Journal of Research on Educational Effectiveness, 3: 56–92, 2010
Copyright © Taylor & Francis Group, LLC
ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345740903277601

The Effects of Within-Class Ability Grouping on Academic Achievement in Early Elementary Years

Takako Nomi
University of Chicago, Chicago, Illinois, USA

Abstract: By incorporating two theoretical frameworks this study examines how school characteristics shape first-grade reading ability-grouping practices, and how this, in turn, affects students' reading achievement. The author uses the data from the Early Childhood Longitudinal Study and applies the propensity-score method to examine whether first-grade ability grouping improves student achievement, whether ability grouping increases achievement inequalities, and whether its effects vary by student initial abilities and/or school contexts. Findings support an argument that ability grouping is an organizational response to problems of diversity in the student body. Schools that use ability grouping are likely to have heterogeneous ability compositions. They are also public, low-performing, low socioeconomic status, and high-minority schools. In these schools, ability grouping has no effects or negative effects, particularly for low-ability students. In contrast, ability grouping may improve achievement for all students in schools with advantageous characteristics, mostly private schools, and may reduce achievement inequalities, because low-ability students benefit the most from this practice.

Keywords: Reading ability grouping, early elementary grades, propensity score method

Ability grouping has been a contentious issue in education. A central question includes whether ability grouping leads to high achievement for all or whether it unfairly limits educational opportunities for disadvantaged students, thus exacerbating existing educational and social inequalities. Historically, ability grouping emerged as an organizational response to problems because of diversity in the student body in terms of cognitive skills, maturity, and linguistic, cultural, and social backgrounds (Barr & Dreeben, 1983). Given these heterogeneities, ability grouping is thought to increase effectiveness in

Address correspondence to Takako Nomi, University of Chicago, Consortium on Chicago School Research, 1313 East 60th Street, Chicago, IL 60637, USA. E-mail: nomit@ccsr.uchicago.edu

organizing instruction and benefit all students because content is taught at the difficulty and pace that is commensurate with past student performance.

In early elementary classrooms, homogenous small-group reading instruction has been a widespread practice. Within-class ability grouping often involves teachers organizing students into small reading groups by their literacy skill levels that are determined by informal assessments, teacher judgment, and/or standardized tests (Schumm, Moody, & Vaughn, 2000). According to a recent national survey, about two thirds of first- through third-grade teachers reported using within-class homogenous grouping and most teachers reported that they use it to meet their students' instructional needs (Chorzempa & Graham, 2006).

The effectiveness of ability grouping, however, remains controversial. Many researchers argue that ability grouping exacerbates achievement inequalities through differential allocation of opportunities-to-learn and social-psychological factors (Barr & Dreeben 1983; Berends, 1994; Cohen, 1997; Eder, 1981; Gamoran, 1986; Hallinan, 1987; Oakes, 1985; Rist, 1970; Rosenbaum, 1976). For example, students in low-ability groups are taught less challenging materials and held to lower expectations by teachers than those in high-ability groups. Students in lower ability groups often spent more time on decoding than students in higher groups (Gambrell, Wilson, & Gantt, 1981). Also, students in lower groups spent more time reading orally than silently, which raised a concern about content coverage, and they were asked fewer comprehension questions and more literal questions than students in higher ability groups (Allington, 1984; Gambrell et al., 1981). Teachers expressed concern that students in lower ability groups spent more time involved in noninstructional activities (Chorzempa & Graham, 2006).

Past studies showed both positive and negative effects of ability grouping on student learning. Most research examined the effect of ability-group *placement* by comparing achievement by students' ability-group placement levels while statistically controlling for factors that were thought to influence both placement and achievement. These studies consistently showed that students in higher ability groups learned more than those in lower groups (Alexander, Cook, & McDill, 1978; Alexander & McDill, 1976; Berends, 1994; Gamoran, 1986, 1987; Gamoran & Berends, 1987; Gamoran & Mare, 1989; Hallinan & Kubitschek, 1999; Hauser, Sewell, & Alwin, 1976; Heyns, 1974; Hoffer, 1992; Lucas, 1999; Pallas, Entwisle, Alexander, & Stluka, 1994; Rosenbaum, 1980; Rowan & Miracle, 1983; Vanfossen, Jones, & Spade, 1987).[1] These

[1] The structure of ability grouping differs between secondary and elementary schools. Secondary schools typically use between-classroom ability grouping that are also differentiated by coursework, whereas elementary schools use within-class ability grouping (Barr & Dreeben, 1983; Hallinan & Kubitschek, 1999; Slavin, 1987). Regardless of these differences, findings of research on ability group placement are similar between secondary and elementary schools.

findings have often been used to support for "detracking,"[2] or nongrouping, because ability grouping seems to widen achievement inequalities by benefiting high-achieving students and hurting low-achieving students.

However, there are both conceptual and methodological limitations to this line of research. First, studies on ability-group placement typically use correlational analyses, which are potentially problematic in drawing causal inferences about ability-group placement as further discussed next. Second, even though studies on ability-group placement are often used to support "detracking," such an argument confuses issues of ability-group placement with issues of ability-grouping use. Most placement studies do not include students who are not ability grouped. Moreover, their analyses do not necessarily address an important policy question: "How would students achieve if they were not ability grouped?"[3]

Other studies specifically address the question about ability-group use by comparing achievement between ability-grouped and ungrouped students. Unlike placement studies, some studies on ability-grouping practices suggest that within-class ability grouping may lead to higher performance than non grouping for all students (Lou, Abrami, Spence, Poulsen, Chambers, & d'Apollonia, 1996; Slavin, 1987). These seemingly contradictory findings may be methodological. In particular, studies that used experimental or matched experimental designs showed positive effects of ability grouping practices for all students.

Although the use of experimental and matched experimental designs is likely to strengthen internal validity, there are also some limitations in prior studies on ability-grouping practices. First, experimental and matched experimental studies on ability-grouping practices did not use nationally representative samples, and their sample sizes were typically small. Thus, their results may not be generalizable to a larger or different study population, and positive ability-grouping effects might have been specific to the school context under study. In addition, most studies on ability-grouping practices focused on upper elementary grades, and few studies, in fact, examined within-class reading ability grouping in early elementary grades, which is when ability grouping is most commonly used.[4] Last, we know little about whether the consequences of "detracking" may vary by student initial abilities and school contexts.

[2] Although the term *detracking* typically refers to a secondary school curricular structure, I use this term to describe an organizational structure for instruction that does not group student by ability.

[3] A few studies (Hoffer, 1992; Tach & Farkas, 2006) examined the effect of ability group *placement* by including ungrouped students. However, their analyses compare the achievement of students who are placed in different ability levels to that of the *average* ungrouped students. It is not clear from these analyses how students placed in different ability groups would perform if they were not ability grouped.

[4] Slavin's (1987) meta-analysis included eight studies on elementary within-class ability grouping. Yet none of them analyzed reading grouping in early grades. A

To addresses these limitations, I apply propensity score methods using the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS–K). The following questions are addressed: (a) Does reading achievement differ between students who are grouped by ability and those who are not? (b) Do the effects of reading ability grouping vary by students' initial abilities? If so, do differential effects lead to greater achievement gaps between high and low achievers? (c) Do the effects of ability grouping vary by schools? (d) Do the effects of ability grouping vary by students' initial abilities *and* schools?

ABILITY GROUPING AND SCHOOL CONTEXTS

This study focuses on the importance of school contexts in understanding the use of ability grouping and how differences in school contexts, in turn, creates variability in its effects on student learning. Two theoretical frameworks may explain how school contexts affect the formation of ability grouping and instructional content, time, and pacing, and how this may produce differential effects of ability grouping on student learning by schools. First, organizational perspectives suggest that the formation of ability grouping is shaped by larger school and classroom contexts, and this, in turn, affects instruction provided to students (Barr & Dreeben, 1983, see also Hallinan & Sorensen, 1983). Barr and Dreeben argued that the structure of instructional groups (e.g., the size, number, and ability compositions) is primarily determined by the proportion of low-achieving students, class size, and cognitive and behavioral heterogeneities, but not by the mean student aptitudes. In particular, ability heterogeneity and the number of low-ability students in classes pose instructional challenges to meet the needs of all students. Thus, teachers tend to divide students into smaller instructional groups when classrooms are large and heterogeneous with many low-skill students. Group ability compositions shape instruction and learning because content and pace is typically tailored to the average students in the group (see, e.g., Beckerman & Good, 1981, Dar & Resh, 1986; Hallinan, 1987; Hallinan & Sorensen, 1983). Likewise, group number and size affect time allocated for particular instructional activities (e.g., supervised vs. unsupervised work), because, for example, teachers spend less time per group in classrooms with many groups.

