The Pennsylvania State University

The Graduate School

College of Education

EDUCATIONAL STRATIFICATION IN EARLY ELEMENTARY SCHOOL:

THE CAUSAL EFFECT OF ABILITY GROUPING ON READING ACHIEVEMENT

A Thesis in

Educational Theory and Policy

by

Takako Nomi

© 2006 Takako Nomi

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2006

The thesis of Takako Nomi was reviewed and approved* by the following:

Regina Deil-Amen
Assistant Professor of Education
Thesis Advisor
Co-Chair of Committee

Joseph Schafer
Associate Professor of Statistics
Co-Chair of Committee

Suet Ling Pong
Professor of Education, Demography, and Sociology

George Farkas
Professor of Sociology, Demography, and Education

Sean F. Reardon
Associate Professor of Education
Special Member

Gerald K. LeTendre
Associate Professor of Education
In Charge of Graduate Program in Educational Theory and Policy

*Signatures are on file in the Graduate School.

ABSTRACT

The purpose of this dissertation is to apply propensity-score methods to examine the causal effects of within-classroom reading ability grouping on student achievement in first grade. Limitations in previous ability grouping research are also critically examined. Four research questions are addressed: 1) Does reading achievement differ between students who are grouped by ability and those who are not? 2) Does the effect of reading ability grouping vary by students' initial abilities? If so, do differential effects contribute to increasing achievement gaps between high and low achievers? 3) Do the effects of ability grouping vary by schools or classrooms? 4) Do the effects of ability grouping vary by students' initial abilities *and* schools or classrooms? The ECLS-K dataset is used, which consists of a nationally representative sample of kindergarteners. Findings suggest that although first-grade reading ability grouping has no significant effects on the average reading achievement for the population as a whole, ability grouping effects are found to vary by student initial abilities and school characteristics, but not classroom characteristics. Ability grouping leads to higher achievement for all students in schools that are least likely to practice ability grouping. In these schools achievement inequalities may be reduced because stronger effects are found among low initial ability students. In contrast, ability grouping leads to lower achievement particularly for students with low and middle initial abilities in schools that are more likely to practice ability grouping. This suggests increasing achievement inequalities in these schools. The schools with positive ability grouping effects are likely to be non-public, smaller, and more homogenous in student cognitive and behavioral characteristics. They also have higher SES, fewer racial minorities, and students with better reading skills than schools with negative ability grouping effects. I argue that school contexts, particularly ability distributions, are key factors both as the determinants and consequences of ability grouping practices. Implications for ability grouping policy to improve student achievement are also discussed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

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

CHAPTER 1

INTRODUCTION:

ABILITY GROUPING IN THE EDUCATION POLICY CONTEXT

Ability grouping, found throughout the U.S. schools, has been a contentious educational issue among educational researchers and practitioners. At issue is whether ability grouping is an effective means of organizing instruction, whether ability grouping leads to high achievement for all, and whether it unfairly limits educational opportunities for disadvantaged students, thus exacerbating existing educational and social inequalities.

Ability grouping is often regarded as an effective organizational practice to deal with a heterogeneous student population who differs not only in prior cognitive skills, but also in linguistic, cultural, and socio-economic backgrounds. Ability grouping is a particularly important policy issue in the contemporarily U.S. schools because student diversities are expected to rise due to a rapid increase in the immigrant population from Latin America and Asia. In particular, Hispanic immigrants are the fastest growing population in the U.S. society today and they have a greater proportion of the school-age population than any other race/ethnic group. Many of these immigrants have limited English proficiencies and come from low socio-economic backgrounds. Thus, American schools may face increasing challenges to provide equitable educational practices to socially and economically disadvantaged students.

Defining Ability Grouping and Tracking

Researchers have often used “ability grouping” and “tracking” interchangeably. However, it is important to clarify the two concepts. “Ability grouping” refers to the practice of assigning students to instructional groups based on their ability. There are different forms of ability grouping and they differ in terms of structural characteristics. In the U.S. secondary schools, ability grouping often occurs between classes, and it often involves curriculum differentiation. Curriculum differentiation takes one of two forms. The first form involves creating distinct instructional programs for students. This overarching program was the original form of “tracking” in American secondary schools, and a typical tracking structure includes Academic, General, and Vocational tracks (Hallinan and Kubitschek 1999).

Tracking was prevalent from the fifties to the seventies. However, it is much less common in contemporary American schools. Lucas (1999) discusses shifts in tracking that occurred in the 1970’s. Explicitly labeled tracks have been replaced by course levels, which vary by the difficulty of course content, the quality of material, and the pace of instruction (e.g., Advanced, Honors, Regular, and Basic groups). While the original form of “tracking” separated students by the programs for the entire day, today’s ability grouping is less rigid and students can be in different levels in different subjects (Lucas 1999).

Even though the formal organizational structure of schools has changed and “tracking” does no longer exist in today’s schools, the curriculum continues to differentiate the academic pathways of students. Students take different courses in different sequences. In addition to student ability previous course taking often influences the assignment of students to future courses (Hallinan and Kubitschek 1999). In this context, researchers have used both “ability grouping” and “tracking” to describe these two types of curriculum differentiation.

Unlike secondary schools, elementary schools commonly group students by different ability levels within the same classroom; this is called "within-class ability grouping". The teacher who uses within-class ability grouping often works with a group of students at a time, while students in other groups engage in seatwork. The curriculum is typically the same for all students in the same classroom. However, students in different groups receive instruction that differs in the pace, the difficulty, and the amount of material.

Ability Grouping and Tracking Debate

Ideologies underlying ability grouping/tracking controversies revolve around equity and excellence in education. Supporters for ability grouping/tracking have argued that heterogeneous grouping would lead to academic mediocrity and undermine excellence in academic achievement. Curricular reforms and school restructuring to promote inclusive classrooms and cooperative learning have often faced strong opposition from teachers and parents (Loveless 1994; Gamoran and Weinstein 1998; Oakes, Quartz, Ryan, and Lipton 2000). Teachers and parents, particularly middle-class parents whose children are more likely to be in higher groups/tracks than other students, have argued that such reforms would inevitably compromise academic excellence because instruction in heterogeneous classrooms will not be rigorous enough for the most able students and take learning opportunities away from these students (Oakes et al. 2000).

In contrast, opponents of ability grouping/tracking have emphasized equity in education (for example, see Bowles and Gintis 1976; Goodlad 1984; Oakes 1985; Rausenbaum 1976). They argue that ability grouping/tracking leads to the inequitable distribution of educational resources because students in lower groups/tracks are often taught less challenging material, are taught by less able teachers, and receive lower academic expectations from teachers.

For example, Cohen (1997a, 1997b) points out that an underlying problem of inequalities in classrooms is status problems. Within classrooms, students are often hierarchically ranked based on academic standings and other social characteristics and differential status is attached to individual students. Ability grouping is an inequitable practice because it creates social interactions among teachers and students in ways that reinforce the distinctions between low-status students and high-status students. This often leads to the inequitable distribution of learning opportunities and contributes to maintaining the existing status order in classrooms.

In order to make educational practices more equitable within classrooms, Cohen (1994) proposes the use of instructional strategies that promote equal status among students. This could be done by, for example, changing teachers' expectations toward students, organizing instruction in heterogeneous small groups, delegating authority to the groups while holding them accountable for their performance, and promoting participation from all students in a group work so that students can learn from one another.

Oakes also argue that ability grouping/tracking is not a neutral organizational practice, but it is discriminatory and socially unjust (Oakes 1994; Oakes et al. 2000). In particular, ability grouping/tracking has disproportionately negative impacts on low-income and minority students. For example, ability grouping/tracking creates within-school segregation by class and race because low-income and minority students are typically overrepresented in lower groups/tracks. Within-school segregation also leads to negative stereotyping and has negative academic and social-psychological effects on low-income and minority students (Eder 1981; Metz 1978; Oaks 1985; Rist 1970).

Proponents of tracking/ability grouping argue that the assignment of students to different track/groups is fair and just because decisions for group/track assignment are primarily based on student ability and mastery skills (see, for example, Rehberg and Rosenthal 1978). However, opponents argue that this is not likely partly because middle-class parents are more likely to use their resources to secure their child's placement in higher groups/tracks than lower-class parents and also because other factors that are unrelated to student ability, such as preference, race, and SES, would affect assignment decisions (Oakes 1985). In addition, unlike their middle-class peers, disadvantaged students and their parents may lack knowledge about courses that are needed for college admission even when they expect to go on college after graduating from high school. In high schools, these students may not be fully informed of consequences of their course taking (McDonough 1997).

Research on ability grouping/tracking is extensive and much research examines its academic consequences. While there are some contradictory findings, Hallian (1994b) summarizes findings of previous research in following ways. First, students in higher groups/tracks learn more than students in lower groups/tracks. Second, ability grouping/tracking does not provide advantages over heterogeneous grouping with respect to students with the middle-range ability level. Third, the quality and quantity of instruction increases with the level of ability groups/tracks (Hallinan 1994b).

These findings seem to suggest that ability grouping/tracking practices exacerbate existing achievement inequalities—high ability students gain more at the expense of low ability students through differential allocations of opportunities-to-learn.

There are, however, important limitations in previous ability grouping/tracking research. First, a majority of these studies are studies on ability group *placement*. These studies address questions as to how students achieve if they are placed in different ability group levels. However, this line of research does not illuminate an important policy question of "detracking", or "ungrouping." In other words, we know little about how students with different ability levels—particularly low and high achieving students—would perform if they attended schools that did not use ability grouping/tracking.

Second, most previous studies have focused on secondary school practices. Ability grouping begins as early as kindergarten. However, less attention has been paid to the role of ability grouping in social stratification and educational inequality in early school years[1]. While neglecting the elementary school context, in status attainment research, for example, researchers have examined the effects of different course taking (e.g., academic courses vs. non-academic courses) in secondary schools on later academic achievement and college entrance (for the review of status attainment research, see Bar and Dreeben 1983, chap. 2).

Of the existing studies on within-class ability grouping in elementary schools, findings are rather equivocal. Studies on ability group placement show that, similar to secondary school studies, students in higher ability groups learn more than students in low ability groups after controlling for their initial ability and background characteristics (Gamoran 1986; Pallas, Entwisle, Alexander, and Stluka 1994; Tach and Farkas forthcoming), and the differences in reading achievement are explained by differences in instruction that children received in these groups (Gamoran 1986 and Pallas et al 1994). In contrast, studies that compare achievement between ability-grouped and ungrouped students suggest that high, medium, and low achievers all learn more in the ability grouped schools than their counterparts in non-ability grouped schools (Slavin 1987). Although these two types of studies address different questions—one addressing a question regarding ability grouping placement and the other addressing a question of ability grouping practices, they lead to different conclusions. The placement research

[1] Although ability grouping is not extensive in kindergarten, McPartland, Coldiron, and Braddock (1987) report that 90 percent of first-grade classrooms use ability grouping.

suggests that ability grouping only benefits students in high ability groups and students in low ability groups suffer from this practice. In contrast, the research on ability grouping practices suggests that ability grouping benefits all students.

From educational policy perspectives, a lack of credible evidence on the causal effect of ability grouping in early elementary grades is a critical gap. During this period children are expected to learn to read and after about third grade children are expected to read to learn (Farkas 1996). Previous research has shown that children start schools with diverse literacy skills mainly because of the differences in family socio-economic backgrounds (Hart and Risley 1995; Lee and Burkam 2002). Deficits in literacy skills in early years are found to have long-term consequences on later academic outcomes (Werner and Smith 1992; Walker, Greenwood, Hart, and Carta 1994). Over the past twenty years, researchers have emphasized the significance of experiences in early school years in shaping future life chances of children (Cunningham and Stanovich 1997; Entwisle and Alexander 1989, 1993; Farkas and Beron 2004). In addition, recent research finds that considerable disparities in achievement growth rates exist within schools in early years (Reardon 2003).

More importantly, current ability grouping research has not fully addressed why previous studies have produced seemingly contradictory findings. The source of the equivocal findings may be either methodological, or ability grouping may not be uniformly beneficial or harmful to students in all schools that practice ability grouping. For example, ability grouping may be effective for some students and its consequences may depend on characteristics of schools where students attend. Few existing studies, however, address such variations in ability grouping effects whether they are explained by ability grouping structures or school contexts, such as the concentration of low-income, minority, and at-risk students.

Aims of the Study

To address knowledge gaps in current ability grouping research, this dissertation uses the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K) and applies propensity score methods (Rosenbaum and Rubin 1983) to examine causal effects of ability grouping in early elementary school years. The ECLS-K contains a nationally representative sample of 21,260 children who attended kindergarten in 1998 and the follow-up studies were conducted when children were in first, third, and fifth grades.

Many previous observational studies of ability grouping examine the consequences of ability group *placement*. Unlike those studies, this dissertation specifically addresses questions regarding a policy of practicing or not practicing ability grouping, that is, whether attending ability grouped classes is beneficial to all students, which group of students might benefit the most from ability grouping practices, whether ability grouping effects vary by characteristics of classrooms or schools, and whether academic gains of one group occur at the expense of other groups.

More specifically, the following four research questions are addressed: 1) Does reading achievement differ between students who are grouped by ability and those who are not ability-grouped? 2) Does reading ability grouping have differential effects by students' initial abilities? If so, do differential effects contribute to increasing achievement gaps between high and low achievers? 3) Do the effects of ability grouping vary by schools or classrooms? 4) Do the effects of ability grouping vary by students' initial abilities *and* schools or classrooms?

This dissertation may make an important contribution to the existing ability grouping research particularly by addressing heterogeneities in ability grouping effects on student achievement by classroom and school contexts. In the United States, schools differ considerably in socio-demographic as well as student ability compositions. Schools also differ in expenditures, climates, and the quality of teachers (for the review of research on the process of between-school stratification, see Riordan 1997, chap. 4). Kozol (1991) depicts striking inequalities in the quality of schooling between public schools in wealthy and impoverished districts.

The debate of ability grouping often surrounds the question of whether ability grouping should or should not be used to improve student academic achievement and whether ability grouping increases within-school inequalities. However, because of the diversity in American schools it is possible that the effect of ability grouping varies by the context of schools or classrooms. For example, ability grouping may lead to higher achievement in some schools/classrooms and lower achievement in others. Similarly, ability grouping may increase achievement inequalities in some schools/classrooms but not in others.

The variability in the effect of ability grouping may be explained by two factors. First, the same education policy may produce different achievement outcomes because of the differences in school or classroom contexts. For example, the practice of ability grouping may be similar across schools/classrooms, but because of the differences in their student compositions, ability grouping may produce differential effects by schools/classrooms. Second, the consequences of ability grouping may differ by schools or classrooms because they implement the policy differently. For example, because schools/classrooms differ in student compositions or teacher experiences, schools/classrooms may practice ability grouping differently. Ability group numbers and group compositions may differ across schools/classrooms. These differences may produce differential effects of ability grouping on student achievement by schools/classrooms. The existing studies, however, have not examined whether or not the effects of ability grouping may vary by classrooms or schools and this question is specifically addressed in this dissertation.

Organization of the Dissertation

This dissertation is organized as follows. Chapter 2 discusses conceptual frameworks of ability grouping to understand how ability grouping influences student academic outcomes. Chapter 3 summarizes previous studies on the consequences of ability grouping on student academic outcomes. Methodological issues are discussed in Chapter 4. The first part of Chapter 4 discusses methodological shortcomings of previous observational studies of ability grouping placement and introduces methods for causal inference, using propensity scores. The last part of Chapter 4 discusses research designs used in this study and propensity score applications to the analysis of causal effects of ability grouping on student achievement. Chapter 5 investigates four research questions addressed in this study. Chapter 6 presents discussions and conclusions.

CHAPTER 2

ABILITY GROUPING AND TRACKING: THEORY AND PRACTICE

Ability Grouping and Tracking: Conceptual Frameworks

Several different conceptual frameworks are relevant in ability grouping/tracking research. Functionalist theories and social reproduction theories are two macro-level theories that offer two distinctive views on the roles and functions of schools in society and the roles that ability grouping/tracking play as school organizational practices in educational stratification. Other conceptual frameworks—pedagogical rationales for ability grouping/tracking, organizational perspectives, and structural perspectives—focus on school organizations and how they influence student learning.

In particular, organizational perspectives and structural perspectives discuss a linkage among school organizations, the formation of ability grouping, instruction, and learning, and structural analyses explain how the effect of ability grouping on student achievement may depend on structural characteristics of ability grouping (e.g., homogeneity and flexibility). These two micro-perspectives help us understand how the effect of ability grouping may depend on school contexts.

Macro-level theories: functionalism and social reproduction and conflict perspectives. Two macro-level theories provide different explanations for the functions and roles of ability grouping/tracking in the process of educational stratification. First, functionalist perspectives suggest that a function of schools in modern meritocratic society is to efficiently and rationally sort students so that the most able and motivated individuals will attain the highest status positions in the future social and occupational ladder. Ability grouping and tracking play key roles in sorting processes. For example, the most talented and motivated individuals are selected for academic programs while others are channeled to vocational paths. Functionalist perspectives assume that ability grouping/tracking practices are effective and fair because abilities and educational potentials of children can and will be objectively and correctly identified, and these objective criteria will serve as the basis for selecting students to different ability groups and tracks (Schafer and Olexa 1971).

In contrast to functionalist perspectives, social reproduction theorists argue that schools are key social institutions in reproducing existing social inequalities (Bourdieu 1976). Similarly, conflict perspectives suggest that schools in capitalist society only serve the needs of capitalists (Bowles and Gintis 1976). For example, through schooling middle and upper-middle class students are defined as more able and they are taught not only skills needed for high skilled occupations but also independence, autonomy, and critical thinking skills. In contrasts, working-class students are typically defined as less able and they are taught obedience and how to be a good worker (Bowles and Gintis 1976; Oaks 1994).

From these perspectives, ability grouping is the key school organizational practice in the reproduction of existing social inequalities. There are two distinctive arguments regarding ability grouping/tracking practices from conflict and social reproduction perspectives. First, they view ability group/track placement as biased and unfair because factors irrelevant to student ability, such as race, gender, and socio-economic backgrounds, always influence placement decisions. In particular, they argue that ability grouping/tracking discriminates against minority and low income students (Oakes 1985). Social reproduction and conflict perspectives also suggest that even when ability grouping/tracking assignment decisions are based on seemingly objective criteria, such as tests scores and grades, ability grouping/tracking can be discriminatory if tests themselves are biased or if tests and grades measure verbal and non-verbal skills that reflect the quality of previous schooling rather than the intellectual potentials for acquiring those skills (Oakes 1985).

