[Working Paper: Please Do Not Cite or Circulate Without the Authors' Permission]

Reading Instruction Time and Homogeneous Grouping in Kindergarten: An Application of the Marginal Mean Weighting Method

Guanglei Hong

Yihua Hong

Ontario Institute for Studies in Education of the University of Toronto

Authors' Note

Guanglei Hong is Assistant Professor, Ontario Institute for Studies in Education of the University of Toronto, Toronto, Ontario, Canada M5S 1V6 (email: ghong@oise.utoronto.ca); Yihua Hong is a research assistant and Ph.D. student, Ontario Institute for Studies in Education of the University of Toronto, Toronto, Ontario, Canada M5S 1V6 (email: yihuahong@oise.utoronto.ca). This research received support from a Social Sciences and Humanities Research Council of Canada Standard Research Grant, a National Academy of Education/Spencer Postdoctoral Fellowship awarded to the first author, and a major research grant entitled "Improving Research on Instruction**:** Models, Designs, and Analytic Methods" funded by the Spencer Foundation. Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies. The authors thank Carl Corter, Janette Pelletier, Stephen Raudenbush, Doug Willms, and Bing Yu for their generous comments and assistance.

Abstract

A kindergartner's opportunities to develop reading skills are likely constrained by the amount of time allocated to reading instruction. In the meantime, the student's engagement in learning tasks will likely increase if the instruction has been adapted to his or her prior ability often through homogeneous grouping. This study investigates whether the grouping effects on kindergartners' reading growth depend on the amount of reading instruction time and the intensity of grouping. To answer our research questions requires causal inferences about concurrent multi-valued instructional treatments. We develop a procedure of applying the marginal mean weighting method to multi-level educational data for this purpose. Results from the Early Childhood Longitudinal Study Kindergarten cohort (ECLS-K) data set lend support to our theoretical hypothesis that, when teachers allocate a substantial amount of time to reading instruction, homogeneous grouping helps kindergartners to gain more in reading. We find no effect of homogeneous grouping when the total amount of reading time is limited.

Grouping students on the basis of their ability has been a controversial issue puzzling elementary reading teachers for decades. Many believe that dividing a class into a number of homogeneous ability groups provides the teacher with more flexibility in adjusting learning objectives and the pace of instruction, and therefore may help the teacher to address a broad range of student needs within a single classroom (Barr & Dreeben, 1983; Gamoran, Nystrand, Berends, & LePore, 1995; Hallinan & Sorensen, 1983; Sorensen & Hallinan, 1986). Others reason that within-class grouping reduces opportunities for the teacher to interact with each group, reduces opportunities for peer interactions across ability levels, and may promote unequal learning opportunities for different groups (Macintyre & Ireson, 2002; Oakes, Gamoran, & Page, 1992; Tach & Farkas, 2006). Extensive reviews and meta-analyses have generated a collection of inconsistent results ranging from negative to positive estimates of the overall effects of within-class homogeneous grouping versus no such grouping (Dawson, 1987; Kulik & Kulik, 1987; Lou, Abrami, Spence, Poulsen, Chambers, & d'Apollonia, 1996; Lou, Abrami, & Spence, 2000; Slavin, 1987). As pointed out by Loveless (1998), the null hypothesis of a zero effect of ability grouping on average would likely be maintained when the treatment has offsetting negative and positive effects across different conditions. In our view, much is yet to be learned about specific instructional conditions under which homogeneous grouping might exert a positive or negative effect on student learning.

Past research has rarely considered the amount of time allocated to reading instruction as a primary condition when examining the grouping effects. This is an important oversight especially when the research focus is at the kindergarten level where we often observe vast differences in reading instruction time across programs. We reason that, in classes that have allocated only a limited amount of time to reading and in the meantime intensively use

homogeneous grouping, a relatively large proportion of the time is perhaps spent on transitioning between different instructional activities as the teacher moves from one group to another, with even less time left for instruction. Hence, the time constraint will likely offset the potential benefit of homogeneous grouping. According to this reasoning, student learning through sustained engagement is more likely to occur when a substantial amount of time has been allocated to reading instruction, and especially when the learning tasks have been tailored to their ability levels. To empirically examine the above reasoning will require studying the interaction effect between reading instruction time and the intensity of homogeneous grouping, the latter defined as the proportion of reading instruction time that students spend in homogeneous groups.

The lack of research evidence on these issues is related to the methodological challenges in studying the causal effects of multiple treatments with non-experimental data. We develop and illustrate a causal inference procedure that applies the marginal mean weighting (MMW) strategy to an evaluation of the effects of concurrent instructional treatments on student learning. To be specific, we compare the population average reading growth of kindergartners across six treatment conditions: no grouping, low-intensity grouping, or high-intensity grouping when a class spends a substantial amount of time on reading instruction; and no grouping, low-intensity grouping, or high-intensity grouping when a limited amount of time is allocated to reading instruction. Our results suggest a positive effect of homogeneous grouping with abundant time on reading instruction. The effect disappears when reading instruction time is limited.

Theoretical Framework

Past Research on Homogeneous Grouping in Elementary Reading Instruction

As a conventional way of differentiating curriculum and instruction in order to accommodate the diverse needs of students, within-class homogeneous grouping has been

examined, challenged, debated, and re-examined for many years (Biemiller, 1993; Hallinan, 2003; Ireson & Hallam, 1999; Loveless, 1998; Sorensen and Hallinan, 1986). In particular, in the 1980s and 1990s, ability grouping in elementary schools and tracking in secondary schools received strong criticisms on ethical and legal grounds (Gamoran, 1992; Oakes, 1985, 1995; Page, 1987; Rosenbaum, 1980; Trimble & Sinclair, 1987; Welner & Oakes, 1996; Wheelock, 1992; Zirkel & Gluckman, 1995). However, past research has not supplied strong empirical evidence in support of a policy of banning within-class homogeneous grouping in the early elementary grades. Below we summarize the findings from the previous literature and examine the existing basis for drawing causal conclusions.

The overall effect of homogeneous grouping, when averaged over a large number of studies, seems to be either slightly positive or negligible. However, among the numerous studies selected and synthesized in the most influential meta-analyses of ability grouping, only a small portion focused on early reading instruction. According to Slavin (1987), dividing a class into two or three ability groups was so widespread in elementary reading instruction prior to the mid-1980s that it was difficult to find non-grouped classes for experimental or quasi-experimental comparisons. Kulik and Kulik (1987) located 19 studies of within-class grouping programs. Only two of these studies examined elementary reading programs for all students, one showing an effect size of 0.29 standard deviations favoring within-class grouping over whole-class instruction, and the other showing an effect size of -0.08. A meta-analysis by Lou, Abrami, Spence, Poulsen, Chambers, and d'Apollonia (1996) reported an average effect size of 0.08 of within-class grouping at the early elementary level, with a larger effect in math than in reading. More recently, McCoach, O'Connell, and Levitt (2006) analyzed the Early Childhood

Longitudinal Study Kindergarten Cohort (ECLS-K) data and reported a higher average reading gain over the kindergarten year in schools with a higher frequency of using ability groups.

The causal validity of the findings from many individual studies has been a major concern. According to Lou and her colleagues' report in their meta-analyses (Lou et al, 1996; Lou, Abrami, & Spence, 2000), as the methodological adequacy of the studies increased, the effect sizes decreased. In most non-experimental studies, statistical adjustment for a very limited number of pretreatment covariates could hardly be relied upon to remove selection bias. In the meantime, controlling for post-treatment covariates that could have been a result of the treatment may introduce additional bias to the treatment evaluation (Rosenbaum, 1987). McCoach et al's (2006) study provided a typical example in which the researchers made linear adjustments for a small number of child-level and school-level covariates when analyzing the effect of within-class grouping. Among the covariates being adjusted for was a summary measure of the principal's perception of kindergarten teachers' success in achieving important educational goals at the end of the year — including challenging high-achieving children, helping low-achieving children, and raising average performance, which arguably could have reflected the impact of the grouping practice. In general, due to the lack of solid empirical evidence, homogeneous grouping has remained controversial.

Instructional Time and Intensity of Grouping

Kindergarten instruction has traditionally prepared children for school life through engaging them in a variety of developmentally appropriate activities (Kilpatrick, 1916). Under the pressure of accountability in the US since the 1980s, schools have increasingly pushed the structured academic curriculum down to the kindergarten level (Jeynes, 2006; Meisels, 1989; Shepard & Smith, 1988). Nonetheless, kindergarten teachers vary in the amount of emphasis

they place on reading instruction. Some spend only a minimal amount of time teaching reading. Others allocate a major block of time to reading instruction everyday. In addition, they take different approaches to accommodating students entering kindergarten with diverse reading-related abilities. Some insist on providing uniform learning opportunities to all students through whole-class instruction; others choose to adopt the convention of placing students in homogeneous ability groups; still others switch flexibly between homogeneous grouping and alternative ways of organizing instruction such as peer tutoring or center-based individual activities (Wiggins, 1994).

In probing the mechanisms through which grouping may influence student learning, some researchers have emphasized the importance of considering the context and quality of instruction (Barr & Dreeben, 1983; Dreeben & Barr, 1988; Gamoran, 1987; Hiebert, 1983). We argue that *reading instruction time* is a primary condition to consider among various contextual factors. This is because time for instruction constrains the learning opportunities available to all students (Anderson, 1984; Bloom, 1974; Carroll, 1963; Millot, 1995). We can divide the total instructional time into academic engaged time and managerial or disengaged time. It is a well-known fact that, when the allocated time is given, the proportion of time in which a student is actively engaged in learning tasks of appropriate difficulty has a more direct impact on his or her learning outcomes (Fisher, Berliner, Fully, Marliave, Cahen, & Dishaw, 1980).

Students tend to engage themselves in stimulating learning tasks that are within the reach of their proximal development. Due to the variation in prior reading ability among kindergartners, some learning materials may exceed the capability of children who have not yet learned the alphabet; while some other tasks may pose little challenges to children who are ready to read and write. Hence, even with abundant time allocated to reading instruction, some students

may lose motivation and become disengaged when there is a mismatch between the uniform tasks and their current ability (Greenwood, Horton, & Utley, 2002). In other words, increasing reading instruction time does not necessarily lead to an increase in students' academic engaged time when there is a lack of adaptation in whole-class instruction.

With a substantial amount of time allocated to reading, instructional differentiation can be carried out through either intensive grouping or flexible switches between different grouping schemes. Both strategies may enable the teacher to flexibly select learning tasks to suit individual needs, which will likely increase academic engaged time and thereby maximizing the learning of all students. Hence, we reason that a student will gain more in reading growth if the kindergarten class spends relatively more time on reading instruction and if homogeneous grouping sustains the student's engagement in meaningful learning tasks matched to his or her current ability.

Reducing instructional time for reading will inevitably limit students' exposure to content materials. We reason that student learning is likely to suffer as a result, especially when students are placed in homogeneous groups on an intensive basis. This is because grouping will require the teacher to spend a relatively large portion of time on simultaneously organizing and managing activities in different groups and on transitioning between groups, with less time left for engaging each group of students in academic tasks for a sustained period. Therefore, the potential benefit of homogeneous grouping will likely vanish under the time constraint.

Based on the above reasoning, we hypothesize that student learning will likely be optimized when students receive a substantial amount of reading instruction time in combination with adaptive instruction through homogeneous grouping. The benefit of instructional time may diminish if the teacher fails to accommodate students' diverse needs in whole-class instruction.

