high-retention school students into seven strata on the basis of the logit of $\hat{q}_1$. As displayed in Table 4, the lowest stratum contains 89 promoted children showing almost no risk of repetition, whereas the highest stratum contains 22 kindergarten retainees showing the highest probability of being retained. In the remaining five strata, we found no significant difference in the distribution of the logit of $\hat{q}_1$ between the promoted children and the retainees. Moreover, under this propensity stratification, 97.2% of the 213 child-level and school-level pretreatment covariates showed no significant difference at the 5% level between the promoted group and the retained group.

Child Propensity of Repetition in Low-Retention Schools, q_0. In parallel, q_0 was estimated through another two-level logistic regression model for children attending low-retention schools only. Retainees in low-retention schools differed from their promoted peers in a number of ways. For example, similar to the trend observed in high-retention schools, kindergartners who had a higher probability of being retained in low-retention schools appeared to have lower achievement in reading, mathematics, and general knowledge, whereas their schools' average pretreatment achievement scores tended to be comparatively high. In addition, retained children were more likely to have fallen behind due to poor health. They were less likely to receive individual tutoring in reading and were more often found in schools in which teacher salaries were low.

Table 4. Distribution of the Logit of Propensity for Kindergarten Retention in High-Retention Schools

	Promoted			Retained		
Stratum	n	Mean	SD	n	Mean	SD
$q_1 = 0$	89	−4.75	.65	0		
$q_1 = 1$	110	−3.43	.25	7	−3.35	.35
$q_1 = 2$	139	−2.51	.28	10	−2.36	.30
$q_1 = 3$	170	−1.25	.46	40	−1.19	.46
$q_1 = 4$	33	.05	.23	40	.09	.26
$q_1 = 5$	19	.87	.32	64	.99	.29
$q_1 = 6$	0			22	2.32	.53

Table 5. Distribution of the Logit of Propensity for Kindergarten Retention in Low-Retention Schools

	Promoted			Retained		
Stratum	n	Mean	SD	n	Mean	SD
$q_0 = 0$	4,062	−8.13	1.21	0		
$q_0 = 1$	3,055	−5.55	.58	9	−5.41	.63
$q_0 = 2$	1,211	−3.90	.36	21	−3.78	.36
$q_0 = 3$	457	−2.83	.23	22	−2.77	.24
$q_0 = 4$	209	−2.09	.21	26	−2.01	.23
$q_0 = 5$	137	−1.39	.19	32	−1.37	.19
$q_0 = 6$	64	−.74	.17	32	−.72	.18
$q_0 = 7$	17	−.23	.08	22	−.21	.09
$q_0 = 8$	12	.56	.47	60	.60	.53
$q_0 = 9$	0			7	2.70	.27

We divided this subsample into 10 strata on the basis of the logit of $\hat{q}_0$. The lowest stratum consisted of 4,062 promoted children whose logit of $\hat{q}_0$ was below -6.40 and did not have counterparts in the retained group. Children in the remaining nine strata, including the seven retainees in the top stratum who did not have counterparts in the promoted group, were considered to be from subpopulation (B). As shown in Table 5, we found no significant difference in the distribution of the logit of $\hat{q}_0$ between the retained group and the promoted group in the middle eight strata. Within-stratum balance was achieved in about 95% of the 213 pretreatment covariates.

Empirical Identification of Low-Retention School Children in Subpopulation (C). Children in subpopulation (C) were unlikely to be retained even under a high retention rate. The likely risk status under a high retention rate of children assigned to low-retention schools is, of course, a counterfactual quantity, the estimation of which requires special assumptions. We assumed that had a low-retention school adopted a high retention rate instead, then its students would have been subject to similar selection criteria as defined by the propensity model specified for high-retention school children. Having estimated the propensity model from the observed data of the high-retention school students, we then used the same model to predict the propensity score for each student enrolled in a low-retention school, and used -3.89 as the cutoff point as before for identifying every child's risk status under a high retention rate. In this way, we identified 1,305 promoted children in 503 low-retention schools who were highly unlikely to be retained even if their schools had chosen to have a high retention rate instead. Also in subpopulation (C) are 89 children in 39 high-retention schools. Balance was achieved between low-retention and high-retention schools in more than 98% of the 224 school-level pretreatment covariates, including the aggregated measures of low-risk children.