Second, social reproduction and conflict theorists argue that ability grouping/tracking only benefits students in higher groups/tracks because it differentially allocates opportunities-to-learn to students.[2] They argue that students in higher groups/tracks learn more at the expense of students in lower groups/tracks by depriving learning opportunities from students in the latter group. Students in occupational tracks or lower ability groups are denied more demanding curricula, they are assigned to low quality teachers, and teachers spend more time on discipline than instruction (Oakes 1985).

Other research suggests that teachers compete for better students (Finley 1984). Learning climates differ between low groups/tracks and high groups/tracks (Metz 1978; Oaks 1985; Page 1991). Page (1991) argues that the "chaos" in lower tracked classrooms is created by ambiguity faced by both students and teachers with regard to classroom topics, learning objectives, and teacher expectations about students as well as their roles as a teacher. These factors would, in turn, lead to lower achievement for students in low ability groups/tracks than students in high ability groups. Moreover, some researchers argue that once children are placed in different groups/tracks, mobility between groups/tracks is rather limited (Kerckhoff and Glenne 1999; Rosenbaum 1976, 1980; Stevenson, Schiller, and Schnieder 1994).

[2] I will refer researchers as social reproduction or conflict theorists if their research supports social reproduction and conflict perspectives although they may not identify themselves as such.

In addition, researchers have argued that the negative effect of ability grouping/tracking on student achievement can occur through social psychological processes (Eder 1981; Rist 1970). The concepts of "self-fulfilling prophecy" (Marton 1948) address reasons why ability grouping/tracking might negatively impact student achievement. For example, ability grouping/tracking affects self-concept and identities by labeling students in lower groups as "dumb", "slow", or "stupid", and labeling others as "smart" and "bright." This contributes to stereotyping. Moreover, teachers treat students based on these expectations and students respond to and internalize such expectations. In other words, teachers' expectations become self-fulfilling prophecies (Rist 1970; Rosenthal and Jacobson 1968). Also, when group/tracking placement is associated with race and SES characteristics of students, ability grouping/tracking creates race and SES segregation within schools, which would, in turn, lead to stereotyping of students based on these student characteristics.

Thus, from social reproduction and conflict perspectives, ability grouping/tracking not only unfairly places students into different academic trajectories, but also unfairly distributes opportunities-to-learn to students based on group/track levels. In addition, if schools provide few opportunities for children to move across different groups/tracks as they progress through grade levels, their academic trajectories would be determined by their initial ability group and track placement. Because educational success is tied with future occupational success, and because ability grouping begins as early as kindergarten, it is important to examine the extent to which ability grouping shapes children's academic achievement in early school years.

Pedagogical rationales for ability grouping. Historically, various forms of ability grouping emerged in response to problems due to the increasing diversity in the student body in classrooms and schools in terms of aptitudes, maturity, and social backgrounds (Barr and Dreeben 1983). The expansion of mass schooling, compulsory secondary schooling, the development of comprehensive school systems, and desegregation all contributed to the increase in pupil diversities particularly in secondary schools (Schafer and Olexa 1971). The expansion of mass schooling led some communities to creating high schools that specialized for talented or troublesome youths. However, most communities created comprehensive high schools where students came from diverse backgrounds and they also differed in social and intellectual maturity and prior mastery skills.

To adapt intellectual and motivational differences among students, comprehensive secondary schools developed both administrative adjustments, such as organizing students into different classes, tracks, and grades, and various intra-classroom methods of organizing instruction (Schafer and Olexa 1971). In elementary school, however, teachers often responded by dividing students into smaller homogeneous groups and using instruction that was geared to students in different group levels, while other teachers responded by using instructional variations adapted to whole-class or seatwork formats (Barr and Dreeben 1983).

A set of pedagogical rationales for ability grouping/tracking suggests that homogeneous ability grouping will increase teachers' effectiveness in organizing classroom instruction. This perspective views all students as benefiting from instruction given in homogeneous groups because content is taught at the difficulty level and pace that is commensurate with past student performance. In contrast, the whole-class instruction is thought of as an ineffective instructional method for teaching diverse students because it typically emphasizes the uniformity of instruction. In addition, secondary school tracking/ability grouping is viewed as beneficial for all students because it meets different needs and interests of individual students (Schafer and

Olexa 1971). Rationales behind a traditional form of tracking are that courses, such as foreign languages and advanced science and mathematics, are thought to need for college education, and they are taught for college-bound students. Non-college bound students, however, are thought to need only basic science and math courses and other technical and vocational courses. These views are particularly problematic in today's society where most high school students aspire to attend college and 80 percent of high school graduates attend college, while many students, particularly disadvantage students, are not aware of the consequences of their course taking patterns (Deil-Amen and Rosenbaum 2002; McDonough 1997).

The issue of how schools and teachers have dealt with the problem of student diversity in different school contexts is an important empirical question. However, two overriding policy questions among education practitioners and researchers have been, first, whether ability grouping/tracking impede learning of socially and economically disadvantaged students, and second, whether individual students learn more when they are taught in ungrouped/un-tracked classes than when they are taught in homogenously composed classes or groups (Oakes 1985; Oakes and Martin 1992)[3]. The following section presents a conceptual framework that explains how within-class ability grouping may, or may not, produce higher student achievement than a whole-class instructional setting.

Organizational perspectives: link between school organizations, instruction, and learning. Many researchers have argued that ability grouping plays a key role in shaping student learning experiences. Bar and Dreeben (1983) present an organizational analysis of learning processes, which focuses on structural characteristics of school systems, their divisions of labor, and how

[3] The analysis in this dissertation focuses on comparisons in achievement between students who are ability grouped within classes and students in heterogeneous classrooms. For a study of heterogeneous groups, see Cohen (1994). For a meta-analysis of heterogeneous classrooms and groups, see Lou et al (1996).

events occurring at each organizational level are linked to one another to produce student learning. The organizational analysis helps us understand how larger school contexts shape the formation of ability grouping and instructional activities, which, in turn, affect student learning.

Organizational perspectives view school systems as hierarchically structured organizations, which consist of school districts, schools, classrooms, and instructional groups. Bar and Dreeben (1983) claim that the hierarchies of organizational levels do not merely represent authority and status distinction. More importantly, the hierarchies of organizational levels represent divisions of labor where resources and tasks are distributed to different levels in school systems in a workable arrangement for the production of knowledge (Bar and Dreeben 1983).

At the highest level, school districts perform managerial functions, which include central finances, hiring school personnel, allocating educational resources to schools, plant maintenance, and supervisions. These activities, however, do not directly relate to tasks for running schools or instructing students (Bar and Dreeben 1983).

The next highest organizational level is the school. Schools main tasks include "the assignment of students to specific teachers, the allocation of learning material to classrooms, the arrangement of a schedule so that all children in the school can be allotted an appropriate amount of time to spend on subjects in the curriculum, and the integration of grades so that work completed in one grade represents adequate preparation for the next" (Bar and Dreeben 1983: 6).

Schools are, however, not the units of instruction. Rather, instructional activities are the properties of classrooms, and teachers are responsible for the direct engagement of students in learning activities (Bar and Dreeben 1983). Teachers deliver instruction, direct the learning activities of children, and bring children into immediate contact with various learning materials.

In elementary schools, teachers often form instructional groups within classrooms to manage instructional activities in order to handle diversities in skills and aptitudes of students. Instructional groups are, thus, considered as sub-organizations within classrooms.

Although instructional groups are a unit of instruction, it is not ability grouping per se that determines student learning. To understand the relationship between ability grouping and student learning, Bar and Dreeben (1983) makes conceptual distinctions between organizations of instruction and instruction that teachers actually provide to students. The organization of instruction is concerned with how teachers arrange the classroom instruction, which may be to teach everyone together in the whole-class setting, or to divide children into small ability groups, or small heterogeneous groups, and work intensively with some groups while having other students work by themselves with little supervision. In contrast, decisions on instruction may include the content, pace, and time spent for instruction. The authors argue that student learning depends on what they are taught, and instructional resources and activities—content coverage (i.e., difficulty) and the pace of instruction in particular—are most important factors that directly affect student learning (Bar and Dreeben 1983).

Both school and classroom contexts are important factors in determining instructional activities that take place in ability grouped classrooms, which, in turn, affect student learning. For example, school organizational characteristics (e.g., size, ability compositions, and demographic characteristics) shape characteristics of classrooms in schools and classroom characteristics, in tern, determine characteristics of ability grouping (e.g., size, number, and discreteness). Characteristics of ability grouping shape instructional activities and student learning.

Bar and Dreeben (1983) discuss several important properties of instructional groups that influence student learning, which include the mean ability level, the number of groups, group size, and their discreteness. Of those, the mean ability level of the group is the most important factor for student learning because it primarily determines the content coverage (i.e., difficulty) and the instructional pace (i.e., the amount of material taught in a given period of time).

The properties of ability groups, such as number, size, and discreetness, depend on characteristics of classrooms, such as the distribution of student aptitudes and class size (Bar and Dreeben 1983). Bar and Dreeben (1983) argue that the teacher decides the number, size, and discreetness of ability grouping based on the number of low achieving children in his/her classroom, heterogeneity of student aptitudes, and class size, but not on the mean student aptitudes.

In addition, the properties of ability grouping influence student learning because they affect the allocation of instructional time in a given group. For example, teachers may spend more time supervising children in the high ability group than those in the low ability group since more difficult material typically requires more time for instruction (Bar and Dreeben 1983). Also, having more groups in classrooms may mean less time for supervised work in each ability group because time spent for supervised work for one group means unsupervised seatwork for other groups. Thus, instructional activities and time allocated for students in a given group depends not only on the mean ability level of that group, but also on the size and ability levels of other groups as well as the number of ability groups in the classroom[4].

From organizational perspectives, whether children learn more in the ability grouped setting than they do in the ungrouped setting is not merely a question of classroom organization

[4] Although Bar and Dreeben (1983) did not discuss the role of teacher's aids or parent volunteers to supervise instruction, these factors would affect the time spent for supervised work for a given group.

because the amount of student learning primarily depends on instructional activities provided by the teacher in each situational setting. Organizational perspectives also suggest that the amount of student learning depends on a larger classroom and school contexts because they influence the formation of ability grouping and instructional activities.

As discussed earlier, proponents of ability grouping argue that within-class ability grouping would be beneficial to all students if teachers could accurately identify student abilities and prior mastery skills, organize instruction accordingly, and adjust grouping and instruction as children master, or fail to master, new skills. In the whole-class instructional setting, because instruction is typically oriented to the average ability students in the class, abler children may find the material too easy or the instructional pace too slow, while less able children might find otherwise.

Opponents of ability grouping, however, argue that ability grouping may not be so desirable, especially for low achieving students, if it leads to more time for unsupervised seatwork, an inattentive and disruptive learning environment, and negative stereotyping. Also, ability grouping would be inappropriate if the ability level of students is incorrectly identified. The whole-class instruction may be beneficial to all students if teachers use instructional variations in the whole-class setting, instead of forming smaller instructional groups.

Many early ability grouping studies have reported the results that support both claims; some studies have found positive effects, other studies have found no effects, and other studies have found negative effects of ability grouping on student achievement (Goldberg, Passow, and Justman 1966; Yates 1996)[5]. While organizational perspectives explain how school structures shape organization of instruction and how this is related to instructional activities to produce

[5] Inconsistencies in findings on ability grouping/tracking effects in early studies are also noted by Sorensen (1970) and Rosenbaum (1984). Also, see Gamoran and Mare (1989) and Lucas (1999).

student learning, they do not adequately explain reasons why some ability grouping/tracking systems may be more productive (i.e., producing higher mean achievement) or more effective (i.e., producing smaller achievement inequality) than others. Inconsistencies in early research findings may be partly attributable to appropriateness and adequateness of the methods used in these studies (Sorensen 1970**).** For example, schools may differ in terms of student compositions, available resources, and learning opportunities. If school characteristics are confounded with ability grouping practices, or if the effect of ability grouping depends on school characteristics, the analysis will produce a biased estimate of ability grouping effects on student achievement without taking into account between-school differences in these school characteristics[6]. Few ability grouping studies, however, consider these factors.

More recently, researchers have argued that ability grouping is not a uniform practice and the effect of ability grouping/tracking on student achievement depends on its structural characteristics, such as flexibility and homogeneity of ability grouping (Gamoran 1992; Sorensen 1970). While Bar and Dreeben (1983) have argued that the structure of within-class ability grouping is primarily the organizational response to the ability distribution of the student body in classrooms, other researchers have maintained that organizational factors other than the student ability distribution in the classroom determine the structure of ability grouping. Differences in the ability grouping structure, in turn, produce differential effects of ability grouping/tracking by schools (Hallinan and Sorensen 1983).

Variable effects of ability grouping and tracking in secondary schools: structural explanation. Several researchers argue that school differences in ability grouping/track structures explain why some ability grouping/tracking systems are more effective than other

[6] This study will use a method that takes into account differences in school characteristics (see Chapter 4 for details).

systems in producing higher student achievement. Sorensen (1970) provided an early structural analysis (for an empirical study, see Gamoran 1992). Although Sorensen's analysis is primarily concerned with secondary school tracking and some key concepts may not be applicable to elementary school ability grouping, it offers insights to understand variable effects of within-class ability grouping on student achievement.

Sorensen (1970) identifies four dimensions on which ability grouping systems may differ between schools. They are; 1) inclusiveness, 2) electivity, 3) selectivity, and 4) scope (also, see Rosenbaum 1976, 1984; Oakes 1985).

1) Inclusiveness

Inclusiveness is defined as the degree of openness to a higher level of education (Sorensen 1970). In other words, inclusive grouping/tracking leads many, rather than few, students to a higher group or track. For example, the inclusive tracking system would include relatively more students in college preparatory curriculums than in non-college preparatory curriculums. Gamoran (1992) argues that since the inclusive tracking system is likely to provide greater learning opportunities for a lager student population than the less inclusive tracking system, it will lead to higher overall achievement and lesser degree of achievement inequalities between high and low achievers.

2) Electivity

Electivity refers to the extent to which student preferences play a role for group assignment. Some researchers have reported that group/track assignment is highly influenced by teachers and school authorities (Ball 1981; Gamoran 1992), while other researchers have reported that many students perceive that they have chosen their courses even when they follow the advice of teachers and school counselors (Jencks, Smith, Acland, Bane, Cohen, Giontis, Heyns, and Michaelson 1972; also, see Gaskell 1985).

Sorensen (1970) suggests that greater electivity increases homogeneity in non-cognitive traits, such as aspirations, motivations, and beliefs, among students in the same groups/tracks. The group homogeneity implies greater differences between groups in these traits, and these differences, in turn, magnify achievement differences between groups/tracks. In contrast, Gamoran (1992) suggests that greater electivity would lead to smaller achievement differences between tracks/groups because students who believe that they selected their own programs are more likely to be motivated and do well academically than students in the less elective ability group/track system

3) Selectivity

According to Sorensen (1970), selectivity refers to the degree of homogeneity created by ability grouping/tracking in terms of student characteristics that are relevant for learning (i.e., cognitive characteristics). He suggests that selective ability grouping/tracking systems create greater differences in student achievement by group/track levels than unselective systems. Gamoran (1992) suggests that schools with greater selectivity leads to greater achievement gaps between groups/tracks. He also suggests that selective schools may have higher overall achievement because teachers can tailor instruction more effectively when the students in a given group are more homogenous.

4) Scope

Sorensen (1970) defines "scope" as the extent to which "a given group of student will be members of the same classroom over time" (362). From Sorensen's perspective, a high degree of scope means that the same group of students takes advanced courses, or regular courses, across different subjects (Gamoran 1992). Oakes (1985) also discusses similar dimensions of tracking structures. She defines four dimensions, which include "extent" (i.e., the proportion of the total number of classes that are tracked), "pervasiveness" (i.e., the number of subjects areas that are tracked), "flexibility" (i.e., whether students are tracked subject by subject or across more than one subject) and "mobility" (the extent to which students move up or down tracks). Rosenbaum (1976) also discusses "mobility" as an aspect of scope[7].

Researchers have argued that a wider scope may create greater distinctions between groups/tracks. For example, if the same group of students takes advanced courses across different subjects, spend a significant portion of school day in the same group, and experience little mobility, then distinctions between these students and students in lower tracks would become more salient. Also, a less flexible ability grouping/tracking system implies that teachers are not likely to adjust group/track assignment according to student cognitive developments, motivational changes, and other developmental changes. Thus, Gamoran (1992) suggests that the tracking/ability grouping system with a wider scope increases achievement inequalities and it also lowers the average achievement level.

Hallinan (1994a) provides somewhat different accounts for school differences in ability grouping/tracking effects in secondary schools. She argues that the effect of ability grouping differs by schools because schools provide differential learning opportunities in a given group/track levels (see Goodlad 1984). Schools differ not only in group/track characteristics,

[7] Also, see Lucas (1999) for further discussion on this concept.

such as the size, number, homogeneity, and student compositions of ability groups/tracks, but also in the quality and quantity of instructions (e.g., instructional time, course availability, teacher's quality). These differences, in turn, affect learning opportunities provided in given groups/tracks.

Hallinan (1994a) also argues that schools differ in assignment processes, and such differences produce differential effects of ability grouping/tracking on student achievement. For example, a student can be placed in different groups/tracks depending on the school that she/he attends. This is because schools differ in ability grouping/tracking structures, flexibility in ability group/track mobility, assignment criteria, and flexibility in course scheduling. However, Hallinan's analysis (1994a) does not clearly explain how differences in assignment processes specifically lead to differential effects of ability grouping/tracking on student achievement by schools (Also, see Hallinan 1991).

The aforementioned studies are primarily concerned with secondary school ability grouping/tracking. In the following section, I will discuss how these concepts may be applicable to elementary school ability grouping.

Variable effects of ability grouping in elementary schools. Structural differences in elementary school ability grouping may also produce differential effects on student achievement (Slavin 1987). In general, Slavin (1987) argues that ability groupings that are rigid and have a wide scope may not be beneficial to student learning, while other forms of ability grouping may facilitate student learning.

Slavin (1987) identifies four ability grouping structures in elementary schools. They are 1) between-class ability grouping for all subjects (i.e., ability grouping that lasts for the entire day), 2) between-class ability grouping for specific subjects, 3) un-graded ability grouping for specific subjects, and 4) within-class ability grouping.

In his meta-analysis, Slavin (1987) compares achievement outcomes of students in schools with various ability grouping structures to those of students in schools without ability grouping. There are two types of analyses. One type of analysis compares the average student achievement and the other type of analysis compares achievement of students who differ in initial ability levels (i.e., low, middle, and high achievers) across different ability grouping structures.