Although organizational perspectives may explain how school contexts shape the organization of instruction and instructional activities, they may not adequately explain why some forms of ability grouping produces higher achievement and/or smaller achievement inequality than others. To explain

meta-analysis by Lou et al. (1996) aggregated studies on various forms of within-class grouping in different subjects from elementary to postsecondary levels. Of those studies, few were, in fact, studies on reading ability grouping in early elementary grades.

school differences in ability grouping effects, Sorensen (1970) provided an early structural analysis by identifying dimensions on which ability grouping may differ between schools. For example, schools may differ in "selectivity," which is defined as the degree of homogeneity created by ability grouping in terms of student cognitive characteristics (Sorensen, 1970).[5] Sorensen argued that more selective ability grouping leads to greater achievement inequalities because it exacerbates distinctions between groups by student abilities. For example, selective ability groups may create more unequal distributions of opportunities-to-learn. Selective ability grouping may also reinforce a self-fulfilling prophesy through adverse social-psychological effects, particularly on low-skill students, because it makes more salient distinctions between groups, which may lead to greater stigmatization.[6]

Ability-group selectivity may be particularly consequential for low-skill students. Selective grouping would put low-achieving students together, where as with unselective groups they would be grouped with higher achieving students than themselves. Because instruction is typically targeted at the average skill level of the group, instruction in a selective group may be slower and learning environment may be more disruptive than an unselective group where higher ability students are at presence. Previous studies indeed suggest that homogeneous grouping leads to lower achievement than heterogeneous grouping among low-ability students (Lou et al., 1996). In addition, having many groups means less supervised instructional time for each group and teacher may spend more time in higher groups because more difficult materials typically require

[5] Other concepts include inclusiveness, electivity, and scope (see also Oakes, 1985; Rosenbaum, 1976, 1984). Sorensen (1970) focused on high school tracking and some concepts may not be applicable to elementary school contexts.

[6] Nomi (2006a), using the ECKS–K data, incorporated these two theoretical frameworks and examined relationships among classroom contexts, ability group structure, and student achievement. Specifically, Nomi (2006a) examined how classroom academic compositions are related to the number of ability groups in classrooms and group academic compositions, and how these aspects of ability group structure are related to students' learning. Results showed that teachers divided students in many small groups when classroom academic compositions were heterogeneous with many low-skill and language minority students. Such classrooms created greater distinctions among students by ability groups than classrooms that used fewer groups. For example, classrooms with five or more groups had much larger between-group differences in incoming reading skills than classrooms with two ability groups although the average class size differed only by one student. Thus, classrooms with many groups tend to create more homogenous ability composition within groups and greater distinctions across groups (i.e., greater selectivity) than classrooms with only a few groups. Furthermore, classrooms with more groups were associated with lower student achievement and greater achievement inequalities in the end of first grade (Nomi, 2006a).

more instructional time (Barr & Dreeben, 1983). This may have more detrimental effects on low-ability students than high-ability students.

School contexts may also affect the *use* of ability grouping. In small schools or schools with homogeneous ability compositions with many high-achieving students, teachers or principals may find it unnecessary to use ability grouping. However, when such schools use ability grouping, they would create less selective groups given the student composition of their school. Such schools may produce higher student performance or smaller achievement inequalities than schools with more selective ability grouping, which are likely to be large and heterogeneous with many low-ability students. Previous studies have not addressed whether school characteristics are related to the use of ability grouping, which needs to be taken into account to make valid causal inferences of ability-grouping effects.

METHODOLOGICAL LIMITATIONS IN ABILITY GROUP PLACEMENT RESEARCH

This study uses the potential outcomes framework (Holland, 1986; Rubin 1978, 2005) as the conceptual framework for causal inference. The key assumption of Rubin's causal model is that, using the language of experiment, participants assigned to treatment and control conditions have potential outcomes in both states, and the causal effect is defined as the difference in potential outcomes between these two states. The definition of a causal effect under this framework makes the stable unit treatment value assumption (SUTVA). SUTVA has two components: (a) one's potential value associated with each treatment is not affected by how and what treatment is assigned to other participants, and (b) there is a single version of each treatment (Rubin, 1986; also see Cox, 1958). The first component of SUTVA means that there is no interaction between units in a way that affects outcomes. This assumption is often problematic in school settings particularly when an intervention is given to students or teachers within schools. For example, the treatment may have "spillover" effects on other students or teachers because they are likely to interact with one another. If potential outcomes are affected due to interactions among students and teachers, the first component of SUTVA does not hold. The second component of SUTVA means that there is no variation in the treatment that is administered to the treatment group and likewise to the control group. This may also difficult to assume in school settings. For example, even though the same intervention may be assigned to teachers or schools, they may implement the intervention differently (Stuart, 2007).

Most ability-grouping research that used observational data applied linear covariate adjustments to estimate the effect of ability-group placement, and there are several limitations in such approaches. First, problems may arise when participants find no comparison groups. Most notably, students' initial abilities differ by ability-group levels. Students in a low-ability group may have few comparison students who are just like them but are, in fact, assigned to a high-ability group. In such cases, to predict what might happen to students in a low-ability group when they are placed in a high-ability group, linear covariate adjustments rely on extrapolation, which does not come from the data but is based on a linear model assumption to compensate for the lack of a viable comparison group.

Second, most ability-group placement research assumes that the benefit from being placed in a high-ability group rather than a low-ability group is the same for low- and high-achieving students (see Gamoran & Mare, 1989; Hoffer, 1992, for discussions). This inference is particularly problematic if few low ability students are found in a high-ability group, or vice versa. The treatment effect may also depend on many other student and school characteristics. However, interaction effects are particularly difficult to estimate using linear regression models as the number of covariates increases.[7]

To address these limitations, some researchers applied propensity score methods and used students who were not ability grouped as a comparison group. Hoffer (1992) and Betts and Shkolnik (2000) compared achievement between students who were placed in different ability levels and their ungrouped counterparts with similar propensities of being placed in a given ability level. However, under the potential outcomes framework, these studies are likely to violate the first component of SUTVA just described. Both studies used data in which students were nested within schools. Yet they did not consider how learning experiences of a particular student may be affected not merely by his or her own placement but the placement of other students. For example, learning climates and peer influences are thought to be more conducive to learning in high-ability groups than low-ability groups (Hallinan, 1987; Page, 1991). Also the pace and difficulty of instruction is primarily determined by ability group compositions (Barr & Dreeben, 1983). This suggests that student achievement is affected not only by his or her own placement but also by how other students are assigned to particular groups. If the assignment of other students influences the outcome of a student the SUTVA does not hold.

One way to avoid violation of the SUTVA is to assign the treatment at the cluster level. By assigning the treatment (e.g., ability grouping) to schools, it is more defensible to assume that the treatment status of one school does not affect potential outcomes of students in other schools although the second component of SUTVA—a single version of the treatment or of the control—may be violated if schools implement ability grouping differently. This study attempt to simulate

[7] One main advantage of using the propensity score method is that it reduces the dimensionality of pretreatment covariates by providing a scalar summary of covariate information.

the cluster randomized trail and methodological assumptions under this study are further discussed next.

APPLICATION OF PROPENSITY SCORE METHODS USING ECLS-K DATA

This study employs propensity score methods to estimate the effect of ability grouping practices (i.e., not the effect of placement) using the ECLS–K data. This analysis has several advantages over previous studies. First, the analysis of ability-grouped versus ungrouped students directly addresses the question of ability grouping practices (i.e., how students would perform if they are ability grouped). Second, this study examines how school contexts affect the use of ability grouping and whether ability grouping effects differ by both student initial abilities and school contexts. Third, this study uses a design to avoid violation of the first component of SUTVA as further discussed next.

Data and Sample

The ECLS–K began in the fall of 1998 with a nationally representative sample of 21,260 children in about 1,200 kindergarten programs. The base year data collection employed a multistage probability sample design. First, 100 Primary Sampling Units—geographic areas which consist of counties or group of counties—were selected. Second, 1,277 public and private schools were sampled within the Primary Sampling Units. Finally, approximately 23 children were randomly sampled within each school (West, Denton, & Reaney, 2001). Of the base sample of 21,260 children, 17,487 were sampled for a longitudinal study and surveyed in fall 1998 (Fall-K), spring 1999 (Spring-K), and spring 2000 (Spring-1st). The fall 1999 assessment (Fall-1st) drew 30% subsample of the longitudinal sample of 17,487 children. This study uses data from spring and fall kindergarten assessments and spring first-grade assessment. The analytic sample consists of 13,512 students.[8]

Ability-Grouping Measures

In the ECLS–K first-grade teacher questionnaire, the classroom teacher of each sampled student was asked about the number of ability groups in child classroom and his or her ability-group placement. Response categories for the

[8] School movers were excluded if their first-grade schools are not among the originally sampled schools.

Table 1. Ability grouping status of students, classrooms, and schools

Student Grouping Status	School Grouping Status			
	Grouped	Ungrouped	Mix[a]	Total
Ability grouped	6,742		2,848	9,590
(Classroom *N*)	(1,470)		(838)	(2,389)
Ungrouped		2,043	1,879	3,922
(Classroom *N*)		(271)	(513)	(817)
Total (School *N*)	6,742	2,043	4,727	13,512
	(451)	(142)	(307)	(900)

[a] Schools with both grouped and ungrouped classrooms.

number of ability groups in child classroom ranged from zero (no grouping) to five or more, and 40% of all classrooms with ability grouping used two to three groups. Response categories for child placement included "not applicable" and Placement Level 1 through 8. These two items were used to cross validate ability-grouping status of students and classrooms, which were then used to construct ability grouping status of schools (Table 1).[9]

The ECLS–K data consist of three types of schools. The first type is defined as "ability-grouped schools" where all students in the same school are in ability grouped classrooms, and 6,742 students attend these schools. The second type of school is defined as "ungrouped schools" where no student in the same school is in ability-grouped classrooms and 2,043 students attend the school of this type. The third type is defined as "mixed schools," which have students in both ability-grouped and ungrouped classrooms. In "mixed schools" 2,848 students are in ability-grouped classrooms and 1,897 students are in ungrouped classrooms. The number of schools identified as grouped, ungrouped, and mixed schools is 451, 142, and 307 schools, respectively.