However, instructional differentiation through high-intensity grouping is unlikely to succeed when the teacher distributes a very limited amount of time to each group of students.

Few evaluations of ability grouping in early elementary reading instruction have quantified the extent to which the grouping effects depend on instructional time or grouping intensity. Typically, researchers viewed measures of instructional contexts and processes as confounders that would require covariance adjustment. For example, McCoach, O'Connell, and Levitt (2006) studied the association between ability grouping and kindergarten reading gain, with full-day versus half-day program as a covariate in a linear additive model. Their results provided no information about whether the association between homogeneous grouping and reading gain differs between full-day programs and half-day programs. There has been an attempt to identify moderators of grouping effects through meta-analysis. In search for factors explaining the variability among a set of 103 effect size estimates extracted from 51 studies of ability grouping versus no grouping in elementary through post-secondary grades (Lou, Abrami, Spence, Poulsen, Chambers, & d'Apollonia, 1996; Lou, Abrami, & Spence, 2000), the researchers tested the homogeneity of effect sizes across settings. They noted that the grouping effect was stronger in studies of high treatment frequency (grouped for more than one period per week) than in studies of low treatment frequency (one period per week or less). However, the statistical results generated from these meta-analyses were summaries over a large number of studies that vary enormously in target population and in methodological adequacy.

We focus this study on examining the effect of intensity of homogeneous grouping in kindergarten as a function of the amount of time allocated to reading instruction. The outcome of interest is children's reading growth over the kindergarten year. To be specific, we regard a kindergarten class as having a major emphasis on reading instruction if more than one hour per

day has been allocated to reading. Students are receiving high-intensity grouping if they spend more than 40% of the reading instruction time in homogeneous groups. They are receiving low-intensity grouping if grouped instruction takes no more than 40% of the time. To test our theoretical hypotheses empirically, we ask: What are the effects of high-intensity within-class homogeneous grouping in comparison with low-intensity grouping or no grouping on kindergartners' reading growth when a relatively high amount of time is allocated to reading instruction? What are the effects of grouping when a relatively low amount of instructional time is spent on reading in kindergarten? What is the effect of reading instruction time on student learning under whole-class instruction? How does the effect of reading instruction time change as the intensity of grouping increases?

Causal Estimands

In the current study, each treatment condition defined by instructional time and intensity of grouping has been assigned at the class level. We use $L0$, $L1$, and $L2$ to denote no grouping, low-intensity grouping, and high-intensity grouping, respectively, when a limited amount of time is allocated to kindergarten reading instruction. The corresponding grouping patterns with a substantial amount of time spent on reading are denoted by $H0$, $H1$, and $H2$, respectively (see the upper panel of Table 1).

Insert Table 1 here

Rubin's causal model (Holland, 1986; Rubin, 1978), when extended to multi-level educational data, typically defines the causal effect of one treatment relative to another as the difference between the respective potential outcomes of a child given the treatment setting (Hong, 2004; Hong & Raudenbush, 2006). In the current study, if a kindergarten class has a

possibility of adopting each of the six treatment conditions, every kindergartener in the class will have six corresponding potential reading outcomes at the end of the year, denoted with $Y_{ij}(z)$ for child i in class j and for $z = L0, L1, L2, H0, H1$, and $H2$. For simplicity, we assume that a child's potential outcome value associated with a certain treatment is relatively stable given the student and teacher composition of the class. This is the so-called "stable-unit-treatment-value assumption" (SUTVA) (Rubin, 1986). In other words, we are not interested in possible changes in the child's potential outcome values if the child has attended a kindergarten class with a different student and teacher composition. Hence, our results are to be generalized to a population of the existing kindergarten classes.

Depending on which treatment has actually been selected for the class, only one of the six potential outcomes can be observed for each child. A *causal estimand* is the population average causal effect of one treatment relative to another. For example, the population average causal effect of high-intensity grouping versus no grouping under high reading time is $E[Y(H2) - Y(H0)]$, where $E[\cdot]$ denotes the expectation taken over all the individuals in the population. The above causal estimand is equivalent to the difference between a pair of population average potential outcomes $E[Y(H2)] - E[Y(H0)]$. Here $E[Y(H2)]$ is the average reading outcome of all kindergarten students in a hypothetical world in which all the kindergarten classes have spent a substantial amount of time on reading with high-intensity grouping; and $E[Y(H0)]$ is the average reading outcome of all students if all the kindergarten classes have spent a high amount of time on reading with no grouping Similarly, we define the population average potential outcomes $E[Y(L0)]$, $E[Y(L1)]$, $E[Y(L2)]$, and $E[Y(H1)]$. Corresponding to our research questions, the causal estimands for the grouping effects under a high amount of time for reading are $E[Y(H2) - Y(H0)]$, the average difference between high-intensity grouping and no grouping, and $E[Y(H1) -$

$Y(H0)]$, the average difference between low-intensity grouping and no grouping. We evaluate the grouping effects under low reading time through estimating $E[Y(L2) - Y(L0)]$ and $E[Y(L1) - Y(L0)]$. The causal estimands for the effects of a high amount of reading instruction time versus a low amount of reading instruction time when there is no grouping, low-intensity grouping, or high-intensity grouping are $E[Y(H0) - Y(L0)]$, $E[Y(H1) - Y(L1)]$, and $E[Y(H2) - Y(L2)]$, respectively (see the lower panel of Table 1).

Marginal Mean Weighting Adjustment for Selection Bias

In an ideal world, optimal combinations of concurrent treatments such as instructional time and intensity of grouping could be identified through a 2×3 factorial randomized experiment. Under randomization, all the six treatment groups would have the same pretreatment composition in expectation. Hence, the observed mean outcome of each treatment group would be an unbiased estimate of the corresponding average potential outcome of the population represented by the sample, that is, $E[Y(z) \mid Z = z] = E[Y(z)]$, for $z = L0$, $L1$, $L2$, $H0$, $H1$, and $H2$. The experimental data can be analyzed through a two-way analysis of variance after taking into account the nesting of students in classrooms and schools. However, with an increasing number of treatments under study, the cost of a factorial experiment would rise in multiples. In addition, researchers have voiced strong concerns about the novelty and artificialities of experimental situations that might have compromised the relevance of the findings to real-world settings (Gamoran, 1987; Hiebert, 1987). Non-experimental surveys of instructional practices in naturalistic circumstances therefore have indispensable value.

Without randomization, causal inference is nonetheless possible if the observed data contain useful information about treatment selection. In this study, whether a kindergarten teacher will spend a great amount of time teaching reading and how intensively the teacher will

use homogeneous grouping can perhaps be predicted by a large number of covariates. An effective way to simplify the statistical adjustment for all the observed covariates that predict a certain treatment is to summarize their information in a unidimensional propensity score, a method initially developed for binary treatments (Rosenbaum and Rubin, 1983). For example, we use $Z = 1$ to indicate an experimental condition and $Z = 0$ for a control condition. Let $Y(1)$ and $Y(0)$ denote an individual's potential outcomes associated with $Z = 1$ and $Z = 0$, respectively. Let $\theta(1)$ denote an individual's propensity of adopting $Z = 1$. The key assumption, also called the "strong ignorability assumption," is that the treatment assignment Z is independent of the potential outcomes $Y(1)$ and $Y(0)$ given the estimated propensity score $\hat{\theta}(1)$ (Rosenbaum & Rubin, 1984). Under this assumption, we have that

$E\{E[Y(1) \mid Z = 1, \hat{\theta}(1)]\} = E\{E[Y(1) \mid \hat{\theta}(1)]\} = E[Y(1)]$ and that

$E\{E[Y(0) \mid Z = 0, \hat{\theta}(1)]\} = E\{E[Y(0) \mid \hat{\theta}(1)]\} = E[Y(0)]$. Hence, we can estimate from the observed data the *conditional treatment effect* for subsets of units that have the same estimated propensity score and then take an average over the distribution of the estimated propensity score. That is, we estimate $E\{E[Y(1) \mid Z = 1, \hat{\theta}(1)] - E[Y(0) \mid Z = 0, \hat{\theta}(1)]\} = E\{E[Y(1) - Y(0) \mid \hat{\theta}(1)]\} = E[Y(1) - Y(0)]$. The analysis can be carried out through matching or stratifying units on the estimated propensity score.

Most propensity score-based causal inference studies have been restricted to evaluations of binary treatments. This is due to the difficulty in simultaneous adjustment for multiple propensity scores when there are multiple treatments (Joffe & Rosenbaum, 1999). In our case, every kindergarten class will have up to six propensity scores corresponding to the six treatment conditions, denoted by $\theta(L0)$, $\theta(L1)$, $\theta(L2)$, $\theta(H0)$, $\theta(H1)$, and $\theta(H2)$. To approximate a factorial

randomized experiment will require identifying subsets of classes that are relatively homogeneous in all the six propensity scores, a procedure hard to implement in practice.[1]

The marginal mean weighting (MMW) method (Hong, 2007), originated from Rosenbaum (1987) and Imbens (2000) and recently applied in epidemiological research (Huang, Frangakis, Dominici, Diette, & Wu, 2005), suggests a viable solution for estimating the causal effects of concurrent treatments and of multi-valued treatments from non-experimental data. Instead of attempting to estimate the average of the conditional treatment effect within subsets of homogeneous units, this alternative strategy directly estimates the population average potential outcome—that is, the *marginal mean outcome*—of a treatment from the observed outcome of the units actually assigned to this treatment under the assumption that the assignment to treatment z is independent of its corresponding potential outcome $Y(z)$ given the observed pretreatment covariates (Imbens, 2000).

Below we use treatment $L0$ as an example. Causal inference is defensible if, among units that are relatively homogeneous in the estimated propensity of adopting $L0$, those that are actually in the $L0$ group are not systematically different from those that are not in terms of the distribution of their prior characteristics. Under this assumption, we have that $E\{E[Y(L0) \mid Z = L0, \hat{\theta}(L0)]\} = E\{E[Y(L0) \mid \hat{\theta}(L0)]\} = E[Y(L0)]$. If all the classes in the $L0$ group had had the same probability of adopting this particular treatment, as would have been the case in a completely randomized experiment, the average observed outcome of the $L0$ group would have been an unbiased estimate of the population average potential outcome associated with $L0$. This would have been equivalent to assigning a weight 1.0 to each class in the $L0$ group. Because the assignment to the $L0$ group was not at random, we need to adjust for selection bias associated with the unequal propensity of adopting $L0$ among the classes in the $L0$ group. Through

comparing the distribution of $\hat{\theta}(L0)$ between the $L0$ group and the entire sample, we assign a higher weight to the classes that are over-represented in the $L0$ group and a lower weight to those under-represented. The weighted sample of the $L0$ group will have the same composition of prior characteristics as that of the entire sample. Hence, the weighted average observed outcome of the $L0$ group will be a consistent estimate of the corresponding population average potential outcome $E[Y(L0)]$.

We apply a similar weighing strategy to every treatment group. In general, for classes that have adopted treatment condition z and have an estimated propensity score $\hat{\theta}(z)$, we assign the weight: $\frac{pr(\theta_z)}{pr\{\theta_z \mid \mathrm{D(z)}=1\}} = \frac{pr(\theta_z)}{pr\{\theta_z, \mathrm{D(z)}=1\}} \times pr\{\mathrm{D(z)}=1\}$. Suppose that each kindergarten class in the target population has a nonzero probability of being assigned to each of the six treatment conditions. After weighting, every treatment group will become representative of the population as represented by the whole sample. We then estimate the causal effects of interest through pair-wise comparisons between the estimated population average potential outcomes.

