

Evaluating Kindergarten Retention Policy:  
Causal Inference for Multi-Level Observational Data

Guanglei Hong and Stephen W. Raudenbush  
University of Michigan

**Please Do Not Cite or Circulate  
Without the Authors' Permission**

### Abstract

The purpose of this study is to extend the potential-outcomes causal framework to encompass multi-level observational data. We handle the multiplicity of potential outcomes associated with each treatment for each unit by invoking the exchangeability assumption. We propose a joint multi-level design for addressing a complex policy issue that involves multiple sets of treatments and is relevant to multiple subpopulations. Using the national Early Childhood Longitudinal Study kindergarten cohort data, we apply the extended framework and adapt the propensity score-based methods to an evaluation of the kindergarten retention policy effects on children's cognitive development. The results show no empirical evidence supporting the practice of retaining children in kindergarten.

Keywords: causal inference; multi-level modeling; policy analysis

## 1. Introduction

Educational research has had relatively little to supply that would inform many of the policy decisions. Government policies are swayed more by political needs than by scientific findings. Local educators rely heavily on instincts and routines. As the federal educational legislation and funding shift toward more emphasis on “scientifically-based research” (Eisenhart & Towne, 2003; National Research Council, 2002; Slavin, 2002), in urgent need are causal inference methods for drawing valid information from well-designed policy studies.

Because modern schooling has a fundamentally multi-level character, almost all the inquires about educational policy or program effects would require causal inferences for multi-level data. The standard practices of allocating children to age-based grade levels and placing them in classrooms for group instruction have been part of the “basic grammar” of schools (Tyack & Cuban, 1995). In order to accommodate diversity in developmental status or academic proficiency of age-grouped children without disrupting the existing organizational structure, grade retention has been one of the devices that further structures the learning opportunities in schools and classrooms. In an extreme case, many schools in this country even allow young children to be retained in kindergarten. A review of the past retention studies reveals some confusion in conceptualizing the causal questions and in drawing causal inferences from multi-level observational data (Hong & Raudenbush, 2003). In particular, there has been little systematic evaluation of the effects of the retention policy not only on the cognitive development of children who are at risk of repeating kindergarten, but also on those who themselves are not at risk but can possibly be grouped with the at-risk children.

Although some major advancements in causal inference theories and methods have been made in statistics, econometrics, and biometrics in the past several decades, defining and estimating treatment effects in a multi-level setting remain challenging. Hence, the multi-level structure of schooling is the focus of methodological as well as substantive inquiries in this study. On the methodological aspect, our goals are to extend the potential-outcomes causal framework and to adapt the propensity score-based causal inference methods to multi-level observational data. The substantive goal is to obtain empirical evidence regarding the causal effects of the kindergarten retention policy on children's cognitive development.

This article is organized as follows. Section 2 describes the research context and clarifies the causal questions about the kindergarten retention policy. Section 3 presents the methodological challenges in defining and estimating the causal effects of the kindergarten retention policy. The sample and data are introduced in section 4. In section 5, we propose a modification of the causal inference framework that facilitates the definition and estimation of kindergarten retention policy effects in multi-level settings. In section 6, we apply the propensity score stratification method to the policy evaluation. Section 7 is a summary of the results.

## **2. Causal Questions**

In the current context of the standards-based accountability movement, ending social promotion is a popular slogan in support of grade retention at most grade levels, including the very first year of schooling (Ellwein & Glass, 1989; Hauser, 1998; Roderick, Bryk, Jacobs, Easton, & Allensworth, 1999). 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. Nonetheless, whether or not the retention policy should be implemented at the kindergarten level has been under heated debate.

Many educators and parents believe that children who have not reached an appropriate stage of social or cognitive development need more time to become mature before progressing to the first grade (Smith & Shepard, 1988). Another popular argument is that retention will benefit the students who have not yet mastered the lower-level facts and skills. It is also expected that changing the peer reference group will increase the self-esteem of kindergarten retainees, which in turn may improve their learning results (Plummer & Graziano, 1987). In contrast, some developmental psychologists argue that, in general, grade retention stifles children's cognitive and social development (Morrison, Griffith, & Alberts, 1997) and stigmatizes these children (Jackson, 1975; Shepard, 1989). Others claim that offering resources to correct problems and adapting instruction to children's existing knowledge are more effective than simply having children repeat an unsuccessful experience (Karweit, 1992; Leinhardt, 1980; Peterson, 1989; Reynolds, 1992; Tanner & Galis, 1997).

Previous retention studies have almost exclusively focused on retention as an individual-level treatment for the retainees. Researchers drew immediate implications for school-level retention policies on the basis of analytic results about the individual-level retention effects. However, because kindergarten retention is carried out in a multi-level school system, its effect may not be confined to the individual children who are retained. Rather, advocates argue that introducing the retention policy may change the instructional structure and climate within a school. When low-achieving students are retained in a

lower grade, a more homogeneous classroom may ease the task for teachers in managing instructional activities (Shepard & Smith, 1988). Meanwhile, children will see grade retention as a punishment for poor performance, and will study harder to avoid being retained (Byrnes, 1989). These arguments are based on the assumption that a student's learning outcome can be affected by the treatments that his or her peers are receiving. Hence, the kindergarten retention policy may have effects on both the retainees and the promoted children. The policy question regarding whether to continue the kindergarten retention practice cannot be fully addressed unless we simultaneously consider the effects of both the individual-level retention treatment and the school-level retention policy.

The decision to retain a child in kindergarten is conditioned by whether the school allows any kindergartners to be retained. We identify three options for each student who is at risk of repeating kindergarten: Being retained in a retention school, being promoted in a retention school, and being promoted in a non-retention school. The latter two treatment options are also available to students not at such a risk.

When we have multiple sets of treatments in a multi-level setting, the causal questions are multi-fold. The first question pertains to the average effect of the school-level kindergarten retention policy on the academic learning of all the children, regardless of their risk status. The second is an inquiry into the causal effect of kindergarten retention as an individual-level treatment under the retention policy on children at risk of repeating kindergarten. The third question is whether the kindergarten retention policy has any direct effect on the at-risk children who are not retained.

### 3. Methodological Challenges

Applying Rubin's (1978) model of causal inference, we will define the retention effect for a student at risk of repeating kindergarten as the difference between the potential outcome resulting from retention and the outcome that would be obtained if the same student were promoted. The effect of the kindergarten retention policy on a student's learning outcome will be defined as the difference between the potential outcome obtained in a retention school and the potential outcome that would be obtained if the school restricted kindergarten retention.

For simplicity, Rubin presented his framework under the stable unit treatment value assumption (SUTVA). It assumes 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 what treatments are received by other experimental units (Rubin, 1986). In essence, Rubin's model is a context-free model that is not directly applicable to multi-level data. Due to the multi-level structure of modern schooling, almost every educational treatment is conducted in an organizational setting that bears an impact on one's potential outcomes. This is partly due to the sharing of and competition for resources within and between organizations (Cook, 1977; Emerson, 1962; Pfeffer & Salancik, 1978), and partly due to agent effects in treatment delivery (Lipsky, 1980; Manski & Garfinkel, 1992). It is quite common to observe interference between students within classrooms and treatment enactment variation across teachers and schools.<sup>1</sup> Hence, conceptually there is a distinct set of potential outcomes associated with each treatment for each unit corresponding to all the possible group compositions, agent allocations, and treatment allocations. This poses a unique problem that is yet to be fully addressed in the

causal inference literature. At times, the contexts can be characterized by one or more additional sets of treatments at various levels of a system. Conceptualizing and estimating the effects of multiple causes in a multi-level setting bring additional challenges and promises to the causal inference methodology.

Causal inferences are especially challenging in non-experimental studies, because the treatment assignment may depend on a variety of pretreatment conditions. Nonetheless, with adequate observational data, researchers may strongly assume that the treatment assignment becomes ignorable given the observed pretreatment covariates. In addition, all the pretreatment information needed for balancing the treatment assignment can be summarized by a unidimensional propensity score when the treatment measure is binary. Rosenbaum and Rubin (1983) made a major contribution by proving that the “treatment assignment and the observed covariates are conditionally independent given the propensity score” (p.44). Hence, when strong ignorability holds, statistical adjustment for the propensity score is usually sufficient for removing the selection bias associated with all the observed pretreatment covariates. In the past two decades, researchers have proposed a number of propensity score-based approaches to causal inferences for observational data under the strong ignorability assumption as well as the SUTVA. These include propensity score matching (Rosenbaum, 1989, 2002; Rubin & Thomas, 1992, 1996), stratification (Rosenbaum, 1991; Rosenbaum & Rubin, 1984), covariance adjustment (Rosenbaum & Rubin, 1983; Rubin & Thomas, 2000), and inverse-probability-of-treatment weighting (IPTW) (Robins, 1997, 2000). Whether and how these techniques work in a variety of multi-level settings are yet to be explored.