Other Measures

The dependent variable is item response theory (IRT)-scaled reading scores in spring first grade, which comes from the direct child cognitive assessment. Numerous school and classroom covariates were examined as candidates to estimate the probability of schools or classrooms using ability grouping. Variable selection was guided by theories and preliminary analyses. Ability grouping is

[9] When there was a disagreement in teacher responses (e.g., teachers responded two for the number of groups in child classroom but responded greater than two for child placement), decision was made based on responses for other students in the same classrooms. However, such disagreement was less than 1%.

considered an organizational response to problems of diversity in the student body (Barr & Dreeben, 1983). Middle-class parents tend to support ability grouping because they perceive that their children may benefit by being placed in higher ability groups (Oakes, Quartz, Ryan, & Lipton, 2000). Schools may use admission and retention policies to regulate characteristics of incoming students. In addition, other variables should be taken into account if they are correlated with both student achievement and school ability grouping practices.

To construct school and classroom contextual variables, approximately 100 student-level variables were aggregated at the classroom and school levels. For each student variable, classroom and school means and standard deviations were computed to capture the average characteristics and heterogeneities of classrooms and schools. Student-level variables include students' socioeconomic and demographic characteristics, cognitive skills, English proficiencies, parental and teacher's ratings on children's cognitive skills and behavior, parental involvement in educational activities at home and expectations for child education, educational resources at home, and preschool experiences. These variables come from teacher and parent questionnaires in spring and fall kindergarten.[10]

Teacher measures come from teacher questionnaires, which include teaching experiences, educational levels, attitudes about teaching, and relationships with other school staff. These measures are also aggregated at the school level by taking the average of these variables. Measures from administrator questionnaires include school size, sector, student demographic characteristics, student mobility, teacher demographic characteristics, teacher absentee problems, school safety, computer availability, school retention and admission policies, and frequencies of PTA meetings and other school activities (the variable list is available upon request).

School Ability Grouping Status and School Characteristics

We first describe the extent to which schools differ in key academic and demographic characteristics by their ability-grouping status (Table 2). Variables on school academic compositions—school average abilities and ability heterogeneities—include the means and standard deviations of the following child specific measures; IRT-scaled spring kindergarten reading scores, teacher ratings on Academic Rating Scales (ARS) on language and literacy skills and child Approaches to Learning. The ARS variable is a composite measure, which consists of several items on child's proficiency in speaking, listening, early reading, and writing, and response categories range 1 (*not yet*), 2 (*beginning*), 3 (*in progress*), 4 (*intermediate*), and 5 (*proficient*). The Approach

[10] These aggregate measures are adjusted for the oversampling of Asian students.

Table 2. Descriptive statistics by school ability grouping status: Selected characteristics

	School Ability Grouping Status					
	Ability Grouped		Ungrouped		Mix	
Variables	*M*	*SD*	*M*	*SD*	*M*	*SD*
M IRT Reading	32.35	5.4	33.51[a]	5.84	31.43[a]	5.1
SD IRT Read	8.67	2.37	8.15[a]	2.64	8.53	2.27
M literacy	3.22	0.74	3.20	1.1	3.10[a]	0.63
SD literacy	0.68	0.19	0.59[a]	0.22	0.68	0.19
M approach to learning	2.96	0.51	2.99	0.71	2.94	0.41
SD approach to learning	0.61	0.14	0.58[a]	0.15	0.64[a]	0.14
% non-English speaking	0.14	0.23	0.07[a]	0.14	0.14	0.22
M SES	0	0.51	0.18[a]	0.48	−0.09[a]	0.48
% Hispanic	0.19	0.27	0.12[a]	0.18	0.19	0.26
% Black	0.15	0.26	0.09[a]	0.21	0.18	0.28
Public school	0.79	0.41	0.44[a]	0.5	0.91[a]	0.29

Note. IRT = item response theory; SES = socioeconomic status.
[a]Statistically different from ability-grouped schools at $p < .05$.

to Learning variable is also a composite variable with six items on child's attentiveness, task persistence, eagerness to learn, learning independence, flexibility, and organization where response categories range from 1 (*never*) to 4 (*very often*). School demographic variables include mean socioeconomic status (SES), which is constructed by taking the school average of child SES,[11] the percentage of students from non-English speaking homes, the percentage of Black and Hispanic students, and a variable on whether school is public or private.

School characteristics differ considerably by school ability-grouping status. In general, ungrouped schools have the most advantageous characteristics, and "mixed schools" are the least advantageous schools. In terms of achievement characteristics, ungrouped schools have the highest average literacy skills and the smallest cognitive and behavioral heterogeneities of all schools. Ungrouped schools also have the highest SES and the lowest percentage of students from non-English-speaking homes and minority students. Also, 44% of ungrouped schools are public schools, compared to 79% and 91% for ability-grouped and mixed schools, respectively. Although many school characteristics

[11] The ECLS–K data provide a composite measure of child SES with a mean of 0 and SD of one, which includes information on parental education and occupation and household income.

in Table 2 are significantly different between "mixed" and ungrouped schools, mixed and ability-grouped schools are similar in the percentage of Hispanic and Black students, the percentage of students from non-English speaking homes, and cognitive heterogeneities.

Substantive differences between ability-grouped/mixed schools and ungrouped schools imply that classroom characteristics are also dissimilar between these two types of schools. Also, additional analyses (results not shown) found that much variation in characteristics between ability-grouped and ungrouped classrooms were, in fact, attributable to between-school differences rather than between-classroom differences. Furthermore, it is rather school factors than classroom factors that predict whether classrooms use ability grouping. Particularly, after differences in school characteristics were taken into account, ability-grouped classrooms in "mixed schools" found almost no comparison classrooms in ungrouped schools and their comparison classrooms were primarily found within "mixed schools." Similarly, classrooms in ungrouped schools found their comparison classrooms mainly from ability-grouped schools and few came from grouped classrooms in mixed schools. Because of a lack of overlap in school characteristics between mixed schools and ungrouped schools, it is not appropriate to make comparisons between classrooms in mixed schools and classrooms in ungrouped schools, and students in ungrouped schools can only be compared to students in grouped schools with similar school characteristics.

Propensity Score Method

Although three types of schools differ in many school characteristics, it is not known from the ECLS–K data whether ability grouping is a classroom- or school-level policy. If ability grouping is a classroom-level policy, one may design a study by viewing classrooms as a unit of treatment assignment. However, because of a lack of comparison groups between classrooms in mixed schools and ungrouped schools with similar school characteristics, I propose a research design that attempts to simulate a cluster randomized experiment where schools are randomly assigned to ability-grouped settings. An advantage of this school-level analysis is that it minimizes the violation of the first component of SUTVA because schoolwide ability grouping in a given school is not likely to influence the achievement of students in other schools. The analysis of school-level ability grouping consists of ability-grouped and ungrouped schools. Although this analysis attempts to generalize the results to the population of the U.S. schools that use schoolwide ability grouping, generalizability is limited because the analytic sample does not necessarily represent this population.[12] The analysis of mixed schools was conducted separately

[12] Different sampling designs (e.g., stratified sampling by school ability grouping status) may be needed to warrant the generalizaiblity to the population. Also because

and results are only briefly mentioned here (full results are available upon request).

However, this analysis is likely to violate the second component of SUTVA—there is only a single version of treatment, because ability-grouping practices are thought to vary by school characteristics as discussed earlier. It should be recognized that ability grouping is not a static practice. Ability grouping varies not only in terms of its structure but also in the amount of time spent in groups and flexibility. Some teachers may combine ability grouping with whole-class instruction (Baumann, Hoffman, Duffy-Hester, & Ro, 2000). Also, some teachers may move students from one group to another as students make progress or lack of it. A limitation of this study is that it does not address how schools implement ability grouping. Instead, this study views ability grouping as the set of practices carried out by schools (Stuart, 2007) and explicitly addresses school characteristics that influence the use of ability grouping in estimating its average effects on student outcomes. In addition, this study makes an "intact schools" assumption where the average treatment effects are generalizable to schools, holding current student membership constant (Hong & Raudenbush, 2007).

Propensity scores are defined as estimated probabilities of schools adopting ability grouping practices. They are specified as a function of observed school-level covariates,

$$\mathrm{P_j} = \mathrm{Prob}(\mathrm{T_j} = 1|\mathrm{W_j}),$$

where $\mathrm{P_j}$ is a propensity score of adopting ability grouping for school *j*, $\mathrm{T_j}$ is the school-level treatment, indicating whether school *j* practices ability grouping, and $\mathrm{W_j}$ is school-level predictors of ability grouping. The propensity scores are estimated using logistic regressions. Propensity score stratification methods are used to construct comparison groups (Rosenbaum & Rubin, 1984). I apply a strategy suggested by Becker and Ichino (2002) and Dehejia and Wahba (2002), which subdivides the sample into strata with an equal propensity score interval (e.g., 0–0.2, 0.2–0.4 . . . . . . 0.8–1.0). When the mean propensity scores are not balanced between the treatment and comparison groups in a given stratum, the sample is further divided in half in that stratum.[13]

The casual effect of ability grouping is defined as the average difference in potential first-grade reading scores between students in ability-grouped schools

this study uses only a subset of the population (i.e., students in schools with and without ability grouping), it did not use weight variables provided by the ECLS–K.