The findings show that schools that practice between-class ability grouping for the entire day *do not* produce higher achievement or greater achievement inequality than schools without ability grouping. However, both un-graded ability grouping and within-class ability grouping produce higher student achievement at all achievement levels than non-grouping (Slavin 1987).

Slavin (1987) argues that the poor performance of the self-contained classes (i.e., between-class ability grouping for the entire day), compared to other forms of ability grouping, is attributable to two factors. First, because self-contained ability grouping is rather rigid and has a wide scope, it creates more salient distinctions between students by group levels. This may increase negative social-psychological effects on students in lower groups. The second factor relates to group homogeneity. Slavin (1987) suggests that self-contained grouping may create greater heterogeneity in some subjects and not other subjects because the ability of students may vary by subject areas. Heterogeneities in student compositions would, in turn, decrease effectiveness in instruction.

Within-class reading ability grouping is the most typical form of ability grouping in elementary schools and is the focus of my dissertation. Slavin (1987) does not discuss whether the effect of within-class ability grouping differs by schools.

Sorensen (1970) and Hallinan's (1994) conceptual frameworks are useful to understand whether within-class ability grouping may differentially affect student achievement depending on school characteristics. For example, school compositions may affect the structure of within-class ability grouping. Hallinan (1984) finds that the size of high ability groups in desegregated schools tend to be larger than that of high ability groups in all-black or all-white schools. A larger group size would increase heterogeneity.

Schools also differ in the number of groups in classrooms. In classrooms with a given number of students, a greater group number implies greater selectivity. For example, the highest group level in the classes with five groups would be more selective and homogenous than the highest group level in the classes with two ability groups with class size being equal. In addition, elementary classrooms may differ in group mobility; some teachers may be more likely to adjust student ability group placement than other teachers during the school year. Schools may also differ in learning opportunities provided for students in a given group level—some schools may have greater learning opportunities (e.g., greater instructional resources and higher teacher qualities) than other schools. Dreeben and Gamoran (1986) find that racial inequality in reading achievement is primarily explained by differences instructional time and content coverage between schools that are attended by white and black children[8].

Within-class ability grouping structures and group homogeneity: different perspectives. The above studies suggest that structures of within-class ability grouping and various school

[8] Their analyses used words learned and standardized reading achievement scores as dependent variables.

characteristics are important factors to consider in understanding how within-classroom ability grouping affects student learning in elementary schools. However, researchers have often ignored differences in structural characteristics of ability groups, such as the size, number, and homogeneity. In particular, researchers have often viewed differences in group numbers across classrooms as analytic problems when comparing the effect of ability group *placement* across different classrooms. Researchers have typically made a common matrix of group levels in order to make the results comparable across different classrooms[9]. This is an important limitation if school or classroom characteristics shape ability grouping numbers and student learning varies by the number, size, and homogeneity of ability groups.

One rationale for ability grouping is to create homogenous groups, which is thought to be more conducive to student learning than heterogeneous groups. Several researchers have theorized how school and classroom characteristics affect ability grouping structures and how these, in turn, affect group homogeneity and student learning. Bar and Dreeben (1983) and Sorensen and Hallinan (1983) both argue that the structural characteristics of classrooms primarily determine the formation of ability groups. However, they have different views on how structural characteristics of classrooms influence group homogeneity.

First, Bar and Dreeben (1983) argue that characteristics of ability grouping (e.g., the size and number) are determined by the distributional characteristics of student abilities in the classroom as well as the class size. They view ability group formation as responsive to classroom and school organizational characteristics. Bar and Dreeben (1983) assume that teachers attempt to create homogenous groups to handle the diversity in the student body, so they will adjust the number and size of the groups depending on the ability distributions of the classrooms. With regard to the group number, Bar and Dreeben (1983) argue that the group

[9] This is done, for example, by categorizing groups in terms of high, medium, and low.

number is greater when the class size is larger, when classrooms are more diverse in student abilities because teachers attempt to increase homogeneity in groups, and when classrooms have a larger proportion of low achieving students because teachers try to better accommodate their instructional needs.

In contrast, Hallinan and Sorensen (1983) view the organizational characteristics of classrooms, such as ability distributions and class size, as constraints on the formation of homogenous ability groups. Hallinan and Sorensen (1983) argue that teachers attempt to equalize the group size across groups and create the number of groups that is considered as the norm regardless of differences in classroom compositional characteristics. They suggest that three groups are the most typical group number across schools. However, because of such teachers' attempts, the organizational characteristics of the classroom (e.g., ability distributions and class size) can impede teachers from creating homogenous groups. In other words, teachers' attempts to create three groups of equal sizes lead some classrooms to have more heterogeneous ability groups than in others because classrooms differ in terms of their size and student compositions. This partly explains why ability grouping/tracking in some schools may produce higher achievement and greater inequalities than that in other schools.

Research has shown that even though the purpose of ability grouping is to create homogeneous groups, considerable heterogeneity exists in elementary ability groups. For example, in the study of first grade ability grouping, Pallas et al. (1994) find that "children with comparative levels of measured academic performance and social backgrounds are often placed in reading groups that rank them anywhere from near the top of their class to near the bottom" (43). However, the source of group heterogeneity is not clear from the findings. It may have resulted from the misplacement of students because the standardized test scores do not accurately reflect student ability or because teachers use criteria other than the measures of cognitive skills, such as attitude and behavior. Alternatively, organizational constraints may have limited the formation of homogenous groups.

How schools differ in the degree of group homogeneity and how such differences relate to student learning are important empirical questions. A meta-analysis by Lou, Abrami, Spence, Poulsen, Chambers, and d'Apollonia (1996) presents studies on small group instruction that compare achievement between students in homogenous ability groups and those in heterogeneous small groups. When the average achievement is compared, the studies generally find positive effects of homogeneous grouping compared to heterogeneous grouping. Of 20 studies examined, 13 studies have produced a higher mean achievement for homogeneously grouped students than for heterogeneously grouped students, one study finds no significant differences, and six studies find lower average achievement for homogeneously grouped students than for heterogeneously grouped students (Lou et al. 1996).

However, when achievement is compared by different ability levels, the results suggest that homogeneous grouping is not beneficial to low ability students in comparison with small heterogeneous grouping. Lou et al. (1996) find that low ability students learn significantly more in heterogeneous small groups than their counterparts in homogenous groups, while middle and high ability students learned more in homogenous groups than comparable students in heterogeneous groups. This may be because the presence of high ability students in heterogeneous groups benefits low ability students because the teacher provides more challenging instruction than he/she would when these students are being grouped with other low ability students, and vise versa for middle and high ability students.

Summary

This chapter focused on theoretical discussions that are relevant to ability grouping research. The macro-level theories—functionalist perspectives and social reproductionist/conflict perspectives—offer two contrasting perspectives on the roles and functions of schools as social institutions, the functions of ability grouping as school organizational practices, and the roles it plays in a larger society. In particular, these two macro theories help us understand the role of ability grouping in educational stratification.

To understand how ability grouping affects student learning, it is important to recognize school contexts in which ability grouping occurs as factors influencing student achievement. In an economically and racially segregated society, students attend schools that differ greatly in demographic characteristics and ability compositions of schools. Ability grouping practices may not be uniform because of these diversities in school characteristics. This dissertation examines how ability grouping effects may vary by school characteristics.

Two micro-level perspectives help us understand how school contexts may shape relationships between ability grouping practices and student achievement. First, organizational perspectives help us understand how schools produce student learning and the role of ability grouping in the production of knowledge. In particular, they focus on school organizational characteristics and how such characteristics are linked to the formation of ability grouping, instructions provided to students, and student learning. Second, structural analyses attempt to explain why ability grouping and tracking effects may vary by characteristics of schools in both secondary and elementary school contexts. The formation of homogenous ability groups may be a factor that affects student learning. The next chapter focuses on empirical literatures on the effect of ability grouping on student achievement.

CHAPTER 3

ACADEMIC CONSEQUENCES OF ABILITY GROUPING AND TRACKING

Research on the academic consequences of ability grouping is extensive. Past studies have examined the extent to which ability grouping/tracking contributes to overall student learning and educational inequalities. While some studies examine effects of between-school ability grouping or tracking in secondary schools, other studies examine those of within-classroom ability grouping. In general, two types of research questions have been addressed. The first research question, which is common to sociological research, is concerned with the effect of ability group or track *placement* on student achievement. The second research question is concerned with the consequences of attending ability grouped classes or schools. This chapter presents empirical findings on academic consequences of ability grouping/tracking by research type and grade levels.

Studies on Ability Grouping/Tracking Placement

A majority of sociological studies have examined academic consequences of ability grouping or track placement in secondary schools (see, for example, Alexander, Cook, and McDill 1978; Gamoran 1987; Hoffer 1992). Secondary school studies typically compare the performance of students in higher tracks/ability groups to those in lower tracks/ability groups. Studies on high school tracking examine achievement differences between those who are in academic or college tracks and those who are in general, non-academic, or vocational tracks. In middle schools, comparisons are made among students in high-, middle-, and low-ability classrooms (Hoffer 1992; Argys, Rees, and Brewer 1996). While fewer studies have examined consequences of ability grouping placement on student achievement in elementary schools, existing studies have examined achievement difference among students in high, middle, and low ability groups within classrooms (Barr and Dreeben 1983; Gamoran 1986; Pallas et al. 1994; Rowan and Miracle 1983; Tack and Farkas 2003).

As discussed earlier, the structures of ability grouping differ between secondary and elementary schools. While secondary schools typically use between-classroom ability grouping and students are often differentiated by curriculum and course taking patterns, in elementary schools students are ability grouped within the same classroom with common curriculum. Regardless of such structural differences, studies consistently show that students in higher tracks/groups learn more and at a faster pace than students in lower tracks/groups after student background and prior performance are taken into account.

These findings seem to be consistent with findings of other studies which suggest that ability grouping helps those in high ability tracks/groups and negatively affects those in low tracks/ability groups. Differential effects of ability grouping by student placement levels can occur through the differential allocation of opportunities-to-learn and social-psychological factors (for example, see Bar and Dreeben 1983; Elder 1981; Gamoran 1987; Hallinan 1987; Oakes 1985; Rist 1970; Rosenbaum 1970). Students in high ability groups/tracks are taught more difficult materials at a faster pace than students in low ability groups/tracks. They are also taught by more qualified and motivated teachers (Alexander et al. 1978; Berends 1994; Friedkin and Thomas 1997; Gamoran 1986, 1987; Gamoran and Mare 1989; Gamoran, Nystrand, Berends, and LePore 1995; Hauser, Sewell, and Alwin 1976; Oakes 1985; Rosenbaum 1980; Schafer and Olexa 1971; Vanfossen, Jones, and Spade 1987).

In contrast, teachers in lower ability groups/tracks have lower expectations, less enthusiasm, and focus more on disciplinary issues than critical thinking (Gamoran and Berends 1987; Gamoran and Page 1992; Gamoran et al. 1995; Lacey 1970; Metz 1978; Rosenbaum 1976; Oakes 1985; Page 1991). Peer influences also affect student learning (Hallinan 1987). For example, grouping low achieving students together may create a disruptive and inattentive environment which is not conducive to learning (Page 1991). In such an environment, students may be less engaged in their schoolwork and teaching may not be effective (Eder 1981; Hallinan 1987; Kellam 1994).

The findings of the placement research have often led to a conclusion that ability-grouping practices widen achievement inequalities. However, there are both conceptual and methodological limitations to this line of research. First, seemingly negative consequences of ability grouping on low ability students have led to a policy implication for “detracking”. For example, such negative consequences will be minimized if classrooms, instead, adopt whole class instruction and create a corporative learning environment. However, because most ability grouping placement studies compare academic achievement only among ability-grouped students and do not include ungrouped students in their analyses, they do not address an important policy question: “How would students achieve if they were not ability grouped?”[10] Second, these placement studies typically use regressions and such analytic methods are potentially problematic in drawing causal inferences about ability grouping placement (this issue is discussed in Chapter 4).

[10] A few studies (Argys, Rees and Brewer 1996; Hoffer 1992; Tack and Farkas Forthcoming) examine the effect of ability group *placement* by including ungrouped students in their analysis. 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 who are placed in different ability groups would perform if they were not ability grouped (also see Kerckhoff 1986).

The second type of ability grouping research has specifically addressed the question of ability grouping policies by comparing achievement between ability-grouped and ungrouped students. Some studies examined between-class ability grouping, while others examined within-class ability grouping. The results depend on types of ability grouping under study. These studies are discussed next.

Studies on Ability Grouping vs. No Grouping

Unlike ability grouping placement research, other studies attempt to understand the consequences of ability grouping practices on student achievement by comparing ability-grouped students and ungrouped students. In addition, these studies differ from placement studies in that researchers have used experimental designs and matched experimental designs. It is also important to note that while placement research finds similar effects of ability grouping regardless of differences in ability grouping structures, the results of studies on ability grouping practices depend on whether grouping involves curricular differentiation and whether grouping takes place within classrooms. The types of ability grouping discussed here include gifted programs, between-class ability grouping, and within-class ability grouping.

There are extensive studies on special programs and grouping arrangement that are designed for students who are identified as gifted or talented (Kulik and Kulik 1982, 1984, 1990; Slavin 1987; Vaughn, Feldhusen, and Asher 1991). They include full-time gifted programs, grouping for acceleration of the curriculum, re-grouping for enriched learning in specific subjects, and enrichment pull-out programs. Grade levels vary from elementary grades to college. These programs or grouping arrangements are often designed superficially to provide certain instruction to high ability students. Generally, findings suggest that they lead to higher achievement for high ability students than their counterparts in the regular classroom arrangements. For example, Kulik and Kulik (1982) point out that in such programs high ability students are likely to benefit from the stimulation provided by other high ability students and from the special curricula.

The positive effects may be attributable to grouping per se, curriculum differentiation, and/or instructional methods used in these programs. It is also likely that a particular grouping arrangement made it possible to use curriculum differentiation and particular instructional methods. A limitation of these studies, however, is that it is not clear what unintended consequences of these programs might have on the learning of other students. That is, greater learning among high ability students in these programs may occur at the expense of the learning of middle and low ability students.

The second type of research analyzes the consequences of between-class ability grouping on average student achievement or student achievement by initial ability levels. Between-class ability grouping in these studies often involves the differentiation of courses by ability levels, which is a prevalent practice in secondary schools. However, these are not special programs that are specifically designed to provide certain instruction for a particular group of students (i.e., gifted students).

Findings of this research show that the *average* achievement does not differ between ability-grouped and non-grouped students (Betts and Shkolnik 2000; Hoffer 1992; Kulik and Kulik 1982, 1987; Slavin 1987, 1990). Some researchers, however, argue that the comparison of the average achievement masks the fact that the effect of ability grouping on student achievement may vary by students' initial ability levels (Hallinan 1990; Hoffer 1992). That is, ability grouping may have positive effects for high-ability students and negative effects for low-ability students, so the net effect of ability grouping on the mean achievement would be zero[11]. The existing studies, however, suggest that in secondary schools, ability grouping has little effects on student achievement at any initial ability levels (Betts and Shkolnik 2000; and Kulik and Kulik 1982; Slavin 1990).

It is important to point out that unlike studies on programs for "gifted" students these studies do not address the effect of curriculum differentiation or particular instructional methods. For example, Slavin (1987) notes that while schools in his meta-analysis often tracked students, many of these studies made comparisons between tracked students in the *same* courses, and they did not compare achievement between students in different tracks. (i.e., comparisons are made between students who are enrolled in the same course of different ability levels, rather than courses that are differentiated by curriculum). In other words, these studies control for different course taking patterns. Thus, these studies do not provide evidence for a lack of the effect of curriculum differentiation or specific instructional practices on student achievement.

The third type of research analyzes the consequences of within-class ability grouping, and this is a central question addressed in this dissertation. Consequences of within-class ability grouping may differ from those of gifted/talented programs and between-classroom ability grouping discussed above because within-class ability grouping differs from these two types of ability grouping in a few important ways. First, within-class ability grouping in elementary schools typically does not involve curriculum differentiation. The importance of curricular differentiation on academic consequences is illuminated by the findings of gifted/talented programs and those of between-class ability grouping discussed above. That is, programs that

[11] Using the LSAY data, Hoffer (1992) presents the OLS and propensity analyses of ability group *placement*. His analysis includes ungrouped students. The findings suggest that ability grouping has positive effects for students in high ability groups and negative effects for students in low ability groups in comparison with the *average* ungrouped students (methodological limitations are discussed in Chapter 4). Betts and Shkolnik (2000) re-analyzed the same data and found that ability grouping has little differential effects for high, average, or low-achieving students.

are designed for gifted/talented students improve student learning for these students while between-class ability grouping has little effects on student achievement when curricular differentiation is taken into account.

Second, the size of instructional groups of within-class ability grouping is small because ability grouping takes place within classrooms rather than between classrooms. In addition, small group arrangement can minimize heterogeneities in student ability compositions, which may allow teachers to organize instruction more effectively.

Prior studies on within-class ability grouping have shown that, unlike all the other studies discussed earlier, the average performance of ability-grouped students is higher than that of their ungrouped counterparts. Also, when student achievement is compared at different initial ability levels (e.g., low, middle, and high achievers), ability-grouped students are found to have higher test scores than students in ungrouped classrooms at all initial ability levels (Lou et al. 1996; Slavin 1987). These positive effects of within-class ability grouping may be explained by rationales of ability grouping discussed in Chapter 2.

In early elementary grades, teachers typically use within-class ability grouping in reading. However, a limitation of current ability grouping research is that while ability grouping research is extensive, very few studies have indeed examined the causal effect of within-class reading ability grouping on student achievement in early elementary years, using nationally representative samples.

For example, Slavin (1987)'s meta-analysis includes eight studies on elementary within-class ability grouping (five randomized studies and three matched experimental studies). Yet, none of them have analyzed reading ability grouping in early grades[12].

[12] All five randomized studies are studies on math ability grouping in upper elementary grades. Of three matched experimental studies, two studies examine math ability grouping (one in upper grade and the other in lower grade)

A meta-analysis by Lou et al. (1996) aggregates studies on various forms of within-class instructional grouping, which include ability grouping, heterogeneous grouping, cooperative learning, and other small group instruction, in various subjects from elementary to post-secondary levels. Although the results find positive effects of both homogenous and heterogeneous grouping for all students compared to whole class instruction, effect sizes vary considerably by subjects and grade levels. For example, the average effect size for reading is smaller than that of math or science. The effect sizes for early elementary grades are smaller than those for any other grade. Because Lou et al. (1996) include few studies on early elementary reading ability grouping, their findings may not be generalizable to this population.