As shown in Hong (2007), Equation (1) is inherently connected to the inverse-probability-of-treatment weight (IPTW) (Robins, 2000). The consistency of the weighted estimate of the treatment effect has been proved by Robins (2000) for single-level data and by Hong & Raudenbush (in press) for multi-level data. However, rather than estimating a weight for each class as a direct function of its estimated propensity of having the treatment actually received (Hernan, Brumback, & Robins, 2000; Robins, 2000; Robins, Hernán, & Brumback, 2000), we opt for a non-parametric approach to estimating Equation (1).[2] Specifically, the marginal mean weighting (MMW) method approximates the distribution of the propensity score for a certain treatment z through dividing the sample into a number of strata on the basis of the

propensity score. The marginal mean outcome is approximately equal to the within-stratum mean outcome of the treated units multiplied by the proportion of units in the stratum and summed over all the strata.[3] We will illustrate our analytic procedure in the next two sections.

Methods

Sample and Measures

We selected data for this study from the Early Childhood Longitudinal study, Kindergarten class of 1998-99 (ECLS_K). The data set contains repeated observations of a nationally representative sample of 21,260 children. Most of the kindergartners were assessed at the beginning (Fall 1998) and the end (Spring 1999) of the school year. The kindergarten reading assessments put a major emphasis on basic skills (including familiarity with print and recognition of letters and phonemes) and minor emphases on various reading comprehension skills. The total reading scale scores encompassing all these domains had a mean of 22.54 with a standard deviation of 8.49 in Fall 1998 and a mean of 32.42 with a standard deviation of 10.35 in Spring 1999. These two waves of assessment scores, equated on a vertical scale, enabled us to assess every child's reading growth over the year (National Center for Education Statistics, 2002). In addition, ECLS-K researchers surveyed the parents and teachers of each sampled child in Fall 1998 and Spring 1999, and collected data from school administrators in Spring 1999.

The two treatments of main interest are reading instruction time and intensity of within-class homogenous grouping. We identified 2,814 kindergarten classes from 990 schools that supplied useful treatment information. The measure of reading instruction time was based on kindergarten teachers' responses to two items in the spring teacher questionnaire: one item asked teachers to rate on a 5-point scale the frequency of reading instruction per week; and the other inquired about the duration of reading instruction on a typical instructional day using a 4-point

scale. To obtain a continuous measure of total reading instruction time per week, we first assigned a middle value to each category in the frequency measure (*never* = 0, *less than once a week* = 0.5, *1-2 times a week* = 1.5, *3-4 times a week* = 3.5, *daily* = 5) and each in the duration measure (*1-30 minutes a day* = 15, *30-60 minutes a day* = 45, *61-90 minutes a day* = 75, *more than 90 minutes a day* = 105). We then multiplied these two scale scores. The product ranges in value from 7.50 to 525 minutes with a mean of 320.81 minutes per week and a standard deviation of 143.20 (see Figure 1). Using 300 minutes per week as a cutoff point, which is equivalent to one hour per day on average, we created a dichotomous measure of reading instruction time. There were about 50% of the kindergarten classes in the "high reading time" category and 50% in the "low reading time" category.

Insert Figure 1 here

To measure the intensity of homogeneous grouping, we computed a ratio of the time spent on ability grouped reading instruction to the total amount of reading instruction time per week. Kindergarten teachers responded to two items in the spring questionnaire about reading ability grouping. One item measured the frequency of within-class homogeneous grouping in reading rated on a 5-point scale ranging from *never* to *daily*. The other measured the duration of reading ability grouping on a typical instructional day, using a 4-point scale ranging from *1-15 minutes per day* to *more than 60 minutes per day*. We assigned a middle value to each category in the frequency measure and an end value to each category in the duration measure. The product of these two variables measured the approximate amount of time a class spent on ability grouped reading instruction per week, ranging in value from 0 to 450 minutes with a mean of about 83 minutes and a standard deviation of 111. The distribution of the intensity of ability grouping,

indicated by the ratio of time for ability grouped reading instruction to total reading instruction time, showed that about a third of the kindergarten classes never used homogeneous ability grouping in reading, about 40% of the classes used grouping less than 40% of the time when teaching reading, and the rest of the classes used grouping on a more intensive basis (see Figure 2). We found that the continuous measures of reading instruction time and intensity of ability grouping were only weakly correlated ($r = 0.241$).

Insert Figure 2 here

Because US teachers tend to have considerable autonomy in determining instructional organization within classrooms (Firestone & Louis, 1999; Lortie, 1975), and because past research has suggested that reading groups are formed primarily on the basis of class size and class composition of students' reading aptitude (Loveless, 1998), we regarded both reading instruction time and within-class reading ability grouping to be mainly results of class-level decisions. We found in the ECLS-K data a considerable amount of within-school variation in treatment adoption among kindergarten classes. About 30% of the schools had only one kindergarten class represented in the sample. In about a third of the schools that had more than one kindergarten class sampled, the classes differed in reading instruction time or in the intensity of within-class homogeneous grouping. This evidence further suggests that the treatments were mostly selected at the class level rather than at the school level. Table 2 shows the frequency distribution of kindergarten classes in the six treatment conditions. The kindergarten classes with treatment information were attended by 16,124 students. Table 3 compares the average reading scores across the six treatment groups.

Insert Table 2 here

Insert Table 3 here

Analytic Procedure

We took seven steps in analyzing the effects of the intensity of within-class homogeneous grouping on kindergartners' reading growth while conditioning on reading instruction time.

1. Identify all the observed pretreatment covariates for each treatment condition. Create missing indicators to capture different missing patterns among categorical covariates; impute missing data in the continuous covariates and outcomes via maximum likelihood estimation.

2. Analyze a multinomial logistic regression model at the class level, estimating five propensity scores per class for five of the six treatment conditions.

3. Empirically identify an analytic sample for causal inference in which every class had a nonzero probability of adopting each of the six treatment conditions.

4. For each of the six treatment conditions, stratify the analytic sample on the basis of the corresponding propensity score and then check within-stratum balance between the treated classes and the untreated classes in the distribution of all the pretreatment covariates.

5. Compute a marginal mean weight for each kindergarten class.

6. Analyze a weighted multi-level growth model with repeated reading assessments at level 1, children at level 2, and classes at level 3, generating estimates of the treatment effects on reading growth. To examine the stability of results, obtain weighted estimates of the treatment effects on the end-of-year reading achievement with children at level 1, classes at level 2, and schools at level 3.

7. Assess the sensitivity of the analytic results to the potential influence of unmeasured

confounders.

We present the details of each step in the Results section. We carried out our analysis with HLM 6.0 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004), SPSS 15.0, and SAS 9.1.

Results

Pretreatment Covariates

In the ECLS-K dataset, we identified 168 pretreatment measures of classroom or school characteristics that showed bivariate associations with reading instruction time or with intensity of ability grouping when the instructional time is given. The pretreatment covariates included class composition of student characteristics, teacher background, school type, and school climate. In general, we found that classes adopting a higher amount of reading instruction time or more intensive homogeneous grouping had a higher concentration of disadvantaged students on average and were more often under the pressure to raise students' academic achievement.

Predictors of reading instruction time. Classes spending more than one hour per day in reading instruction tended to have a higher proportion of minority students, students with Limited English Proficiency (LEP), students who were above the age for kindergarten, and students who were repeating their kindergarten year. Although classes in full-day kindergarten programs spent more time on reading instruction on average when compared with those in half-day programs, we observed a considerable amount of variation in reading instruction time within each type of programs. We noticed that classes in public schools or schools with no selective criteria for kindergarten enrollment tended to allocate more time to reading instruction on average. Such schools usually had more students from low-income families, were relatively large in size, and often had inadequate educational resources. Teachers spending more time on reading

instruction often observed more disruptive classroom behaviors among their students. They tended to have less control of curriculum, receive a lower salary, and report a lower level of job satisfaction. We also found a higher amount of time spent on reading instruction by teachers who had a higher education level or were certified in elementary/early childhood education. In addition, these teachers tended to place a higher emphasis on academic and behavioral readiness for school and on the importance of helping parents teach child to read. Such classes were more often found in schools in which principals were evaluated by students' standardized test scores and especially by raising the performance level of low-achieving students.

Predictors of within-class homogeneous grouping. In agreement with the previous literature, our results showed that the practice of ability grouping was predicted by class composition of student characteristics. To be specific, classes with a higher proportion of students with special needs—including gifted, disabled, or LEP—were more likely to resort to ability grouping. Teacher ethnicity (e.g. being Hispanic or Latino), training (e.g. in bilingual education), and experience (e.g. years of teaching ESL) mirrored the demographic composition of the classes they taught, and thus also predicted more intensive usage of homogeneous grouping. Classes with homogeneous grouping were more often found in public schools, schools with no selective criteria for kindergarten enrollment, schools located in low-income communities with a relatively high proportion of minority students, large schools, and schools with inadequate educational resources. However, we also found that, in schools where more intensive homogeneous grouping was adopted, standardized test scores and especially the performance level of low-achieving students had a greater influence on principal evaluation. These schools tended to set up as their major goal to facilitate kindergarten children's progress in language and number skills. Although teachers who practiced intensive homogeneous grouping

often expressed a lower level of job satisfaction, they also tended to emphasize more on child preparation for school and on assessing child performance relative to national or state standards. Meanwhile, we found that teachers who had a relatively higher education degree and those who had more experience of teaching at a higher grade level in the past showed a higher tendency of using homogeneous grouping. Interestingly, among classes that spent relatively less time on reading instruction, we found that teachers using homogenous grouping placed more emphasis on the importance of evaluating individual child achievement relative to the rest of the class. Contrary to the widely held belief that ability grouping is related to differential standards, such teachers were less likely to report that they applied different evaluation standards to students based on their talent, when compared with teachers not using grouping in the "low-reading-time" category. We did not find these distinctions between grouped classes and non-grouped classes in the "high-reading-time" category.

Retrospectively predicted reading pretest scores. The pretreatment covariates of particular interest were class mean and class standard deviation of children's reading ability at kindergarten entry. Because the Fall reading assessment occurred about two months into the kindergarten year on average, and because the assessment time differed among the students, we retrospectively predicted every child's reading pretest score at the time of kindergarten entry through analyzing a three-level polynomial growth model with the repeated reading assessments at level 1, children at level 2, and kindergarten classes at level 3. We estimated the growth parameters for subpopulations of children defined by age, socio-economic status, gender, and ethnicity. Within each subpopulation, every child had a child-specific intercept at time 0 and a child-specific linear growth rate. We then aggregated to the class level the mean and standard deviation of child reading pretest score at time 0 (see the Appendix for details).

Propensity Score Estimation

Under the assumption that no unobserved covariates were confounded with the treatment given the observed covariates (Rosenbaum & Rubin, 1983; Imbens, 2000), we specified a multinomial logistic regression model for estimating the conditional probability that a kindergarten class would adopt each of the six treatment conditions. The propensity scores, denoted by $\theta_j(z)$ for class j, $z = L1, L2, H0, H1$, and $H2$, with $L0$ as the reference group, were estimated as functions of the observed pretreatment covariates, W_j:

$$\theta_j(z) = \Pr(Z_j = z \mid W_j),$$
$$\log\left(\frac{\theta_j(z)}{\theta_j(L0)}\right) = f_z(W_j). \quad (1)$$

The six estimated propensity scores summed up to 1.0 for every kindergarten class. We had employed a step-wise procedure to reduce the number of covariates in the propensity models. The multinomial logistic regression model contained, in total, 68 covariates, 20 quadratic terms, and 9 missing indicators.