The small number of low-risk students attending high-retention schools limited our statistical power to detect the effect of attending a high-retention school on low-risk children. Therefore, we undertook an additional analysis using a less stringent cutoff point in defining low risk; specifically, we chose −3.00 as the new cutoff point. The expanded sample included 199 promoted children enrolled in 46 high-retention schools and 2,610 promoted children enrolled in 625 low-retention schools. Again, we were able to achieve balance in >98% of the covariates.

3.2 Causal Effect Estimation

Retention Effect on Reading for Low-Retention School Children in Subpopulation (A). The first estimand defined in (10) is equivalent to $\delta_{Z0} = E[Y_A(1,0) - Y_A(0,0)|V = 0, q_0]$. After removing the children identified to be at almost no risk for repetition under a low retention rate (i.e., those in the $q_0 = 0$ stratum), the subsample included 5,162 promoted children and 231 retainees from 714 low-retention schools. The raw difference in reading between the retainees and the promoted children was −19.27. The within-stratum mean differences between the two groups, as shown in Table 6, were much smaller. Although there appeared to be a decrease in the retention effect on reading as low-retention school children's propensity of repetition increased, such a trend was not obvious in the middle part of the distributions of the two treatment groups (i.e., from the second to the sixth strata), where the data were less sparse. We made a similar observation in comparing the bivariate distribution of the reading outcome and the logit of $\hat{q}_0$ between the retained group and the promoted group (Fig. 1). Under our assumptions, the vertical distance between the two loess lines approximated the conditional effect of retention versus promotion for any given value of the logit of $\hat{q}_0$.

The standard deviations in Table 6 were estimated with no consideration of the clustering of students within schools. To take into account the clustered nature of the sample and to examine the variation of the retention effect across the high-retention schools, we analyzed a two-level model, as shown here for child i in school j. We computed the proportion of retainees in each stratum as a sample weight, as recommended by Pfefferman, Skinner, Homes, Goldstein, and Rasbash (1998),

$$Y_{ij} = \gamma_0 + u_j + (\delta_{Z0} + \Delta_{Z0j})z_{ij} + \sum_{g=1}^{8} \gamma_g q_{0gij} + \gamma_9(\text{Logit}\_\hat{q}_0)_{ij} + e_{ij}; \quad (13)$$

$$\begin{pmatrix} u_j \\ \Delta_{Z0j} \end{pmatrix} \sim \text{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_u & \tau_{u\cdot Z0} \\ \tau_{u\cdot Z0} & \tau_{Z0} \end{pmatrix}\right]; \qquad e_{ij} \sim \text{N}(0, \sigma^2).$$

Here q_{0gij}, $g = 1, \ldots, 8$, are dummy indicators for eight of the nine propensity strata that subclassify the at-risk children in low-retention schools. We made additional adjustments for $(\text{Logit}\_\hat{q}_0)_{ij}$ to remove residual within-stratum bias. In the random part of the model, u_j is the school-specific random effect on the reading outcome, whereas Δ_{Z0j} represents the school-specific increment to the retention effect. The foregoing model allowed us to estimate the variation of the retention effect across the low-retention schools, denoted by τ_{Z0}, as well as the covariance between the school-specific potential outcome of promotion and the school-specific retention effect, $\tau_{u\cdot Z0}$. The analytic results are presented in Table 7.