From a methodological point of view, the above studies on within-classroom ability grouping would warrant greater internal validity than correlational studies because they employed experimental and matched experimental designs. In contrast, the external validity in such studies is limited because samples do not typically represent a larger population. In addition, the existing studies do not address the extent to which ability grouping effects differ by school or classroom contexts and what factors might explain such variations.

These limitations are addressed in this dissertation. First, this study uses the Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (National Center for Educational Statistics 2000), which contains a nationally representative sample of kindergarteners. Unlike previous observational studies, which typically rely on ordinary least squares regression, this study applies the propensity-score stratification (Rosenbaum and Rubin 1984) to analyze the causal effects of within-classroom ability grouping in first grade. Unlike previous correlational studies on ability grouping, this study may yield more credible evidence for the causal effects of ability grouping on student achievement. The next chapter discusses

---

and one study examines reading ability grouping in fourth grade.

methodological limitations of previous ability-grouping placement studies and summarizes methodologies of causal inferences applied to this study.

CHAPTER 4

DATA AND METHODS

To understand how ability grouping affects student achievement, most researchers to date have used observational data to examine the effect of ability grouping *placement* on student achievement. The first section of this chapter critically examines the methodological shortcomings of such studies in understanding causal effects of ability grouping. A propensity score approach is proposed to examine the causal effect of ability grouping (not placement). The latter sections discuss data and analytic methods used in this dissertation.

Methodologies for Causal Inference

Methodological limitations in past non-experimental studies on ability grouping. The conceptual framework for causal inference applied in this study is based on the potential outcomes framework (Holland 1986; Rubin 1978; Rubin 2005). The key assumption of the potential outcomes framework is that, using the language of experiment, subjects assigned to treatment and control groups have potential outcomes in both states (Winship and Morgan 1999). The causal effect is defined as the difference in potential outcomes between these two states; a potential outcome when a subject is given the treatment (Y^t) and when the same subject is not given that treatment (Y^c). Formally, this is expressed as:

$$\delta = Y_i^t - Y_i^c \tag{1}$$

However, because these two treatment conditions cannot be given to the same subject simultaneously we will never observe a causal effect of the treatment for any one subject. Holland (1986) calls this as the “fundamental problem of causal inference.”

This is a problem of missing data, and one way to deal with it is to conduct a randomized experiment. Randomization of subjects into treatment and control conditions creates two groups that are, in expectation, equivalent in both observed and unobserved characteristics; the only difference is that one receives the treatment and the other does not. Members of the two groups have equal probabilities of receiving the treatment. This condition allows us to estimate the *average* causal effect, that is, the average differences in potential outcomes between treatment and control groups, without bias. The standard estimator for the average treatment effect is estimated as:

$$E(\overline{\delta}) = E(\overline{Y}^{t}_{i\in T}) - \mathrm{E}(\overline{Y}^{c}_{i\in C}), \tag{2}$$

where $\overline{\delta}$ is the average treatment effect, $\overline{\mathrm{Y}}^{\mathrm{t}}$ is the average outcome for the treatment group, and $\overline{\mathrm{Y}}^{\mathrm{c}}$ is the average outcome for the control group. The definition of a causal effect under the potential outcomes framework makes an important assumption, called the stable unit treatment value assumption (SUTVA). SUTVA states that one’s potential value associated with each treatment is not affected by how the treatments are assigned and what treatment is received by other subjects (Rubin 1986; also see Cox 1958).

When a randomized experiment is not feasible, researchers often have to rely on observational studies. Under such circumstances, the strong ignorablility assumption must hold to estimate an unbiased treatment effect. This assumption states that given observed covariates, the potential outcomes are independent of treatment assignment. This is an issue of selection bias, which is a potential problem of observational studies. In true experiments, the potential outcomes and treatment assignment are independent due to randomization. In other words, no factor other than chance determines who receives the treatment. However, in non-randomized studies where researchers have no control over the treatment assignment, treatment and control groups may differ prior to the treatment, and omitted variables may bias an estimate of the treatment effect if these variables are associated with the treatment assignment and the outcomes.

Various methods can be used to estimate causal effects using observational studies. They include linear covariate adjustments, instrumental variable methods, difference-in-difference methods, regression discontinuities, and propensity score methods[13] (for example, see Angrist, Imbens, and Rubin 1996; Dynarski 2003; Kreuger 1999; Rosenbaum 2002; Shadish et al. 2002). In ability grouping research, most observational studies have applied linear covariate adjustments and researchers typically use the ordinary least square (OLS) regression to estimate the effect of ability grouping placement (for example, see Alexander et al. 1978; Gamoran 1987; Pallas et al. 1994).

There are, however, several limitations in such approaches. First, problems of the OLS regression may arise when subjects find no comparison groups. Ability grouping placement research shows that student characteristics—most notably student initial abilities—differ by ability group levels. It is possible that students in a given ability group have few comparison groups who have similar characteristics, but are, in fact, assigned to a different ability level. For example, students in a high ability group may find few students who are just like them in a low 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, the OLS method relies largely or in part on

[13] These methods adjust for selection bias while estimating the treatment effect. One can also incorporate measured covariates into the design of the study through, for example, matched sampling (Shadish, Cook, and Campbell 2002).

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, the regression model assumes that the average treatment effect is constant for the population[14]. This means that, for example, the OLS estimates predict that students in a low ability group would gain as much as high achieving students if they are placed in a high ability group (see Gamoran and Mare 1989; Hoffer 1992 for discussions). In addition, the treatment effect may depend on many characteristics other than student initial ability levels. The assumption of a constant treatment effect needs to be empirically tested. Similarly, it is possible that if many low achievers are moved from a low group to a high group, the benefit of the high-group placement may diminish. These factors are not considered in conventional placement research.

Third, linear covariate adjustments may only include a limited number of covariates to avoid a problem of multi-collinearity or a loss of precision by reducing the degrees of freedom. Also, it becomes more difficult to examine interaction effects between the treatment and covariates as the number of covariates increase in the model.

To address these methodological shortcomings in conventional ability grouping placement research, some researchers have used propensity score methods (Betts and Shkolnik 2000; Hoffer 1992)[15]. These studies also include students who are not ability grouped as a comparison group. In studies of between-classroom ability grouping placement in middle schools, Hoffer (1992) and Betts and Shkolnik (2000) use ordered probit models to estimate the propensity scores of students being placed in different ability group levels (e.g., low, middle, and high ability levels). First, these authors use a sample of students who attend schools using ability

[14] This is true only if the model has no interactions between the treatment and covariates.
[15] Propensity-score methods are further discussed below

grouping to estimate the propensity scores of ability group placement. Then, the probit coefficients are used to calculate predicted probabilities of ability group placement for students who attend schools without ability grouping. After estimating the propensity scores of ability group placement for all students, students are subdivided into quintiles based on the estimated propensity scores[16]. Then, test scores of students who are placed in different ability levels are compared to those of ungrouped students within a given stratum.

However, a problem in studies by Hoffer (1992) and Betts and Shkolnik (2000) is that under the potential outcomes framework, they are likely to violate the SUTVA. Both studies use the Longitudinal Study of American Youth (LSAY) data in which students are nested within classrooms and schools. However, these authors did not take into account how learning experiences of a particular student may be influenced by the placement of other students in the same schools. For example, previous studies suggest that learning climates and peer influences in high ability groups are more conducive to learning than those in low ability groups (Hallinan 1987; Page 1991). Also, a purpose of ability grouping is to provide instruction that is commensurate with student ability and the pace and difficulty of instruction is primarily determined by the ability level of the group (Bar and Dreeben 1983). This indicates that the academic performance of a student is determined not only by his/her own ability group placement, but also by how other students in the same classes or schools are assigned to ability groups. If the assignment of other students influences the outcome of a student by changing the average ability level of the group or changing the context of learning or peer relationships, the SUTVA is not likely to hold.

[16] Betts and Shkolnik (2000) used propensity score quartiles.

The advantages of the propensity score methods. Unlike previous studies, this study employs propensity score methods to estimate the causal effect of ability grouping practices (i.e., not the effect of placement). No other studies to my knowledge have yet examined the causal effect of within-class ability grouping practices in early elementary years using a nationally representative sample[17]. The analysis of ability-grouped versus ungrouped students not only directly addresses the question of ability grouping policies (i.e., how students achieve if they are ability-grouped) and but also avoids the violation of the SUTVA. In addition, heterogeneity in the treatment effect can be examined.

Advantages of using propensity score methods in non-randomized studies have been written elsewhere (for example, see D'Agostino 1998; Dehejia and Wahba 2002; Morgan and Harding 2005; Rosembaum and Rubin 1984, 1985). First, let X be the vector of pre-treatment variables and T be the treatment status. A main advantage of propensity score method under the potential outcomes framework is that propensity scores reduce the dimensionality of X. For example, it will be difficult to find a comparison group with an exact value on X for each treated unit as the number of variables increase. If all *n* variables are categorical variables the number of possible values for a vector X will be 2^n.

While traditional methods of bias adjustment, such as matching and covariate adjustment can use only a limited number of covariates, propensity score methods do not have this limitation because they provide a scalar summary of the covariate information (D'Agostino 1998). To reduce the dimensionality of X, Rosenbaum and Rubin (1983) propose to use propensity scores—the probabilities of receiving the treatment, given X. The estimation of propensity

[17] Tach and Farkas (forthcoming) attempt to do this analysis using a covariate adjustment approach. However, their studies do not take into account the fact that schools that use ability grouping and those that do not use ability grouping are different in many characteristics, such as average student achievement, race, and SES compositions. While their analyses control for classroom characteristics, omitting these school-level variables may bias the estimate of ability grouping effects.

scores is written as $P(X) = Pr(T=t|X)$ where $t=1$ if subjects receive the treatment and $t=0$ otherwise.

Rosenbaum and Rubin (1983) show that if the treatment assignment is independent conditional on X, then it is also independent conditional on the propensity score, P(X). This is to say that if the treated and untreated units are balanced on the propensity score, they are also, in expectation, balanced on X. Then, the treatment effect is defined as the difference in the outcomes between the treated and untreated groups, given the propensity score, P(X). Although the estimated treatment effect is conditional only on the observed covariates, if one can measure many of the covariates that are thought to be related to the treatment assignment, then one can be fairly confident to say that the estimated treatment effect is approximately the unbiased treatment effect (D'Agostino 1998).

Second, unlike linear covariate adjustment approaches, propensity score approaches allow us to examine the extent to which the treated and untreated units overlap with respect to observed characteristics. Before estimating the treatment effect, it is important for researchers to investigate the extent to which the treatment groups have viable control groups who overlap on observed covariates. This is done by examining the region of common support— the overlap of propensity scores, P(X) between the treated and untreated units.

Limiting causal inferences to the region of common support may result in discarding some observations. In such a case, the resulting estimate is informative only for those who are equivalent with respect to observed treatment selection variables (Morgan and Harding 2005). However, this may be substantively more meaningful because it allows more precise interpretation of the treatment effect. For example, in evaluation research, researchers are typically interested in estimating the treatment effect for those who received the treatment. In such cases, it would be more informative to select a non-treated subject for each treated subject who is equivalent in X, and discard all subjects that do not have comparison groups.

Third, propensity score methods can address the heterogeneity of treatment effects. For example, the treatment effect may vary by the level of X, or (equivalently) by the level of P(X). Similarly, the treatment effect for the treated group and that for the untreated group may not be the same. For example, to evaluate an effect of a particular program, one can estimate the causal effect for those who are already in the program. However, this estimate may not be what we might expect if we expand the program to those who are not likely to participate in the program. Propensity score methods can examine the treatment effect for this population as long as they find comparison groups who overlap on the propensity scores.

Fourth, the treatment effect can be semi-parametrically estimated by using propensity score matching or stratification methods, which rely on weaker assumptions than parametric methods for their validity of the treatment effect. For example, propensity score matching and stratification methods make no assumption regarding the distribution of the dependent variable (Y). Also, in parametric models, the incorrect specification of functional forms of X in relation to Y biases the estimate of the treatment effect, while estimators based on matching and stratification do not assume linearity between the dependent variable (Y) and covariates (X).

Propensity score methods may face the same difficulties of parametric methods because propensity scores are typically estimated parametrically (e.g., logistic regressions). The previous literature has shown mixed findings with regard to the sensitivity of the estimated treatment effect to the misspecification of the propensity scores. Some researchers argue that the estimated treatment effects may be sensitive to the model misspecification; however, the sensitivity may depend on the violation of the strong ignorability assumption (Smith and Todd 2005). Other researchers argue that the estimated treatment effects are quite robust to misspecified propensity models (Dehejia and Wahba 1999; Drake 1993; Jalan and Ravallion 2003; Levine and Painter 2003; Zhao 2005).

Methods of basis reduction using propensity scores. In observational studies, propensity scores are often used to reduce bias and increase precision (D'Agostino 1998). D'Agostino (1998) discusses the three most common techniques that use propensity scores. They are matching, stratification (or subclassification), and covariance (regression) adjustment. Propensity score matching methods are a technique used to select a control subject who is matched to a treated subject who has a similar propensity score. While it is difficult to match a control subject to a treated subject on all covariates that are needed to be controlled, propensity score matching methods solve this problem by allowing an investigator to control for many covariates simulataneously by matching on a single scalar variable (D'Agostino 1998). Matching is particularly useful, for example, when there is a relatively small group of the treated subjects and a much larger group of subjects who are not exposed to the treatment (Rosenbaum and Rubin 1985). In some cases, the treated subjects may also have very different characteristics than the control subjects. In such circumstances, matching allows an investigator to select the control subjects for the treated subjects who are similar in observed characteristics, while discarding all the other control subjects who are not comparable.

Although it is beyond a scope of this dissertation to discuss various matching methods in details, matching methods used in previous studies include nearest available matching on the estimated propensity scores, Mahalanobis metric matching including the propensity scores, nearest available Mahalanobis metric matching within calipers defined by the propensity scores, and kernel matching (see D’Agostino 1998; Rosenbaum and Rubin 1985; Heckman, Ichimura, and Todd 1997, 1998).

Stratification is a technique that controls for systematic differences between the treated and control groups by grouping subjects into strata based on observed characteristics, so that within a stratum, the distribution of the covariates between the treated and control groups are equivalent. As in matching, stratification on observed covariates becomes more difficult as the number of covariates increases. The propensity score is useful because it is a scalar summary of all the observed covariates and stratification on the propensity score alone can balance the distributions of the covariates in the treated and control groups (D’Agostino 1998; Rousenbaum and Rubin 1984). Rousenbaum and Rubin (1984) also show that stratification on the propensity scores balance all covariates that are used to estimate the propensity scores, and five strata remove 90 percent of the bias (Cochran 1968; Rousenbaum and Rubin 1984).

The third approach is to use covariance (regression) adjustments with propensity scores (D’Agostino 1998). However, D’Agostino (1998) states that covariate adjustments need to be performed with a caution because, as Rubin (1979) shows, this approach may increase the expected squared bias if the covariance matrices in the treated and untreated groups are unequal. In addition, covariance adjustment is problematic when the variance in the treated and untreated groups is very different (D’Agostino 1998). Propensity-score matching and stratification may be preferable to covariate adjustment because their estimators are non-parametric, which rely on fewer assumptions than regression estimators.

This dissertation uses propensity score stratification methods because, as shown below, the sample sizes are fairly similar between the treated groups (i.e., ability-grouped schools and classrooms) and control groups (i.e., ungrouped schools and classrooms). Also, these two groups are found to have a sufficient overlap on the estimated propensity scores.

Application of Propensity Score Methods to Within-class Reading Ability Grouping Studies using ECLS-K Data.

Data. This study uses data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (National Center for Educational Statistics, 2000). ECLS-K began in the fall of 1998 with a nationally representative longitudinal sample of 21,260 children in about 1,200 kindergarten programs. Of all, 17,487 were subsampled for a longitudinal study and were assessed in reading, math, and general knowledge in the fall of 1998 (Fall-K), the Spring of 1999 (Spring-K), and the Spring of 2000 (Spring-1st). Additionally, a 30 percent subsample of the original longitudinal sample was drawn and assessed in the fall of 1999 (Fall-1st). The same children were assessed in the spring of third grade (2002) and fifth grade (2004). This study will only use data from kindergarten and first grade assessments.

The ECLS-K base year data collection (1998) employed a multistage probability sample design. First, 100 Primary Sampling Unites (PSU)—geographic areas which consist of counties or group of counties—were selected. These PSU were selected with probability proportional to size, using the number of five year olds as the basic measure of the size (see chap 4. in the ECLS-K user guide manual for details). In the second stage, 1,277 public and private schools with kindergarten programs were sampled within the Primary Sampling Units. These schools were selected using probability proportional to a weighted measure of size and the number of kindergartners enrolled was used as the measure of size. Finally, approximately 23 children were sampled within each of the sampled schools (West, Denton, and Germino-Hausken, 2000). It is noted that Asian Pacific Islanders (API) were the only subgroup which needed to be oversampled to meet the study's precision goals. To oversample API, two independent sampling strata were formed with one stratum comprising schools without API and the other stratum comprising of schools with API. Within each stratum, students were selected using equal probability systematic sampling, using a higher rate for the Asian Pacific Islanders stratum.

The ECLS-K assessments consist of parent questionnaires, teacher questionnaires, which are given to classroom teachers, and school questionnaires, which are given to school administrators. This study uses data from kindergarten and first grade assessments. School movers are excluded from the sample if their first-grade schools were not among the originally sampled schools[18]. The base sample of this study consists of 13,512 first graders in the fall 1999 assessment.

Ability grouping measures. An advantage of using the ECLS-K dataset is that it includes a sample of ungrouped students which can be used as a comparison group. In the ECLS-K first-grade teacher questionnaires, the classroom teacher of each sampled student was asked about the number of ability groups in the child classroom and the ability-group placement of the child. These two items are used to construct the ability grouping status of students, classrooms, and schools, which are summarized in Table 4.1.

Table 4.1. Ability Grouping Status of Students, Classrooms, and Schools

Student Grouping Status	School Grouping Status: Grouped	Ungrouped	Mix*	Total
Ability grouped	6,742		2,848	9,590
(# classrooms)	(1,551)		(838)	(2,389)
Ungrouped		2,043	1,879	3,922
(# classrooms)		(304)	(513)	(736)
Total	6,742	2,043	4,727	13,512
(# of schools)	(451)	(142)	(307)	(899)

*schools with both grouped and ungrouped classes

[18] In other words, school movers were kept in the sample if they moved to a school in the base sample. A dummy variable is created to indicate these students.