[13] Alternatively, one can divide the sample into quintile. When the propensity model is property specified this method removes 90% of the bias associated with the treatment assignment (Cochran, 1968; Imbens, 2004; Rosenbaum & Rubin, 1984). Preliminary analyses used both stratification methods and the pattern of results were similar (results are available upon request).

and those in ungrouped schools, given the propensity of schools practicing ability grouping. In estimating the effects of ability grouping, covariate adjustments are used to reduce error variance in the outcome and gain precisions of the estimates (Krueger & Zhu, 2004; Rubin, 1974). School-level covariates include the logit of the estimated propensity scores, school mean kindergarten reading scores constructed by taking the school average of child IRT-scaled kindergarten reading scores, and a dichotomous variable on school sector (1 = private and 0 = public). Student covariates include a dichotomous variable on gender (1 = male, 0 = female); age in months; its square term; a dichotomous variable indicating whether a child changed the school; IRT-scaled spring kindergarten reading scores; their square terms; a set of dummy variables on child race distinguishing White (omitted category), Black, Hispanic, Asian, and other; a child SES variable (a composite variable provided by the ECLS–K data; *M* = 0, *SD* = 1); and assessment dates. Including these students' covariates would also help reduce bias if they are associated with school ability grouping status after propensity score adjustment on school characteristics.

To examine the first research question—whether reading achievement differs between ability-grouped and ungrouped students, the following model is estimated.

$$Y_{ij} = \delta T_j + \sum_{p=1}^{P} (\beta_p X_p)_{ij} + \sum_{m=1}^{M} (\gamma_m D_m)_j + \sum_{n=1}^{N} (\gamma_{M+n} W_n)_j + e_{ij} + u_j, \qquad (1)$$

where Y_{ij} is the first-grade reading scores for student *i* in school *j*, δ is the effect of ability grouping T_j ($T_j = 1$ if school *j* uses ability grouping and $T_j = 0$ otherwise), X_{pij} is a vector of student covariates where p = 1,2 . . . P, D_{mj} is a set of dummy variables indicating propensity stratum *m* for school *j* where m = 1, 2 . . . M, W_{nj} is a vector of school covariates where n = 1,2 . . . N, and $e_{ij,}$ and u_j are, respectively, student- and school-level error terms.

The second analysis examines whether ability grouping has differential effects by student initial ability levels. The initial ability levels are measured by dividing the students into three ability levels—low, middle, and high, based on the overall distributions of spring kindergarten reading scores, and a set of dummy variables are created accordingly. The effects of ability grouping on low, middle, and high achievers are estimated as

$$Y_{ij} = \beta_1(\mathrm{Low})_{ij} + \beta_2(\mathrm{Middle})_{ij} + \beta_3(\mathrm{High})_{ij} + \sum_{p=1}^{P} \beta_{3+p} X_{pij} + e_{ij},$$

$$\beta_1 = \delta_1 T_j + \sum_{m=1}^{M} (\gamma_{1m} D_m)_j + \sum_{n=1}^{N} (\gamma_{1(M+n)} W_{1n})_j + u_{1j},$$

$$\beta_2 = \delta_2 T_j + \sum_{m=1}^{M} (\gamma_{2m} D_m)_j + \sum_{n=1}^{N} (\gamma_{2(M+n)} W_{2n})_j + u_{2j},$$

$$\beta_3 = \delta_3 T_j + \sum_{m=1}^{M} (\gamma_{3m} D_m)_j + \sum_{n=1}^{N} (\gamma_{3(M+n)} W_{3n})_j + u_{3j}, \tag{2}$$

where Y_{ij} is the first-grade reading score for student *i* in school *j*; δ_1, δ_2, and δ_3 are the effects of ability grouping T_j for low, middle, and high initial ability students, respectively; X_{pij} is a vector of student covariates where p = 1,2 . . . . P; D_{mj} is a set of dummy variables indicating propensity stratum *m* for school *j* where m = 1, 2 . . . M; W_n is a vector of school covariates where n = 1,2 . . . . N; and e_{ij} and $u_{\cdot j}$ are, respectively, student- and school-level error terms. In Equation 2, dummy variables indicating child initial ability levels (low, middle, and high) do not have an omitted category. Thus, students' first-grade reading scores are estimated separately at each ability level (β_1, β_2, and β_3) and ability grouping effects are also estimated separately at each ability level ($\delta_{1,}$ $\delta_{2,}$ and δ_3).

To formally test differential effects of ability grouping by student initial ability levels, a hypothesis $\delta_1 = \delta_{2,} = \delta_3 = 0$ is tested. Differences in the magnitude of ability-grouping effects for low, middle, and high achievers indicate how ability grouping affects achievement inequality. For example, if high-ability students in ability-grouped schools have higher reading scores than their counterparts in ungrouped schools whereas low-ability students in ability-grouped and ungrouped schools have similar test scores, this suggests that ability grouping widens achievement inequality relative to no grouping because high-ability students learn more in the ability grouped setting.

The third analysis examines how ability grouping effects vary by school characteristics defined by the strata. This is done by estimating stratum-specific effects by interacting the treatment variable and propensity strata dummy variables. This is expressed as

$$Y_{ij} = \sum_{m=1}^{M} [\delta_m (D_m * T)_j] + \sum_{p=1}^{P} (\beta_p X_p)_{ij} + \sum_{m=1}^{M} (\gamma_m D_m)_j + \sum_{n=1}^{N} (\gamma_{M+n} W_n)_j + e_{ij} + u_j, \tag{3}$$

where Y_{ij} is the first-grade reading scores for student *i* in school *j*, δ_m is the effect of school-level ability grouping T_j for stratum *m* where m = 1, 2 . . . M, D_{mj} is a set of dummy variables indicating propensity stratum *m* for school *j* where m = 1, 2 . . . M, X_{pij} is a vector of student covariates where p = 1,2, . . .

P, W_n is a vector of school covariates where n = 1,2 . . . . N, and e_{ij} and u_j are, respectively, student- and school-level error terms.

The final analysis examines how the effects of ability grouping on low, middle, and high achievers vary by school characteristics defined by the strata. This is done by comparing achievement between ability-grouped students and ungrouped students at each initial ability level within each stratum. This is written as

$$Y_{ik} = \beta_1(\mathrm{Low})_{ij} + \beta_2(\mathrm{Middle})_{ij} + \beta_3(\mathrm{High})_{ij} + \sum_{p=1}^{P} \beta_{3+p}X_{pij} + e_{ij},$$

$$\beta_1 = \sum_{m=1}^{M} [\delta_{1m}(D_m * T_j)] + \sum_{m=1}^{M} (\gamma_{1m}D_m)_j + \sum_{n=1}^{N} (\gamma_{1(M+n)}W_{1n})_j + u_{1j},$$

$$\beta_2 = \sum_{m=1}^{M} [\delta_{2m}(D_m * T_j)] + \sum_{m=1}^{M} (\gamma_{2m}D_m)_j + \sum_{n=1}^{N} (\gamma_{2(M+n)}W_{2n})_j + u_{2j},$$

$$\beta_3 = \sum_{m=1}^{M} [\delta_{3m}(D_m * T_j)] + \sum_{m=1}^{M} (\gamma_{3m}D_m)_j + \sum_{n=1}^{N} (\gamma_{3(M+n)}W_{3n})_j + u_{3j}, \quad (4)$$

where Y_{ij} is the first-grade reading score for student *i* in school *j*; δ is the effect of ability grouping T_j, which is estimated for each achievement group in the strata *m* where m = 1, 2 . . . M; D_{mj} is a set of dummy variables indicating propensity stratum *m* for school *j* where m = 1, 2 . . . M; X_{pij} is a vector of student covariates where p = 1,2 . . . P, W_n is a vector of school covariates where n = 1,2 . . . . N; and e_{ij} and $u_{.j}$ are, respectively, student- and school-level error terms.

Similar to Equation 2, dummy variables on students' initial ability levels (low, middle, and high) do not have an omitted category; thus, first-grade reading achievement is estimated separately at each ability level. For each initial ability level, the effects of school ability grouping ($\delta_{1m,}$ $\delta_{2m,}$ and δ_{3m}) are estimated within each propensity stratum, which is represented by the interaction between the stratum dummy variable D and the treatment indicator T.

For students in each stratum, a formal hypothesis is tested to see whether ability grouping has differential effects by student initial ability levels ($\delta_{1m} = \delta_{2m,} = \delta_{3m} = 0$ for m^{th} stratum). This analysis illuminates whether the effects of ability grouping on achievement inequalities vary by school characteristics that are defined by the stratum D_m. For example, to see whether ability grouping increases achievement inequalities between low and high initial ability students, I test a hypothesis, $\delta_{1m} = \delta_{3m,}$ within each stratum.

Additional Causal Assumptions

To estimate treatment effects without bias, propensity score methods assume that the covariates used in the propensity model are not affected by the treatment assignment. A limitation of this study is that we do not know when schools or teachers, in fact, adopted ability grouping. This is a common limitation in survey research where researchers cannot control the timing of policy implementation. For example, in the study of kindergarten retention policies, using the ECLS–K data, Hong and Raudenbush (2005) used covariates measured in kindergarten to predict propensities of schools adopting retention policies.

It is likely that schools had been using ability grouping before the ECKL–K data collection. Prior research suggests that ability grouping is primarily a school or classroom response to their student characteristics (Barr & Dreeben, 1983). By using variables measured in kindergarten to predict whether schools practice ability grouping in first grade, this study assumes that the kindergarten covariates measure the average characteristics of incoming students in the schools, and kindergarten student samples provide unbiased estimates of the student characteristics that led to the initial adoption of ability grouping. Also, kindergarten characteristics are assumed to be unaffected by first-grade ability grouping practices.