The methodological questions that we intend to address in the current study include the following:

(1) How shall we define the causal effect in a multi-level setting when assumptions about the stability of potential outcomes do not hold?

(2) How shall we estimate the causal effects of multiple sets of treatments at multiple levels, especially when the effect of one set of treatments is conditioned by another set of treatments? How shall we apply the propensity score-based methods to causal inferences for multi-level observational data?

#### **4. Sample and Data**

We selected data from the Early Childhood Longitudinal Study Kindergarten cohort (ECLS-K), which contains repeated observations of a nationally representative sample of students, their families, teachers, and schools over the kindergarten and first-grade years. Most of the students were observed in the fall and the spring of the kindergarten year and then in the spring of the treatment year. In addition, a random subsample of students was observed in the fall of the treatment year. As illustrated in Figure 1, our analytic sample included 471 kindergarten retainees and 10,255 promoted students in 1,080 retention schools, and 1,117 promoted students in 141 non-retention schools.

-----  
Insert Figure 1 about here  
-----

The outcome variables were reading and math scale scores calibrated via item response theory (IRT). The test scores of each subject obtained from up to four 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 to compare the reading and math achievement, respectively, of students from different grade levels. Table 1 presents the sample size, mean, and standard deviation for each of the four rounds of reading and math assessment scores. On average, the students gained about 13 points in reading and 10 points in mathematics during the treatment year.

-----  
Insert Table 1 about here  
-----

## 5. Extended Framework of Causal Inference

In this section, we propose a modification of Rubin’s causal inference framework by invoking the exchangeability assumption. We develop the potential outcome models, redefine the causal effects, and discuss a hypothetical joint multi-level experimental design appropriate for addressing the causal questions regarding kindergarten retention. We then apply the concept of principal stratification to joint multi-level observational data.

### 5.1 Potential Outcomes and Causal Effects under Exchangeability

*Treatment settings.* When the experimental units are clustered to receive individual-level or cluster-level treatments, the composition of units and agents along with other contextual factors within each cluster constitutes a unique local environment in which the treatment takes place. We call each local environment “a treatment setting.”<sup>2</sup> Some treatment settings may enhance the treatment effect, while in some other settings,

the treatment effect may be weakened by certain adverse factors. Consequently, the value of one's potential outcome associated with each treatment is likely to change as the treatment setting shifts.

Suppose that the causal question is how much an at-risk child will achieve through a year of retention in kindergarten, in comparison with the outcome of a year of study in the first grade. We use  $Z = 1$  to represent the retention treatment, and  $Z = 0$  for the promotion treatment. The respective potential outcomes of child  $i$  in a given treatment setting  $s$  are denoted by  $Y_{is}^{(Zis=1)}$  and  $Y_{is}^{(Zis=0)}$ . Here the treatment setting may consist of one's school environment, class composition, and teacher allocation. The child will demonstrate a stable outcome value only when the treatment and the treatment setting are both given. The causal effect of kindergarten retention for child  $i$  in treatment setting  $s$  can be defined as the difference between the two potential outcomes:

$$\Delta_{is} = Y_{is}^{(Zis=1)} - Y_{is}^{(Zis=0)}. \quad (1)$$

Under this modified framework, the causal question will be relevant not only to a population of individuals, but also to a population of treatment settings for these individuals. In essence, the population average causal effect,  $\delta$ , is the expected value of the setting-specific treatment effect for all the individuals over all the possible treatment settings:

$$\delta = E[E(\Delta_{is} | s)] = E[Y^{(1)}] - E[Y^{(0)}]. \quad (2)$$

*Potential outcome models and exchangeability.* Assuming additivity and linearity, we specify a pair of potential outcome models for child  $i$  in treatment setting  $s$ :

$$\begin{aligned} Y_{is}^{(Z_{is}=0)} &= \gamma + v_s + e_{is}, \\ Y_{is}^{(Z_{is}=1)} &= \gamma + \delta + v_s + \Delta_{vs} + e_{is} + \Delta_{eis}. \end{aligned} \quad (3)$$

In the first model above,  $\gamma$  is the population average potential outcome associated with promotion.  $v_s$  is the incremental effect of treatment setting  $s$ , and  $e_{is}$  is the increment effect of child  $i$  in treatment setting  $s$ . In the second model,  $\delta$  is the population average retention effect,  $\Delta_{vs}$  is the setting-specific increment to the retention effect, and  $\Delta_{eis}$  is the child-specific increment to the retention effect.

If a scientific theory about the causal relationships is elaborate, the theory should specify the aspects of treatment settings that condition the treatment effects, and those specific contextual conditions should be explicitly included in the theoretical model of causal inference. If we have no prior knowledge indicating a clear distinction between the treatment settings, we resort to the Bayesian perspective by assuming that the joint distribution of the incremental effects of one treatment setting is exchangeable with that of any other setting in the population. Similarly, we assume that the joint distribution of the incremental effects of a child is exchangeable with that of any other child in the population of children at risk of repeating kindergarten (Lindley, 1972; Lindley & Smith, 1972). By definition, the setting-specific increments and the child-specific increments are mutually independent.

$$\begin{aligned} \begin{pmatrix} v_s \\ \Delta_{vs} \end{pmatrix} &\sim \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{v00} & \tau_{v01} \\ \tau_{v10} & \tau_{v11} \end{pmatrix} \right); \\ \begin{pmatrix} e_{is} \\ \Delta_{eis} \end{pmatrix} &\sim \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_e^2 & \sigma_{e\Delta} \\ \sigma_{e\Delta} & \sigma_\Delta^2 \end{pmatrix} \right). \end{aligned} \quad (4)$$

By assuming variance homogeneity within each treatment, we leave room for heterogeneity of variance across the treatments. Bryk and Raudenbush (1988) found the presence of heterogeneous variance across the treatment groups in a randomized study a strong indication that the treatments have differential effects on individual units. When this is the case, we modify the exchangeability assumption to one of exchangeability within a subpopulation of individuals or a subpopulation of treatment settings. In our current application, retention schools and non-retention schools are two distinct types of treatment settings for the individual-level treatments. Meanwhile, children at risk of repeating kindergarten and those at no such risk represent two different subpoulations that may respond differently to the kindergarten retention policy.

*Causal effects of kindergarten retention.* We use  $D = 1$  and  $D = 0$  to indicate the kindergarten retention policy and the non-retention policy, respectively. The potential outcome of a child at risk of repeating kindergarten is denoted by  $Y_{AR}$ , while that of a child at no such risk is denoted by  $Y_{NR}$ . If we focus on the causal effect of the kindergarten retention policy only, there is a pair of potential outcomes for child  $i$  attending school  $j$ .

$$\begin{aligned} Y_{ij}^{(Dj=0)} &= \gamma + u_j + e_{ij}; \\ Y_{ij}^{(Dj=1)} &= \gamma + \delta_D + u_j + \Delta_{D,uj} + e_{ij} + \Delta_{D,eij}. \end{aligned} \quad (5)$$

The average effect of the kindergarten retention policy, the policy effect on the at-risk subpopulation, and that on the subpopulation of children not at risk of repetition are defined as the following:

$$\begin{aligned}
\delta_D &= E[Y^{(D=1)} - Y^{(D=0)}]; \\
\delta_{D.AR} &= E[Y_{AR}^{(D=1)} - Y_{AR}^{(D=0)}]; \\
\delta_{D.NR} &= E[Y_{NR}^{(D=1)} - Y_{NR}^{(D=0)}].
\end{aligned} \tag{6}$$

In order to disentangle the effect of the school-level retention policy and that of the individual-level retention treatment for the at-risk subpopulation, we define three potential outcomes corresponding to the three treatment options for at-risk child  $i$  attending school  $j$ :

$$\begin{aligned}
Y_{ij}^{(Z_{ij}=0, D_{j=0})} &= \gamma + u_j + e_{ij}, \\
Y_{ij}^{(Z_{ij}=0, D_{j=1})} &= \gamma + \delta_{D0} + u_j + \Delta_{D0.uj} + e_{ij} + \Delta_{D0.eij}, \\
Y_{ij}^{(Z_{ij}=1, D_{j=1})} &= \gamma + \delta_{Z1} + \delta_{D0} + u_j + \Delta_{Z1.uj} + \Delta_{D0.uj} + e_{ij} + \Delta_{Z1.eij} + \Delta_{D0.eij}.
\end{aligned} \tag{7}$$