The ECLS-K data consist of three types of schools. The first type of school 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 students in the same school are 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. There are 4,727 students in the third type of schools. Of those, 2,848 students are in ability-grouped classrooms and 1,879 students are in ungrouped classrooms. The number of schools for grouped, ungrouped, and mixed schools is 415, 127, and 296 schools, respectively.

Other measures. The dependent variable of this study is reading achievement in spring of first grade. This variable comes from the ECLS-K direct child cognitive assessment and is measured by IRT scales. Independent variables include student, classroom, and school characteristics. Student-level variables include demographic characteristics of students and their families, parental involvement, students' preschool experiences, their cognitive skills, their English proficiencies, their behavioral characteristics, and parental and teacher's ratings on children's cognitive skills and behaviors. All of the student-level variables come from teacher and parent questionnaires in spring and fall kindergarten. As discussed below, many of child variables are used to construct aggregate characteristics of classrooms and schools.

Classroom-level variables include teachers' experiences, classroom demographic characteristics, student behavioral characteristics, distributional characteristics of children's cognitive skills in the beginning of the first-grade school year. Some classroom-level variables come from teacher questionnaires and other variables are aggregated using student-level variables by teacher ID numbers (Appendix A)

School-level variables include school demographic characteristics, enrollment numbers, school types, school admission policies, school climates, characteristics of teachers, demographic characteristics of students and their families, and distributional characteristics of children's cognitive skills and behaviors. Some school-level variables come from administrator questionnaires and other variables are aggregated student variables by school ID numbers. All aggregate measures of school characteristics are adjusted for the over-sampling of Asian students.

Using the language of experimental design, ability grouping in first grade is the "treatment" in my study. This treatment is measured in two different ways: classroom-level and school-level ability grouping. The first treatment views ability grouping as a school-level practice, indicating whether or not the school practices ability grouping. The second treatment views ability grouping as a classroom-level policy, indicating whether the classroom teacher uses ability grouping. In order to estimate the effect of ability grouping, I will design the study in two different ways, one using school ability grouping as a treatment measure and the other using classroom ability grouping as a treatment measure.

Propensity score stratification methods. It is not clear from the data whether ability grouping is a classroom-level policy or school-level practice. If ability grouping is a classroom-level policy, one may design a study by viewing classrooms as a unit of treatment assignment. Similarly, if ability grouping is a school-level practice, it is appropriate to design a study by viewing schools as a unit of treatment assignment. It is also possible that ability grouping is a school policy for some schools, while for others teachers decide whether they use ability grouping.

Given these limitations, this study mimics two different research designs. The first design attempts to simulate a cluster randomized experiment where schools are randomly assigned to ability-grouped settings. The analysis excludes schools that have both ability-grouped and ungrouped classrooms (i.e., "mixed" schools). The resulting sample consists of 8,785 students in 593 schools. The second design attempts to simulate an experiment where classrooms are randomly assigned to ability-grouping settings. The second analysis only uses schools that have both ability-grouped and ungrouped classrooms (i.e., "mixed" schools)[19].

Research design one: schools as a unit of treatment assignment. In the first research design, propensity scores are defined as estimated probabilities of schools adopting ability grouping practices, and they are specified as a function of the observed-school level covariates,

$$P_k = \text{Prob}(T_k = 1 | W_k), \tag{3}$$

where P_k is a propensity score of adopting ability grouping for school *k*, T_k is the school-level treatment, indicating whether school *k* practices ability grouping, and W_k is school-level predictors of school ability grouping.

The selection of school-level covariates to estimate a school propensity of adopting ability grouping is guided by theories and preliminary analyses. Previous research suggests that the purpose of ability grouping is to deal with problems of diversity in the student body so that teachers can organize instruction more effectively (Bar and Dreeben 1983). Also, middle class parents tend to support ability grouping policies because their children are more likely to be

[19] Originally, it was proposed to use the whole sample and use both classroom- and school-level covariates to predict the probability of classrooms adopting ability grouping. However, preliminary analyses find that such an analysis is not feasible (see below for discussions).

placed in higher ability groups and may benefit more from such practices (Oakes et al., 2000). Various school practices may serve as a function of controlling the student body (e.g., admission and retention policies), and these practices may be important to consider. In addition, I examine whether variables that are suggested by previous research to influence student achievement are also correlated with the ability grouping status of the school.

In the preliminary analysis, 107 variables and 13 sets of dummy variables are found to have significant bivariate associations with the school ability grouping status. These variables include behavioral and cognitive characteristics of students, student and school demographic characteristics, school sizes, school types, teacher's race, teacher's salary, school admission policies, and school climates (see Appendix for variable descriptions and descriptive statistics). All these measures come from kindergarten questionnaires. The propensity scores are estimated using a logit model and step-wise logistic regressions are performed to determine variables to be included in the propensity model (See Hong and Raudembush 2005).

After estimating a school propensity of adopting ability grouping, ability-grouped and ungrouped schools were matched using two different stratification methods. I use two stratification methods because the cut points used in each method are somewhat arbitrary. Depending on the distribution of the estimated propensity scores (e.g., the skewedness of the estimated propensity scores), the resulting stratifications may considerably differ in sample sizes and the range of propensity scores in the strata. Using two different stratification strategies allows me to examine the extent to which the estimates of ability grouping effects are sensitive to how strata are defined.

The first strategy is to stratify the sample into quintiles based on the estimated propensity scores. This creates five strata of equal sample sizes. Previous literatures have suggested that using quintiles removes 90 % of the bias associated with the treatment assignment (Cochran 1968; Imbens 2004; Rosenbaum and Rubin 1984)

The second stratification method employs a strategy suggested by Becker and Ichino (2002) and Dehejia and Wahba (2002). This strategy begins with subdividing the sample into strata with an equal propensity score interval (e.g., 0-0.2, 0.2-0.4…...0.8-1.0). Then, the mean propensity scores are compared between the treatment and comparison groups in each stratum. If they are not balanced in a given stratum, the sample is further divided in half in that stratum. Both stratification strategies are used to estimate the average treatment effects and results are compared with one another.

The casual effect of ability grouping (δ) is defined as the average difference in potential first-grade reading scores between students in ability-grouped schools and those in ungrouped schools, given the propensity of schools practicing ability grouping. This is expressed as:

$$\delta = E(Y_{ik} \mid P_k, T_k = 1) - E(Y_{ik} \mid P_k, T_k = 0),$$

where

$$P_k = \mathrm{Prob}(T_k = 1 \mid W_k),$$

and

$$\ln(P_k/(1-P_k)) = W_k\beta, \tag{4}$$

where δ is the treatment (ability grouping) effect, T_k is the school-level treatment, indicating whether school *k* practice ability grouping, Y_{ik} is the first-grade reading score for student *i* in school *k*, P_k is a propensity of schools *k* having ability grouping, and W_k is school-level predictors of school ability grouping practices.

The population average effect of ability grouping on first-grade reading achievement is obtained by taking the weighted average of the stratum-specific effect, weighted by the proportion of the population falling into each stratum. Because students are nested within schools, the two-level hierarchical linear model is used to answer the research questions addressed in this dissertation.

Additionally, covariate adjustments are combined to propensity score stratification methods. Krueger and Zhu (2004) and Rubin (1974) suggest that including key variables as additional covariates are appropriate to gain precisions of the estimates. Also, child-level covariates are included to reduce bias that is associated with student characteristics since matching is based on school characteristics.

For these purposes, the following student-level and school-level covariates are included. Student-level covariates include child gender, age at the time of first-grade assessment, its square term, a variable indicating whether a child changed the school prior to first grade, a variable indicating whether a child repeated kindergarten, spring kindergarten reading scores, its square term, child race, SES, and assessment dates. The dates for both the spring kindergarten and first grade are also controlled since assessments were not conducted on the same date.

Including student kindergarten reading scores is particularly important partly because it would reduce bias due to unobserved variables that may affect the probability of a school adopting ability grouping. Kindergarten reading achievement is the most important predictor of first grade reading achievement, and including this variable as a linear adjustment would reduce bias due to omitted variables if the effects of these unobserved factors on first-grade achievement are mediated by their effects on kindergarten reading achievement. School-level covariates include the logit of the estimated propensity scores, the school mean of kindergarten reading scores, and school type (i.e., public schools).

To examine the first research question—whether reading achievement differs between ability-grouped students and ungrouped students—I will use the following HLM model to estimate the average ability grouping effect (δ) on reading achievement for student *i* in school *k*,

$$\mathrm{Y}_{\mathrm{ik}} = \delta T_k + \sum_p (B_p X_{pik}) + \gamma_1 (Public)_k + \gamma_2 (Mean\_read)_k + \gamma_3 (Log\_prop)_k + \sum_m \gamma_{3+m} D_{mk} + e_{ik} + v_k , \tag{5}$$

where δ is the treatment (school-level ability grouping) effect, Y_{ik} is the first-grade reading score for student *i* in school k, T_k is the treatment condition of school *k*, X_{pik} is student-level covariates where p=1,2…p, D_{mk} is a set of dummy variables indicating propensity stratum *m* for school *k*, and $e_{ik,}$ and v_k are, respectively, student- and school-level error terms.

The second analysis examines whether ability grouping has differential effects by student initial ability levels. Measures of the initial ability levels are constructed by dividing the students into three ability levels (i.e., low, meddle, and high achievers) and a set of dummy variables are created accordingly. The causal effects of ability grouping on low, medium, and high achievers are estimated under the model,

$$\begin{aligned}
\mathrm{Y}_{\mathrm{ik}} &= \beta_1 (Low)_{ik} + \beta_2 (Middle)_{ik} + \beta_3 (High)_{ik} + \sum_p \beta_{3+p} X_{pik} + e_{ik} , \\
\beta_1 &= \delta_1 T_k + \gamma_{11} (Public)_k + \gamma_{12} (Mean_read)_k + \gamma_{13} (Log_prop)_k + \sum_m (\gamma_{1(3+m)} D_{mk}) + v_k , \\
\beta_2 &= \delta_2 T_k + \gamma_{21} (Public)_k + \gamma_{22} (Mean_read)_k + \gamma_{23} (Log_prop)_k + \sum_m (\gamma_{2(3+m)} D_{mk}) + v_k , \\
\beta_3 &= \delta_3 T_k + \gamma_{31} (Public)_k + \gamma_{32} (Mean_read)_k + \gamma_{33} (Log_prop)_k + \sum_m (\gamma_{3(3+m)} D_{mk}) + v_k ,
\end{aligned} \tag{6}$$

where δ is the treatment (school-level ability grouping) effect, Y_{ik} is the first-grade reading score for student *i* in school *k*, T_k is the treatment condition of school *k*, X_{pik} is student-level covariates for student *i* in school *k* where p=1,2….p, D_{mk} is a set of dummy variables indicating propensity stratum *m* for school *k*, and e_{ik} and v_k are, respectively, student- and school-level error terms.

This is slopes-as-outcomes model where the relationships between the student initial achievement levels and first-grade achievement are allowed to vary by schools. The effects of ability grouping are estimated for students with low, middle, and high initial ability levels (δ_1, δ_2, and δ_3, respectively). To formally test whether ability grouping has differential effects by student initial ability levels, I will test the hypothesis $\delta_1 = \delta_{2,} = \delta_3 = 0$.

To examine the extent to which ability-grouping affects achievement inequality, I will compare differences in the magnitude of ability grouping effects for low, middle, and high achievers. For example, to test whether ability grouping widens achievement inequality between low and high initial ability students, I will test the hypothesis $\delta_1 = \delta_3$. If high ability students in ability-grouped schools have higher reading scores than their counterparts in ungrouped schools while 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. Similar hypotheses tests will be conducted to test the differences in ability grouping effects between low and middle initial ability students as well as between middle and high initial ability students.

To examine the third research question—whether ability grouping effects vary by schools—I will estimate stratum-specific causal effects by interacting the treatment variable and propensity strata dummy variables. In other words, the effect of ability grouping is estimated for each stratum. This analysis illuminates the extent to which ability grouping effects vary by school characteristics that are defined by the stratum D_m. The statistical model is written as:

$$\mathrm{Y_{ik}} = \sum_m [\delta_m (D_{mk} * T_k)] + \sum_m (\gamma_m D_{mk}) + \sum_p (B_p X_{qik}) + \gamma_{m+1} (Public)_k + \gamma_{m+2} (Mean\_read)_k$$
$$+ \gamma_{m+3} (Log\_prop)_k + e_{ik} + v_k, \tag{7}$$

where δ_m is the treatment (school-level ability grouping) effect for stratum *m*, Y_{ik} is the first-grade reading score for student *i* in school *k*, T_k is the treatment condition of school *k*, D_{mk} is a set of dummy variables indicating propensity stratum *m* for school *k*, X_{pik} is student-level covariates where p=1,2,…p, and e_{ik} and v_k 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. This is done by estimating the interaction effects between the ability grouping variable and stratum dummy variables at each initial ability level of students. In other words, this compares achievement between ability-grouped students and ungrouped students at each initial ability level within each stratum. This is written as follows:

$$\mathrm{Y_{ik}} = \beta_1 (Low)_{ik} + \beta_2 (Middle)_{ik} + \beta_3 (High)_{ik} + \sum_p \beta_{3+p} X_{pik} + e_{ik},$$
$$\beta_1 = \sum_m [\delta_{1m} (D_{mk} * T_k)] + \sum_m (\gamma_{1m} D_{mk}) + \gamma_{1(m+1)} (Public)_k + \gamma_{1(m+2)} (Mean\_read)_k + \gamma_{1(m+3)} (Log\_prop)_k + v_k,$$
$$\beta_2 = \sum_m [\delta_{2m} (D_{mk} * T_k)] + \sum_m (\gamma_{2m} D_{mk}) + \gamma_{2(m+1)} (Public)_k + \gamma_{2(m+2)} (Mean\_read)_k + \gamma_{2(m+3)} (Log\_prop)_k + v_k,$$
$$\beta_3 = \sum_m [\delta_{3m} (D_{mk} * T_k)] + \sum_m (\gamma_{3m} D_{mk}) + \gamma_{3(m+1)} (Public)_k + \gamma_{3(m+2)} (Mean\_read)_k + \gamma_{3(m+3)} (Log\_prop)_k + v_k, \tag{8}$$

where δ is the treatment (school-level ability grouping) effect, Y_{ik} is the first-grade reading score for student *i* in school *k*, T_k is the treatment condition of school *k*, D_{mk} is a set of dummy variables indicating propensity stratum *m* for school *k*, X_{pik} is student-level covariates where p=1,2…p, and e_{ik} and v_k are, respectively, student- and school-level error terms.

For students in each stratum, a formal hypothesis test is conducted 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 also illuminates whether the effects of ability grouping on achievement inequalities depend on 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, $\delta_1 = \delta_{3,}$ within each stratum. Hypothesis tests are similarly conducted to examine achievement gaps between low and middle initial ability students and between middle and high initial ability students within each stratum.

Research design two: classrooms as a unit of treatment assignment. The second research design attempts to simulate an experiment where classrooms are randomly assigned to ability group settings. It only uses schools that have both ability-grouped and ungrouped classrooms. Analytic strategies in this design are similar to those in the first research design. I first estimate probabilities of classrooms adopting ability grouping by using classroom-level covariates. Then, the sample is subdivided into strata, and the average causal effects are estimated.

The propensity scores are defined as estimated probabilities of classrooms adopting ability grouping, and they are specified as a function of the observed classroom-level covariates,

$$P_j = \text{Prob}(T_j = 1 \mid C_j), \qquad (9)$$

where P_j is a propensity score of adopting ability grouping for classroom *j*, T_j is the classroom-level treatment, indicating whether classrooms *j* practices ability grouping, and C_j is classroom-level predictors of classroom ability grouping. In estimating the probability of classrooms using ability grouping, all classroom covariates are regarded as fixed even though classrooms are nested within schools. This is because preliminary analyses find little evidence for intra-class correlations.

The process of variable selections to estimate a classroom propensity of adopting ability grouping is similar to the school-level analysis discussed above. Classroom variables examined here include class size, demographic compositions of students, student ability distributions, student behavioral characteristics, and teacher characteristics. In the preliminary analysis, 16 variables and 2 sets of dummy variables are found to have bivariate associations with the classroom ability grouping status (see Appendix C for variable descriptions and descriptive statistics). The propensity scores are estimated using a logit model and step-wise logistic regressions are performed to determine variables to be included in the propensity model.

After estimating a classroom propensity of adopting ability grouping, ability-grouped and ungrouped schools were subdivided into strata, using the two stratification methods discussed above. The first method stratifies the sample into quintiles based on the estimated propensity scores and the second method 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 between the treatment and comparison groups are not balanced in a given stratum, the sample was further divided in half in that stratum. Both stratification strategies are used to estimate the average treatment effects and results are compared with one another.

The casual effect of classroom ability grouping on first-grade achievement is defined as the average difference in potential first-grade reading scores between students in ability-grouped classrooms and those in ungrouped classrooms with the same propensity for practicing ability grouping. This is expressed as:

$$\delta = E(Y_{ij} \mid P_j, T_j = 1) - E(Y_{ij} \mid P_j, T_j = 0),$$

where

$$P_j = \text{Prob}(T_j = 1 \mid C_j),$$

and

$$\ln(P_j/(1-P_j)) = C_j\beta, \tag{10}$$

where δ is the treatment (ability grouping) effect, T_j is the classroom-level treatment, indicating whether classroom *j* practices ability grouping, Y_{ij} is the reading score in first grade for student *i* in classroom *j*, P_j is a propensity of classroom *j* having ability grouping, and C_j is classroom-level predictors of classroom ability grouping.

In all analyses, propensity score stratification is combined with covariate adjustments, using both student- and classroom-level covariates. Student-level covariates include child gender, age, a variable indicating whether a child repeated kindergarten, a variable indicating whether a child changed schools prior to first grade, spring kindergarten reading scores, child race, SES, and assessment dates. Classroom-level covariates include the logit of the estimated propensity scores and classroom means of kindergarten reading scores.

The average treatment effect of ability grouping on reading achievement is obtained by taking the weighted average of the stratum-specific causal effects, weighted by the proportion of the population falling into each stratum. This is estimated by using the two-level HLM model,

$$Y_{ij} = \delta T_j + \sum_p (\beta_p X_{pij}) + \gamma_1 (Mean\_read)_j + \gamma_2 (Log\_prop)_j + \sum_m \gamma_{2+m} D_{mj} + e_{ij} + u_j, \tag{11}$$

where δ is the treatment (ability grouping) effect, Y_{ij} is the first-grade reading score for student *i* in classroom *j*, T_j is the treatment condition of classroom *j* in school *k*, D_{mj} is a set of dummy variables indicating propensity stratum *m* for classroom *j*, and e_{ij} and u_j are, respectively, student-and classroom-level error terms. Because preliminary analyses suggest that school-level error terms do not vary by schools they are set as zero.