RESULTS

Estimation of the Propensity of a School to Adopt Ability Grouping

The preliminarily analysis examined bivariate associations between school characteristics and school ability grouping status. Overall, 107 variables and 13 sets of dummy variables were significantly associated with school ability grouping at the probability level of .05 (see Nomi, 2006b, for a list of all variables).[14] As described further next, schools using ability grouping had, in general, more disadvantageous characteristics than schools without ability grouping. It is also noted that cognitive heterogeneities were more important predictors of ability grouping when they were measured by teacher ratings than test scores or parent ratings. In addition, there were more variables on cognitive and behavioral heterogeneities than their average characteristics that predicted the use of ability grouping. These findings support a claim that ability grouping is a school (or classroom) response to student diversity.

To estimate the propensity of a school adopting ability grouping, covariates were selected through a stepwise logistic regression (see Hong & Raudenbush, 2005). The handling of missing covariates followed a method used by

[14] A total of 246 covariates are tested.

Rosenbaum (1986), who used mean imputations for the treated and control groups accordingly.[15] The final propensity model used 18 variables, three quadratic terms, and four sets of dummy variables. School academic compositions include the mean kindergarten IRT scores in reading and its square term, the mean kindergarten IRT scores in general knowledge, the mean ARS scores on literacy, the percentage of children who were "not yet" or "beginning" to be able to use complex sentence structure based on teacher ratings, and the percentage of children whose parents responded that their child's ability to articulate is better than other students of his or her age. School ability and behavioral heterogeneities were measured by the standard deviations of four students' variables—reading IRT scores, ARS scores on literacy, teacher ratings on the child's ability to use strategy to read unfamiliar words, and teacher ratings on the child's ability to use a computer for various goals.[16]

School socioeconomic and demographic variables included the percentage of students whose mothers did not have a high school diploma, the percentage of students who have ever been on Aid to Families with Dependent Children, the percentage of students who were eligible for free lunch, the percentage of students whose mothers' occupational prestige scores were not applicable, the mean number of residential moves and its square term, the percentage of Hispanic students and its square term, a variable indicating whether schools have limited English proficiency students, and the percentage of students who failed the Oral Language Development Scales.[17] One school admission policy variable—whether kindergartens require SAT scores—was also included. Four sets of dummy variables were the percentage of students who spoke a non-English language at home (0% as an reference category, 0.1–7.0%, 7.1–25.0%, and above 25%), school type (private, nonregular public, and public schools as a reference category), regions (West, South, Midwest, and East as a reference category), and the total school enrollment numbers (0–149 students as a

[15] Although mean imputation may produce biased estimates and underestimate standard errors, this would be less problematic in estimating propensity scores because the purpose here is to achieve covariate balance rather than to produce unbiased estimates. However, a concern should be raised when covariates have many missing cases. In this study, of the 21 covariates in the final propensity model 4 covariates have 5 to 10% missing cases, 3 covariates have 10 to 15% missing cases, and 14 covariates have less than 5% missing cases. Also, more than 95% of all missing dummy indicators were balanced between ability-grouped and ungrouped schools within each stratum at the probability level of .05.

[16] Response categories for the child's strategy to solve math problems and the use of a computer are 1 (*not yet*), 2 (*beginning*), 3 (*in progress*), 4 (*intermediate*), and 5 (*proficient*).

[17] The Oral Language Development Scales is a screen test given to children who had a non-English-language background to determine whether they understood English well enough to receive the direct cognitive assessments in English.

Table 3. Estimated propensity scores and logit of estimated propensity scores by school ability grouping status

	Ability-Grouped Schools			Ungrouped Schools		
	M	*SD*	Range	*M*	*SD*	Range
Prop Scores	0.83	0.17	.13–1.0	0.55	0.25	0.02–1.0
Logit Prop Scores	2.07	1.37	−1.87–6.34	0.25	1.43	−4.00–4.96

reference category, 150–299 students, 300–499 students, 500–749 students, and 750 or more students).

The mean estimated propensity scores for ability-grouped and ungrouped schools were, respectively, .83 and .55 (Table 3). The range was from .13 to 1.0 for ability-grouped schools and from .02 to 1.0 for ungrouped schools. The means of the logit of the estimated propensity scores for ability-grouped and ungrouped schools were, respectively, 2.07 and .25. The range was from –1.87 to 6.34 for ability-grouped schools and from –4.00 to 4.96 for ungrouped schools.

Results of Propensity Score Stratifications: Differences in School Characteristics by Propensity Strata

The result of propensity score stratification was shown in Table 4 where the highest stratum (Stratum 6) consists of schools that were most likely to use ability grouping and the lowest stratum (Stratum 1) consists of schools that

Table 4. Propensity score stratification

	Propensity Scores		Ability-Grouped Schools		Ungrouped Schools	
Stratum	*M*	Range	School *N*	Student *N*	School *N*	Student *N*
6	0.91	0.80–1.0	309	4, 599	23	325
5	0.75	0.70–0.80	64	956	20	306
4	0.65	0.60–0.70	32	514	22	324
3	0.51	0.40–0.60	27	406	33	510
2[a]	0.32	0.20–0.40	18	258	28	374
1[a]	0.11	0.02–0.20	1	9	16	204
Total			451	6, 742	142	2, 043

[a] The subsequent analyses combine the bottom two strata.

were least likely to use ability grouping. In Stratum 1, there was only 1 ability-grouped school and 16 ungrouped schools, and the data were too sparse to estimate the stratum specific effect of ability grouping. Instead of discarding these 17 schools, they were combined with Stratum 2 to increase the sample sizes in the low end of the propensity score distribution.[18] Propensity score stratification greatly reduced observed differences between schools with and without ability grouping; both overall balance and within-stratum balance were achieved more than 95% for all covariates at the probability level of .05 (see Appendixes A and B for overall and within-stratum covariate balance on selected school characteristics).[19]

School characteristics systematically differed by propensity strata (Table 5). Generally, schools with high likelihoods of using ability grouping had disadvantageous characteristics. For example, schools in higher propensity strata were more likely to be public and had lower mean SES, a higher proportion of minority students, lower literacy skills, and more heterogeneous literacy skills than schools in lower propensity strata. In addition, administrators in schools that were likely to use ability grouping reported greater problems of teacher and student absenteeism and more negative school climates, including school safety problems (results not shown).

In comparison, schools with low likelihoods of using ability grouping were private schools, were homogeneous in students' cognitive and behavioral characteristics, had small enrollment numbers, and used various kindergarten admission policies, such as requiring admission tests, academic records, recommendations, child interviews, and advising to delay school entry based on standardized tests. This suggests that ungrouped schools regulate their student body through various admission processes.

These systematic differences between ability-grouped and ungrouped schools are substantively important. Prior research viewed ability grouping as within-school stratification processes. However, the aforementioned results showed some evidence for stratification between ability-grouped and ungrouped schools. Generally, schools that use ability grouping have more disadvantageous characteristics and greater diversity than schools without such practices. Few studies, however, recognize these between-school differences.

[18] This did not change the results of the estimates of ability grouping effects in the subsequent analyses.

[19] Balance was checked for all school-level covariates that were found to have bivariate associations with school ability grouping status (107 variables and 17 sets of dummy variables). Results are available upon request. In addition, we examined whether students' characteristics were balanced between ability-grouped and ungrouped schools within stratum (results available upon request). Results showed that many key covariates, such as kindergarten reading scores, academic rating scales, and SES, were balanced within stratum. However, race characteristics were not balanced in the Stratum 2 and 3, and race variables were included in the outcome models.

Table 5. Descriptive statistics on selected school characteristics by propensity strata

	Stratum *M* (*SD*)				
Variables	1	2	3	4	5
Public School	0.11	0.33	0.50	0.63	0.93
	(0.32)	(0.48)	(0.50)	(0.49)	(0.25)
M SES	0.35	0.26	0.25	0.19	– 0.12
	(0.36)	(0.45)	(0.44)	(0.53)	(0.49)
% Black	0.06	0.08	0.08	0.13	0.17
	(0.17)	(0.18)	(0.17)	(0.25)	(0.27)
% Hispanic	0.11	0.09	0.13	0.10	0.23
	(0.17)	(0.15)	(0.19)	(0.14)	(0.29)
M IRT Read	34.95	33.82	33.86	32.80	31.82
	(5.57)	(5.86)	(5.88)	(6.16)	(4.97)
SD IRT Read	7.76	8.42	8.58	8.29	8.90
	(3.25)	(2.34)	(2.36)	(2.59)	(2.36)
M ARS Literacy	2.90	2.83	2.81	2.65	2.49
	(0.45)	(0.48)	(0.47)	(0.46)	(0.44)
SD ARS Literacy	0.53	0.56	0.61	0.62	0.71
	(0.23)	(0.20)	(0.17)	(0.20)	(0.18)
No. enrolled in school	204.83	333.23	309.13	381.28	529.17
	(117.5)	(227.71)	(181.32)	(209.30)	(226.37)
Require SAT for admission	0.52	0.25	0.23	0.08	0.05
	(0.50)	(0.44)	(0.42)	(0.26)	(0.21)

Note. SES = socioeconomic status; IRT = item response theory; ARS = Academic Rating Scales.

Consequently previous studies have not addressed how the effect of ability grouping may vary by school characteristics.