From Equation 7, we derive the estimand of the causal effect of kindergarten retention under the retention policy and the estimand of the direct policy effect for children at risk of repetition:

$$\delta_{Z1.AR} = E[Y_{AR}^{(D=1, Z=1)} - Y_{AR}^{(D=1, Z=0)}];$$

$$\delta_{D0.AR} = E[Y_{AR}^{(D=1,Z=0)} - Y_{AR}^{(D=0,Z=0)}]. \quad (8)$$

### 5.2. A Joint Multi-Level Experimental Design

A randomized experiment ensures the ignorability of the treatment assignment, and hence the internal validity of the estimation results. We expect to obtain unbiased estimates of the above causal effects through a joint multi-level randomized experiment, which is a combination of a cluster randomized trial and a multi-site randomized trial. Hypothetically, schools can be assigned at random to the kindergarten retention policy with a probability  $Q$ , where  $Q = \Pr(D = 1)$ . Then within each retention school, the at-risk children can be assigned at random to the kindergarten retention treatment with a probability  $q$ , where  $q = \Pr(Z = 1 | D = 1)$ .

In multi-level experimental designs, a subtle question is whether individuals need to be randomly allocated to clusters prior to the treatment assignment. This will be necessary if the population average causal effect of a treatment is defined as the expected value of all the unit-specific causal effects over all the possible treatment settings. However, due to certain constraints in reality, some of the treatment settings may have little or no chance of occurrence. The average treatment effect over all the theoretically possible settings may be different from the average treatment effect over all the realistic settings. Such a difference will indicate a lack of generalizability of the former. For this reason, we favor a design of sampling intact treatment settings over a design of creating new treatment settings through random allocation of units to clusters.

Note that a joint multi-level experimental design is different from a factorial experimental design. In a joint multi-level design, the probability of the second treatment

assignment,  $Z$ , is endogenous to the first treatment assignment,  $D$ , while in a factorial design, each experimental unit is assigned to a combination of a factor level of  $D$  and a factor level of  $Z$ . If we apply a factorial design to the current example, each at-risk child will be assigned at random to one of three treatments with some known probabilities. Let  $P_{11} = \Pr(Z = 1, D = 1)$ ,  $P_{01} = \Pr(Z = 0, D = 1)$ , and  $P_{00} = \Pr(Z = 0, D = 0) = 1 - P_{11} - P_{01}$ . Note that schools are no longer intact under a factorial design. These two types of designs have different treatment assignment mechanisms, which will have important consequences when the data are non-experimental.

### 5.3 Propensity Scores and Principal Stratification for Joint Multi-Level Observational Data

A joint multi-level observational design represents a reality in which policies selected by intact organizations provide the contexts for individual-level treatments. The latter are then made available to or are sought out by individuals within the organizations. Here we place particular emphasis on the organizations' role in choosing organization-level policies and allocating individual-level treatments. In the current example, once a school has adopted a school-level program, the school composition is assumed intact. In other words, we assume that either there is no individual mobility or such mobility is not associated with the school-level policy.

We use  $X$  to denote observed individual-level pretreatment covariates,  $U_X$  for unobserved individual-level covariates;  $W$  and  $U_W$  denote observed and unobserved cluster-level covariates, including the first two moments of  $X$  and  $U_X$  aggregated to the cluster level. In observational data, a school's propensity of adopting the kindergarten retention policy is a function of the school-level pretreatment covariates only, while a

student's propensity of kindergarten repetition is a function of the student-level pretreatment covariates given which school the student attends and given the school-level policy regarding kindergarten retention. For school  $j$  and student  $i$  attending this school, the respective propensities are:

$$\begin{aligned} Q_j &= \Pr(D_j = 1 \mid W_j, U_{W_j}); \\ q_{ij} &= \Pr(Z_{ij} = 1 \mid D_j = 1, X_{ij}, U_{X_{ij}}, j). \end{aligned} \tag{9}$$

Because the individual-level retention treatment is an intermediate outcome of the school-level retention policy, Frangakis and Rubin's (2002) notions of principal stratification and principal causal effects are relevant here. A basic principal stratification with respect to  $Z$  partitions the population such that, within each stratum, all students have the same potential probability of retention under the retention policy. Under principal stratification,  $D$  and  $Z$  are independent of the potential outcomes. A principal effect is defined as a comparison of the potential outcomes within a principal stratum.

Table 2 displays three types of children who constitute three principal strata. Here  $q$  represents a child's propensity of retention under the kindergarten retention policy, while  $q'$  represents the probability of retention under the non-retention policy. Children in principal stratum 1 will always be retained under the kindergarten retention policy; those in principal stratum 2 will probably be retained under the retention policy; while those in principal stratum 3 will never be retained under the retention policy. Hence, only those in the first two principal strata are at risk of repeating kindergarten. The probability of retention under the non-retention policy is always zero, regardless of which principal

stratum a child belongs to. For children in principal stratum 1, the kindergarten retention effect,  $E[Y^{(Z=1,D=1)} - Y^{(Z=0,D=0)}]$ , is a sum of the retention treatment effect under the retention policy,  $\delta_{ZI.AR}$ , and the direct effect of the retention policy,  $\delta_{D0.AR}$ . We can define  $\delta_{ZI.AR}$  and  $\delta_{D0.AR}$  separately for children in principal stratum 2. Only the average effect of the kindergarten retention policy,  $\delta_{D.NR}$ , is defined for children in principal stratum 3.

The ultimate challenge is that, from the observed data, we can only estimate a child's propensity of retention under the actual policy that the school has adopted. Because the probability of retention is zero for all the children attending non-retention schools, it is impossible to identify the principal stratum membership and to estimate the retention effect for the non-retention school students, unless we make additional assumptions.

## 6. Propensity Score-Based Causal Inferences

In this section, we make a distinction between the causal effects of kindergarten retention that are estimable and those that are not estimable from observational data. For the estimable effects, we illustrate an application of propensity score stratification in combination with covariance adjustment.

### 6.1 Can We Estimate the Average and Differential Effects of the Kindergarten Retention Policy?

*Average policy effects.* The average effects of the kindergarten retention policy are estimable if we can assume that the school-level policy adoption is ignorable given the observed school-level pretreatment covariates. Given the rich information contained in the ECLS-K data, the strong ignorability assumption seems plausible. Every school's

propensity of adopting the kindergarten retention policy was specified as a function of the observed school-level pretreatment covariates,  $W$ :

$$\hat{Q}_j = \Pr(D_j = 1 \mid W_j);$$

$$\text{Logit}(\hat{Q}_j) = \ln\left(\frac{\hat{Q}_j}{1 - \hat{Q}_j}\right) = f(W_j). \quad (10)$$

In the ECLS-K data set, we identified more than two hundred pretreatment covariates that predicted a school's adoption of the kindergarten retention policy. In general, the retention schools appeared to have notable advantages over the non-retention schools. The school-level propensity model included 29 covariates and the quadratic terms for 4 of these 29 variables (see Appendix A for a list of the covariates).<sup>3</sup>

We divided the sample of non-retention schools into seven strata on the basis of the logit of  $\hat{Q}$ , and identified the retention schools within each stratum (see Table 3). No matches were found for the ten non-retention schools that were least likely to adopt the kindergarten retention policy. These ten non-retention schools would not contribute to the estimation of the kindergarten retention policy effect. Hypothesis testing through analyzing a general linear model confirmed that balance was achieved in the distributions of the logit of the propensity score between the retention schools and the non-retention schools in the remaining strata.

We examined the unadjusted mean difference in the reading and the math outcomes between retention school students and non-retention school students. On average, students enrolled in the retention schools scored 4.02 points higher in reading ( $t$

= 9.488,  $p < .001$ ) and 2.74 points higher in mathematics ( $t = 9.771$ ,  $p < .001$ ), when compared with the non-retention school students at the end of the treatment year. Because the retention schools showed some pretreatment advantages over the non-retention schools, we should not view these unadjusted results as evidence in support of the kindergarten retention policy.

We next examined the effects of the kindergarten retention policy on the reading and math outcomes within each stratum. The results are displayed in Tables 4 and 5. On average, students attending retention schools did not seem to be better off in reading and math learning, when compared with their counterparts in the non-retention schools.