Analytic Sample

We compared, between each treatment group and the rest of the sample, the distribution of the logit of the estimated propensity of being assigned to that particular treatment condition. This procedure enabled us to empirically identify classes in each treatment group that did not have matches in the rest of the sample, or classes in the rest of the sample that would almost never be assigned to a particular treatment. In total, about 956 classes showed an almost zero probability of being assigned to one or more of the treatment conditions and therefore would not contribute to our causal inference. Hence, we restricted our analysis to the remaining sample of 1858 classes in 740 schools attended by 10,189 sampled students. In this analytic sample, we

found about 16% of the classes having low reading time with no grouping, 16% having low reading time with low-intensity grouping, and 13% having low reading time with high-intensity grouping. The proportions of classes having no grouping, low-intensity grouping, or high-intensity grouping under high reading time were 16%, 28%, and 11%, respectively.

We compared the 10,189 kindergartners in our analytic sample with the nationally representative sample of 21, 260 kindergarteners. The two samples showed no statistically significant differences in average age, gender composition, proportion of black children, and proportion of Asian children. Nonetheless, our analytic sample appeared to have some slight differences when compared with the entire ECLS-K sample in that the former had a smaller proportion of white children, a larger proportion of Hispanic children, and hence a smaller proportion of children from English-speaking families. In addition, with a higher percentage of free-lunch students represented in the analytic sample, the average socio-economic status of the children in the analytic sample was slightly lower than those in the full sample (see Table 4).

Insert Table 4 here

Propensity Score Stratification

For each of the six treatment conditions, we divided the analytic sample into either five or six strata on the basis of its corresponding logit of the estimated propensity score (see Tables 5-10). According to Cochran (1968), stratifying a sample into five subclasses typically removes at least 90% of the bias associated with a pretreatment covariate. We then examined, under each stratification, whether the classes in the focal treatment condition and those not in this treatment condition had the same composition of pretreatment characteristics within each stratum. We found within-stratum balance in 95~98% of the 168 pretreatment covariates except for the $H0$

group in which nearly 6% of the pretreatment covariates showed statistically significant differences between the *H*0 group and the rest of the sample. Further analysis indicated that, among the ten imbalanced covariates for *H0,* only five of them (i.e. about 3% of the pretreatment covariates) were significant predictors of reading growth and would actually confound the treatment effect estimates.

Insert Tables 5-10 here

Marginal Mean Weight

Our next step was to construct a weighted sample for each treatment condition. For classes in treatment group *z* and in stratum *s*, the marginal mean weight was computed as the following:

$$MMW = \frac{n_s}{n_{z,s}} \times \Pr(Z = z). \quad (2)$$

Here n_s denotes the number of classes in stratum *s*; $n_{z,s}$ denoted the number of classes in stratum *s* that were actually assigned to the treatment condition; $\Pr(Z = z)$ is the proportion of classes in treatment group *z* (Hong, 2007). The estimated marginal mean weight, ranging from 0.18 to 4.57, had a mean of 1.0 and a standard deviation of 0.89. Table 11 illustrated the construction of the weighted sample of *L*0. For example, the kindergarten classes in stratum 1 had a relatively low propensity of adopting *L*0. These classes had a lower representation in the *L*0 group when compared with its representation in the whole sample. For the 38 classes in the *L*0 group in this stratum, the estimated marginal mean weight was 3.04. Hence, the weighted *L*0 group had 115.66 classes in stratum 1. The kindergarten classes in a higher stratum had a relatively higher representation in the *L*0 group and thus received a relatively lower weight. As a result, the weighted sample of the *L*0 group across all the strata resembled the entire analytic sample as if

the classes had been assigned at random to *L*0. Following the same logic, after weighting, all the six treatment groups became comparable in their distributions of the observed pretreatment characteristics.

Insert Table 11 here

Model-Based Estimation of Treatment Effects

Model specification. A weighted three-level growth model generated the results for comparing the population average potential reading growth rates across the six treatment conditions. For simplicity, we specified a linear growth model at level 1 for the Fall and Spring reading scale scores of child *i* in class *j*:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}(\text{Dur\_K})_{tij} + e_{tij}, \quad e_{tij} \sim N(0,\ \sigma_{et}^2). \tag{3}$$

The reading pretest score at kindergarten entry π_{0ij} and the reading growth rate π_{1ij} varied randomly among children at level 2 and classes at level 3. The level-2 model was simply

$$\pi_{0ij} = \beta_{00j} + r_{0ij},$$
$$\pi_{1ij} = \beta_{10j} + r_{1ij};$$

$$\begin{pmatrix} r_{0ij} \\ r_{1ij} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\pi 00} & \tau_{\pi 01} \\ \tau_{\pi 10} & \tau_{\pi 11} \end{pmatrix}\right]. \tag{4}$$

Applying the estimated marginal mean weight for the kindergarten classes at level 3, we analyzed a saturated model for both class mean reading pretest and class mean reading growth rate:

$$\beta_{00j} = \gamma_{001}(\text{DM\_}L)_j + \gamma_{002}(\text{DM\_}L1)_j + \gamma_{003}(\text{DM\_}L2)_j + \gamma_{004}(\text{DM\_}H)_j + \gamma_{005}(\text{DM\_}H1)_j + \gamma_{006}(\text{DM\_}H2)_j + u_{00j},$$
$$\beta_{10j} = \gamma_{101}(\text{DM\_}L)_j + \gamma_{102}(\text{DM\_}L1)_j + \gamma_{103}(\text{DM\_}L2)_j + \gamma_{104}(\text{DM\_}H)_j + \gamma_{105}(\text{DM\_}H1)_j + \gamma_{106}(\text{DM\_}H2)_j + u_{10j};$$
$$\begin{pmatrix} u_{00j} \\ u_{10j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\beta 00} & \tau_{\beta 00.10} \\ \tau_{\beta 10.00} & \tau_{\beta 10} \end{pmatrix}\right] \tag{5}$$

Here $(DM\_L)_j$ is a dummy indicator taking a value of 1 if class j was in the low-reading-time category and 0 otherwise; $(DM\_H)_j$ takes a value of 1 if the class was in the high-reading-time category and 0 otherwise; (DM_L1), (DM_L2), (DM_H1), and (DM_H2) are dummy indicators for the treatment groups $L1$, $L2$, $H1$, and $H2$, respectively. We used $L0$ as the reference group for classes of low reading time, and $H0$ as the reference group for those of high reading time. We estimated the model through pseudo-maximum likelihood that produced consistent estimates of the variance components and therefore of the causal effects of interest (Hong & Raudenbush, in press). Given our large sample size of classes, we based our hypotheses testing on robust standard errors.

Unweighted analysis. The first panel in Table 12 compares the average reading pretest scores and the average reading growth rates across the six treatment groups without weighting. Children attending kindergarten classes with high reading time and high-intensity grouping ($H2$) appeared to have the lowest average pretest score (18.83 – 0.84 = 17.99); while those in classes with low reading time and low-intensity grouping ($L1$) had the highest average pretest score (18.82 + 0.13 = 18.95). According to hypotheses testing, the mean difference between these two groups was significantly different from zero (mean difference = -0.96, SE = 0.49, t = -1.97, $p <$ .05), indicating a lack of equivalence among the treatment groups at kindergarten entry. Due to the pretreatment differences among the treatment groups, the unweighted mean differences in reading growth would likely be biased estimates of the treatment effects.

Insert Table 12 here

Weighted analysis. The results of analyzing the weighted growth model as specified in Equations (3), (4), and (5) are listed in the second panel of Table 12. After weighting, the

average reading pretest scores showed no statistically significant differences among all six treatment groups. Under the assumption that the treatment assignment was independent of the unmeasured covariates given the observed pretreatment covariates, the weighted sample of each treatment group would provide a consistent estimate of the population average potential growth rate, that is, the average reading growth we would have observed had the whole population of kindergarten classes represented by our analytic sample been assigned to that particular treatment condition. We found that, under low reading time, the estimated average monthly reading growth rates were 1.55 if there was no grouping (*L*0), 1.58 for low-intensity grouping (*L*1), and 1.52 for high-intensity grouping (*L*2). Hypothesis testing showed no statistically significant differences between low-intensity grouping (*L*1) and no grouping (*L*0) ($\hat{\gamma}_{102} = 0.02$, $SE = 0.05$, $t = 0.54$), between high-intensity grouping (*L*2) and no grouping (*L*0) ($\hat{\gamma}_{103} = -0.03$, $SE = 0.05$, $t = -0.58$), or between high-intensity grouping (*L*2) and low-intensity grouping (*L*1) (contrast = -0.05, $SE = 0.05$, $\chi^2 = 1.12$, $df = 1$, $p = .29$).

The estimated average reading growth rates under high reading time were 1.58 for no grouping (*H*0), 1.69 for low-intensity grouping (*H*1), and 1.73 for high-intensity grouping (*H*2). Multivariate hypothesis testing showed no significant difference between the latter two (contrast = -0.04, $SE = 0.05$, $\chi^2 = 0.44$, $df = 1$, $p > .50$). According to our estimation, experiencing low-intensity grouping (*H*1) rather than no grouping (*H*0) would raise a kindergartner's reading growth by about 0.99 over nine months ($\hat{\gamma}_{105} = 0.11$, $SE = 0.04$, $t = 2.76$, $p < .01$). Having high-intensity grouping (*H*2) rather than no grouping (*H*0) during the same period would raise reading growth by 1.35 ($\hat{\gamma}_{106} = 0.15$, $SE = 0.06$, $t = 2.64$, $p < .01$).

We found no statistically significant difference between high reading time with no grouping (*H*0) and low reading time with no grouping (*L*0) (contrast = 0.02, $SE = 0.04$, $t = 0.55$,

df = 1, *p* > .50). However, the largest contrast existed between high reading time with high-intensity grouping (*H*2) and low reading time with high-intensity grouping (*L*2) (contrast = 0.21, *SE* = 0.05, *t* = 4.46, *p* < .001). The difference in average reading growth accumulated over a 9-month period was 0.21 × 9 = 1.89, equivalent to 13.04% of the average reading growth over the school year, or close to a whole month of reading growth for a typical kindergartner.

To illustrate, we have graphed in Figure 3, for a typical kindergartner in the population represented by our analytic sample, the six potential reading growth trajectories that the child would display under the six corresponding treatment conditions. The estimated average reading pretest score at kindergarten entry was 18.46, which we used as the starting point for all the six potential trajectories.