Model Based Estimates of Ability Grouping Effects on First-Grade Reading Achievement

Does Reading Achievement Differ Between Students in Ability-Grouped Schools and Those in Ungrouped Schools? The first question was addressed by estimating the average effect of ability grouping on first-grade reading scores. There was no significant difference in average achievement between students in ability-grouped schools and those in ungrouped schools (Table 6). The estimated effect of ability grouping is –.17 ($p > .1$). This result provided little evidence that ability grouping leads to higher average student achievement.

Do the Effects of Reading Ability Grouping Vary by Students' Initial Abilities? If so, Do Differential Effects Contribute to Increasing Achievement Gaps Between

Table 6. Model-based estimation of the average effects of school ability grouping on first-grade reading achievement

Est.	*SE*
–0.17	0.41

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

High and Low Achievers? Some researchers have argued that although ability grouping may have no effects on the average achievement, its effect may depend on student initial abilities (Hallinan, 1990; Hoffer, 1992). To examine this claim, reading achievement was compared between students in ability-grouped and ungrouped schools at low, middle, and high achievement levels. The results showed no significant difference in first-grade reading achievement between the two groups at any ability levels (Table 7). The average effects for low, middle, and high achievers were, respectively, –.43, .26, and –.61 ($p > .10$). These findings rendered little support for a claim that all students benefit from ability grouping regardless of their initial ability levels. Neither did it support claims that ability grouping led to higher achievement for high initial ability students while hurting low initial ability students. Also, ability grouping did not increase overall achievement inequalities.

Do the Effects of Ability Grouping Vary by Schools? The next analysis examined whether the effects of ability grouping on first-grade reading achievement differed by school characteristics that were defined by the propensity stratum. Results showed that ability grouping increased first-grade reading achievement for students in schools that were less likely to practice ability grouping (Table 8). First-grade scores for ability-grouped students in stratum one schools were 3.55 points higher than their counterparts in ungrouped schools ($p < .01$). The

Table 7. Model-based estimation of the average effects of school ability grouping on first-grade reading achievement by initial ability levels

Initial Ability Levels					
Low		Middle		High	
Est.	SE	Est.	SE	Est.	SE
−0.43	0.71	0.26	0.54	−0.61	0.44

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 8. Model-based estimation of the average effects of school ability grouping on first-grade reading achievement by propensity score strata

Stratum	Est.	SE	Effect Size
5	−0.93	0.82	0.08
4	−2.54**	0.9	0.22
3	−1.07	0.97	0.09
2	1.33	1.02	0.12
1	3.55**	0.97	0.31

Note. Effect sizes are based on student standard deviations from the unconditional hierarchical linear model ($SD = 11.44$).

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

effect size was .31, suggesting that ability grouping had small to moderate effects on student achievement in schools that were not likely to use ability grouping. Ability grouping also led to higher achievement by 1.33 in stratum two schools, although this was not statistically significant. These findings provided partial support for the findings of previous experimental and matched experimental studies, which suggested that ability grouping led to higher average achievement (Lou et al., 1996, Slavin, 1987).

As discussed earlier, schools with low likelihoods of using ability grouping had higher initial ability students, more homogenous student cognitive skills, higher mean SES, and fewer minority students and non-English speakers. Private and smaller schools were also unlikely to practice ability grouping. In fact, schools in Stratum 1 were primarily private schools; almost 90% of all schools in Stratum 1 were private schools, and all seven public schools—four regular and three nonregular public schools—in this stratum did not use ability grouping. It was in this type of schools where students benefited the most from ability grouping.[20]

In contrast, ability grouping might lead to lower reading achievement in schools with more disadvantaged characteristics. Students in schools with ability grouping in Stratum 4 had lower first-grade reading scores than those in ungrouped schools by 2.54 points ($p < .01$) with a small effect size of .22. These schools were characterized as having lower mean kindergarten test scores, more heterogeneous cognitive skills, lower SES, and a higher percentage of minority students. They were likely to be public and have large enrollments. Ability

[20] In Stratum 2, although the coefficient was not statistically significant, I examined whether the effects of ability grouping differed by school sector. Results showed that the effects were similar; both private and public schools produced small positive effects, which were not statistically significant.

grouping might lead to lower average achievement in this type of schools.[21] However, this was only suggestive because ability grouping showed no effect in schools in the Stratum 5—those that were most likely to use ability grouping. Although the direction of the effect was negative, it did not reach statistical significance.

Do the Effects of Ability Grouping Vary by Students' Initial Abilities and Schools? Earlier findings showed that the effects of ability grouping did not vary by student initial ability levels. The next analysis examined whether ability grouping effects on the achievement of low, middle, and high initial ability students might vary by schools attended by these students. To conduct this analysis, it is important to make sure that students' characteristics are balanced for each ability level within each stratum; because propensity score analyses were conducted based on school characteristics, this might not produce covariate balance at the student level when students were further subclassified by their incoming ability levels. To address this concern, Appendix C showed students' incoming ability distribution by school ability grouping status and propensity stratum. Appendix D showed within-stratum balance on selected students' characteristics by student initial ability levels. Overall, key students' characteristics, including incoming ability levels, are similar between ability-grouped and ungrouped schools within each stratum.

Results showed that ability grouping might lead to lower achievement at low and middle initial ability levels in schools that were likely to use ability grouping, and negative effects might be the strongest for low initial ability students (Table 9). The average effects of ability grouping were negative in stratum four schools, and the negative effects were significantly greater for low initial ability students than middle and high initial ability students ($p < .01$).

In contrast, among schools that were least likely to practice ability grouping, ability grouping led to higher achievement at *all* initial ability levels. In ability-grouped schools first-grade reading scores for low, middle, and high initial ability students were, respectively, 5.43, 3.99, and 2.18 points higher than reading scores of their counterparts in ungrouped schools ($p < .01, p < .01$, and $p < .05$, respectively). These results provided partial support for the findings by Slavin (1987) and Lou et al. (1996), which suggested that ability grouping led to higher achievement at all initial ability levels. Additional analyses suggested that the positive effects were significantly greater for low initial ability students than those for middle and high initial ability students. This suggests that ability grouping was particularly beneficial for low achievers in this type of schools.

The schools with the most advantageous characteristics were predicted to be least likely to practice ability grouping. However, these schools produced

[21] It is also noted that private schools in stratum four produced negative and statistically significant results.

Table 9. Model-based estimation of the average effects of school ability grouping on first-grade reading achievement by initial ability levels and propensity strata

Stratum	Low Est. (SE)	Middle Est. (SE)	High Est. (SE)
5	−1.30	−0.52	−1.42
	(0.93)	(1.12)	(.86)
4	−4.80**	−2.93**	−1.25
	(1.54)	(1.07)	(1.03)
3	−0.66	−0.76	−1.65
	(1.95)	(1.19)	(.89)
2	1.81	2.68*	−0.05
	(1.77)	(1.15)	(.98)
1	5.43**	3.99**	2.18*
	(1.68)	(1.35)	(.97)

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

higher achievement for all when they practiced ability grouping. In these schools, these findings provided little evidence for a claim that ability grouping increased achievement inequalities in comparison to ungrouped schools with similar school characteristics. On the contrary, ability grouping might *reduce* achievement inequalities in these schools because low initial ability students were more likely to benefit from ability grouping than higher initial ability students. In contrast, schools that were more likely to use ability grouping had more disadvantageous characteristics, and in these schools achievement inequalities might *increase* when they practiced ability grouping. This was because ability grouping led to lower achievement particularly among low initial ability students, while students with high initial ability were less likely to be affected by such practices. These results were only tentative, however, because ability grouping effects, while negative, did not reach statistical significance in schools that were most likely to use it.[22]

[22] Additional analyses examined the effect of classroom-level ability grouping using “mixed” schools. They attempted to simulate an experiment where classrooms were randomly assigned to an ability-group setting and propensity scores were estimated using classroom-level covariates. Student achievement was then compared between ability-grouped and ungrouped classrooms. None of the analyses showed significant effects of classroom ability grouping in “mixed schools.” There were, however, major limitations in the ECLS–K data to conduct classroom-level analyses, including that (a) students were not the representative sample of students in classrooms, but of students in schools, and (b) the sample size was relatively small per classroom (3.7 students per classroom). Thus, we were not able to reliably measure classroom characteristics to estimate the propensity scores of classroom ability grouping status. Full results are available upon request.

DISCUSSION

This study addresses important issues that have not been examined in previous research. Particularly, this study highlights the significance of school contexts. First, school contexts shape the use of ability grouping. School characteristics and heterogeneous student compositions in particular explain why some schools use ability grouping whereas others do not. Second, in many schools ability grouping does not benefit or hurt student learning. However, ability grouping matters in some school settings and its effect seems to depend on the characteristics of schools. The schools that showed positive ability grouping effects were primarily private and small schools with homogenous ability compositions. These schools were attended by students from advantaged backgrounds—high SES, White, two-parent families, and high initial cognitive skills. Although these schools typically do not practice ability grouping, when they do, they use it in such a way to benefit all students. In contrast, in schools that serve students from disadvantaged backgrounds, ability grouping does not improve student achievement. More important, in schools that showed negative effects, the adverse effects were stronger among low-skilled students. These schools are characterized as having lower SES, higher percentage of minority students, larger enrollment, and greater cognitive heterogeneities than other schools. It is also noted that although many of these schools were public schools, an additional analysis showed that private schools with similar disadvantageous characteristics also produced negative effects of ability grouping.