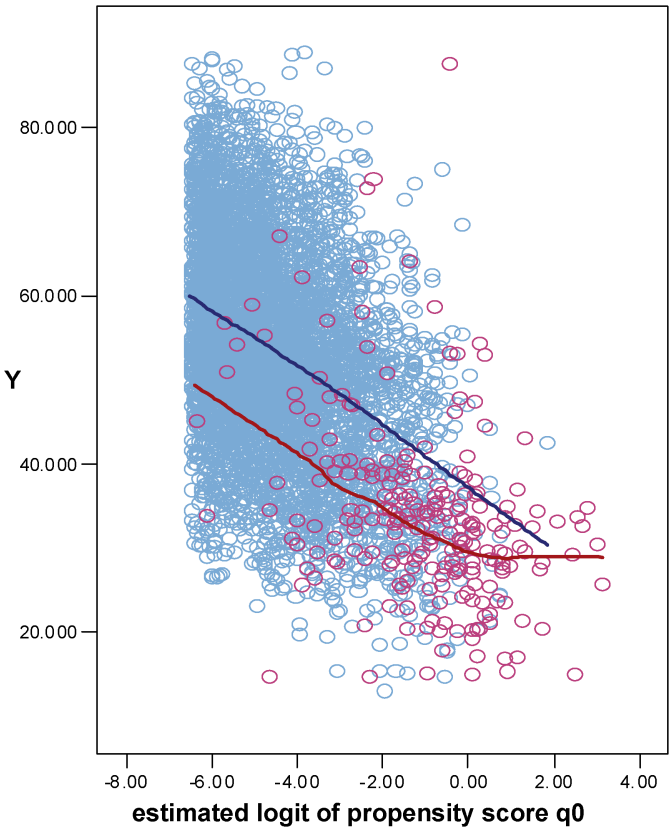

Figure 1. Association Between the Logit of Propensity of Retention and the Retention Effect on Reading for Low-Retention School Children in Subpopulation (A) (Z: ◯ promoted; ◯ retained).

The adjusted within-stratum mean difference between the retained and promoted students was $\hat{\delta}_{Z0} = -8.18$, with a standard error of .94 and a 95% confidence interval of (−10.02, −6.34). This point estimate is almost three-fifths of a standard deviation of the reading outcome. We found statistically significant variation in the retention effect across the low-retention schools ($\hat{\tau}_{Z0} = 38.17$, $\chi^2 = 310.32$, $df = 169$, $p < .001$). Assuming the school-specific retention effect to be normally distributed, this effect would range from −20.29 to 3.93 among 95% of the low-

Table 6. Within-Stratum Distribution of the Reading Outcome of At-Risk Students Attending Low-Retention Schools

	Promoted			Retained			Mean
Stratum	n	Mean	SD	n	Mean	SD	difference
$q_0 = 1$	3,051	56.83	11.48	9	45.00	14.63	−11.83
$q_0 = 2$	1,207	51.19	11.41	21	41.11	11.73	−10.08
$q_0 = 3$	457	48.58	11.98	22	39.58	12.49	−9.00
$q_0 = 4$	209	44.43	10.96	26	36.96	10.82	−7.48
$q_0 = 5$	137	42.46	12.01	32	33.86	7.73	−8.60
$q_0 = 6$	64	39.87	12.19	32	32.59	13.45	−7.28
$q_0 = 7$	17	38.66	12.75	22	33.53	8.20	−5.13
$q_0 = 8$	12	32.74	9.67	60	29.27	8.32	−3.47
$q_0 = 9$	0			7	28.80	6.81	
Total	5,154	53.57	12.37	231	34.30	11.14	−19.27

Table 7. Causal Effect of Retention versus Promotion on Retainees' Reading Outcome in Low-Retention Schools

Fixed effect	*Coefficient*	*SE*	*t*
Promoted at-risk children intercept, γ_0	51.90	.30	170.48
Retention effect in low-retention schools, δ_{Z0}	−8.18	.94	−8.71
Propensity stratum 1, γ_1	−2.01	3.91	−.52
Propensity stratum 2, γ_2	−3.18	3.61	−.88
Propensity stratum 3, γ_3	−2.58	3.49	−.74
Propensity stratum 4, γ_4	−4.42	3.38	−1.31
Propensity stratum 5, γ_5	−4.56	3.31	−1.38
Propensity stratum 6, γ_6	−4.01	3.35	−1.20
Propensity stratum 7, γ_7	−2.82	3.42	−.83
Propensity stratum 8, γ_8	−4.40	3.14	−1.40
Logit of propensity score, γ_9	−2.85	.30	−9.38

Random effect	*Variance*	*df*	χ^2	*p value*
School mean, u_j	50.43	169	1053.92	<.001
School retention effect, Δ_{Z1j}	38.17	169	310.32	<.001
Correlation between u_j and Δ_{Z1j}	−.377			
Level-one effect, e_{ij}	77.21			

retention schools. The estimated correlation between school-specific retention effect and school mean outcome was −.38, indicating a more severe negative effect of retention on the reading outcome of retainees in low-retention schools in which the average reading achievement was higher. Result of additional analysis supported our earlier observation that the retention effect did not depend on the propensity of repetition.