Insert Figure 3 here

Observing no statistically significant differences in average reading growth either among *L*0, *L*1, and *L*2, or between *H*2 and *H*1, we estimated a parsimonious model by combining each of these two clusters of treatment groups. Using *H*0 as the reference category, the fixed-effect part of the class-level model in Equation (5) became the following:

$$\begin{aligned}\beta_{00j} &= \gamma_{000} + \gamma_{001}(\mathrm{DM\_}L)_j + \gamma_{002}(\mathrm{DM\_}H12)_j + u_{00j},\\ \beta_{10j} &= \gamma_{100} + \gamma_{101}(\mathrm{DM\_}L)_j + \gamma_{102}(\mathrm{DM\_}H12)_j + u_{10j}.\end{aligned} \tag{6}$$

Here $(\mathrm{DM\_}H12)_j$ took a value of 1 if class *j* was either in *H*1 or *H*2 and 0 otherwise. By analyzing this model, we intended to make inference about a new causal estimand

$$\delta_{H12-0} = E\left[Y(H1 \text{ or } H2) - Y(H0)\right]. \tag{7}$$

Here δ_{H12-0} denotes the effect of low-intensity grouping or high-intensity grouping (*H*1 or *H*2) relative to no grouping (*H*0) under high reading time. A likelihood ratio test indicated that the

parsimonious model was as sufficient as the saturated model ($\chi^2 = 4.73$, $df = 6$, $p > .50$). The results showed that, with more than one hour of reading instruction per day, within-class homogeneous grouping, regardless of its intensity (*H*1 or *H*2), was expected to raise reading growth by 1.08 over the year ($\hat{\gamma}_{102} = 0.12$, $SE = 0.04$, $t = 3.14$, $p < .01$) when compared with no grouping (*H*0). This amounted to 7.45% of the average yearly reading growth, which was equivalent to a typical kindergartner's reading growth in about two-thirds of a month. We found no statistically significant difference between low reading time regardless of grouping (*L*0, *L*1, or *L*2) and high reading time with no grouping (*H*0) ($\hat{\gamma}_{101} = -0.02$, $SE = 0.04$, $t = -0.64$, $p > .50$).

Stability analysis. In our analytic sample, about 40% of the schools had only one kindergarten class sampled, and about 60% of the schools had no more than two kindergarten classes. To assess the stability of our conclusions about the treatment effects across different model specifications, we alternatively estimated the treatment effects on the end-of-year reading achievement with students at level 1, classes at level 2, and schools at level 3. We applied the marginal mean weight at level 2 where the class average reading achievement at the end of the 9-month kindergarten year was regressed on the treatment indicators. To increase precision in estimation, we made covariance adjustment for the empirical Bayes estimate of child pretest score obtained from estimating Equation (A4), and denoted it with $(Y\_\text{pre})_{ijk}$ for child *i* attending kindergarten class *j* in school *k*. In addition, we centered the lapse of time from kindergarten entry on the end of the 9-month school year for the Spring reading assessment, and denoted it with $(\text{Dur\_K} - 9)_{ijk}$. The three-level saturated model was specified as follows:

Level 1

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(Y\_\text{pre})_{ijk} + \pi_{2jk}(\text{Dur\_k} - 9)_{ijk} + e_{ijk};$$

Level 2

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{DM\_}L1)_{jk} + \beta_{02k}(\text{DM\_}L2)_{jk} + \beta_{03k}(\text{DM\_}H0)_{jk} + \beta_{04k}(\text{DM\_}H1)_{jk} + \beta_{05k}(\text{DM\_}H2)_{jk} + r_{0jk},$$
$$\pi_{1jk} = \beta_{10k},$$
$$\pi_{2jk} = \beta_{20k};$$

Level 3

$$\beta_{00k} = \gamma_{000} + u_{00k},$$
$$\beta_{0ck} = \gamma_{0c0}, \text{ for } c = 1, ..., 5, \quad (8)$$
$$\beta_{d0k} = \gamma_{d00}, \text{ for } d = 1, 2;$$
$$e_{ijk} \sim N(0, \sigma^2); \qquad r_{0jk} \sim N(0, \tau_\pi); \qquad u_{00k} \sim N(0, \tau_u).$$

By equating the average pretest scores at kindergarten entry across all the six treatment groups through weighting and covariance adjustment, we expect that a comparison between any two treatment groups of their end-of-year reading achievement would approximately replicate the corresponding between-group comparison of 9-month reading growth.

The weighted analysis of this alternative three-level model showed that, under low reading time, no grouping (*L*0) seemed to have led to a higher level of reading achievement at the end of the kindergarten year when compared with low-intensity grouping (*L*1) ($\hat{\gamma}_{010}$ = -0.07, *SE* = 0.35, *t* = -0.19, *p* > .50) or high-intensity grouping (*L*2) ($\hat{\gamma}_{020}$ = -0.48, *SE* = 0.40, *t* = -1.20, *p* = .23), although the differences were not statistically significant. When using *H*0 as the reference category, we found that low-intensity grouping under high reading time (*H*1) displayed a higher end-of-year reading achievement (contrast = 0.71, *SE* = 0.31, *t* = 2.28, *p* < .05). Similarly, high-intensity grouping under high reading time (*H*2) seemed to have led to a higher reading achievement (contrast = 0.81, *SE* = 0.40, *t* = 2.04, *p* < .05) in comparison with no grouping (*H*0). In the meantime, no statistically significant difference was detected between high reading time with no grouping (*H*0) and low reading time with no grouping (*L*0) ($\hat{\gamma}_{030}$ = -0.11, *SE* = 0.32, *t* = -0.35). Yet the largest contrast existed between high reading time with high-intensity grouping (*H*2) and low reading time with high-intensity grouping (*L*2) (contrast = 1.18, *SE* = 0.45, *t* =

2.64, $p < .01$). As before, we found no statistically significant differences in end-of-year reading achievement among *L*0, *L*1, and *L*2 or between *H*1 and *H*2 (results not tabulated). We therefore merged each of these two clusters of treatment conditions in our subsequent analysis. Grouping under high reading time, regardless of intensity (*H*1 or *H*2), continued to show a significant positive effect on reading achievement (contrast = 0.76, *SE* = 0.29, $t = 2.58$, $p < .01$) in comparison with no grouping (*H*0). In addition, the latter (*H*0) showed no benefit when compared with low reading time (*L*0, *L*1, or *L*2) (contrast = 0.04, *SE* = 0.28, $t = 0.13$, $p > .50$). In general, the growth model analysis and the cross-sectional analyses both suggested a positive effect of low-intensity or high-intensity grouping when compared with no grouping under high reading time. Neither analysis detected any statistically significant effect of grouping under low reading time.

Comparison with causal inferences for a single treatment. Our analysis was guided by the theory that the grouping effects may depend on the amount of reading instruction time and the intensity of grouping. We suspected that evaluating ability grouping only as a binary treatment, which has been typically the case in most causal inference studies, might generate uninformative or even misleading results. To reveal this point, we estimated the effect of grouping versus no grouping on kindergarten reading achievement without taking into consideration reading instruction time and the intensity of grouping. The class-level model in Equation (8) was modified as follows:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{DM\_Z12})_{jk} + r_{0jk} \qquad (9)$$

Here $(\text{DM\_}Z12)_j$ is a dummy indicator taking a value of 1 if class *j* had adopted either low-intensity grouping or high-intensity grouping. The result indicated that within-class homogeneous grouping had no effect on end-of-year reading achievement ($\hat{\gamma}_{010} = 0.28$, *SE* =

0.22, $t = 1.28$, $p = .20$). However, this results was unstable, as a parallel analysis of reading growth showed a small positive effect of grouping versus no grouping (contrast = 0.07, SE = 0.03, $t = 2.59$, $p < .01$) equivalent to about 4.36% of the average reading growth rate.

Ignoring reading instruction time, we also estimated the effect of the intensity of ability grouping by changing the class-level model to the following:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{DM\_Z1})_{jk} + \beta_{02k}(\text{DM\_Z2})_{jk} + r_{0jk}. \quad (10)$$

Here $(\text{DM\_}Z1)_j$ took a value of 1 if class j had adopted low-intensity grouping and 0 otherwise; $(\text{DM\_}Z2)_j$ took a value of 1 if class j had adopted high-intensity grouping and 0 otherwise. In comparison with the practice of no grouping, the results showed no significant effect of either low-intensity grouping ($\hat{\gamma}_{010} = 0.39$, $SE = 0.24$, $t = 1.66$, $p = .10$) or high-intensity grouping ($\hat{\gamma}_{020} = 0.06$, $SE = 0.30$, $t = 0.21$, $p > .50$). A parallel analysis of growth model, instead, showed a small positive effect of infrequent grouping (contrast = 0.08, $SE = 0.03$, $t = 2.77$, $p < .01$) and no effect of frequent grouping (contrast = 0.05, $SE = 0.04$, $t = 1.30$, $p = .19$). These results were hard to interpret in theory and disguised the fact that grouping made no difference under low reading time and that, in contrast, high-intensity and low-intensity grouping were more effective than no grouping under high reading time.

Sensitivity Analysis

The weighted estimation of the treatment effects would be unbiased under the assumption that the assignment to each treatment condition was independent of the unobserved confounders given the observed covariates. We assessed the extent to which our causal conclusions would be altered by additional adjustment for potential unmeasured confounders, the omission of which would introduce a bias comparable to that of the most important observed covariates (Lin, Psaty, & Kronmal 1998; Rosenbaum 1986, 2002). The parsimonious class-level model as shown in

Equation (6) contrasted $H0$ with the combination of $H1$ and $H2$. Suppose that an unmeasured class-level covariate U_H had a mean difference $U_{H12} - U_{H0}$ between classes in $H1$ or $H2$ and those in $H0$. Also suppose that the bivariate association ϕ_H between U_H and the potential reading growth was linear. To remove from $\hat{\gamma}_{102}$ the potential bias associated with U_H, we computed new estimates

$$\delta^*_{H12-0} = \gamma_{104} \pm \phi_H (U_{H12} - U_{H0}). \quad (11)$$

Because our list of pretreatment covariates was relatively comprehensive, we generated plausible values for $U_{H12} - U_{H0}$ from computing the mean differences between the combination of $H1$ and $H2$ groups and the $H0$ group in the observed pretreatment covariates W_m, m = 1, …, 168. To generate plausible values for ϕ_H, we selected all the classes in the high-reading-time category and regressed, under marginal mean weighting, class mean reading growth rate β_{10j} on each of the observed covariates one at a time,[4]

$$\beta_{10j} = \phi_0 + \phi_1 (\text{DM\_}L2)_j + \phi_{Hm} W_{mj} + \varepsilon_j, \; m = 1, \ldots, 168. \quad (12)$$

Among the observed pretreatment covariates, school safety rate was the strongest observed confounder. After adjusting for either a positive or a negative hypothetical confounding effect of the same magnitude, our new estimates of δ_{H12-0} were 0.11 and 0.13. Both 95% confidence intervals (0.03, 0.19) and (0.05, 0.21) were still above zero. Hence, our conclusion about the positive effect of $H1$ and $H2$ relative to $H0$ did not seem highly sensitive to the possible influence of unmeasured covariates comparable to the strongest confounders observed.

Conclusion

In this study we examined the effects of within-class homogeneous grouping on kindergartners' reading growth and its dependence on reading instruction time. Our results have

suggested that, when a substantial amount of time was devoted to reading instruction in kindergarten, within-class homogeneous grouping was expected to better improve kindergartners' reading growth when compared with no grouping. This seemed to be true regardless of how intensively a teacher groups students in reading instruction. The effect size was about 7.45% of the average reading growth over 9 months or equivalent to a typical kindergartner's reading growth in two-thirds of a month. The 95% confidence interval for the effect size was (2.58%, 12.33%). This conclusion was insensitive to the impact of plausible unmeasured confounders. In contrast, when reading instruction time was limited, our results suggested no effect of homogeneous grouping on kindergartners' reading growth when compared with no grouping.

When homogeneous grouping was never used in a class, we detected no advantage of allocating one hour or more per day to reading instruction in comparison with no more than one hour per day. The effect size was about 1.24% of the average 9-month reading growth with a confidence interval of (-3.63%, 6.12%). However, when teachers adopted homogeneous grouping in reading with a high intensity, we found the largest benefit of spending more time on reading. The effect size was 11.78% of the yearly reading growth or about one month of reading growth with a confidence interval of (1.45%, 20.91%). All the above conclusions were stable under an alternative model specification.