Variability in ability grouping effects may be explained by the fact that schools, or more specifically classrooms in certain schools, practice ability grouping differently. I argue that a key to understanding how ability grouping produces student achievement lies in the context of schools and how teachers use ability grouping in classrooms within a certain school context. School characteristics differ considerably between schools that typically use ability grouping and schools without such practices. It is reasonable to assume that classroom characteristics also differ between these two types of schools. Consequently, ability grouping experiences are also likely to be different. For example, classrooms in schools that typically use ability grouping would be heterogeneous with many low-ability students, compared to ability-grouped classrooms in schools that are not likely to use such practices. Because of these differences, classrooms in these schools may differ in ability group numbers and "selectivity"—the degree of group cognitive homogeneity and cognitive distinctions by groups (Sorensen, 1970).

Ability grouping is thought to promote student learning because instruction is given at the difficulty level and pace that is commensurate with the skill levels of students. However, having many groups may not be beneficial because teachers would spend less time on each group. In addition, as discussed earlier, having more groups in a classroom with a given class size would create more selective grouping, and this may lead to lower achievement and greater achievement inequalities by reinforcing self-fulfilling prophecies and more unequal allocations of opportunities-to-learn (Gamoran, 1992; Sorensen, 1970).

An additional analysis shows some suggestive evidence for this claim. Among schools with ability grouping, classrooms in schools with higher likelihoods of using ability grouping tend to use more groups than classrooms in schools that are not likely to use ability grouping (see Appendix E). For example, in stratum four schools which showed negative effects of ability grouping, 71% of their classrooms use four groups or more. In contrast, in Stratum 1 schools (i.e., those that are not likely to use grouping) where students benefited the most from ability grouping, about 60% of classrooms use two to three groups. Of interest, the average class size was similar between the two types of schools; schools in Stratum 4 had the average class size of 21 students, compared to 22 students in Stratum 1 schools. This suggests that typical schools with ability grouping (i.e., disadvantageous schools with many low-ability students and heterogeneous compositions) tend to use more groups of smaller sizes, and such groups would be more homogeneous and have greater distinctions across groups (i.e., more selective ability grouping). In comparison, atypical ability-grouped schools (i.e., advantageous schools) use fewer groups with a larger size, and such groups would be less selective.

School and classroom ability compositions may be particularly important in understanding why ability grouping is most consequential for low ability students. As discussed earlier, school and classroom ability compositions and the proportion of low-ability students in particular may directly affect ability group compositions for low-ability students. For example, in classrooms with only a few low-ability students, these students are more likely to be grouped with peers who have higher abilities than themselves when there are only two or three groups in their classes. In comparison, low-ability students would be grouped with other low-ability students in classrooms that are attended by many low-ability students and use four or five groups. This suggests that in schools that are unlikely to use ability grouping, when they do low-ability students are likely to be grouped with higher ability students because of the school characteristics previously discussed. In such a case, the level of instruction would be more challenging than instruction given in a group with many low-ability students. Thus, it is in this context where ability grouping may be particularly beneficial for lower ability students.

In comparison, in schools that are more likely to use ability grouping, low-ability students are likely to be grouped with other students like them because these classrooms have many low-ability students and heterogeneous classroom compositions. Because their classrooms typically have many groups, teachers are likely to spend less time for each group. Moreover, teachers may spend more time instructing children in higher ability groups because more difficult material typically requires more instructional time (Barr & Dreeben, 1983). These factors may explain the negative ability grouping effects for students attending these schools.

These claims are only suggestive because this study did not directly analyze relationships between ability group structures and student learning. In addition, there may be other differences, such as the use of reading teacher aids and flexibility of ability grouping to explain why ability grouping produced higher achievement in schools with advantaged characteristics. Ability-grouped schools with advantageous characteristics might have additional resources to hire teacher aids, or teachers might change students' group assignment as they make progresses, in comparison with ability-grouped schools with disadvantaged characteristics. Alternatively, there may also be differences in instruction across schools that do not use ability grouping. These are topics for future research.

Within-class ability grouping is a key organizational practice in elementary classrooms. Thus, for future study, it is important to further investigate how school and classroom compositions shape ability group structures, such as number, size, and homogeneity; how schools differ in instructional resources; and how group structures and instructional resources shape instructional activities. Such studies will help us understand under what conditions ability grouping may promote or hinder student learning.

ACKNOWLEDGMENTS

This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation and the National Center for Education Statistics of the Institute of Education Sciences (U.S. Department of Education) under NSF Grant #REC-0310268. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies.

REFERENCES

Alexander, K., Cook, M., & McDill, E. L. (1978). Curriculum tracking and educational stratification: Some further evidence. *American Sociological Review*, *43*, 47–66.

Alexander, K. L., & McDill, E. L. (1976). Selection and allocation within schools: Some causes and consequences of curriculum placement. *American Sociological Review*, *41*, 963–980.

Allington, R. L. (1984). Content coverage and contexual readin in reading groups. *Journal of Reading Behavior*, *16*, 85–96.

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

Baumann, J. F., Hoffman, J. V., Duffy-Hester, A., & Ro, J. M. (2000). The first reading: Yesterday and today: U.S. Elementary reading instruction practices reported by teachers and administrators. *Reading Research Quarterly, 35*, 338–377.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal*, *2*, 358–377.

Beckerman, T., & Good, T. (1981). The classroom ratio of high- and low-aptitude students and its effect on achievement. *American Educational Research Journal*, *78*, 317–327.

Berends, M. (1994). Educational stratification and students' social bonding to school. *British Journal of Sociology of Education*, *16*, 327–351.

Betts, J. R., & Shkolnik, J. L. (2000). The effects of ability grouping on student achievement and resource allocation in secondary schools. *Economics of Education Review*, *19*, 1–15.

Chorzempa, B. F., & Graham, S. (2006). Primary-grade teachers' use of within-class ability grouping in reading. *Journal of Educational Psychology*, *98*, 529–541.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrika*, *24*, 205–213.

Cohen, E. G. (1997). Understanding status problems: Sources and consequences. In E. G. Cohen & R. A. Lotan (Eds.), *Working for equity in heterogeneous classrooms* (pp. 61–76). New York: Teachers College Press.

Cox, D. R. (1958). *Planning of experiment*. New York: Willy.

Dar, Y., & Resh, N. (1986). Classroom intellectual composition and academic achievement. *American Educational Research Journal*, *23*, 357–374.

Dehejia, R. H., & Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *Review of Economic and Statistics*, *84*, 151–161.

Eder, D. (1981). Ability grouping as a self-fulfilling prophecy: A micro-analysis of teacher-student interaction. *Sociology of Education*, *54*, 151–162.

Gambrell, L. B., Wilson, R. M., & Gantt, W. N. (1981). Classroom observations of task-attending behaviors of good and poor readers. *Journal of Education Research*, *74*, 400–404.

Gamoran, A. (1986). Instructional and institutional effects of ability grouping. *Sociology of Education*, *59*, 185–198.

Gamoran, A. (1987). The stratification of high school learning opportunities. *Sociology of Education*, *60*, 135–155.

Gamoran, A. (1992). The variable effects of high school tracking. *American Sociological Review*, *57*, 812–828.

Gamoran, A., & Berends, M. (1987). The effects of stratification in secondary schools: Synthesis of survey and ethnographic research. *Review of Educational Research*, *57*, 415–435.

Gamoran, A., & Mare, R. D. (1989). Secondary school tracking and educational inequality: Compensation, reinforcement, or neutrality? *American Journal of Sociology*, *94*, 1146–1183.

Hallinan, M. T. (1987). Ability grouping and student learning. In M. T. Hallinan (Ed.), *Social organization of schools* (pp. 41–69). New York: Plenum.

Hallinan, M. T. (1990). The effects of ability grouping in secondary schools: A response to Slavin's best-evidence synthesis. *Review of Educational Research*, *60*, 501–504.

Hallinan, M. T., & Kubitschek, W. N. (1999). Curriculum differentiation and high school achievement. *Social Psychology of Education*, *3*, 41–62.

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

Hauser, R., Sewell, W., & Alwin, D. (1976). High school effects on achievement. In W. Sewell, R. Hauser & D. Featherman (Eds.), *Schooling and achievement in American society* (pp. 309–341). New York: Academic Press.

Heyns, B. (1974). Social selection and stratification within schools. *American Journal of Sociology*, *79*, 1434–1451.

Hoffer, T. B. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis*, *14*, 205–227.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, *81*, 945–970.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, *27*, 205–224.

Hong, G., & Raudenbush, S. W. (2007). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*, 901–910.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, *86*, 4–29.

Krueger, A. B., & Zhu, P. (2004). Another look at the New York city school voucher experiment. *American Behavioral Scientist*, *47*, 658–698.

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*, 423–458.

Lucas, S. R. (1999). *Tracking inequality: Stratification and mobility in American high school*. New York: Teachers College Press.

Nomi, T. (2006a). *The variable effects of within-class ability grouping: The effect of the group number on reading achievement in first grade*. Paper presented at the American Educational Research Association, San Francisco.

Nomi, T. (2006b). *Educational stratification in early elementary school: The causal effect of ability grouping on reading achievement*. The Pennsylvania State University.

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven: Yale University Press.

Oakes, J., Quartz, K. H., Ryan, S., & Lipton, M. (2000). *Becoming good American schools: The struggle for civic virtue in education reform*. San Francisco, CA: Jossey-Bass.

Page, R. N. (1991). *Lower track classroom: A curricular and cultural perspective*. New York: Teachers College Press.