Observing no systematic variation in the retention policy effect in each subject area across the strata, we tentatively assumed a constant treatment effect, and adopted a model-based estimation approach as shown in the following equation. We combined propensity score stratification with covariance adjustment for the logit of the propensity score to remove the remaining bias, if there was any, within the strata.

$$Y_{ij} = \gamma + \delta_D D_j + \gamma_2 (\text{dur\_f})_{ij} + \gamma_3 (\text{logit\_}\hat{Q})_j + \sum_{h=1}^6 \beta_h M_{hj} + u_j + e_{ij};$$

$$u_j \sim N(0, \tau); e_{ij} \sim N(0, \sigma^2). \quad (11)$$

For simplicity, variance homogeneity was assumed at both the individual level and the school level, after some preliminary analyses suggested no variance heterogeneity.

Because the last round of the assessment was not conducted on the same date for all the sampled students,  $(\text{dur\_f})_{ij}$  indicates the length of time that student  $i$  in school  $j$  had spent in the treatment since the beginning of the school year. School  $j$ 's estimated logit of

propensity of adopting the kindergarten retention policy is denoted by  $(\text{logit\_}\hat{Q})_j$ . The  $M$  dummy series represent six of the seven school-level propensity strata.

The results are summarized in Table 6. After adjusting for the selection bias at the school level, we found the average effects of the kindergarten retention policy negligible. The estimated policy effects were  $-0.24$  in reading with a standard error of  $0.86$ , and  $-0.14$  in mathematics with a standard error of  $0.55$ .

*Differential policy effects.* In order to identify the children at risk of repeating kindergarten, we estimated an individual-level propensity score for the retention school students. For student  $i$  from pretreatment school  $j$ , the estimated propensity score,  $q_{ij}$ , is a function of pretreatment personal characteristics  $X_{ij}$ , classroom characteristics  $F_{ij}$ , school characteristics  $W_j$ , and the empirical Bayes estimate of the residual random effect of school  $j$ , denoted by  $u_j^*$ .<sup>4</sup>

$$\hat{q}_{ij} = \Pr(Z_{ij} = 1 \mid D_j = 1, X_{ij}, F_{ij}, W_j, u_j^*). \quad (12)$$

The propensity model included 39 predictors and seven quadratic terms (see Appendix B for a list of the predictors). We used the logit of the above propensity score as an index to indicate the extent to which a student was at risk of repeating kindergarten. The students in the retained group showed a minimum value of  $-6.06$  in this index. There were 3,087 promoted children in the retention schools whose logit of the estimated propensity was even lower than this value. Because these children did not have counterparts in the retained group, they were identified as at little or no risk of kindergarten repetition.

The non-retention school students' risk status could not be directly estimated. Nonetheless, we could expect that, had a non-retention school changed its policy, some of their students would have been retained in kindergarten. Moreover, we assumed that the students would have been subject to some similar selection criteria for retention assignment as approximated by the above propensity model. In other words, a student who was at a high risk of kindergarten repetition in a retention school would have found him- or herself in a similar risk status in a comparable non-retention school, should the latter have changed its policy. Having estimated the propensity model from the observed data of the retention school students, we used the same model to predict the propensity score for each student enrolled in a non-retention school. In this way, we identified 820 non-retention school students who would be at risk of repeating kindergarten under the retention policy, and 297 children who would be at little or no risk of repetition.

For descriptive information, we compared the unadjusted average reading and math outcomes between the at-risk students attending retention schools and those attending non-retention schools. The mean differences were 3.85 in reading and 2.87 in mathematics. Between the retention school children and the non-retention school children who were at little or no risk of repeating kindergarten, the mean differences were 3.76 in reading and 1.91 in mathematics. All these differences were statistically significant.

We modified the analytic model represented in Equation 11 by including the binary indicator for a child's risk status as an additional covariate that interacted with the policy indicator. The estimated effects of the kindergarten retention policy on a child at little or no risk of repetition were  $-0.87$  in reading and  $-0.80$  in mathematics. For an at-risk child, the estimated policy effects were  $-0.30$  in reading and  $-0.05$  in mathematics.

None of these estimates was statistically significant (see Table 6). Attending a retention school showed no effect on reading and math learning, regardless of one's risk status.

### 6.2 Can We Estimate the Kindergarten Retention Effect under the Retention Policy for the At-Risk Subpopulation?

*Unestimable causal effect.* The effect of the individual-level retention treatment under the kindergarten retention policy,  $\delta_{ZI,AR}$ , is defined for both retention school students and non-retention school students at risk of repetition:

$$\begin{aligned}\delta_{ZI,AR} &= E[Y_{AR}^{(Z=1,D=1)} - Y_{AR}^{(Z=0,D=1)}] \\ &= E[Y_{AR}^{(Z=1,D=1)} - Y_{AR}^{(Z=0,D=1)} \mid D = 1] \Pr(D = 1) \\ &\quad + E[Y_{AR}^{(Z=1,D=1)} - Y_{AR}^{(Z=0,D=1)} \mid D = 0] \Pr(D = 0).\end{aligned}\tag{13}$$

However, in a non-experimental study, the retention school students may not be comparable to the non-retention school students. We cannot assume that  $E[Y_{AR}^{(Z=1,D=1)} - Y_{AR}^{(Z=0,D=1)} \mid D = 1]$  and  $E[Y_{AR}^{(Z=1,D=1)} - Y_{AR}^{(Z=0,D=1)} \mid D = 0]$  are equal except for children from the same basic principal stratum. Unable to empirically stratify the at-risk children attending non-retention schools on the basis of their potential propensity of retention under the retention policy, we cannot estimate the expected values of their counterfactual outcomes,  $E[Y_{AR}^{(Z=1,D=1)} \mid D = 0]$  and  $E[Y_{AR}^{(Z=0,D=1)} \mid D = 0]$ . Therefore,  $\delta_{ZI,AR}$  is not estimable.

*Estimable causal effect.* Nonetheless, we can estimate the retention effect under the retention policy for at-risk children attending retention schools, denoted by  $\delta_{ZI,AR}^*$ :

$$\delta_{ZL,AR}^* = E[Y_{AR}^{(Z=1,D=1)} - Y_{AR}^{(Z=0,D=1)} | D = 1]. \quad (14)$$

We have empirically estimated the individual-level propensity of retention under the retention policy,  $q$ , for the retention school students (see Equation 12). Under strong ignorability, retention assignment  $Z$  is independent of the potential outcomes,  $Y_{AR}^{(Z=1,D=1)}$  and  $Y_{AR}^{(Z=0,D=1)}$ , for retention school students with the same value of  $\hat{q}$ . Therefore, we expect to obtain an unbiased estimate of  $\delta_{ZL,AR}^*$  through statistical adjustment for  $\hat{q}$ .

We divided the at-risk students in the retention schools into fifteen strata on the basis of the logit of  $\hat{q}$  (see Table 7). The top stratum contained eight students who had no matches in the promoted group. Within each of the remaining fourteen strata, the retained students and the promoted students demonstrated similar distributions of the observed covariates. The result of hypothesis testing showed no statistically significant difference between these two groups over all the strata.

With no statistical adjustment for selection bias, the mean differences between the 471 kindergarten retainees and the 7,168 promoted at-risk students in retention schools were  $-18.51$  in reading and  $-11.06$  in mathematics at the end of the treatment year. Given the pretreatment imbalance between these two groups, these mean differences were likely to be negatively biased.

The within-stratum mean differences in the reading and math outcomes between the retained students and the at-risk promoted students in the retention schools are displayed in Tables 8 and 9. The mean differences in reading ranged from  $-0.96$  to  $-12.14$ , while that in mathematics ranged from  $1.57$  to  $-7.98$ . After a year of treatment, the

kindergarten retainees generally reached a lower achievement status in both reading and mathematics than did their comparable promoted peers.