In summary, within-class homogeneous grouping seems to bring benefit to student learning when reading instruction receives a major emphasis in kindergarten. With abundant time spent on reading, instructional differentiation can be carried out through either intensive grouping or flexible switches between different grouping schemes. Both strategies may facilitate instructional adaptation to individual needs and hence are more effective than no grouping.

When reading instruction time is limited, homogeneous grouping shows no advantage when compared with no grouping perhaps due to a relatively large portion of time spent on managing multiple groups and on transitioning between group activities. These results are consistent with our hypotheses that, during the kindergarten year, student learning will likely be optimized when students receive a substantial amount of reading instruction time in combination with adaptive instruction through homogeneous grouping. The positive effect of homogeneous grouping will likely disappear when only a limited amount of time is available for teaching reading. In addition, according to our hypothesis, when there is a lack of adaptation to students' diverse needs in whole-class instruction, increasing reading instruction time may not lead to an increase in students' academic engaged time and therefore may not improve student learning. We indeed found evidence that the potential benefit of increasing reading instruction time tends to diminish when the teacher never uses homogeneous grouping. In contrast, the benefit of increasing instructional time is maximized in classes with high-intensity grouping.

We have found that kindergarten teachers in public schools with a larger concentration of low-income and minority students and under a greater pressure to raise student performance tended to use within-class homogeneous grouping more often. We have also observed a higher amount of reading instruction time on average in similar types of classes and schools. As suggested by the results of our causal inference, these teaching efforts in combination appear to be generally fruitful in helping kindergartners to read. Hence, if a kindergarten teacher intends to enhance student learning by allocating more time to reading instruction, the increased instructional time will be most effectively used when students are engaged in learning tasks tailored to their ability levels. In fact, the results have shown that, without flexible adaptation to individual needs, the increase of instructional time will bring little benefit to student learning.

These results have important implications especially for schools striving for raising the performance of a large number of children from disadvantaged backgrounds. The evidence generated from this study will also help to inform the ongoing debate about ability grouping. Our findings have clearly indicated that, universally abolishing homogeneous grouping, a policy endorsed by the opponents to grouping in many school districts, could have inadvertently harmed the kindergartners' reading growth.

Most of the previous research typically singled out ability grouping as an isolated practice without taking into account the inter-play between instructional time and the intensity of grouping. We have used our data to reveal the potential consequences when researchers took over-simplistic approaches to studying instruction. As we have shown, ignoring reading instruction time as an important contextual factor would lead to analytic results of little theoretical or practical significance.

Because instruction is a multifaceted intervention program, every single element needs to operate in concert with other parts of the program to produce a joint impact. The effects of various elements therefore cannot be assumed additive. In studying the effects of homogeneous grouping, we find it useful to conceptually distinguish between instructional settings and grouping processes. *Instructional settings* refer to organizational structure, resources, and climate that provide a context for ability-grouped or non-grouped instruction. Measures of instructional settings include time allocation to a certain subject, class size, class composition, and teacher knowledge. Within the same context, ability-grouped classes may vary substantially in their instructional processes, representing different treatment versions under the same label. Besides the intensity of grouping, measures of *grouping processes* also include how teachers differentiate across ability groups in instructional pace, content coverage, pedagogical methods, and teacher-

student interactions. Rather than simply assuming a constant effect of homogeneous grouping across different settings, or assuming a constant effect of reading instruction time across different grouping processes, researchers can provide more useful knowledge for practitioners by identifying an optimal combination of instructional setting and grouping process.

The effects of instructional time and grouping may not be the same for students at different ability levels. For example, we suspect that low-achieving children are most likely to suffer when they are not provided with sufficient time to learn. Moreover, these effects will be carried over into the later schooling years. A first-grade teacher's decision about whether to divide a class into homogeneous groups or, in a first-grade class that has adopted homogeneous grouping, the teacher's decision about group placement will largely be shaped by the results of her students' learning experience in kindergarten. Hence, in order to improve our knowledge base and to better inform educational practice, researchers will need to study the differential effects and the cumulative effects of homogeneous grouping over years across a wide range of time-varying instructional settings, grouping processes, and students' growing abilities.

In the past, the multidimensionality of classroom instruction has posed a major challenge to causal inferences of instructional effects. Causal inference studies with non-experimental data have been mostly restricted to evaluations of binary treatments. This study has introduced a new statistical adjustment method for evaluating multi-valued treatments and for testing theories about the inter-dependence among concurrent treatments in multi-level educational data. Future research will explore the potential of the MMW method for investigating more complex causal questions about instruction effects on student learning.

References

Anderson, L. W. (1984). Time and school learning (ed.). London: Croom Helm.

Barr, R., & Dreeben, R. (1983). *How schools work*. Chicago: University of Chicago Press.

Biemiller, A. (1993). Lake Wobegon revisited: On diversity and education. *Educational Researcher, 22*, 7-12.

Bloom, B. (1974). Time and learning. *American Psychologist*, *29*, 682-688.

Carroll, J. (1963). A model of school learning. *Teachers College Record, 64*, 723-733.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 295-313.

Dawson, M. M. (1987). Beyond ability grouping: A review of the effectiveness of ability grouping and its alternatives. *School Psychology Review, 16*(3), 348-369.Dreeben, R., & Barr, R. (1988). The formation and instruction of ability groups. *American Journal of Education*, 34-64.

Firestone, W. A. & Louis, K. S (1997). Schools as cultures. In Joseph Murphy & Karen Seashore Louis (Eds.), *Handbook of research on educational administration: A project of the American Educational Research Association* (2nd ed.). San Francisco, CA: Jossey-Bass publishers. 297-323.

Fisher, C. W., Berliner, D. C., Fully, N. N., Marliave, R. S., Cahen, L. S., & Dishaw, M. M. (1980). Teaching behaviors, academic learning time and student achievement: An overview. In Carolyn Denham and Anne Lieberman (Eds.), *Time to learn: A review of the Beginning Teacher Evaluation Study* (pp.7-32). Washington D.C.: National Institute of Education.

Gamoran, A. (1987). Organization, instruction, and the effects of ability grouping:

Comment on Slavin's "best-evidence synthesis." *Review of Educational Research*, *57*(3), 341-345.

Gamoran, A. (1992). Is ability grouping equitable? *Educational Leaderhsip*, *52*(2), 11-17.

Gamoran, A., Nystrand, M., Berends, M., & LePore, P. C. (1995). An organizational analysis of the effects of ability grouping. *American Educational Research Journal, 32*(4), 687-715.

Greenwood, C. R., Horton, B. T., & Utley, C. A. (2002). Academic engagement: Current perspectives in research and practice. School Psychology Review, 31(3), 328-349.

Hallinan, M. T. (2003). Ability grouping and student learning. *Brookings Papers on Education Policy, 2003*, 95-124.

Hallinan, M. T. & Sorensen, A. B. (1983). The formation and stability of instructional groups. *American Sociological Review*, *48*, 838-851.

Hernán M.A., Brumback B., and Robins J.M. (2000) Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11:561-570.

Hiebert, E. H. (1983). An examination of ability grouping for reading instruction. *Reading Research Quarterly, 18*(2), 231-255.

Hiebert, E. H. (1987). The context of instruction and student learning: An examination of Slavin's assumptions. *Review of Educational Research*, *57*(3), 337-340.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Hong, G. (2004), "Causal Inference for Multi-Level Observational Data with Application to Kindergarten Retention," unpublished Ph.D. dissertation, University of Michigan, School of Education.

Hong, G. (2007). Marginal mean weighting adjustment for selection bias. Invited presentation at the Education Workshop. Chicago, IL: University of Chicago.

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

Hong, G., & Raudenbush, S. W. (In press). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*.

Huang, I-C., Frangakis, C., Dominici, F., Diette, G. B., and Wu, A. W. (2005). Approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Services Research, 40*, 253-278.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*, 706-710.

Ireson, J., & Hallam, S. (1999). Raising standards: Is ability grouping the answer? *Oxford Review of Education, 25*(3), 343-358.

Jeynes, W. H. (2006). Standardized tests and Froebel's original kindergarten model. *Teachers College Record*, *108*(10), 1937-1959.

Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology*, *150*, 327-333.

Kilpatrick, W. H. (1916). *Froebel's kindergarten principles critically examined*. New York: Macmillan.

Kulik, J. A., & Kulik, C. L. (1987). Effects of ability grouping on student achievement. *Equity and Excellence, 23*, 22-30.

Lortie, D. (1975). *Schoolteacher: A sociological analysis*. Chicago: University of Chicago Press.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research, 66*(4), 423-458.

Lou, Y., Abrami, P. C., & Spence, J. C. (2000). Effects of within-class grouping on student achievement: An exploratory model. *Journal of Educational Research, 94*(2), 101-112.

Loveless, T. (1998). The tracking and ability grouping debate. *The Fordham Report*. Washington, DC: The Fordham Foundation.

Macintyre, H., & Ireson, J. (2002). Within-class Ability Grouping: placement of pupils in groups and self-concept. *British Educational Research Journal, 28*(2), 249-263.

McCoach, D. B., O'Connell, A. A., & Levitt, H. (2006). Ability grouping across kindergarten using an Early Childhood Longitudinal Study. *The Journal of Educational Research*, *99*(6), 339-346.

Meisels, S. J. (1989). High-stakes testing in kindergarten, *Educational Leadership*, *46*(7), 16-22.

Millot, B. (1995). Economics of educational time and learning. In Martin Carnoy (Ed.), *International Encyclopedia of Economics of Education*. Oxford: Pergamon-Elsevier. 353-358

National Center for Education Statistics. (2002). *Early Childhood Longitudinal Study-Kindergarten class of 1998-99 (ECLS-K), Psychometric Report for Kindergarten Through First*

Grade, NCES 2002-05, by Donald A. Rock and Judith M. Pollack, Educational Testing Service, Elvira Germino Hausken, project officer. Washington, DC.

Oakes, J., Gamoran, A., & Page, R. N. (1992). Curriculum differentiation: Opportunities, outcomes, and meanings. In P. W. Jackson (Ed.), *Handbook of Research on Curriculum* (570-608). Washington, DC: American Educational Research Association.

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.

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

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*, 387-394.

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). "Statistics and causal inference: Comment: Which ifs have causal answers," *Journal of the American Statistical Association, 81*(396), 961-962.

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. (1987a). Ability grouping and student achievement in elementary schools: A best-evidence systhesis. *Review of Educational Research*, *57*(3), 293-336.

Sorensen, A. B., & Hallinan, M. T. (1986). Effects of ability grouping on growth in academic achievement. *American Educational Research Journal*, *23*(4), 519-542.

Tach, L. M., & Farkas, G. (2006). Learning-related behaviors, cognitive skills, and ability grouping when schooling beings. *Social Science Research*, *35*(4), 1048-1079.

Wiggins, R. A. (1994). Large group lesson/small group follow-up: Flexible grouping. *The Reading Teacher*, *47*(6), 450-460.

Zirkel, P. A., & Gluckman, I. B. (1995). It's the Law: Ability grouping. *Principal*, *75*(1), 62-63.