Retention Effect on Reading for High-Retention School Children in Subpopulation (AR). The second causal estimand, as defined in (11), is equivalent to $\delta_{Z1} = E[Y_{AR}(1,1) - Y_{AR}(0,1)|V = 1, q_1]$. After excluding the high-retention school children identified to be at low risk of repetition under a high retention rate (i.e., those in the $q_1 = 0$ stratum), 654 children remained in this subsample, including 471 promoted students and 183 retainees from 57 high-retention schools. The results paralleled those for the low-retention schools. In general, the within-stratum mean differences in reading between the retainees and the promoted children were smaller than the raw difference, −15.96 (not tabulated). Nonetheless, the kindergarten retainees did not perform as well on average as their counterparts in the promoted group in most of the propensity strata except for the $q_1 = 1$ stratum, in which there were only seven retainees. Again, there was no clear pattern suggesting any dependence of the retention effect on the propensity of repetition. Hence a weighted two-level linear model similar to that specified in (13) seemed applicable, except that the current model included five dummies to represent five of the six propensity strata.

The adjusted within-stratum mean difference between the retained and promoted students was $\hat{\delta}_{Z1} = -8.86$, with a standard error of 1.38 and a 95% confidence interval of $(-11.56, -6.16)$. Again, there was statistically significant variation in the retention effect across the high-retention schools ($\hat{\tau}_{Z1} = 42.58$, $\chi^2 = 122.05$, $df = 49$, $p < .001$). Assuming the school-specific retention effect to be normally distributed, this effect would range from −21.65 to 3.93 among 95% of the high-retention schools. The correlation between school-specific retention effect and school mean outcome was −.02, indistinguishable from 0.

Causal Effect of High Retention Rate Relative to Low Retention Rate on Reading for Children in Subpopulation (C). The third estimand, δ_{V0} as defined in (12), was estimated through statistical adjustment for a school's propensity of adopting a high retention rate. Although the raw difference in the average reading outcome was −3.08 favoring the low-retention school children, in five of the seven strata in which the reading outcome was observed for children from both treatment groups, the mean difference in reading seemed to fluctuate around 0, ranging from −5.89 to 3.56.

The first two strata, in which no subpopulation (C) children were sampled in high-retention schools, were combined into one in our subsequent analysis. Again, to take into account the clustered nature of the data, we specified a two-level regression model,

$$Y_{ij} = \gamma_0 + u_j + \delta_{V0} v_j + \gamma_1(\text{Logit}\_\hat{Q})_j + \sum_{h=2}^{6} \gamma_h Q_{hj} + e_{ij},$$

$$u_j \sim \text{N}(0, \tau); \qquad e_{ij} \sim \text{N}(0, \sigma^2). \quad (14)$$

The estimate of δ_{V0} was .62, with a standard error of 2.03 and a 95% confidence interval of $(-3.36, 4.60)$. No statistically significant difference in the reading outcome was detected between high-retention school children and low-retention school children, perhaps due to the lack of statistical power given that only 89 children from high-retention schools were identified to be in subpopulation (C).

As mentioned earlier, our strategy to increase power was to reanalyze the retention rate effect for an expanded sample that included 2,610 children promoted in low-retention schools and 199 children promoted in high-retention schools from the first two strata of q_1. The new estimate was .67, with a standard error of 1.53 and a 95% confidence interval of $(-2.33, 3.67)$. The magnitude of the estimated effect of retention rate for low-risk children remained negligible.