Pallas, A. M., Entwisle, D. R., Alexander, K. L., & Stluka, F. M. (1994). Ability-group effects: Instructional, social, or institutional? *Sociology of Education*, *67*, 27–46.

Rist, R. (1970). Social class and teacher expectations: The self-fulfilling prophecy in ghetto education. *Harvard Educational Review*, *40*, 411–451.

Rosenbaum, J. E. (1976). *Making inequality: The hidden curriculum of high school tracking*. New York: Wiley.

Rosenbaum, J. E. (1980). Social implications of educational grouping. *Review of Research in Education*, *7*, 361–401.

Rosenbaum, J. E. (1984). The social organization of instructional grouping. In L. P. Peterson, L. C. Wilkinson & M. T. Hallinan (Eds.), *The social context of instruction:*

Group organization and group processes, (pp. 53–67). Orlando, FL: Academic Press.

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

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

Rowan, B., & Miracle, A. W., Jr. (1983). Systems of ability grouping and the stratification of achievement in elementary schools. *Sociology of Education*, *56*, 133–144.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

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

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

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–331.

Schumm, J. S., Moody, S. W., & Vaughn, S. (2000). Grouping for reading instruction: Does one size fits all? *Journal of Learning Disabilities*, *33*, 477–488.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research*, *57*, 293–336.

Sorensen, A. B. (1970). Organizational differentiation of students and educational opportunity. *Sociology of Education*, *43*, 355–376.

Stuart, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher*, *36*, 187–198.

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

Vanfossen, B. E., Jones, J. D., & Spade, J. Z. (1987). Curriculum tracking and status maintenance. *Sociology of Education*, *60*, 104–122.

West, J., Denton, K., & Reaney, L. M. (2001). *The kindergarten year: Findings from the early childhood longitudinal study, kindergarten class of 1998–99*. Washington D.C: National Center for Educational Statistics.

APPENDIX A

Table A1. Overall differences between schools with and without ability grouping before and after adjustment: Selected school characteristics

	No Adjustment		After Adjustment	
Variables	Diff	*T*	Diff	*T*
Direct Cognitive Assessments				
M IRT Read	−1.15	−2.18	.26	.44
M IRT General Knowledge	−2.20	−4.91	−.10	.50
SD IRT Read	.52	2.24	.08	.28
SD IRT General Knowledge	.26	2.03	.10	.73
Teacher ratings				
SD ARS Literacy	.09	4.87	.00	.20
SD ARS General knowledge	.11	4.53	−.02	−.8
SD Approach to Learning	.03	2.27	−.02	−1.89
SD Interpersonal behavior	.04	3.05	−.01	−.39
SD Externalizing behavior	.04	2.52	−.01	−.41
Demographics				
M SES	−.18	3.62	.06	1.06
% non-English speaking	.06	3.27	.02	.76
% Hispanic	.07	3.07	.01	.47
% Black	.06	2.54	.01	0.41
% Teacher Black	3.01	3.05	1.03	0.89
% Teacher Hispanic	2.17	2.67	1.17	1.21
No. of enrollment	145	6.01	12	.62
Public school	0.35	8.42	−0.01	−.28

Note. Chi-square statistics. IRT = item response theory; ARS = Academic Rating Scales; SES = socioeconomic status. Full results are available upon request.

APPENDIX B

Table B1. Within-stratum differences between schools with and without ability grouping before and after adjustment: Selected school characteristics

Variables	Stratum	Diff	*T*
Direct Cognitive Assessments			
M IRT Read	5	−0.06	−0.05
	4	1.06	0.68
	3	−0.84	−0.51
	2	0.22	0.14
	1	1.02	0.65
M IRT General Knowledge	5	−0.48	−0.48
	4	0.24	0.21
	3	−1.10	−0.88
	2	0.71	0.65
	1	0.19	0.2
SD IRT Read	5	−0.07	−0.14
	4	−0.01	−0.02
	3	0.52	0.81
	2	0.22	0.36
	1	−0.19	−0.23
SD IRT General knowledge	5	−0.33	−1.17
	4	0.13	0.38
	3	0.56	1.83
	2	0.51	1.50
	1	−0.10	−0.29
Teacher ratings			
SD ARS literacy	5	0.02	0.54
	4	−0.00	−0.07
	3	0.02	0.63
	2	−0.02	−0.47
	1	0.01	0.11
SD ARS general knowledge	5	−0.01	−0.21
	4	−0.00	−0.07
	3	0.68	0.84
	2	−0.09	−1.52
	1	−0.07	−0.91
Demographics			
M SES	5	0.04	0.38
	4	0.05	0.37
	3	0.06	0.46
	2	0.06	0.47
	1	0.10	0.94

(Continued on next page)

Table B1. Within-stratum differences between schools with and without ability grouping before and after adjustment: Selected school characteristics *(Continued)*

Variables	Stratum	Diff	*T*
% Black	5	0.02	0.35
	4	−0.00	−0.02
	3	0.06	1.28
	2	−0.05	−1.05
	1	0.03	0.7
% Hispanic	5	0.09	1.49
	4	0.00	0.04
	3	−0.73	−1.40
	2	−0.03	−0.79
	1	0.03	0.64
Public school	5	0.07	1.28
	4	0.04	0.32
	3	0.07	0.55
	2	−0.13	−1.09
	1	−0.15	−1.87

Note. IRT = item response theory; ARS = Academic Rating Scales.

APPENDIX C

Table C1. Students' ability distribution by school ability grouping status and propensity stratum

	Ungrouped Schools						Ability-Grouped Schools					
	Low		Mid		High		Low		Mid		High	
Stratum	Student *N*	School *N*	Student *N*	School *N*	Student *N*	School *N*	Student *N*	School *N*	Student *N*	School *N*	Student *N*	School *N*
5	110	20	107	22	97	20	1,413	294	1,447	298	1,481	287
4	93	16	101	18	100	18	273	53	297	60	355	56
3	68	17	129	21	122	21	130	29	167	32	213	31
2	137	24	165	33	201	31	107	23	134	25	162	25
1	115	28	219	39	225	40	53	16	95	18	117	19

Note. Within each stratum, Chi-square tests were conducted to examine whether student ability distribution was significantly different between schools with and without ability grouping. In all strata, balance was achieved at the .05 probability level.

APPENDIX D

Table D1. Within-stratum covariate balance by students' initial ability levels: selected students' characteristics

		Student Ability Levels					
		Low		Middle		High	
Variables	Stratum	Diff	*Z*	Diff	*Z*	Diff	*Z*
Kindergarten IRT reading scores	5	−0.34	0.76	0.07	.32	1.37	1.43
	4	0.15	.30	−0.19	−.57	1.5	1.02
	3	−0.53	−.93	−0.11	−.35	0.24	.18
	2	1.25	2.24	−0.70	−2.78	−0.26	−.22
	1	0.52	1.03	−0.07	−.16	0.05	.04
ARS literacy	5	−0.04	−0.48	0.04	0.46	0.07	0.74
	4	0.42	0.44	0.17	1.43	0.18	1.60
	3	0.25	2.44	0.07	0.63	0.24	1.84
	2	0.30	2.46	0.21	2.20	0.18	1.42
	1	0.02	0.13	−0.04	−0.32	−0.11	−0.72
Approach to learning	5	0.03	0.3	−0.08	−1.00	0.30	0.49
	4	−0.01	−0.14	−0.4	−0.35	0.11	1.41
	3	0.08	0.65	0.05	0.47	0.10	1.21
	2	0.16	1.38	0.06	0.55	0.17	0.16
	1	0.03	0.19	−0.02	−0.2	−0.07	−0.74
SES	5	0.11	1.09	−0.03	−0.28	0.03	0.23
	4	0.06	0.40	0.03	0.21	−0.04	−0.3
	3	0.10	0.56	0.00	0.07	0.18	1.33
	2	0.03	0.23	0.11	0.63	−0.00	−0.02
	1	0.12	1.11	0.31	0.17	0.16	1.43
		Diff in%	χ^2	Diff in%	χ^2	Diff in%	χ^2
Black	5	2.65	0.38	2.67	0.46	−2.30	0.51
	4	1.39	0.08	0.73	0.04	−3.20	1.40
	3	5.33	1.32	6.65	5.64	4.20	2.49
	2	−9.83	4.85	−3.60	2.10	−4.00	3.29
	1	7.90	3.24	0.54	0.05	0.70	0.16
Hispanic	5	10.28	6.08	5.60	2.19	2.19	2.19
	4	2.27	0.37	2.05	0.45	0.56	0.56
	3	−25.59	19.74	−0.84	0.05	1.54	1.54
	2	1.4	0.11	−1.30	0.20	2.19	2.19
	1	2.2	0.13	2.40	0.44	1.88	1.88

Note. IRT = item response theory; ARS = Academic Rating Scales; SES = socioeconomic status.

APPENDIX E

Table E1. Frequency distribution: Number of groups per classrooms by propensity strata

	Propensity Strata									
	1		2		3		4		5	
No. of Groups in Class	Class *N*	%	Class *N*	%	Class *N*	%	Class *N*	%	Class *N*	%
2	6	25.0	8	20.0	1	1.9	15	10.3	77	7.0
3	9	37.5	11	27.5	19	35.2	27	18.5	322	29.0
4	4	16.7	11	27.5	23	42.6	54	37.0	409	36.8
5 or more	5	20.8	10	25.0	11	20.4	50	34.3	304	27.3
N	24	100	40	100	54	100	146	100	1, 112	100
Missing *N*	0		2		2		0		90	