Appendix

Retrospective Prediction of Reading Pretest Scale Scores

At level 1, we modeled the reading assessment score Y_{tij} of child *i* in class *j* at time *t* as a polynomial function of $(\text{Dur\_K})_{tij}$ denoting the lapse of time between the beginning of the kindergarten year and the assessment:

$$Y_{tij} = \sum_{q=0}^{3} \pi_{qij} (\text{Dur\_K})_{tij}^{q} + e_{tij}, \quad e_{tij} \sim N(0, \sigma_{et}^2). \tag{A1}$$

This model used linear, quadratic, and cubic terms to fully capture the nonlinear features of an individual growth trajectory. The intercept π_{0ij} represented the child's extrapolated reading pretest score when the time lapse was zero. Using reliability information (λ_t) supplied in the ECLS-K documentation (National Center for Education Statistics, 2002) along with the estimated variance of reading scale score in each wave (σ_{Yt}^2), we computed the measurement error variance σ_{et}^2 by applying the formula $\sigma_{et}^2 = (1 - \lambda_t)\sigma_{Yt}^2$. The estimated error variances were 5.32 for the Fall reading assessment and 5.94 for the Spring assement.

Based on our preliminary analysis, we fixed the quadratic and cubic slopes π_{2ij} and π_{3ij} at level 2 and level 3. In the level 2 model,

$$\pi_{qij} = \beta_{q0j} + \sum_{h=1}^{7} \beta_{qhj} X_{hij} + r_{qij}, \text{ for } q = 0, 1;$$

$$\pi_{qij} = \beta_{q0j} + \sum_{h=1}^{7} \beta_{qhj} X_{hij}, \text{ for } q = 2, 3; \tag{A2}$$

$$\begin{pmatrix} r_{0ij} \\ r_{1ij} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\pi 00} & \tau_{\pi 01} \\ \tau_{\pi 10} & \tau_{\pi 11} \end{pmatrix} \right].$$

To increase precision in prediction, the above level-2 model estimates the growth parameters for

subpopulations of children defined by age, socio-economic status, gender, and ethnicity (including white, black, Hispanic, Asian, and other ethnic groups). We used X_{hij}, $h = 1, \ldots, 7$, to denote these demographic measures. The level-3 model has 32 equations corresponding to the 32 coefficients at level 2:

$\beta_{q0j} = \gamma_{q00} + u_{q0j}$, for $q = 0, 1$;

$\beta_{q0j} = \gamma_{q00}$, for $q = 2, 3$;

$\beta_{qhj} = \gamma_{qh0}$, for $q = 0, \ldots, 3$; $h = 1, \ldots, 7$;

$$\begin{pmatrix} u_{00j} \\ u_{10j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\beta 00} & \tau_{\beta 00.10} \\ \tau_{\beta 10.00} & \tau_{\beta 10} \end{pmatrix}\right]. \quad \text{(A3)}$$

An empirical Bayes estimate of the reading pretest score for child *i* attending kindergarten class *j*, specified as

$$\pi_{0ij} = \gamma_{000} + \sum_{h=1}^{7} \gamma_{0h0} X_{hij} + u_{00j} + r_{0ij}, \quad \text{(A4)}$$

could be obtained from the level-2 residual file. This extrapolated pretest score had a mean of 19.53 and a standard deviation of 8.11. We then computed the mean and standard deviation of reading pretest for each kindergarten class.

Footnotes

[1] Simply including five of the six propensity scores as covariates along with the treatment indicators in a multiple regression model is not recommendable, because the assumptions of linearity and additivity are often unwarranted. Alternatively, if we stratify a sample on five of the six propensity scores, with five strata in each dimension, there will be as many as $5^5 = 3{,}125$ strata. Unless the sample is sufficiently large, data in most strata will be too sparse to allow for within-stratum estimation of all the causal estimands.

[2] Hong (2007) showed some relative strengths of the MMW method in reducing bias. In particular, IPTW estimation of treatment effects could be biased when units with no counterfactual information in the data receive a nonzero weight, or when the propensity model fails to represent a nonlinear relationship between pretreatment covariates and treatment assignment. The MMW method, due to its non-parametric approach, has a built-in procedure to avoid these problems. Moreover, because MMW is estimated as a ratio of the sample sizes within each stratum rather than as a direct function of the estimated propensity score itself, the MMW estimate of treatment effect remains robust even when a nonlinear propensity model is mis-specified as a linear one.

[3] Hong (2007) derived the following result:

$$
\begin{aligned}
E\{Y(z)\} &\approx E\left[E\{Y(z) \mid D(z)=1, s\}\right] \\
&= \sum_{s=1}^{K} E\{Y(z) \mid D(z)=1, s\} pr(s) \\
&= \sum_{s=1}^{K} \left(\sum_{i \in z,s} \frac{Y_i}{n_{z,s}} \right) \frac{n_s}{N} \\
&= \frac{1}{N} \sum_{s=1}^{K} \left\{ \sum_{i \in z,s} \left(\frac{n_s}{n_{z,s}} \right) Y_i \right\} ,
\end{aligned}
$$

where $D(z)$ is a dummy indicator that takes on a value of 1 if the unit is in treatment group z and 0 otherwise; $s = 1, \ldots, K$ indicates one of the K strata on the basis of the propensity score for treatment z; $pr(s)$ is the proportion of units in stratum s; n_s is the number of units in stratum s; $n_{z,s}$ is the number of units in stratum s who were actually assigned to treatment z; and N is the total sample size.

[4] We had obtained an empirical Bayes estimate of the class average reading growth rate from the level-3 residual file of the saturated model (see Equation 6).

Table 1

Potential Outcomes and Causal Estimands

Treatments	Labels	Potential Outcomes
$L0$	Low reading time with no grouping	$Y(L0)$
$L1$	Low reading time with low-intensity grouping	$Y(L1)$
$L2$	Low reading time with high-intensity grouping	$Y(L2)$
$H0$	High reading time with no grouping	$Y(H0)$
$H1$	High reading time with low-intensity grouping	$Y(H1)$
$H2$	High reading time with high-intensity grouping	$Y(H2)$

	Causal Estimands
Grouping Effects	$E[Y(H2) - Y(H0)]$; $E[Y(H1) - Y(H0)]$; $E[Y(L2) - Y(L0)]$; $E[Y(L1) - Y(L0)]$.
Time Effects	$E[Y(H0) - Y(L0)]$; $E[Y(H1) - Y(L1)]$; $E[Y(H2) - Y(L2)]$.

Table 2

Distribution of Kindergarten Classes in Six Treatment Conditions

		Reading Instruction time	
		Low	High
Intensity of Reading Ability Grouping	No Grouping	505	446
	Low-intensity Grouping	424	769
	High-intensity Grouping	375	295

Table 3

Descriptive Statistics of Reading Scaled Scores across the Six Treatment Conditions

	Fall 1998			Spring 1999		
Treatment conditions	*n*	*M*	*SD*	*n*	*M*	*SD*
No grouping, low reading time (*L0*)	3120	22.56	8.49	3492	31.87	10.11
Low-intensity grouping, low reading time (*L1*)	2432	22.97	8.83	2656	32.46	10.68
High-intensity grouping, low reading time (*L2*)	1523	22.21	8.55	1793	32.05	10.93
No grouping, high reading time, (*H0*)	2308	22.54	8.48	2600	32.37	10.11
Low-intensity grouping, high reading time (*H1*)	3191	22.39	8.13	3517	33.00	10.14
High-intensity grouping, high reading time (*H2*)	1059	22.20	8.58	1220	32.83	10.41
Total	13633	22.54	8.49	15278	32.42	10.35

Table 4

Comparison between Full sample and Analytic Sample

Student demographic features	Full Sample	Analytic Sample
Average age (months)	65.49	65.48
Proportion of girls	.49	.50
Proportion of white	.55	.54
Proportion of black	.15	.15
Proportion of Asian	.06	.07
Proportion of Hispanic	.18	.19
Proportion speaking English at home	.85	.84
Proportion of children receiving free/reduced lunch	.31	.33
Average socio-economic status	.01	-.01

Table 5

Balance in Logit of the Propensity Score for No Grouping and Low Reading Time

	No grouping, low reading time ($L0 = 1$)			Other treatment conditions ($L0 = 0$)		
Stratum	*N*	*M*	*SD*	*N*	*M*	*SD*
1	38	-3.00	0.44	688	-3.08	0.43
2	37	-2.07	0.19	342	-2.08	0.17
3	73	-1.39	0.24	348	-1.43	0.23
4	110	-0.58	0.27	167	-0.60	0.27
5	38	0.30	0.23	17	0.26	0.21
Total	296	-1.16	1.02	1562	-2.19	0.96

Table 6

Balance in Logit of the Propensity Score for Low-Intensity Grouping and Low Reading Time

	Low-intensity grouping, low reading time ($L1 = 1$)			Other treatment conditions ($L1 = 0$)		
Stratum	*N*	*M*	*SD*	*N*	*M*	*SD*
1	34	-3.01	0.47	686	-3.11	0.52
2	66	-1.95	0.21	453	-1.98	0.21
3	74	-1.27	0.22	295	-1.25	0.21
4	58	-0.65	0.14	93	-0.67	0.14
5	58	-0.06	0.26	41	0.00	0.28
Total	290	-1.26	0.95	1568	-2.21	0.98

Table 7

Balance in Logit of the Propensity Score for High-Intensity Grouping and Low Reading Time

	High-intensity grouping, low reading time ($L2 = 1$)			Other treatment conditions ($L2 = 0$)		
Stratum	*N*	*M*	*SD*	*N*	*M*	*SD*
1	39	-3.22	0.66	883	-3.27	0.66
2	57	-1.92	0.26	415	-1.93	0.24
3	48	-1.25	0.16	219	-1.26	0.16
4	48	-0.70	0.13	65	-0.73	0.14
5	48	0.02	0.35	36	-0.06	0.33
Total	240	-1.37	1.11	1618	-2.48	1.06

Table 8

Balance in Logit of the Propensity Score for No Grouping and High Reading Time

	No grouping, high reading time ($H0 = 1$)			Other treatment conditions ($H0 = 0$)		
Stratum	*N*	*M*	*SD*	*N*	*M*	*SD*
1	24	-3.03	0.39	466	-3.13	0.41
2	50	-2.20	0.24	477	-2.17	0.23
3	111	-1.41	0.24	414	-1.43	0.23
4	39	-0.81	0.11	86	-0.80	0.11
5	50	-0.38	0.14	80	-0.38	0.15
6	34	0.29	0.26	23	0.16	0.19
Total	308	-1.23	0.93	1550	-2.05	0.92

Table 9

Balance in Logit of the Propensity Score for Low-Intensity Grouping and High Reading Time

	Low-intensity grouping, high reading time ($H1 = 1$)			Other treatment conditions ($H1 = 0$)		
Stratum	*N*	*M*	*SD*	*N*	*M*	*SD*
1	16	-2.60	0.27	246	-2.69	0.30
2	54	-1.86	0.22	349	-1.83	0.21
3	94	-1.12	0.20	348	-1.15	0.20
4	41	-0.70	0.06	81	-0.69	0.06
5	191	-0.23	0.22	239	-0.25	0.22
6	122	0.58	0.29	77	0.51	0.29
Total	518	-0.48	0.87	1340	-1.33	0.94

Table 10

Balance in Logit of the Propensity Score for High-Intensity Grouping and High Reading Time