Sensitivity Analyses. The analytic results are valid only under the assumption of strongly ignorable treatment assignment given the observed covariates. For the estimate of each causal effect that was significantly different from 0, we examined whether our conclusion would be altered by additional adjustment for unmeasured confounders, the omission of which would create a bias comparable to that of the most important observed covariates (Lin, Psaty, and Kronmal 1998; Rosenbaum 1986, 2002). In our multilevel context, hidden bias may originate from the individual level, the cluster level, or both. Hence we assumed that there might exist a student-level unmeasured composite, U_X, and a school-level unmeasured composite, U_W, comparable to the most important student-level and school-level observed covariates. The impact of the omission of U_X and U_W would depend on their associations with the treatment assignment, represented by $E[U_{W1}] - E[U_{W0}]$ and $E[U_{X1}] - E[U_{X0}]$, and their associations with the outcome, represented by π_W and π_X. Because our original estimate of the retention effect was negative, to test sensitivity, we constructed a worst-case scenario in which the potential confounding effects of U_X and U_W were both positive. By taking into account these hypothetical unmeasured confounders, we created a new estimate $\hat{\delta}^* = \hat{\delta} + \pi_W(E[U_{W1}] - E[U_{W0}]) + \pi_X(E[U_{X1}] - E[U_{X0}])$.

The original estimate, $\hat{\delta}$, was considered insensitive to the violation of strong ignorability if the confidence interval for $\hat{\delta}^*$ was strictly negative.

In studies of treatment effects on student learning, once the lagged outcome variables are controlled, additional covariates typically have comparatively little of the variation in the outcome (Bloom 2005). For this reason, the observed individual-level covariate that demonstrated the second strongest association with the learning outcome—kindergarten teacher's report of a child's approach to learning—provided a plausible reference value for π_X. We applied a similar criterion in choosing a value for π_W. The student-level and school-level covariates that showed the strongest associations with the retention assignment provided the empirical basis for setting the values of $E[U_{X1}] - E[U_{X0}]$ and $E[U_{W1}] - E[U_{W0}]$. With additional adjustment for the hypothetical unmeasured confounders, the 95% confidence interval for the new estimate of δ_{Z1} was $(-6.95, -1.54)$ and did not contain 0 or any positive values. Hence if an unmeasured confounder were to explain away the negative effect of kindergarten retention in high-retention schools, then that variable would need to be a stronger confounder than the strongest measured confounder other than the pretest. In contrast, the 95% confidence interval for the new estimate of δ_{Z0} was $(-2.44, 1.24)$ and did contain 0. Thus an unmeasured confounder comparable to the strongest measured confounder other than the pretest would be sufficient for altering our conclusion about the negative effect of kindergarten retention in low-retention schools.

Kindergarten Retention Effects on Math Learning. The math results showed a similar pattern to that in reading. The estimate of the retention effect under a low retention rate for retainees attending low-retention schools was -4.79 [standard error (SE) = .63]. The retention effect under a high retention rate for retainees attending high-retention schools was estimated to be -5.56 (SE = 1.11). Both effects varied significantly across schools. The estimated effect of a high retention rate versus a low retention rate was .81 (SE = 1.12) for children in subpopulation (C) and 1.01 (SE = 0.71) for the expanded sample of low-risk children, a negligible effect in both cases. Similar to the results in reading, refutation of the conclusion about the negative effect of kindergarten retention in high-retention schools would require an unmeasured confounder stronger than the strongest measured confounder other than the pretest. Our conclusion about the negative effect of kindergarten retention in low-retention schools could be altered by an unmeasured confounder comparable to the strongest measured confounder other than the pretest.

4. CONCLUSION

Our key aim was to relax the SUTVA to investigate new questions about school assignment and peer effects. Under our assumptions, our results indicated that kindergarten retainees attending high-retention schools would have achieved more in reading and math during the treatment year had these children instead been promoted. Similar results held for retainees in low-retention schools. Although we inferred that the retention effect varied significantly across schools, the estimated school-specific retention effects were negative in a great majority of the schools. In addition, the data showed no evidence that children at low risk of repetition benefited from a school's high retention rate. Therefore, we found no empirical support for the kindergarten retention policy. These conclusions are based on the following assumptions.