In order to generate a more conclusive answer to the causal question, we specified a two-level analytic model below. Let  $Y_{ij}$  denote the observed outcome of at-risk child  $i$  in school  $j$ ,

$$Y_{ij} = \gamma + \delta_{ZL,AR}^* Z_{ij} + \gamma_1 (\text{Logit}_{\hat{q}})_{ij} + \sum_{s=2}^{15} \alpha_s L_{sij} + \gamma_2 (\text{dur}_f)_{ij} + u_j + \Delta_{Z,uj} Z_{ij} + e_{ij};$$

$$\begin{pmatrix} u_j \\ \Delta_{Z,uj} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_u & \tau_{u,Zu} \\ \tau_{u,Zu} & \tau_{Zu} \end{pmatrix} \right); e_{ij} \sim N(0, \sigma^2). \quad (15)$$

Here  $L_{sij}$ ,  $s = 2, \dots, 15$ , are dummy indicators for fourteen of the fifteen propensity strata that sub-classify the at-risk students. We made additional adjustment for  $(\text{Logit}_{\hat{q}})_{ij}$ , the logit of a student's estimated propensity of being retained. As before,  $(\text{dur}_f)_{ij}$  indicate a student's duration in the treatment year up to the time of the last round of reading assessment. Each school provided a treatment setting for kindergarten retention. The above model enabled us to estimate the variation of the retention effect across the retention schools, denoted by  $\tau_{Zu}$ , along with the covariance of the retention effect and a school's average performance, denoted by  $\tau_{u,Zu}$ .

The estimation results for the reading outcome are presented in Table 10. An at-risk child promoted in a retention school was expected to achieve 53.99 in reading near the end of the treatment year. If the child was retained, his or her reading achievement was expected to be 9.01 points lower, with a standard error of 0.68 and a 95% confidence interval of (-10.42, -7.76). This estimated kindergarten retention effect on reading in

retention schools amounted to about two-thirds of a standard deviation of the reading outcome. There was a statistically significant variation in the retention effect across the retention schools ( $\hat{\tau}_{Zu} = 18.83$ ,  $\chi^2 = 280.30$ ,  $df = 230$ ,  $p < 0.05$ ). With a normal distribution of the retention effect assumed for the population of retention schools, the retention effect on reading would range from -17.52 to -0.50 among 95% of the retention schools. The correlation between school-specific retention effect and school mean outcome was -0.27, indicating that kindergarten retention had a more severe negative effect in retention schools in which the average reading achievement was higher.

We also investigated the dependence of the kindergarten retention effect on a child's severity of retention risk. For students in the stratum 12 or above, their odds of being retained were estimated to be more than 1.22. Almost 30% of the kindergarten retainees were found in this high-risk category. The rest of the population under consideration was at a moderate risk of repetition. Further analysis showed a statistically significant difference in the kindergarten retention effect on the reading outcome between high-risk children and moderate-risk children (coefficient = 6.95, standard error = 2.33,  $t = 2.98$ ). A moderate-risk child was expected to lose 9.59 points in reading with a standard error of 0.70 if they were retained rather than being promoted. Meanwhile, a high-risk child's expected loss in reading associated with retention was estimated to be only 2.64 with a standard error of 2.19, which was not significantly different from zero.

We reported in Table 11 the estimated effect of kindergarten retention on the math learning of retention school students. If an at-risk child was retained rather than being promoted in a retention school, the math achievement of this child was expected to be 5.89 points lower with a standard error of 0.50. This estimated deficiency, with a 95%

confidence interval of (-6.97, -5.01), was also about two-thirds of a standard deviation of the math outcome. The retention effect on math varied across the retention schools ( $\hat{\tau}_{Zu} = 12.77$ ,  $\chi^2 = 289.14$ ,  $df = 230$ ,  $p < .01$ ), ranging from -12.89 to 1.11 among 95% of the schools. The correlation between school-specific retention effect and school mean outcome was 0.40, indicating that kindergarten retention had a more severe negative effect in retention schools in which the average math achievement was lower.

Further analysis showed a statistically significant difference in the retention effect on math learning between high-risk children and moderate-risk children (coefficient = 3.58, standard error = 1.62,  $t = 2.22$ ). The retention effect for the moderate-risk children was estimated to be -6.20 with a standard error of 0.52; while for the high-risk children, the estimate of the retention effect was -2.62 with a standard error of 1.56. The latter showed no statistical significance.

In general, the kindergarten retention effects for children attending retention schools were negative in both reading and mathematics. Although the retention effect estimates showed a relatively smaller magnitude for high-risk children, these estimates lacked precision. As shown in Tables 8 and 9, when we controlled the student-level propensity scores, there were very few promoted students in retention schools that could be matched to the high-risk retainees. The sparseness of the data hampered our ability to make a certain conclusion about the diminishing retention effects. Nonetheless, there was evidence that the kindergarten retention effects varied across the retention schools. The retention effect on reading tended to be more striking in retention schools with a higher reading performance; while the retention effect on math achievement tended to be more detrimental in schools with a lower performance in mathematics.

*Sensitivity analysis.* We accept the validity of the analytic results only to the extent that the strong ignorability assumption holds. For the estimate of the kindergarten retention effect in retention schools in each subject area, we examined its sensitivity to any possible departure from the strong ignorability assumption. The purpose was to see if the general conclusion based on hypotheses testing would be altered by additional adjustment for some unmeasured confounders comparable to the most important covariates observed (Lin, Psaty, & Kronmal, 1998; Rosenbaum, 1986; Rosenbaum, 2002).

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 in reading and mathematics. Using Equation 16, we computed a new estimate of the retention effect in each subject area,  $\delta_{ZO,AR}^{**}$ , that adjusted for  $U_W$  and  $U_X$  in addition to the adjustment for the estimated propensity score.

$$\delta_{ZO,AR}^{**} = \delta_{ZO,AR}^* - \pi_W (E[U_{W1}] - E[U_{W0}]) - \pi_X (E[U_{X1}] - E[U_{X0}]). \quad (16)$$

We identified, among the school-level covariates, the one that showed the largest standardized mean difference between the treatment groups, and used its unstandardized mean difference as the hypothetical value of  $E[U_{W1}] - E[U_{W0}]$ . The hypothetical value of  $E[U_{X1}] - E[U_{X0}]$  was obtained in a similar way from the student-level covariates. Dependent on the circumstances, the student-level and the school-level covariates that had either the largest or next to the largest standardized regression coefficient in predicting the observed outcome provided the empirical basis for setting the values of  $\pi_W$

and  $\pi_X$ .<sup>5</sup> We then computed the bounds for the estimates of the kindergarten retention effects with additional adjustment for the hypothetical student-level and school-level unmeasured confounders. The results are summarized in Table 12. The 95% confidence intervals for the upper bounds of the re-adjusted retention effect estimates do not contain zero or any positive values. Therefore, we concluded that our estimation results were not sensitive to omission of unmeasured confounders.

### 6.3 Can We Estimate the Direct Effect of the Kindergarten Retention Policy on the At-Risk Subpopulation?

The estimand of the direct effect of the kindergarten retention policy is the expected difference between the potential outcomes associated with promotion in a retention school and that associated with promotion in a non-retention school.

$$\begin{aligned}\delta_{D0,AR} &= E[Y^{(Z=0,D=1)} - Y^{(Z=0,D=0)}] \\ &= E[Y^{(Z=0,D=1)} - Y^{(Z=0,D=0)} \mid D = 1] \Pr(D = 1) \\ &\quad + E[Y^{(Z=0,D=1)} - Y^{(Z=0,D=0)} \mid D = 0] \Pr(D = 0)\end{aligned}\quad (16)$$

In a non-experimental study,  $E[Y^{(Z=0,D=1)} - Y^{(Z=0,D=0)} \mid D = 1]$  and  $E[Y^{(Z=0,D=1)} - Y^{(Z=0,D=0)} \mid D = 0]$  cannot be assumed equal except for children from the same basic principal stratum. Because we are unable to empirically estimate the expected values of the counterfactual outcomes,  $E[Y_{AR}^{(Z=0,D=1)} \mid D = 0]$  and  $E[Y_{AR}^{(Z=0,D=0)} \mid D = 1]$ ,  $\delta_{D0,AR}$  is not estimable.

## 7. Conclusion

We modified Rubin's causal framework by replacing the single potential outcome under a given treatment with a potential outcome model, in which one's potential outcome is a function of the treatment settings at various levels. We have shown that, when SUTVA is relaxed in a multi-level setting, exchangeability becomes a useful assumption for simplifying the multiplicity of potential outcomes. In this way, we take into account interference between units within an organization and treatment enactment variation across agents. Moreover, the exchangeability assumption enables us to specify a distribution of the treatment effect over a population of treatment settings.

We focused on the multi-level situations in which intact organizations choose the organization-level policies that provide the contexts for individual-level treatments. A joint multi-level design approximates the causal problems involving multiple sets of treatments conducted simultaneously at multiple levels of a system. This is true especially when the treatment assignment at one level depends on the treatment assignment at another level. However, the propensity score-based techniques show limitations in analyzing joint multi-level observational data. Due to the counterfactual nature of an individual's potential propensity of receiving the individual-level treatment under the alternative organization-level policy, our ability to implement a principal stratification is constrained.