Stratum	High-intensity grouping, high reading time ($H2 = 1$)			Other treatment conditions ($H2 = 0$)		
	N	*M*	*SD*	*N*	*M*	*SD*
1	28	-3.64	0.56	856	-3.69	0.62
2	40	-2.19	0.28	463	-2.26	0.27
3	50	-1.42	0.18	197	-1.45	0.18
4	63	-0.68	0.27	120	-0.74	0.27
5	25	0.30	0.26	16	0.16	0.22
Total	206	-1.44	1.19	1652	-2.77	1.15

Table 11

Marginal Mean Weighted Sample of Kindergarten Classes with Low Reading Time and No Grouping (*L*0)

Stratum	*Unweighted Sample*			MMW	*Weighted Sample*
	L0 = 1	*L0* = 0	Total		*L0* = 1
1	38	688	726	3.04	115.66
2	37	342	379	1.63	60.38
3	73	348	421	0.92	67.07
4	110	167	277	0.40	44.13
5	38	17	55	0.23	8.76
Total	296	1562	1858	---	296

Table 12

Saturated Model Estimation of Treatment Effects without Weighting versus with Marginal Mean Weighting

Fixed Effect	Unweighted			Weighted		
	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*
Reading pretest						
Intercept for no grouping, low reading time (*L0*), γ_{001}	18.82	0.29	64.23***	18.82	0.32	58.51***
Low-intensity grouping, low reading time (*L1*), γ_{002}	0.13	0.43	0.31	-0.16	0.46	-0.35
High-intensity grouping, low reading time (*L2*), γ_{003}	-0.47	0.45	-1.04	-0.59	0.51	-1.14
Intercept for no grouping, high reading time (*H0*), γ_{004}	18.83	0.32	58.31***	18.86	0.34	56.00***
Low-intensity grouping, high reading time (*H1*), γ_{005}	-0.23	0.40	-0.58	-0.19	0.41	-0.47
High-intensity grouping, high reading time (*H2*), γ_{006}	-0.84	0.48	-1.76	-0.58	0.54	-1.06
Reading growth						
Intercept for no grouping, low reading time (*L0*), γ_{101}	1.54	0.03	52.25***	1.55	0.03	48.46***
Low-intensity grouping, low reading time (*L1*), γ_{102}	0.03	0.04	0.64	0.02	0.05	0.54
High-intensity grouping, low reading time (*L2*), γ_{103}	-0.01	0.05	-0.20	-0.03	0.05	-0.58
Intercept for no grouping, high reading time (*H0*), γ_{104}	1.60	0.03	50.55***	1.58	0.03	50.55***
Low-intensity grouping, high reading time (*H1*), γ_{105}	0.11	0.04	2.76**	0.11	0.04	2.76**
High-intensity grouping, high reading time (*H2*), γ_{106}	0.12	0.05	2.27*	0.15	0.06	2.64**
Random Effect	Variance	*df*	χ^2	Variance	*df*	χ^2
Level 3						
Class mean reading pretest, u_{00j}	13.91	1852	4207.03***	8.35	1852	4122.13***
Class mean reading growth rate, u_{10j}	0.12	1852	3418.89***	0.06	1852	3341.05***
Class-level correlation between pretest and growth		0.10			0.07	
Level 2						
Student reading pretest, r_{0ij}	50.61	8331	51193.11***	51.92	8331	51193.81***
Student reading growth rate, r_{1ij}	0.50	8331	23372.50***	0.52	8331	23373.54***
Student-level correlation between pretest and growth		-0.12			-0.12	

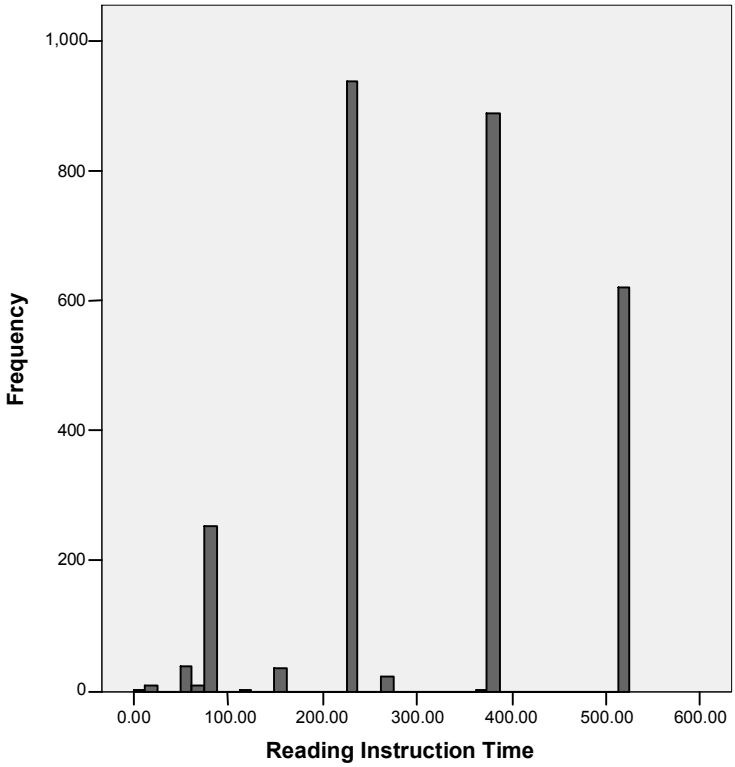
* $p < .05$; ** $p < .01$; *** $p < .001$

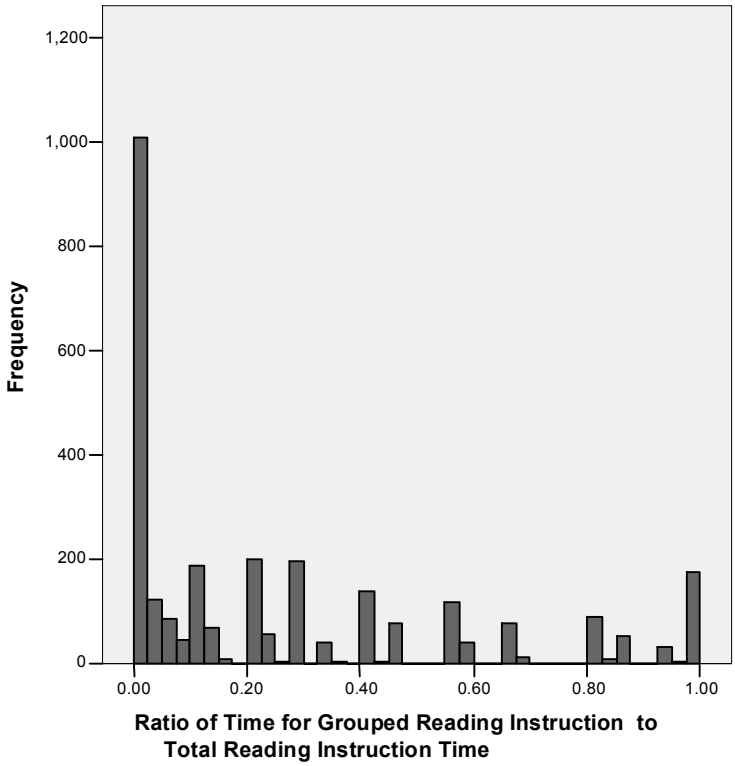
Figure Captions

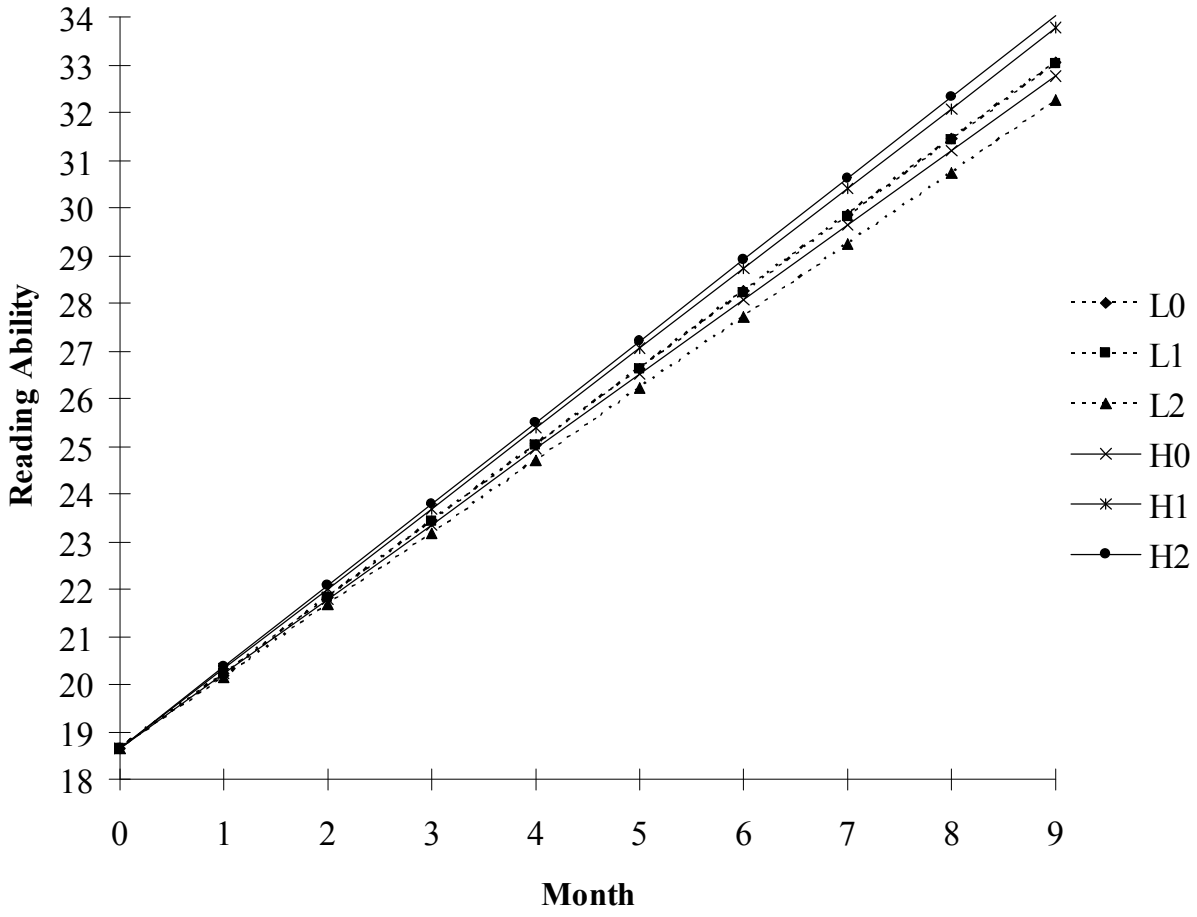
Figure 1. Distribution of Reading Instruction time.

Figure 2. Distribution of the Ratio of Grouped Instruction Time to Total Amount of Reading Instruction Time

Figure 3. Population Average Potential Reading Growth Trajectories Under Six Treatment Conditions


1,000
800
600
400
200
0
Frequency
0.00
100.00
200.00
300.00
400.00
500.00
600.00
Reading Instruction Time


1,200
1,000
800
600
400
200
0
Frequency
0.00
0.20
0.40
0.60
0.80
1.00
Ratio of Time for Grouped Reading Instruction to Total Reading Instruction Time



Note:

*L*0: Low reading time with no grouping;

*L*1: Low reading time with low-intensity grouping;

*L*2: Low reading time with high-intensity grouping;

*H*0: High reading time with no grouping;

*H*1: High reading time with low-intensity grouping;

*H*2: High reading time with high-intensity grouping.