To examine the second research question—whether ability grouping has differential effects by student initial ability levels—the causal effect on first-grade reading achievement is estimated at each students initial ability level (i.e., low, middle, and high initial ability) by using the following two-level HLM model. Preliminary analyses suggest that the classroom slopes do not vary across schools, thus all level-two slopes are set as fixed. The effect of ability grouping for low, middle, and high initial ability students are, respectively, δ_1, δ_2, and δ_3. The model is written as:

$$\begin{aligned}
Y_{ij} &= \beta_1 (Low)_{ij} + \beta_2 (Middle)_{ij} + \beta_3 (High)_{ij} + \sum_p \beta_{3+p} X_{pij} + e_{ij}, \\
\beta_1 &= \delta_1 T_j + \gamma_{11} (Mean_read)_j + \gamma_{12} (Log_prop)_j + \sum_m (\gamma_{1(2+m)} D_{mj}) + u_j, \\
\beta_2 &= \delta_2 T_j + \gamma_{21} (Mean_read)_j + \gamma_{22} (Log_prop)_j + \sum_m (\gamma_{2(2+m)} D_{mj}) + u_j, \\
\beta_3 &= \delta_3 T_j + \gamma_{31} (Mean_read)_j + \gamma_{32} (Log_prop)_j + \sum_m (\gamma_{3(2+m)} D_{mj}) + u_j,
\end{aligned} \tag{12}$$

where δ is the treatment (ability grouping) effect, Y_{ij} is the first-grade reading score for student *i* in classroom *j*, T_j is the treatment condition of classroom *j*, X_{pij} is classroom-level covariates where p=1,2,…p, D_{mj} is a set of dummy variables indicating propensity stratum *m* for classroom *j*, and e_{ij} and u_j are, respectively, student- and classroom-level error terms. To formally test whether ability grouping has differential effects by student initial ability levels, I will test the hypothesis $\delta_1 = \delta_2 = \delta_3 = 0$.

To examine the extent to which ability grouping affects achievement inequality, I will compare differences in the magnitude of ability grouping effects among low, middle, and high achievers. To test whether ability grouping widens achievement inequality between low and high initial ability students, I will test the hypothesis $\delta_1 = \delta_3$. If the ability grouping effect is greater for high initial ability students than for low initial ability students, this indicates that ability grouping widens inequality between these two groups because high ability students learn more in the ability grouped setting. Similar hypothesis tests will be conducted to test the differences in ability grouping effects between low and middle initial ability students as well as between middle and high initial ability students.

The next analysis examines the third research question—whether the ability grouping effects vary by classrooms. This is done by estimating stratum-specific causal effects by introducing interaction terms between the classroom ability grouping variable and propensity strata dummy variables. This analysis illuminates the extent to which ability grouping effects vary by classroom characteristics that are defined by the stratum D_m. The following two-level HLM model is estimated:

$$\mathrm{Y}_{\mathrm{ij}} = \sum_m \left[\delta_m (D_{mj} * T_j)\right] + \sum_m (\gamma_m D_{mj}) + \sum_p (\beta_p X_{pij}) + \gamma_{m+1}(Mean\_read)_j + \gamma_{m+2}(Log\_prop)_j + e_{ij} + u_j \,, \tag{13}$$

where δ_m is the treatment (ability grouping) effect for stratum *m*, Y_{ij} is the first-grade reading score for student *i* in classroom *j*, T_j is the treatment condition of classroom *j*, D_{mj} is a set of dummy variables indicating propensity stratum for classroom *j*, X_{pij} is student-level covariates where p=1,2,…p, and e_{ij} and u_j are, respectively, student- and classroom-level error terms.

The final analysis examines whether the effects of ability grouping on reading achievement for low, middle and high achievers vary by classroom characteristics. This is done by estimating the interaction effects between the ability grouping variable and stratum dummy variables at each initial ability level of students. This analysis also illuminates whether the effects of ability grouping on achievement inequality vary by classroom characteristics, which is defined by the stratum D_m. The following two-level HLM model is estimated:

$$\begin{aligned}
\mathrm{Y_{ij}} &= \beta_1(Low)_{ij} + \beta_2(Middle)_{ij} + \beta_3(High)_{ij} + \sum_p \beta_{3+p} X_{pij} + e_{ij}, \\
\beta_1 &= \sum_m \left[\delta_{1m}(D_{mj} * T_j)\right] + \sum_m (\gamma_{1m} D_{mj}) + \gamma_{1(m+1)}(Mean_read)_j + \gamma_{1(m+2)}(Log_prop)_j + u_j, \\
\beta_2 &= \sum_m \left[\delta_{2m}(D_{mj} * T_j)\right] + \sum_m (\gamma_{2m} D_{mj}) + \gamma_{2(m+1)}(Mean_read)_j + \gamma_{2(m+2)}(Log_prop)_j + u_j, \\
\beta_3 &= \sum_m \left[\delta_{3m}(D_{mj} * T_j)\right] + \sum_m (\gamma_{3m} D_{mj}) + \gamma_{3(m+1)}(Mean_read)_j + \gamma_{3(m+2)}(Log_prop)_j + u_j,
\end{aligned} \tag{14}$$

where δ is the treatment (ability grouping) effect, Y_{ij} is the first-grade reading score for student *i* in classroom *j*, T_j is the treatment condition of classroom *j*, D_{mj} is a set of dummy variables indicating propensity stratum *m* for classroom *j*, X_{pij} is student-level covariates where p=1,2…..p, and $e_{ij,}$ and u_j are, respectively, student- and classroom-level error terms. For students in each stratum, a formal hypothesis test is conducted to see whether ability grouping has differential effects by student initial ability levels ($\delta_{1m} = \delta_{2m,} = \delta_{3m} = 0$ for m[th] stratum).

The above analysis also illuminates whether the effects of ability grouping on achievement inequality vary by classroom characteristics that are defined by D. For example, to see whether ability grouping widens achievement inequality between low and high initial ability students, I conduct a hypothesis test, $\delta_1 = \delta_{3,}$ within each stratum. Achievement inequality is similarly examined between low and middle initial ability students and between middle and high initial ability students within each stratum.

Causal Assumptions

It is important to note that propensity score methods assume that all covariates are not affected by the treatment assignment. A limitation of this study is that we do not know when schools or classrooms, in fact, adopted ability grouping practices. 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) also use covariates measured in kindergarten to predict propensities of schools adopting retention policies.

It is possible that schools or teachers had been using ability grouping before the ECKL-K data collection began. Prior research suggests that ability grouping is primarily a school or classroom response to their student characteristics (Bar and Dreeben 1983). To predict propensity scores of schools or classrooms practicing ability grouping in first grade, this study use variables that are measured in kindergarten. In doing so, it is assumed that ability grouping is a school or classroom response to the type of students who typically enter the schools or classrooms.

Limitations of the Study

There are several limitations in this dissertation. While this study examines how student learning differs between students in ability-grouped classrooms and those in ungrouped classrooms, schools practice different kinds of instructional grouping. They include gifted/talented programs, special education programs, pull-out programs (such as reading recovery programs or programs for students with limited English proficiency), and heterogeneous small grouping. Also, teachers employ different pedagogical practices, such as cooperative learning and individualized instruction. These aspects of organizational practices and instructional activities are beyond the scope of this dissertation and will not be addressed.

Second since classrooms are nested within schools, the classroom analysis of ability grouping, which use “mixed schools”, may violate SUTVA if using ability grouping in a given classroom, or a teacher using ability grouping in a given classroom, affects student achievement in other classrooms in the same schools. In multi-level settings where the experimental units are clustered to receive individual-level or cluster-level treatments, Hong and Raudenbush (2003) show that SUTVA can be relaxed by identifying the level at which the treatment is assigned and explicitly representing the treatment effects and random effects at each level in the model. While assuming that the effects of unit interferences are random events, Hong and Raudenbush (2003) show that the cluster specific random effects incorporate the effects of interference between units, the effects of agents who are giving the treatment (e.g., teachers), and all other site-level contextual effects that interact with the treatment effects (Hong and Raudenbush, 2003). Thus, the treatment effects are consistently estimated.

In comparison to classroom-level analyses of ability groupoing, the unit interference is unlikely in the analysis of school-level ability grouping. It is reasonable to assume that the use of ability grouping in a given school does not affect the achievement of students in other schools unless prior knowledge suggests otherwise.

Third, because ability grouping takes place within classrooms, some may argue that it is desirable to use the whole sample and estimate the causal effect of classroom ability grouping on student achievement. In such an analysis, one would design a study by viewing classrooms as the unit of treatment assignment. Since classrooms differ in classroom and school characteristics, both classroom- and school-level covariates need to be used to predict the propensity of a classroom using ability grouping.

However, preliminary analyses suggest that when both school- and classroom-level covariates are used as predictors of classroom ability grouping, the propensity score analysis is not feasible because of a lack of comparison groups. This means that school-level covariates are highly predictive of whether classrooms in a given school use ability grouping, and most of the ability-grouped classrooms in ability-grouped schools and ungrouped classrooms in ungrouped schools do not have comparison groups that have similar school characteristics.

This indicates that when all classrooms within the same school use ability grouping, or when no classrooms within the same school use ability grouping, ability grouping may be indeed a school-wide decision. Even when schools do not have a formal ability grouping policy, it is possible that first-grade teachers in the same school have consensus as to whether or not they use ability grouping. As shown in Chapter 5, ungrouped schools differ from both ability-grouped schools and mixed schools in many school characteristics. This suggests the importance of school characteristics in understanding the effect of ability grouping on student achievement. School-level analyses will help us understand how the effects of ability grouping may depend on school characteristics.

Other limitations stem from available datasets. The Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (National Center for Educational Statistics 2000) used in this dissertation has rather limited measures on ability grouping structural characteristics. While the number of groups is available, information, such as group size and group homogeneity, is not known. Also the data do not provide information as to whether teachers change group structures during school years or whether students are moved to different group levels.

In addition, the ECLS-K data are not best suited for classroom-level analyses because sampled students are not the representative of students in classrooms, but that of students in schools. Also, in many first-grade classrooms, sample sizes are small. The average sample size per classroom is 3.7 students and the standard deviation is 2.27. Approximately 16 percent of the total classrooms have only one student. These factors limit the analysis in the second research design particularly because many predictors of classroom ability grouping are aggregate characteristics of students (e.g., student ability distributions and demographic characteristics). Because of these limitations, the results of the second research design must be interpreted with a caution.

The ECLS-K dataset, however, is the first longitudinal data that contain a nationally representative sample of young children. It provides rich information on characteristics of students, their families, classrooms, schools, and instructional practices. Regardless of these limitations, this dissertation will make a significant contribution to the existing ability grouping research.

CHAPTER 5

RESULTS

Research Design One: Ability Grouping as a School-Level Practice

As discussed in Chapter 4, the analyses in the first research design use schools that practice ability grouping and schools without ability grouping. The sample sizes of ability-grouped schools and ungrouped schools are, respectively, 451 and 142. The results are presented in the following order: The first result presents predictors of schools adopting ability grouping. The second result shows the estimated propensity scores of schools adopting ability grouping. Third, the results of two propensity score stratifications are shown. The next result shows systematic differences in school characteristics between schools that are likely to use ability grouping and schools that are not likely to use ability grouping. This is done by comparing school characteristics by propensity strata. The fifth result presents descriptive statistics on the first-grade reading scores between ability-grouped and ungrouped schools by strata. This provides rough estimates of the average effects of ability grouping. Finally, the results of the model based estimates of ability grouping effects are presented. The analyses in the final results attempt to answer the four research questions addressed in this dissertation.

Predictors of a school adopting ability grouping. The first analysis examines bivariate associations between school characteristics and school ability grouping status (i.e., ability grouping is used school wide). The results show that 107 variables and 13 sets of dummy variables have significant bivariate relationships with school ability grouping status at the probability level of .05 (Appendix A). Overall, schools with ability grouping have more disadvantageous characteristics than schools without ability grouping. For example, ability-grouped schools have lower mean scores in various cognitive assessments than ungrouped schools. Also, ability-grouped schools are more heterogeneous in student cognitive skills and behavioral characteristics than ungrouped schools. In general, the results show that heterogeneities in student cognitive and behavioral characteristics are more strongly associated with school ability grouping status than the means of cognitive and behavioral characteristics. This is particularly the case when the measures of student cognitive skills and behavior come from teachers' assessments of their students. The above results provide some support for a claim that ability grouping is a school (or classroom) response to diversity in the student body, particularly when diversity is measured using teacher ratings on student behavior and cognitive skills.

In addition, the results suggest that students in ability-grouped schools are more socially and economically disadvantaged than students in ungrouped schools. Ability-grouped schools have lower mean SES and they have a greater percentage of minority students (both Blacks and Hispanics), students who come from single-parent families, students on welfare, students who participated in Head Start, Limited English Proficiency (LEP) students, students who speak non-English language at home, and students who did not pass the Oral Language Development Scales than do schools that do not practice ability grouping. In addition, administrators in ability-grouped schools are more likely to report the problems of teacher and student absenteeism and school safety problems. However, ability-grouped schools have more instructional resources; they have more computers than ungrouped schools.

Teacher characteristics are also different between schools with ability grouping and without ability grouping. Ability-grouped schools have a greater percent of Black and Hispanic teachers. Also, the teachers' annual base salaries are higher in ability-grouped schools than they are in ungrouped schools.

Compared to ability-grouped schools, ungrouped schools are likely to be private schools and they have small enrollment sizes. Schools with no ability grouping tend to use various kindergarten admission policies, such as requiring admission tests, SAT scores, academic records, recommendations, child interviews, and advising to delay school entry based on standardized test scores. This suggests that, compared to ability-grouped schools, ungrouped schools are likely to regulate their student body through various admission processes. Also, teachers in ungrouped schools are more likely to report positive school climates than teachers in ungrouped schools.

These differences between ability-grouped and ungrouped schools are substantively important. First, most ability grouping studies focus on within-classroom stratification processes by viewing ability grouping as an organizational practice that contributes to maintaining the existing achievement inequalities. However, the above results show evidence for between-school stratification between ability-grouped schools and ungrouped schools. Generally, schools that use ability grouping have more disadvantaged characteristics than schools that do not use ability grouping.

Few studies recognize this between-school stratification. For example, in comparing achievement between ability-grouped students and ungrouped students, previous non-experimental studies do not take into account systematic differences in school characteristics that these students attend. This may result in biased estimates of ability grouping effects (for example, see Hoffer 1992; Pallas et al. 1994; Tack and Farkas Forthcoming).

Second, although experimental studies on ability grouping may yield greater internal validity than non-experimental studies, the external validity of experimental studies is often limited. School sample sizes in experimental studies are small, and the estimate of ability-grouping effects may be specific to the type of schools that are sampled for these studies. Consequently, previous experimental studies do not address the extent to which the effect of ability grouping may vary by school characteristics.

Estimation of the propensity of a school to adopt ability grouping. To estimate the propensity of a school adopting ability grouping, the selection of covariates is determined as follows. First, by following a method used by Hong and Raudenbush (2005), a stepwise logistic regression is performed to select variables to be included in the initial propensity model. Second, the propensity scores are estimated using these covariates. Third, schools are subdivided into strata based on the estimated propensity scores by using two stratification methods—quintile stratification and interval-based stratification. Then, the balance is checked for all 107 covariates and 13 sets of dummy variables within each stratum. If the balance is not met for covariates that are not included in the initial propensity model, those variables are added to the initial propensity model. With the inclusions of these covariates, the propensity scores are re-estimated and this process is repeated until a satisfactory balance is achieved for all covariates within each stratum. For some covariates, higher order terms are added to improve the balance on these covariates.

The final propensity model includes 17 variables, three quadratic terms of the 17 variables, and four sets of dummy variables (Appendix A). Of the 17 variables, three variables—mean reading scores, mean general knowledge scores, and the standard deviation of reading scores—are school aggregate characteristics of student ability. They are constructed using the child direct cognitive assessments in spring kindergarten. Other measures of student literacy skills include the percent of students who failed the Oral Language Development Scales in fall kindergarten, the mean Academic Rating Scale scores on literacy in fall kindergarten, and the standard deviation of the Academic Rating Scale scores on literacy in spring kindergarten. School socio-demographic variables are the percent of students whose mothers do not have a high school diploma, the percent of students who have ever been on AFDC, the percent of students who are on the school free lunch program, the percent of students whose mothers' occupational prestige scores are not applicable, the mean number of residential moves, the percent Hispanic students, and a variable indicating whether schools have LEP students. One variable of school admission policies—a variable indicating whether kindergartens require SAT scores—is also included. Four sets of dummy variables are the percent of students who speak non-English language at home, school type (private, non-regular public, and public schools being an omitted category), the regions (West, South, Midwest, and East as an omitted category), and the total school enrollment numbers (see Appendix A for dummy coding).

The handling of missing predictors follows a method used by Rosenbaum (1986) who used mean imputations for the treated and control groups accordingly. Dummy variables are constructed to indicate missing cases for all covariates with missing values and these dummy variables are included in the propensity models when appropriate. Of the 17 covariates in the final propensity model, 10 covariates have missing cases of less than 5 percent, four covariates have missing cases between 5 and 10 percent, and three covariates have missing cases between 10 and 15 percent. The four sets of dummy variables included in the final propensity model have missing cases of less than 5 percent (See Appendix A).

Table 5.1 presents descriptive statistics of the estimated propensity scores and their logit transformations by school ability grouping status. The mean propensity scores for ability-grouped schools and ungrouped schools are, respectively, .83 and .55. For ability-grouped schools, the estimated propensity scores range from .13 to 1.0 and those for ungrouped schools range from .02 to 1.0. The means of the logit of the estimated propensity scores for ability-grouped and ungrouped schools are, respectively, 2.07 and .25. Their ranges are from -1.87 to 6.34 for ability-grouped schools and from -4.0 to 5.0 for ungrouped schools.