We have used the kindergarten retention study to illustrate:

(1) how to conceptualize the causal effect in a multi-level setting where there can be a multiplicity of potential outcomes, and how to define the causal effects for multiple sets of treatments in multi-level settings,

(2) how to identify the causal effects that are estimable and the causal effects that are not estimable in a non-experimental study, and

(3) how to use propensity score stratification to adjust for selection bias in multi-level observational data.

Our results suggested that, on average, kindergarten retainees attending retention schools would have achieved a significantly higher level of learning in both reading and mathematics during the treatment year if promoted. Hence, kindergarten retention as an individual-level treatment tends to impede young children's cognitive development. Meanwhile, the kindergarten retention policy made no difference in children's reading and math learning, regardless of their risk status. So far we have found no empirical evidence supporting the practice of routinely using kindergarten retention as a general solution to the difficulties experienced by young children.

### References

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, *104*(3), 396-404.

Byrnes, D. A. (1989). Attitudes of students, parents, and educators toward repeating a grade. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 108-131). Philadelphia: Falmer Press.

Cook, K. S. (1977). Exchange and power in networks of inter-organizational relations. *Sociological Quarterly*, *18*, 62-82.

Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on "scientifically based" educational research. *Educational Researcher*, *32*(7), 31-38.

Ellwein, M. C., & Glass, G. V. (1989). Ending social promotion in Waterford: Appearances and reality. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 151-173). London: The Falmer Press.

Emerson, R. (1962). Power-dependence relations. *American Sociological Review*, *27*, 31-40.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*, 21-29.

Hauser, R. M. (1998). *Should we end social promotion? Truth and consequences*. Paper presented at the Conference of the Harvard Civil Rights Project on Civil Rights and High Stakes Testing, New York.

Hong, G., & Raudenbush, S. W. (2003). *What if these kindergartners were not retained? The effect of kindergarten retention versus promotion on children's literacy*

*growth*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Jackson, G. B. (1975). The research evidence on the effect of grade retention. *Review of Educational Research*, 45(3), 613-635.

Karweit, N. L. (1992). Retention policy. In M. Alkin (Ed.), *Encyclopedia of educational research* (pp. 114-118). New York: Macmillan.

Leinhardt, G. (1980). Transition rooms: Promoting maturation or reducing education? *Journal of Education Psychology*, 72, 55-61.

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.

Lindley, D. V. (1972). *Bayesian Statistics, A Review*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34, 1-41.

Lipsky, M. (1980). *Street-level bureaucracy : dilemmas of the individual in public services*. New York: Russell Sage Foundation.

Manski, C. F., & Garfinkel, I. (1992). *Evaluating welfare and training programs*. Cambridge, MA: Harvard University Press.

Morrison, F. J., Griffith, E. M., & Alberts, D. M. (1997). Nature-nurture in the classroom: Entrance age, school readiness, and learning in children. *Developmental Psychology*, 33(2), 254-262.

National Research Council (2002). *Scientific research in education*. R. Shavelson & L. Towne (Eds.), Committee on Scientific Principles for Educational Research. Washington, DC: National Academy Press.

Pfeffer, J., & Salancik, G. R. (1978). *The external control of organizations: A resource dependence perspective*. New York: Harper & Row.

Plummer, D. L., & Graziano, W. G. (1987). Impact of grade retention on the social development of elementary school children. *Developmental Psychology*, 23(2), 267-275.

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

Reynolds, A. J. (1992). Grade retention and school adjustment: An explanatory analysis. *Educational Evaluation and Policy Analysis*, 14(2), 101-121.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent variable modeling and applications to causality* (pp. 69-117). New York: Springer Verlag.

Robins, J. M. (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.

Roderick, M., Bryk, A. S., Jacobs, B. A., Easton, J. Q., & 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(3), 207-224.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408), 1024-1032.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of Royal Statistical Society, B*, 53(3), 597-610.

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

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387), 516-524.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1), 34-58.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396), 961-962.

Rubin, D. B., & Thomas, N. (1992). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*, 79(4), 797-809.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52, 249-264.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustment for prognostic covariates. *Journal of the American Statistical Association*, 95(450), 573-585.

Shepard, L. A. (1989). A review of Research on kindergarten retention. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 64-78). London: The Falmer Press.

Shepard, L. A., & Smith, M. L. (1988). Escalating academic demand in kindergarten: Counterproductive policies. *The Elementary School Journal*, 89(2), 135-145.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31(7), 15-21.

Smith, M. L., & Shepard, L. A. (1988). Kindergarten readiness and retention: A qualitative study of teachers' beliefs and practices. *American Educational Research Journal*, 25(3), 307-333.

Tanner, C. K., & 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(2), 107-114.

Tyack, D., & Cuban, L. (1995). *Tinkering toward Utopia*. Cambridge, MA: Harvard University Press.

Zill, N., Loomis, L. S., & 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.

## Appendix A

### Propensity Model Predictors of Kindergarten Repetition Policy

1. Suburban school
2. Public school
3. Spring, treatment year more than 50% minority students
4. Spring, K school welcoming parents to observe
5. Spring, K school offering orientation programs for parents
6. Spring, K school offering adult literacy or basic education for parents
7. Spring, K school having adequate computer lab
8. Spring, K school having third grade
9. Spring, K children with disability not served
10. Spring, K percentage of white teachers
11. Spring, K principal evaluated by raising performance
12. Spring, K school's emphasis on fine and gross motor skills
13. Spring, K school being successful in teaching language and number skills
14. Spring, K school administrator's experience of teaching ESL program
15. Spring, K school administrator's highest level of education
16. Spring, K school installed metal detectors
17. Spring, K fire alarms observed in school
18. Spring, K number of fighting children observed in school
19. Spring, K classroom order observed in school
20. Spring, K people observed congregating near school
21. Number of current students classified as gifted/talented in the pretreatment year

22. Continual teaching experience of pretreatment teachers for the current students
23. ESL teaching experience of pretreatment teachers for the current students
24. Current students' reading language arts curriculum received in the pretreatment year
25. Spring, treatment year principal having taken courses in early childhood education
26. Current students' aggregated reading IRT scale score in spring, K
27. Current students' aggregated measure of activity in structured play in spring, K
28. Current students' aggregated class size in fall, K
29. Current students' aggregated measure of school outreach to parents in fall, K

**Appendix B**

## Propensity Model Predictors of Kindergarten Retention in Retention Schools

1. Spring, K reading scale score and its quadratic term
2. Fall, K teacher report of kid's approaches
3. Fall, K teacher report of kid's self control
4. Spring, K teacher report of kid's math score
5. Spring, K teacher report of kid's interpersonal skills and its quadratic term
6. Spring, K teacher report of kid's falling behind due to health
7. Spring, K teacher report of kid's receiving Title I service in reading
8. Spring, K teacher rating of kid's language skills
9. Spring, K teacher rating of kid's science/social skills and its quadratic term
10. Spring, K teacher report of parents' coming for informal meetings
11. Spring, K teacher report of kid's assignment to the lowest reading group
12. Fall, K parent report of kid's self control
13. Spring, K parent paid tuition
14. Kid's age at kindergarten entry
15. Fall, K % Hispanics in class and its quadratic term
16. Fall, K % boys in class
17. Fall, K class size and its quadratic term
18. Fall, K number of retainees in class
19. Fall, K instructional hours per day
20. Fall, K teacher's preschool teaching experience
21. Spring, K teacher's number of conferences with parents

22. Spring, K class no use of computer equipments
23. Spring, K teacher visited homes before year beginning
24. Spring, K teacher gave parents orientation at year beginning
25. Spring, K math content coverage and its quadratic term
26. Spring, K school's frequency of fundraising
27. Spring, K school had kindergarten level
28. Spring, K school had grade 6 and beyond
29. Spring, K school number of ESL/bilingual teachers
30. Spring, K teacher lowest annual base salary
31. Spring, K teacher highest annual base salary
32. Spring, K developing language and number skills as school goal and its quadratic term
33. Spring, K principal evaluated by support for staff
34. Spring, K principal's experience in teaching pre-K/Headstart and kindergarten
35. Spring, K principal had courses in child development
36. Spring, K school had problem with gangs
37. Spring, K children with weapons in school
38. Spring, K school security measures observed
39. Spring, K decorated hallways in school

### **Author Note**

The authors have received substantial benefits from the stimulating exchanges with Susan Murphy, Ben Hensen, Natalya Verbitsky, and participants in the Hierarchical Linear Modeling class at the University of Michigan in Fall Term 2003, including Daniel Almirall, Andres Martinez, and many others.