Specification of Peer Effects. Under our dichotomous characterization of the effect of the ensemble of treatment assignments within a school, the potential outcome $Y_{ij}(\mathbf{z}_j) = Y_{ij}[z_{ij}, v(\mathbf{z}_j)]$, so that the effect of a child's peers on that child's outcomes operates strictly through the scalar function $v(\mathbf{z}_j) = 1$ if the school has a "high" retention rate versus $v(\mathbf{z}_j) = 0$ if it does not. One might explore a more refined scalar function (e.g., the proportion of students retained) or a multidimensional function (e.g., the fraction of disruptive children retained and the mean IQ of those retained). Richer specifications can be generated and tested with the ECLS–K data and with minor modifications of our causal inference framework.

Intact Schools. We considered the data as being generated by a two-stage experiment in which intact schools were first assigned to retention rates and then at-risk students were assigned to be retained within schools. Suppose, however, that parents were able to foresee the school-level or child-level treatments and chose schools accordingly for their children. The assumption of assignment of intact schools to treatments would then become untenable. In this case, information about school choice and mobility would be required in modeling treatment assignment.

No Interference Between Schools. Our model allows for the potential outcomes of each student to be affected by the identities and treatment assignments of children attending the same school. However, we assumed that the identities and treatment assignments of students attending other schools do not affect the potential outcomes of children in a given school. Although plausible, one can construct a scenario in which such an assumption is false. For example, suppose that a retained child attending school A lives next door to a promoted child attending school B, leading to a stigmatization of the retained child. In this case, peer effects would operate through neighborhoods as well as schools and would contradict our assumption of no interference between schools. Regression models for the joint effects of neighborhoods and schools were considered by Raudenbush (1993) and Goldstein (2003), although their models did not identify causal effects.

Strongly Ignorable Treatment Assignment. We assumed that school assignment to high versus low retention rates was ignorable given observed school-level covariates and that student assignment to be retained was ignorable given observed school-level covariates, observed student-level covariates, school retention rates, and school-specific posterior-expected random effects. We found in the ECLS–K dataset comparatively rich measurement of pretreatment covariates. Each of our estimated propensity scores balanced >95% of these pretreatment covariates. Nonetheless, the validity of the strong ignorability assumption is subject to further scrutiny and debate on scientific grounds. A sensitivity analysis suggested that the estimated negative effect of retention in high-retention schools

could be refuted only in the presence of unmeasured confounders stronger than any of the observed covariates other than the pretest.

An additional concern involves the *timing of the treatment*. We assumed that the treatment begins when a child is assigned to be retained or promoted. However, such treatment assignments may be endogenous to the outcomes of the past year's school-level retention rate. For example, kindergartners attending high-retention schools may work harder if they are fearful of being retained. A possible solution would be to use longitudinal data on schools and to model the school-level retention rate as a time-varying treatment. This would require a richer dataset than ECLS–K and a further extension of the causal framework for multilevel, longitudinal data.

Finally, our inference of no effect of the retention rate on the low-risk students required identification of children in subpopulation (C). We achieved this identification by assuming that if the low-retention schools were to adopt a high-retention policy, then the assignment of students to be retained would follow a similar process as in the high-retention schools. This assumption, although seemingly reasonable, cannot be checked with available data.

[Received August 2004. Revised October 2005.]

REFERENCES

Bloom, H. S. (2005), "Randomizing Groups to Evaluate Placed-Based Programs," in *Learning More From Social Experiments: Evolving Analytic Approaches*, ed. H. S. Bloom, New York: Russell Sage Foundation, pp. 115–172.

Ellwein, M. C., and Glass, G. V. (1989), "Ending Social Promotion in Waterford: Appearances and Reality," in *Flunking Grades: Research and Policies on Retention*, eds. L. A. Shepard and M. L. Smith, Philadelphia: Falmer Press, pp. 151–173.

Goldstein, H. (2003), *Multilevel Statistical Models* (3rd ed.), London: Arnold.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991), *Fundamentals of Item Response Theory*, Newbury Park, CA: Sage.