Table 5.1. Estimated Propensity Scores and Logit of Estimated Propensity Scores by School Ability Grouping Status

	Ability-Grouped Schools			*Ungrouped Schools*		
	Mean	SD	Range	Mean	SD	Range
Prop scores	.83	.17	.13-1.0	.55	.25	.02-1.0
Logit prop scores	2.07	1.37	-1.87-6.34	.25	1.43	-4.00-4.96

The distributions of the estimated propensity scores and the logit of the estimated propensity scores are shown in Figure 5.1 and Figure 5.2, respectively. Figure 5.1 shows that most ability-grouped and ungrouped schools have comparison groups across the range of their estimated propensity scores. Approximately 96.6 percent of all schools fall within the region of common support, which is defined as the range where the propensity scores of ability-grouped and ungrouped schools overlap. Only 11 ability-grouped schools and 9 ungrouped schools fall outside of the region of common support. Figure 5.1 also shows that the distribution of ability-grouped schools is relatively sparse with the propensity scores less than .2.

A good overlap on propensity scores justifies the use of propensity score stratification methods as a strategy to remove bias in this study.

Figure 5.1. Distribution of the Estimated Propensity Scores by School Ability Grouping Status (Treated = Ability-grouped schools. Untreated = Ungrouped schools)

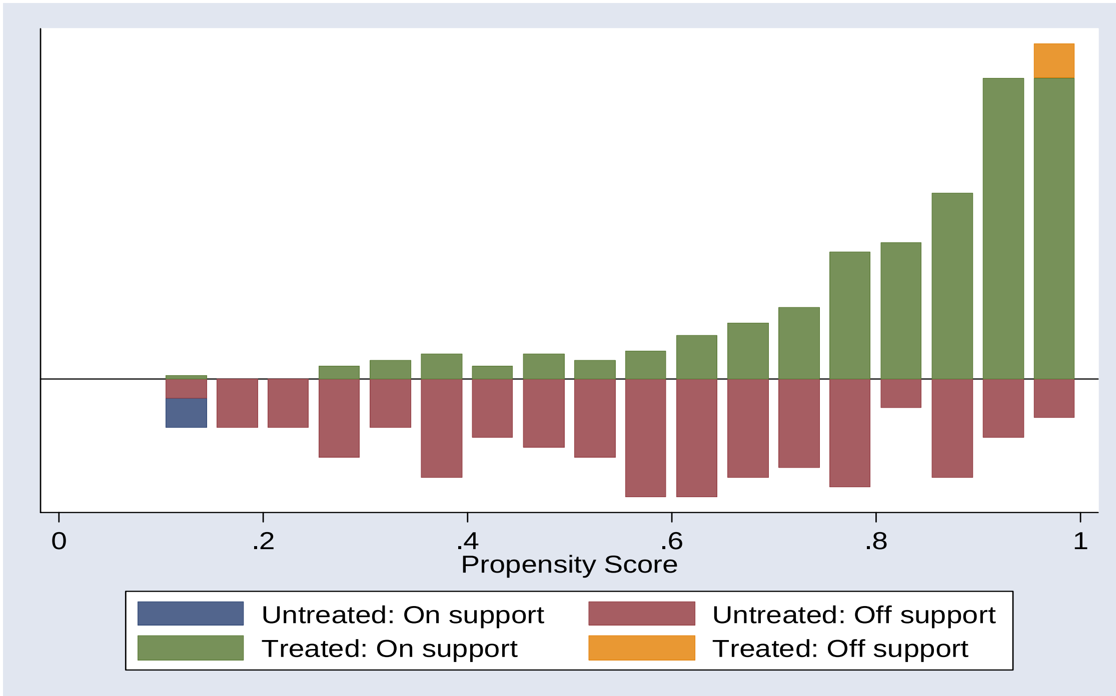
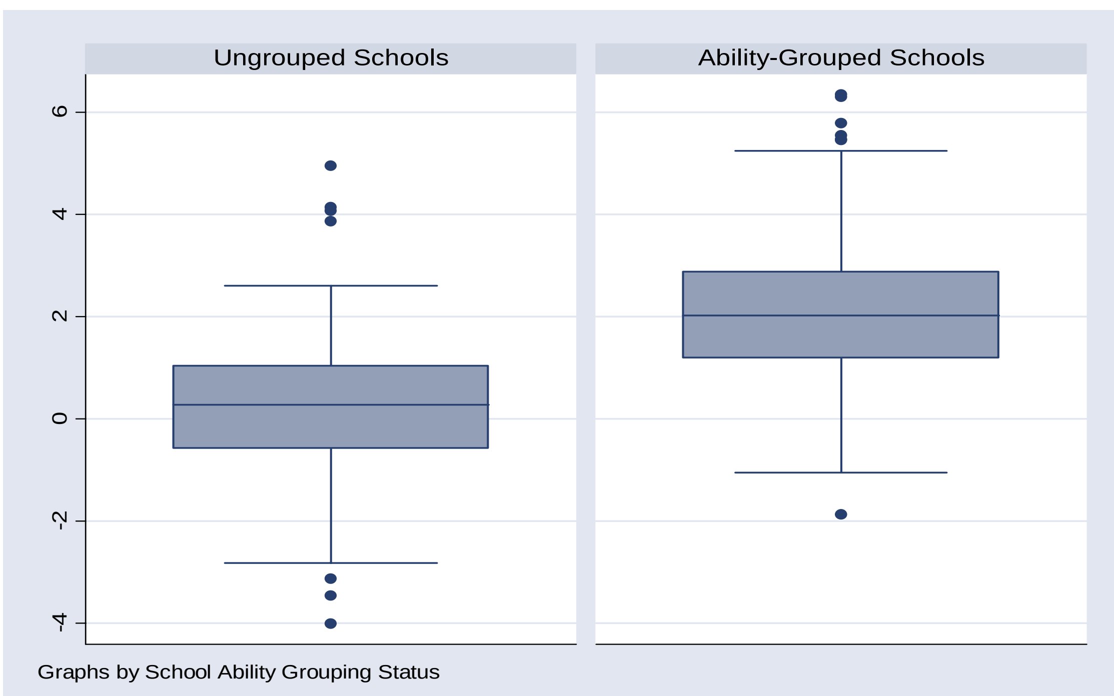


Figure 5.2. Distribution of the Logit of the Estimated Propensity Scores by School Ability Grouping Status



Results of propensity score stratifications. As discussed in Chapter 4, two stratification strategies are used in this study. The first strategy is to stratify the sample into quintiles based on the estimated propensity scores. This creates five strata of equal sample sizes. The second strategy creates stratification using the propensity score intervals. The strategy for the interval-based stratification is described as follows. First, the sample is stratified into five strata of equal propensity score intervals. Second, the mean estimated propensity scores are compared between ability-grouped and ungrouped schools within each stratum. When significant differences are found in the mean estimated propensity scores in a given stratum, that stratum is further divided into half. The resulting stratification consists of six strata.

The results of two propensity score stratifications are shown in Table 5.2.1 and Table 5.2.2. Table 5.2.1 presents a stratification using quintiles and Table 5.2.2 presents a stratification based on the propensity score intervals. In both tables higher strata indicate greater propensities of schools adopting ability grouping.

Table 5.2.1. Propensity Score Stratification by Quintiles

	Total Prop. Scores		*Ability-Grouped Schools*		*Ungrouped Schools*	
Stratum	Mean	Range	N Schools	N Students	N Schools	N Students
5	.97	.95-1.0	114	1,623	4	52
4	.92	.89-.95	109	1,667	10	152
3	.83	.78-.88	104	1,559	14	188
2	.69	.59.-78	80	1,247	39	611
1	.38	.02-.59	44	646	75	1,040
		Total	451	6,742	142	2,043
		Prop score mean	.83	NA	.55	NA
		SD	.17		.25	

Table 5.2.2. Propensity Score Stratification based on the Propensity Score Intervals

	Total Prop. Scores		*Ability-Grouped Schools*		*Ungrouped Schools*	
Stratum	Mean	Range	N Schools	N Students	N Schools	N Students
6	.91	.80-1.00	309	4,599	23	325
5	.75	.70-.80	64	956	20	306
4	.65	.60-.70	32	514	22	324
3	.51	.40-.60	27	406	33	510
2*	.32	.20-.40	18	258	28	374
1*	.11	.02-.20	1	9	16	204
		Total	451	6,742	142	2,043
		Prop score mean	.83	NA	.55	NA
		SD	.17		.25	

*in subsequent analyses, the bottom two strata are combined to create a single stratum

The two stratification results differ in terms of the sample sizes and the range of propensity scores that each stratum contains. The quintile stratification creates five strata with equal sample sizes and each stratum contains 119 or 118 schools. In comparison, when schools are stratified using propensity score intervals, the number of schools varies greatly by strata. For example, the highest stratum (stratum six) have 331 schools, while the bottom stratum (stratum one) contains only 17 schools. Also, in the interval-based stratification, the highest stratum contains 304 of 451 ability-grouped schools while only 23 of 152 ungrouped schools are observed in this stratum. In the bottom stratum (stratum one), only one school is an ability-grouped school and 16 schools are ungrouped schools.

The interval-based stratification, however, divides the sample more evenly in terms of the estimated propensity scores than the quintile stratification. Table 5.2.2 shows that the interval of the propensity scores is .2 in all strata except for stratum four and stratum five. In comparison, the propensity scores of the quintile stratification range from .02 to .6 in the bottom stratum (stratum one), while those in the top stratum (stratum five) range only from .95 to .99. In subsequent analyses, the results of the two stratification methods are compared with one another.

It is noted that in the interval-based stratification method, the bottom stratum has only one ability-grouped school and 16 ungrouped schools. In this stratum, the data are too sparse to estimate the stratum specific effect of ability grouping. Instead of discarding these 17 schools from the analysis, they are combined to the second lowest stratum (stratum two) to increase the sample sizes in the low end of the propensity score distribution. After combining the bottom two strata, balance is checked for the mean estimated propensity scores and covariates.

The preliminary analysis suggests that when the bottom two strata are combined the mean propensity scores become significantly different between ability-grouped and ungrouped schools at the probability level of .05. Regardless of this difference, no schools will be excluded from the sample to improve the balance of the mean propensity scores for the following reasons. First, even when the mean propensity scores are not balanced between ability-grouped and ungrouped schools, the covariate balance is sufficiently met by combining the schools in the bottom two strata. Second, a preliminary analysis suggests that even when some schools are excluded from the sample to improve the balance on the mean propensity scores this does not affect the balancing of covariates. Neither does it affect the results of the subsequent analyses[20]. Third, the estimated propensity scores in the lowest or highest end of the propensity distribution are often unreliable and may not reflect the true propensity scores. For these reasons, 17 schools in stratum one and 46 schools in the stratum two in the interval stratification are merged to create a single stratum. The resulting sample size of the lowest stratum is 62 schools, of which 44 schools are ability-grouped schools and 19 schools are ungrouped schools.

---

[20] In a preliminary analysis, schools that fall outside of the region of common support are excluded to improve the balance of the mean propensity scores. This results in excluding 9 ungrouped schools. This produces a sufficient balance of both the mean propensity scores and covariates. However, because the covariate balance is little changed by the exclusion of these 9 schools (i.e., the covariate balance is met whether or not these 9 schools are excluded from the sample), these schools are kept in the sample for subsequent analyses. Also, the exclusion of these 9 schools does not change the results of all subsequent analyses.

After schools are subclassified using two stratification strategies, statistical tests are conducted to examine within-stratum differences in covariates between ability-grouped and ungrouped schools. The results suggest that the overall balance is met for all covariates at the probability level of .05 in both stratification methods (See Appendix B)[21]. Within each stratum, over 95 percent of covariates are balanced at the probability level of .05 in both stratification methods.

Differences in school characteristics by propensity strata. The purpose of the analysis in this section is to examine the extent to which school characteristics differ systematically between schools that are predicted to use ability grouping and schools that are predicted not to use ability grouping. This is done by describing differences in school characteristics by propensity strata. Table 5.3.1 and Table 5.3.2 present descriptive statistics on selected school characteristics by propensity strata based on quintile stratification and interval stratification, respectively. In both tables, lower strata (e.g., stratum one) consist of schools that are less likely to practice ability grouping, whereas higher strata (e.g., stratum six) consist of schools that are more likely to practice ability grouping.

The school characteristics examined here include a variable indicating whether schools are public, school socio-demographic characteristics (mean SES, percent black students, and percent Hispanic students), the percent of students who failed the Oral Language Development Scales, mean literacy skills (mean IRT reading scores and mean ARS literacy scores), and heterogeneities in literacy skills (the standard deviation of IRT reading scores and the standard deviation of ARS literacy scores).

[21] The overall covariate balance is defined as the average within-stratum differences in the means of covariates between ability-grouped and ungrouped schools

Table 5.3.1. Descriptive Statistics on Selected School Characteristics by Propensity Strata using Quintile Stratification (the numbers in parenthesis are standard deviations)

	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5
Public School	.21	.57	.81	.95	.97
	(.41)	(.50)	(.39)	(.22)	(.16)
Mean SES	.30	.21	.02	-.06	-.26
	(.42)	(.48)	(.53)	(.43)	(.49)
Pct Black	.07	.09	.16	.17	.20
	(.18)	(.20)	(.28)	(.27)	(.28)
Pct Hispanic	.10	.12	.13	.17	.37
	(.16)	(.17)	(.20)	(.24)	(.35)
K2 Mean IRT Read	34.40	33.34	31.60	31.60	32.18
	(5.83)	(5.84)	(5.49)	(4.29)	(5.36)
K2 SD IRT Read	7.98	8.31	8.95	8.45	9.24
	(2.73)	(2.55)	(2.35)	(2.24)	(2.51)
Pct. failed OLDS	.01	.02	.04	.05	.19
	(.03)	(.07)	(.10)	(.11)	(.23)
Mean ARS literacy (K2)	2.84	2.74	2.6	2.48	2.45
	(.48)	(.48)	(.47)	(.43)	(.40)
SD ARS literacy (K2)	.53	.61	.69	.70	.76
	(.22)	(.18)	(.18)	(.18)	(.18)

Table 5.3.2 Descriptive Statistics on Selected School Characteristics by Propensity Strata using Interval Stratification (the numbers in parenthesis are standard deviations)

	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5
Public School	.11	.33	.50	.63	.93
	(.32)	(.48)	(.50)	(.49)	(.25)
Mean SES	.35	.26	.25	.19	- .12
	(.36)	(.45)	(.44)	(.53)	(.49)
Pct Black	.06	.08	.08	.13	.17
	(.17)	(.18)	(.17)	(.25)	(.27)
Pct Hispanic	.11	.09	.13	.10	.23
	(.17)	(.15)	(.19)	(.14)	(.29)
K2 Mean IRT Read	34.95	33.82	33.86	32.80	31.82
	(5.57)	(5.86)	(5.88)	(6.16)	(4.97)
K2 SD IRT Read	7.76	8.42	8.58	8.29	8.90
	(3.25)	(2.34)	(2.36)	(2.59)	(2.36)
Pct. failed OLDS	.01	.01	.01	.03	.10
	(.03)	(.03)	(.03)	(.08)	(.18)
Mean ARS literacy (K1)	2.90	2.83	2.81	2.65	2.49
	(.45)	(.48)	(.47)	(.46)	(.44)
SD ARS literacy (K2)	.53	.56	.61	.62	.71
	(.23)	(.20)	(.17)	(.20)	(.18)

The results generally suggest that schools in higher strata (i.e., schools that are more likely to use ability grouping) have more disadvantageous characteristics than schools with lower propensity scores (schools that are less likely to use ability grouping). Table 5.3.1 and Table 5.3.2 show that schools in higher propensity strata have a higher proportion of public schools, lower mean SES, a higher proportion of minority students, lower mean literacy skills, and more heterogeneous literacy skills than schools in lower propensity strata. These results are consistent with earlier findings on the bivariate associations between school characteristics and school ability grouping status.

Because schools differ in many school characteristics systematically by the levels of propensity scores, when these schools use ability grouping, its effects may also vary by school characteristics. Heterogeneities in ability grouping effects are examined in the subsequent analyses.

Comparison between students in ability-grouped schools and ungrouped schools: descriptive results. In order to see rough estimates of treatment effects within strata, Table 5.4.1 and Table 5.4.2 present descriptive statistics on first-grade reading scores for students in ability-grouped schools and those in ungrouped schools by propensity strata. The last column of these two tables shows a result of a simple mean comparison of student reading scores in first grade.

Table 5.4.1. Descriptive Statistics on Unweighted Mean Reading Scores in First Grade by School Ability Grouping Status and Propensity Strata (Quintile Stratification)

	Ability-Grouped Schools			*Ungrouped Schools*			
Stratum	N student	Mean	SD	N student	Mean	SD	Mean diff
5	1,623	55.27	13.81	52	51.44	15.3	3.83
4	1,667	55.8	12.64	152	57.95	12.04	-2.15
3	1,559	55.74	13.55	188	55.05	14.46	.69
2	1,247	58.57	12.86	611	59.9	11.71	-1.33
1	646	60.03	11.64	1,040	58.12	12.84	1.91
Total	6,742	56.58	13.19	2,043	58.19	12.79	-1.61

Table 5.4.2. Descriptive Statistics on Unweighted Mean Reading Scores in First Grade by School Ability Grouping Status and Propensity Strata (Interval Stratification)

	Ability-Grouped Schools			*Ungrouped Schools*			
Stratum	N student	Mean	SD	N student	Mean	SD	Mean diff
5	4,599	55.58	13.31	325	55.78	13.56	-.19
4	956	57.58	13.53	306	58.69	12.37	-1.11
3	514	59.18	12.18	324	60.35	12.14	-1.17
2	406	58.6	12.05	510	57.15	13.43	1.44
1	267	62.05	10.42	578	57.15	13.42	4.9
Total	6,742	56.58	13.19	2,043	58.19	12.79	-1.61

The results show that, overall, the average first-grade reading scores for students in ability-grouped schools are 56.58 and those for students in ungrouped schools are 58.19. The difference between the two groups is -1.61 points. Within-stratum comparisons suggest that when schools are stratified using propensity score intervals, ability-grouped students have lower achievement than their counterparts in ungrouped schools for schools that are more likely to use ability grouping (i.e., schools in higher propensity strata). However, ability-grouped students seem to have higher first-grade reading scores than ungrouped students in schools that are less likely to practice ability grouping (i.e., schools in lower propensity strata). This pattern is less clear when the strata are defined using propensity quintiles.

The results of the quintile stratification (Table 5.4.1) show that the average first-grade reading scores for students in ability-grouped schools and those in ungrouped schools are, respectively, 60.03 and 58.12 in stratum one, 58.57 and 59.90 in stratum two, 55.74 and 55.05 in stratum three, 55.80 and 57.95 in stratum four, and 55.27 and 51.44 in stratum five. In the interval-based stratification (Table 5.4.2), the average first-grade reading scores for students in ability-grouped schools and those in ungrouped schools are, respectively, 62.05 and 57.15 in stratum one, 58.60 and 57.15 in stratum two, 59.18 and 60.35 in stratum three, 57.58 and 58.69 in stratum four, and 55.58 and 55.57 in stratum five.