This research was based on the first author's dissertation, supported by a grant from the American Educational Research Association which receives funds for its "AERA 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 NSF 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. The University of Michigan International Institute provided a doctoral fellowship during the preparation of the dissertation proposal in Fall Term 2002. Financial support also came from the Study of Instructional Improvement Project. Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies.

### Notes

<sup>1</sup>Rubin (1978) argued that different versions of a treatment should be considered as different treatments. This is important advice to follow when there are qualitative or quantifiable distinctions between treatment versions under the same treatment label. However, in observational studies of some social practices, the number of versions of a treatment can be as many as the number of agents who deliver it. If such variation cannot be readily identified and perhaps scaled along certain dimensions, the convention has been to leave it in the error term.

<sup>2</sup>In a broad sense, if cluster composition, treatment assignment, and agent allocation in some neighboring clusters show direct or indirect impact on one's potential outcomes, those factors should be included in defining one's treatment setting. For the purpose of simplicity, we assume independence between clusters. Hence, we restrict the attention to a unit's immediate surrounding, that is, the cluster in which the unit is located.

<sup>3</sup>In this application, we viewed the kindergarten retention policy and the retention treatment as time-invariant treatments. In the ECLS-K sample, principals reported on their schools' retention policies in two consecutive years. About 20% of the schools changed their policies between these two years. Because it was ambiguous when the policy became operational for the cohort of students under study, we chose to focus on the current year's policy. All the student characteristics and school characteristics measured in the previous year, except for the previous year's retention policy, were pretreatment covariates for a school's current policy. If, instead, the retention decisions were made under the previous year's policy, we would need to conceptualize the causal

problems differently, and would have a time-varying joint multi-level design. The previous year's policy would predict a student's learning outcome at the end of the previous year, both variables would predict the student's retention treatment in the current year, and all these three variables, along with the current year's policy, would predict the student's learning outcome at the end of the current year.

<sup>4</sup>Ideally, the propensity model that predicts  $Z$  should include the fixed effects of all the pretreatment schools represented by dummy indicators, along with the student-level pretreatment covariates. Having saturated the propensity model at the school level, we expect to remove all the selection bias associated with the pretreatment schools. However, because kindergarten retention was a rare event in most schools, in the ECLS-K sample, the number of retainees was smaller than the number of pretreatment schools. Hence, there were not enough degrees of freedom to estimate the fixed effect of each school. An alternative way of controlling the school effects would be through within-school matching. However, a child was selected to repeat the kindergarten year, usually because he or she demonstrated more learning difficulty or behavioral problems, in comparison with other students in the same school. Therefore, among those who attended the same school and were considered in no need of the retention treatment, we could hardly find an exact match for the retainee. The subsequent estimation of the within-school retention effect would rely heavily on linear extrapolation. Moreover, if we restricted to within-school matching, we would make no use of many potential good matches from other schools. Our strategy was to view school  $j$  as an element randomly drawn from a population of pretreatment schools given the observed school-level characteristics,  $W$ . Statistical adjustment for a comprehensive list of  $W$  would capture

most of the school-level selection bias; while additional adjustment for the residual random effect of each pretreatment school would help remove the remaining school-related bias. For a discussion of the empirical Bayes estimate of the random effect, see Raudenbush and Bryk (2002).

<sup>5</sup>In studies of treatment effects on student learning, once the pretest score has been adjusted for, it is usually impossible to find an observed or unobserved pretreatment covariate that can rival the pretest score. In this circumstance, we chose the observed covariate that demonstrated the second largest absolute standardized regression coefficient in predicting the learning outcome.

Table 1

Descriptive Statistics of the Four Rounds of Assessment

## A. Reading

<i>Reading IRT scale score</i>	<i>n</i>	<i>Mean</i>	<i>SD</i>
Fall, Kindergarten year	12,645	23.36	8.76
Spring, Kindergarten year	13,023	33.56	10.95
Fall, treatment year	4,024	39.33	12.91
Spring, treatment year	13,442	56.30	13.64

## B. Mathematics

<i>Math IRT scale score</i>	<i>n</i>	<i>Mean</i>	<i>SD</i>
Fall, Kindergarten year	13,121	20.12	7.24
Spring, Kindergarten year	13,268	28.30	8.71
Fall, treatment year	4,068	33.43	9.46
Spring, treatment year	13,440	43.78	9.00

Table 2

Principal Stratification and Potential Outcomes

Principal Stratum	Post-treatment probability of kindergarten repetition		Potential outcomes	
	$D = 1$	$D = 0$	$D = 1$	$D = 0$
1	$q = 1$	$q' = 0$	$Y^{(Z=1,D=1)}$	$Y^{(Z=0,D=0)}$
2	$0 < q < 1$	$q' = 0$	$Y^{(Z=1,D=1)}, Y^{(Z=0,D=1)}$	$Y^{(Z=0,D=0)}$
3	$q = 0$	$q' = 0$	$Y^{(Z=0,D=1)}$	$Y^{(Z=0,D=0)}$

Table 3

Balance of the Logit of the Propensity Score for the Kindergarten Retention Policy

Stratum	Retention Schools			Non-retention Schools		
	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>
<i>M</i> = 0	---	---	---	10	-2.34	0.92
<i>M</i> = 1	11	-0.38	0.25	18	-0.57	0.36
<i>M</i> = 2	14	0.12	0.08	14	0.04	0.12
<i>M</i> = 3	113	0.93	0.32	43	0.93	0.35
<i>M</i> = 4	162	1.80	0.23	28	1.81	0.27
<i>M</i> = 5	156	2.45	0.16	14	2.42	0.21
<i>M</i> = 6	624	4.27	2.10	14	3.51	0.68

Table 4

Within-stratum Mean Difference in Reading Achievement between Retention School and Non-retention School Students

Stratum	Retention			Non-retention			Mean Diff
	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>	
0	0	---	---	154	47.77	13.47	---
1	97	47.88	14.42	153	51.87	10.90	-4.00
2	123	54.73	12.55	73	49.77	14.49	4.96
3	778	51.43	13.53	286	53.12	13.84	-1.69
4	1350	54.19	13.49	201	52.20	13.48	1.99
5	1422	56.19	13.43	109	55.53	12.83	0.65
6	6910	57.98	13.21	141	57.65	13.06	0.33

Table 5

Within-stratum Mean Difference in Math Achievement between Retention School and Non-retention School Students

Stratum	Retention			Non-retention			Mean Diff
	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>	
0	0	---	---	154	36.70	8.51	---
1	97	37.99	9.42	153	41.82	8.30	-3.83
2	123	42.76	8.21	73	38.98	8.90	3.78
3	778	40.30	9.01	286	41.23	9.06	-0.93
4	1349	42.21	9.02	201	41.27	9.02	0.94
5	1421	44.00	8.99	109	43.40	8.29	0.60
6	6910	44.93	8.65	141	45.50	8.85	-0.57

Table 6

Kindergarten Retention Policy Effects

	Reading		Mathematics	
	<i>Coefficient</i>	<i>95% C.I.</i>	<i>Coefficient</i>	<i>95% C.I.</i>
Average policy effect	-0.24	(-1.93, 1.45)	-0.14	(-1.22, 0.94)
Differential policy effect				
At-risk children	-0.30	(-2.06, 1.46)	-0.05	(-1.11, 1.01)
Children at little risk	-0.87	(-3.05, 1.31)	-0.80	(-2.21, 0.61)

Table 7

Balance of the Logit of the Propensity Score for Kindergarten Retention in RetentionSchools

Stratum	Retention Policy					
	Retained			Promoted		
	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>
<i>L</i> = 1	9	-5.42	0.40	3,054	-5.38	0.39
<i>L</i> = 2	12	-4.25	0.24	1,661	-4.27	0.25
<i>L</i> = 3	14	-3.42	0.20	983	-3.50	0.20
<i>L</i> = 4	12	-2.98	0.10	323	-2.97	0.10
<i>L</i> = 5	24	-2.57	0.17	441	-2.55	0.16
<i>L</i> = 6	23	-2.15	0.07	154	-2.14	0.07
<i>L</i> = 7	47	-1.75	0.15	213	-1.77	0.15
<i>L</i> = 8	48	-1.24	0.14	143	-1.27	0.14
<i>L</i> = 9	46	-0.83	0.11	85	-0.86	0.10
<i>L</i> = 10	48	-0.46	0.12	49	-0.47	0.11
<i>L</i> = 11	47	-0.01	0.11	35	-0.04	0.15
<i>L</i> = 12	49	0.53	0.22	16	0.56	0.17
<i>L</i> = 13	45	1.16	0.20	8	1.14	0.17
<i>L</i> = 14	39	1.99	0.33	3	1.80	0.17
<i>L</i> = 15	8	3.70	1.09	0	---	---