Hauser, R. M. (1998), "Should We End Social Promotion? Truth and Consequences," presented at the Conference of the Harvard Civil Rights Project on Civil Rights and High-Stakes Testing, New York.

Jackson, G. B. (1975), "The Research Evidence on the Effect of Grade Retention," *Review of Educational Research*, 45, 613–635.

Karweit, N. L. (1992), "Retention Policy," in *Encyclopedia of Educational Research*, ed. M. Alkin, New York: Macmillan, pp. 1114–1117.

Leinhardt, G. (1980), "Transition Rooms: Promoting Maturation or Reducing Education?" *Journal of Education Psychology*, 72, 55–61.

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998), "Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies," *Biometrics*, 54, 948–963.

Morrison, F. J., Griffith, E. M., and Alberts, D. M. (1997), "Nature-Nurture in the Classroom: Entrance Age, School Readiness, and Learning in Children," *Developmental Psychology*, 33, 254–262.

Peterson, P. L. (1989), "Alternatives to Student Retention: New Images of the Learner, the Teacher and Classroom Learning," in *Flunking Grades: Research and Policies on Retention*, eds. L. A. Shepard and M. L. Smith, London: Falmer Press, pp. 174–201.

Pfefferman, D., Skinner, C. J., Homes, D. J., Goldstein, H., and Rasbash, J. (1998), "Weighting for Unequal Selection Models in Multilevel Models," *Journal of the Royal Statistical Society*, Ser. B, 60, 23–40.

Plummer, D. L., and Graziano, W. G. (1987), "Impact of Grade Retention on the Social Development of Elementary School Children," *Developmental Psychology*, 23, 267–275.

Raudenbush, S. (1993), "A Crossed Random Effects Model for Unbalanced Data With Applications in Cross-Sectional and Longitudinal Research," *Journal of Educational Statistics*, 18, 321–349.

Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., and du Toit, M. (2005), *HLM6: Hierarchical Linear and Nonlinear Modeling*, Lincolnwood, IL: Scientific Software International.

Reynolds, A. J. (1992), "Grade Retention and School Adjustment: An Explanatory Analysis," *Educational Evaluation and Policy Analysis*, 14, 101–121.

Roderick, M., Bryk, A. S., Jacobs, B. A., Easton, J. Q., and Allensworth, E. (1999), *Ending Social Promotion: Results From the First Two Years*, Chicago: Chicago Consortium on School Research.

Rosenbaum, P. R. (1986), "Dropping Out of High School in the United States: An Observational Study," *Journal of Educational Statistics*, 11, 207–224.

——— (2002), *Observational Studies* (2nd ed.), New York: Springer-Verlag.

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

Rubin, D. B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.

——— (1986), "Comment: Which Ifs Have Causal Answers?" *Journal of the American Statistical Association*, 81, 961–962.

——— (1990), "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies," *Statistical Science*, 5, 472–480.

Shepard, L. A. (1989), "A Review of Research on Kindergarten Retention," in *Flunking Grades: Research and Policies on Retention*, eds. L. A. Shepard and M. L. Smith, London: Falmer Press, pp. 64–78.

Shepard, L. A., and Smith, M. L. (1988), "Escalating Academic Demand in Kindergarten: Counterproductive Policies," *The Elementary School Journal*, 89, 135–145.

Smith, M. L., and Shepard, L. A. (1988), "Kindergarten Readiness and Retention: A Qualitative Study of Teachers' Beliefs and Practices," *American Educational Research Journal*, 25, 307–333.

Tanner, C. K., and Galis, S. A. (1997), "Student Retention: Why Is There a Gap Between the Majority of Research Findings and School Practice?" *Psychology in the Schools*, 34, 107–114.

Verbitsky, N., and Raudenbush, S. W. (2004), "Causal Inference in Spatial Settings," in *Proceedings of the Social Statistics Section*, American Statistical Association, pp. 2369–2374.

Zill, N., Loomis, L. S., and West, J. (1997), *The Elementary School Performance and Adjustment of Children Who Enter Kindergarten Late or Repeat Kindergarten: Findings From National Surveys* (Statistical Analysis Report NCES 98-097), Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.