Table 8

Within-Stratum Mean Difference in Reading in Retention Schools

	<i>Z</i> = 1, <i>D</i> = 1			<i>Z</i> = 0, <i>D</i> = 1			<i>Mean Diff</i>
	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>	
<i>L</i> = 1	9	47.41	11.73	3044	59.22	11.64	-11.80
<i>L</i> = 2	12	47.25	15.50	1654	55.62	11.30	-8.37
<i>L</i> = 3	14	42.66	15.92	977	52.49	11.95	-9.83
<i>L</i> = 4	12	45.24	17.32	321	52.45	12.53	-7.21
<i>L</i> = 5	24	44.07	14.43	440	50.05	12.15	-5.97
<i>L</i> = 6	23	36.63	13.88	153	48.77	12.67	-12.14
<i>L</i> = 7	47	37.41	10.92	211	45.89	12.40	-8.48
<i>L</i> = 8	48	39.79	15.37	143	46.00	13.37	-6.21
<i>L</i> = 9	45	37.78	12.27	85	43.28	11.52	-5.50
<i>L</i> = 10	48	35.00	9.77	49	42.18	13.34	-7.18
<i>L</i> = 11	47	34.30	9.05	35	40.45	13.08	-6.16
<i>L</i> = 12	49	32.06	9.31	16	33.03	14.33	-0.96
<i>L</i> = 13	45	31.09	11.96	8	35.27	8.25	-4.18
<i>L</i> = 14	38	33.87	9.15	3	40.47	8.60	-6.60
<i>L</i> = 15	8	32.84	7.71	0	---	---	---

Table 9

Within-Stratum Mean Difference in Reading in Retention Schools

	<i>Z</i> = 1, <i>D</i> = 1			<i>Z</i> = 0, <i>D</i> = 1			<i>Mean Diff</i>
	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>	
<i>L</i> = 1	9	38.88	9.22	3044	45.56	7.58	-6.67
<i>L</i> = 2	12	40.87	11.14	1654	43.48	7.70	-2.61
<i>L</i> = 3	14	35.12	12.65	976	41.59	8.48	-6.48
<i>L</i> = 4	12	35.62	12.68	321	41.77	8.50	-6.15
<i>L</i> = 5	24	35.64	9.20	440	39.42	8.44	-3.78
<i>L</i> = 6	23	31.05	9.70	153	39.02	9.41	-7.98
<i>L</i> = 7	47	33.37	9.44	211	37.19	9.70	-3.81
<i>L</i> = 8	48	34.49	10.43	143	36.97	10.09	-2.48
<i>L</i> = 9	45	34.98	10.95	85	34.43	9.17	0.55
<i>L</i> = 10	48	30.76	8.07	49	34.56	10.37	-3.80
<i>L</i> = 11	47	30.97	9.56	35	35.63	8.98	-4.66
<i>L</i> = 12	49	28.34	9.53	16	26.77	12.71	1.57
<i>L</i> = 13	45	27.04	10.24	8	30.96	8.22	-3.92
<i>L</i> = 14	38	30.10	8.58	3	37.45	5.66	-7.36
<i>L</i> = 15	8	27.49	11.53	0	---	---	---

Table 10

Kindergarten Retention Effect on Reading Achievement in Retention Schools

Fixed Effect	<i>Coefficient</i>	<i>SE</i>	<i>t</i>
Retention school promoted at-risk kid intercept, $\gamma$	53.99	0.28	192.15
Retention effect in retention schools, $\delta_{ZLAR}^*$	-9.01	0.68	-13.27
Student logit of propensity score, $\gamma_l$	-3.77	0.39	-9.67
Student propensity stratum 2, $\alpha_2$	-0.35	0.52	-0.67
Student propensity stratum 3, $\alpha_3$	-0.73	0.81	-0.90
Student propensity stratum 4, $\alpha_4$	0.23	1.09	0.21
Student propensity stratum 5, $\alpha_5$	0.15	1.21	0.13
Student propensity stratum 6, $\alpha_6$	-0.87	1.45	-0.59
Student propensity stratum 7, $\alpha_7$	-1.29	1.54	-0.84
Student propensity stratum 8, $\alpha_8$	-0.59	1.77	-0.34
Student propensity stratum 9, $\alpha_9$	-0.02	1.98	-0.01
Student propensity stratum 10, $\alpha_{10}$	-0.89	2.20	-0.40
Student propensity stratum 11, $\alpha_{11}$	-1.19	2.41	-0.49
Student propensity stratum 12, $\alpha_{12}$	-0.58	2.67	-0.22
Student propensity stratum 13, $\alpha_{13}$	1.52	2.97	0.51
Student propensity stratum 14, $\alpha_{14}$	6.36	3.36	1.89
Student propensity stratum 15, $\alpha_{15}$	12.44	5.35	2.32
Student duration in treatment year, $\gamma_3$	0.66	0.44	1.51

(table continues)

Random effect	<i>Variance</i>	<i>df</i>	$\chi^2$	<i>p value</i>
School mean, $u_j$	55.02	230	1863.46	.000
School retention effect, $\Delta_{z.uj}$	18.83	230	280.30	.013
Correlation between school mean and school retention effect			-0.27	
Level-1 effect, $e_{ij}$	88.22			

Table 11

Kindergarten Retention Effect on Mathematics Achievement in Retention Schools

Fixed Effect	<i>Coefficient</i>	<i>SE</i>	<i>t</i>
Retention school promoted at-risk kid intercept, $\gamma$	42.32	0.19	227.95
Retention effect in retention schools, $\delta_{ZLAR}^*$	-5.89	0.50	-11.79
Student logit of propensity score, $\gamma_I$	-2.53	0.28	-9.18
Student propensity stratum 2, $\alpha_2$	-0.00	0.37	-0.01
Student propensity stratum 3, $\alpha_3$	0.01	0.57	0.02
Student propensity stratum 4, $\alpha_4$	0.81	0.77	1.06
Student propensity stratum 5, $\alpha_5$	0.02	0.85	0.03
Student propensity stratum 6, $\alpha_6$	-0.40	1.04	-0.39
Student propensity stratum 7, $\alpha_7$	-0.58	1.09	-0.53
Student propensity stratum 8, $\alpha_8$	0.19	1.25	0.15
Student propensity stratum 9, $\alpha_9$	0.16	1.40	0.11
Student propensity stratum 10, $\alpha_{10}$	-0.56	1.56	-0.36
Student propensity stratum 11, $\alpha_{11}$	0.60	1.70	0.35
Student propensity stratum 12, $\alpha_{12}$	-0.85	1.89	-0.45
Student propensity stratum 13, $\alpha_{13}$	-0.16	2.11	-0.07
Student propensity stratum 14, $\alpha_{14}$	3.66	2.40	1.53
Student propensity stratum 15, $\alpha_{15}$	7.91	3.95	2.00
Student duration in treatment year, $\gamma_3$	0.29	0.30	0.97

(table continues)

Random effect	<i>Variance</i>	<i>df</i>	$\chi^2$	<i>p value</i>
School mean, $u_j$	23.32	230	1408.06	.000
School retention effect, $\Delta_{z.uj}$	12.77	230	289.14	.005
Correlation between school mean and school retention effect			0.40	
Level-1 effect, $e_{ij}$	44.08			

Table 12

Sensitivity Analyses for Stratification Estimates of Kindergarten Retention Effects inRetention Schools

	$E[U_{X1}-U_{X0}]$	$\pi_X$	$E[U_{W1}-U_{W0}]$	$\pi_W$	$\delta_{Z0}^{**}$	95% CI of the upper bound
<i>Reading</i>						
	-1.02	1.95	-0.06	2.03	(-11.12, -6.90)	(-8.23, -5.57)
	-1.02	1.95	0.07	0.64	(-11.04, -6.98)	(-8.31, -5.65)
<i>Math</i>						
	-1.02	0.83	-0.06	2.04	(-6.86, -4.92)	(-5.90, -3.94)
	-0.51	1.55	-0.05	-0.70	(-6.72, -5.06)	(-6.04, -4.08)

Figure 1. ECLS-K Sample of Students and Schools in Each Treatment Group