These simple mean comparisons of student reading score presented above, however, does not take into account sample size differences between ability-grouped and ungrouped schools nor the clustering of students in schools. These issues are dealt with in the following analyses.

Model based estimates of ability grouping effects on first-grade reading achievement. The following section presents the results of the model based estimates of ability grouping effects on first-grade reading achievement. Hierarchical linear models (Raudenbush and Bryk 2002) are used in the analyses (see Chapter 4 for statistical models). To reiterate the following four research questions are addressed in this dissertation. 1) Does reading achievement differ between students who are grouped by ability and those who are not ability-grouped? 2) Does reading ability grouping have differential effects by students' initial abilities? If so, do differential effects contribute to increasing achievement gaps between high and low achievers? 3) Do the effects of ability grouping vary by schools? 4) Do the effects of ability grouping vary by students' initial abilities *and* schools?

The results of four analyses are presented below, which include the average ability grouping effect for the overall population (Equation 5), the average ability grouping effect by student initial ability levels (Equation 6), the average ability grouping effect by propensity score strata (Equation 7), and the average ability grouping effect by student initial ability levels and propensity score strata (Equation 8).

The model based estimate of the treatment effect is better than a simple comparison of the means of the first-grade reading scores between students in ability-grouped schools and those in ungrouped schools that are shown in Table 5.4.1 and Table 5.4.2. For example, HLM models take into account the clustering of students in schools. HLM models also take into account differences in student sample sizes between ability-grouped and ungrouped schools. Furthermore, propensity score stratification methods can be combined with covariate adjustments at both student and school levels to improve precision and remove additional bias.

The population in the school-level analyses is defined as the U.S. schools that use ability grouping at school levels. However, the generalizability of the results is somewhat limited because the analytic sample does not necessarily represent the population defined above. The ECLS-K was designed to include a nationally representative sample of kindergarteners and the sample of children who had not attended kindergarten was added to the original sample in the first-grade data collection. In order for the results to be generalizable to the population of ability-grouped schools, different sampling designs (e.g., stratified sampling by school ability grouping status) may be needed. Because of this limitation, the results of the analyses are only roughly generalizable to the population defined above.

The average treatment effects of ability grouping for the overall population. To address the first research question, the first analysis examines the average effect of ability grouping on first-grade reading achievement for the population as whole (Table 5.5). The results show that the two stratification methods yield similar outcomes. Both results suggest no significant differences in average first-grade reading achievement between students in ability-grouped schools and those in ungrouped schools. The average causal effect is -.18 (SE=.41) for the quintile stratification and -.21 (SE=.44) for the interval-based stratification. These results provide little evidence that overall ability grouping practices lead to higher mean student achievement.

Table 5.5. Model Based Estimation of Average Effects of School Ability Grouping on First-Grade Reading Achievement

Stratification by Quintiles		*Stratification by Propensity Intervals*	
Coefficient	SE	Coefficient	SE
-.09	.41	-.17	.41

Significant level: ***p<.001, ** p<.01, * p<.05, + p<.10

Differential effects of ability grouping by initial ability levels. The analysis in this section attempts to answer the second research question, "Does reading ability grouping have differential effects by students' initial abilities?" If so, do differential effects contribute to increasing achievement gaps between high and low achievers?"

Some researchers argue that while ability grouping may have no effects on the average student achievement, the effect of ability grouping may depend on student initial abilities (Hallinan 1990; Hoffer 1992). In order to examine this claim, I divide the sample into three initial ability levels—high, middle, and low achievers—based on the spring-kindergarten reading scores. Dummy variables are constructed to indicate the three initial ability levels and reading achievement is compared between students in ability-grouped schools and those in ungrouped schools at each ability level (Table 5.6).

The results from the two stratification methods in Table 5.6 show that ability grouping has no significant effects on first-grade reading achievement at any ability level. The average ability grouping effects for low, middle, and high achievers are, respectively, -.35, .26, and -.51 (p>.10) in the quintile stratification. The comparative figures for the propensity interval stratification are - .43, .26, and -.61 (p>.10). These findings render little support for a claim that all students benefit from ability grouping regardless of their initial ability levels. Neither does it support claims that ability grouping leads to higher achievement for high initial ability students, while hurting achievement for low initial ability students. The results suggest that ability grouping has little effects on overall achievement inequalities.

Table 5.6. Model-based Estimation of the Average Effects of School Ability Grouping on First-Grade Reading Achievement by Initial Ability Levels

	Stratification by Quintiles			*Stratification by Propensity Intervals*		
	Low achievers	Middle achievers	High achievers	Low achievers	Middle achievers	High achievers
Coeff.	-.35	.26	-.51	- .43	.26	- .61
SE	.70	.53	.43	.71	.54	.44

Significant level: ***p<.001, ** p<.01, * p<.05, + p<.10

Heterogeneity in the effects of ability grouping by schools. The next analysis examines whether ability grouping effects may vary by the type of schools students attend. This is done by estimating the stratum-specific effect of ability grouping on first-grade reading achievement.

Table 5.7 displays the results of the two stratification methods. First, the results of the quintile stratification show that ability grouping increases first-grade reading achievement for students in schools that are less likely to practice ability grouping. In stratum one with lower propensity scores of schools using ability grouping, students in ability-grouped schools score 1.77 points higher than their counterparts in ungrouped schools (<.05). The results of the interval stratification also show that students in ability-grouped schools in stratum one have higher achievement than students in ungrouped schools by 3.55 points (p<.01). Ability grouping also leads to higher achievement by 1.33 in stratum two although this is not statistically significant. These findings provide partial support for the findings of previous experimental and matched experimental studies, which suggest that ability grouping leads to higher average achievement (Lou et al. 1996, Slavin 1987).

Table 5.7. Model-based Estimation of the Average Effects of School Ability Grouping on First-Grade Reading Achievement by Propensity Score Strata

	Stratification by Quintiles			*Stratification by Propensity Intervals*		
Stratum	Prop. Score	Coefficient	SE	Prop. Score	Coefficient	SE
5	.95-1.0	1.7	1.2	.80-.99	-.93	.82
4	.89-.95	-2.34*	1.07	.70-.79	-2.54**	.9
3	.78-.88	-.38	.94	.60-.69	-1.07	.97
2	.59.-78	-1.47*	.71	.40-.59	1.33	1.02
1	.02-.59	1.77*	.73	.10-.39	3.55**	.97

Significant level: ***p<.001, ** p<.01, * p<.05, + p<.10

In addition, recall that the propensity scores of schools in the bottom two strata of the interval-based stratification are equivalent to those of schools in the bottom stratum of the quintile stratification. The estimated propensity scores of these schools are less than .6. The schools in stratum one in the interval stratification are least likely to practice ability grouping, and the results show that ability grouping improves student achievement particularly in this type of schools.

Earlier analyses, which examined relationships between school characteristics and school ability grouping status, suggest that schools that are less likely to practice ability grouping have higher student achievement, more homogenous student cognitive and behavioral characteristics, higher mean SES, a lower percent of minority students, and a lower percent of non-English speakers at home. Private schools and schools with smaller enrollment sizes are also less likely to practice ability grouping. It is in these schools where students benefit the most from ability grouping.

In contrast, the results of this analysis suggest that ability grouping leads to lower reading achievement for schools that are more likely to practice ability grouping. The predicted propensity scores of schools in stratum two in the quintile stratification range from .59 to .78. They are roughly equivalent to the propensity scores of schools in stratum three and four in the interval-based stratification. In these schools, the results show that ability grouping leads to lower achievement. In the quintile stratification, reading scores of ability-grouped students in the stratum two schools are 1.47 points lower than those of their counterparts in ungrouped schools ($p<0.05$). The results from the interval-based stratification also show that ability-grouped students in schools in stratum three and four have lower first-grade reading scores than their counterparts in ungrouped schools by 1.07 points and 2.54 points, respectively, although only the coefficient for the stratum four schools is statistically significant at the probability level of .01.

Similarly, the estimated propensity scores of the schools in stratum three, four, and five in the quintile stratification are roughly equivalent to those of the schools in stratum five in the interval-based stratification. Their propensity scores range from .8 to 1.0. The results of the interval-based stratification suggest that ability grouping has a negative, but negligible effect for students attending schools with propensity scores above .8 (i.e., students in schools in stratum five). However, the results of the quintile stratification show that ability-grouped students in the stratum four schools have lower reading scores than their ungrouped counterparts by 2.34 points ($p<.01$). It is noted that the sample size of ungrouped schools with the propensity scores above .8 is only 23 compared to over 300 ability-grouped schools, and results may be somewhat sensitive to how stratification is defined.

Earlier analyses show that schools that are more likely to practice ability grouping have lower mean kindergarten test scores, more heterogeneous student cognitive and behavioral characteristics, lower mean SES, and a higher percentage of minority students than schools that are less likely to adopt ability grouping. Also they are more likely to be public and have larger enrollments. The above findings suggest that ability grouping leads to lower average achievement in these types of schools. Possible explanations as to why the causal relationships vary by school characteristics are discussed in Chapter 6.

Heterogeneity in the effects of ability grouping by student initial ability levels and schools. Earlier analyses find that the overall effects of ability grouping do not vary by student initial ability levels (see Table 5.6). To answer the fourth research question, "Do the effects of ability grouping vary by students' initial abilities *and* schools?", the next analysis examines whether ability grouping effects on the achievement of low, middle, and high initial ability students may vary by characteristics of schools that these students attend. This is done by estimating the stratum specific effects of ability grouping by student initial ability levels (Table 5.8).

Table 5.8. Model-based Estimation of the Average Effects of School Ability Grouping on First-Grade Reading Achievement by Initial Ability Levels and Propensity Strata

	Stratification by Quintiles				*Stratification by Propensity Intervals*			
Stratum	Prop. Scores	Low Coeff.	Middle Coeff.	High Coeff.	Prop. Scores	Low Coeff.	Middle Coeff.	High Coeff.
5	.95-1.0	1.37	-.62	.04	.80-.99	-1.30	-.52	-1.42+
		(1.45)	(1.88)	(1.73)		(0.93)	(1.12)	(.86)
4	.89-.95	-3.08*	-2.50+	-2.73*	.70-.79	-4.80**	-2.93**	-1.25
		(1.32)	(1.33)	(1.17)		(1.54)	(1.07)	(1.03)
3	.78-.88	-1.67	2.14	-.33	.60-.69	-.66	-.76	-1.65+
		(1.51)	(1.69)	(1.17)		(1.95)	(1.19)	(.89)
2	.59.-78	-2.05	-1.90	-.98	.40-.59	1.81	2.68*	-.05
		(1.34)	(.81)	(.72)		(1.77)	(1.15)	(.98)
1	.02-.59	2.91*	3.09***	.54	.10-.39	5.43**	3.99**	2.18*
		(1.32)	(.88)	(.73)		(1.68)	(1.35)	(.97)

Significant level: ***p<.001, ** p<.01, * p<.05, + p<.10
Standard errors are in parenthesis

The results in Table 5.8 show that, in general, for students who attend the types of schools that typically use ability grouping, ability grouping leads to lower achievement at all initial ability levels, and negative effects are particularly strong for low initial ability students. For example, the results of the propensity quintile stratification show that in the stratum four schools with the propensity scores ranging from .89 to .95, differences in first-grade achievement between ability-grouped and ungrouped students are -3.8 points for low initial ability students, -2.5 points for middle initial ability students, and -2.7 points for high initial ability students ($p<.05$, $p<.1$, and $p<.5$, respectively).

Similarly, the results of the interval-based stratification show that for students attending the stratum four schools with the propensity scores ranging from .70 to .79, low initial ability students in ability-grouped schools have particularly lower achievement than their counterparts in ungrouped schools. First-grade reading scores of low, middle, and high initial ability students in the stratum four schools are lower than their counterparts in ungrouped schools by, respectively, -4.80, -2.93, and -1.25 ($p<.01$, $p<.01$, and $p>.1$, respectively). Additional analyses suggest that the negative effects of ability grouping for low initial ability students is significantly greater than those for middle and high initial ability students ($p<.01$). The results of the propensity interval stratification also show that high initial ability students in ability-grouped schools in stratum three and stratum five have somewhat lower achievement than their counterparts in ungrouped schools and these results are statistically significant only at the probability level of .1.

In contrast, the results of the quintile stratification show that for schools that show positive ability grouping effects on reading achievement (i.e., schools in stratum one), the positive effects are found among low and middle initial ability students, but not among students with high initial abilities. In ability-grouped schools in stratum one, which have the propensity scores of less than .6, reading achievement for low and middle initial ability students are approximately 3.0 points higher than their counterparts in ungrouped schools.

The results of the interval-based stratification, however, show that for schools that are least likely to practice ability grouping, ability grouping leads to higher achievement at *all* initial ability levels. In ability-grouped schools with the lowest propensity interval (i.e., propensity scores between .1 and .4), first-grade reading scores for low, middle, and high initial ability students are, respectively, 5.43, 3.99, and 2.18 points higher than reading scores for their counterparts in ungrouped schools ($p< .01$, $p<.01$, and $p<.05$, respectively). These results provide partial support for the findings by Slavin (1987) and Lou et al (1996), which suggest that ability grouping leads to higher achievement at all initial ability levels. Additional analyses suggest that positive effects of ability grouping are significantly greater for low initial ability students than those for middle and high initial ability students. This suggests that ability grouping is particularly beneficial for low achievers in this type of schools.

The schools that have the most advantageous characteristics are predicted to be least likely to practice ability grouping. However, these schools produce higher achievement for all when they practice ability grouping. In these schools, the above findings provide little evidence to suggest that ability grouping increases initial achievement inequalities in comparison to ungrouped schools with similar school characteristics. On the contrary, ability grouping may *reduce* achievement inequalities in these schools because low initial ability students are more likely to benefit from ability grouping than high initial ability students.

In contrast, the propensity score models predict that schools that have more disadvantageous characteristics are more likely to practice ability grouping, and in these schools achievement inequalities may *increase* when they practice ability grouping. This is because ability grouping leads to lower achievement particularly among low initial ability students, while students with high initial ability are less likely to be affected by such practices. This is illuminated by the results of the propensity score interval stratification, which suggest that ability grouping is more likely to have detrimental effects on the achievement of low initial ability students than that of middle and high initial ability students. However, the above conclusions are rather tentative because this pattern is less clear from the results of the propensity quintile stratification.

Research Design Two: Ability Grouping as a Classroom-Level Practice

The second research design examines the effect of ability grouping on first-grade reading achievement by regarding classrooms as being randomly assigned to an ability-grouped setting. This analysis uses the sample of schools that have classrooms with and without ability grouping in the same schools (i.e., “mixed schools”).

Some may argue that the analysis of classroom ability grouping should include the whole sample, and both classroom- and school-level covariates can be used to predict a classroom propensity to adopt ability grouping. However, preliminary analyses have found that when the whole sample is included and both classroom- and school-level covariates are used as predictors of classroom ability grouping, the propensity score analysis is not feasible because of a lack of comparison groups. In other words, school-level covariates are found to be highly predictive of whether classrooms in a given school use ability grouping, and most of the ability-grouped classrooms in ability-grouped schools and ungrouped classrooms in ungrouped schools do not have comparison groups that have similar school characteristics. Because of this difficulty, the analyses of classroom-level ability grouping use only the sample of “mixed schools”.

However, the fact that classrooms do not have viable comparison groups when school-level covariates are used as predictors of classroom ability grouping status may imply that it may be school characteristics, rather than classroom characteristics, that are the key to understanding how ability grouping produces particular student achievement. In other words, even though ability grouping is a classroom-level practice, school factors may be more important than classroom factors in understanding what determinants the practice of classroom ability grouping. Similarly, school factors may be more important than classroom factors in understanding the consequences of ability grouping. To explore this possibility, the classroom-level analyses will begin with describing how characteristics of "mixed schools" differ from those of ability-grouped and ungrouped schools.

Characteristics of "mixed schools". School characteristics examined here include school type and student socio-demographic, cognitive, and behavioral characteristics (Table 5.9). The results in Table 5.9 show that ungrouped schools are the most advantageous schools and "mixed schools" are the least advantageous schools of the three types of schools. For example, "mixed schools" have the lowest mean IRT reading scores, the lowest mean Academic Rating Scale scores on literacy, and the lowest mean SES of all schools.

In addition, 91 percent of "mixed schools" are public schools, while comparative figures for ability-grouped and ungrouped schools are, respectively, 79 percent and 44 percent. The fact that only a few private schools are "mixed schools" indicates that at least for private schools the use (or non-use) of ability grouping may be a school-wide decision.

In Table 5.9 greater differences are observed between "mixed schools" and ungrouped schools than between "mixed schools" and ability-grouped schools in many school characteristics. For example, while most of the covariates in Table 5.9 are significantly different between "mixed" and ungrouped schools, "mixed" and ability-grouped schools are similar in the percent of Hispanic and Black students, the percent of students from non-English speaking homes, and the level of cognitive heterogeneities.

Table 5.9. Descriptive Statistics by School Ability Grouping Status: Selected Characteristics

	School Ability Grouping Status		
Variables	Ability Grouped	Non-Ability Grouped	Mix
Spring-K Mean Read	32.35	33.51[a]	31.43[a]
	(5.40)	(5.84)	(5.10)
Spring-K SD Read	8.67	8.15[a]	8.53
	(2.37)	(2.64)	(2.27)
Mean ARS Literacy	3.22	3.20	3.10[a]
	(.74)	(1.10)	(.63)
SD ARS Literacy	.68	.59[a]	.68
	(.19)	(.22)	(.19)
Mean Approach to Learning	2.96	2.99	2.94
	(.51)	(.71)	(.41)
SD Approach to Learning	.61	.58[a]	.64[a]
	(.14)	(.15)	(.14)
PCT non-English speaking	.14	.07[a]	.14
	(.23)	(.14)	(.22)
Mean SES	.00	.18[a]	-.09[a]
	(.51)	(.48)	(.48)
Pct Hispanic	.19	.12[a]	.19
	(.27)	(.18)	(.26)
Pct Black	.15	.09[a]	.18
	(.26)	(.21)	(.28)
Public School	.79	.44[a]	.91[a]
	(.41)	(.5)	(.29)

[a] Statistically different from ability-grouped schools at $p<.05$

Substantive differences between "mixed schools" and ungrouped schools imply that characteristics of ungrouped classrooms in "mixed schools" are also dissimilar to characteristics of ungrouped classrooms in ungrouped schools. As shown in the following section, when the whole sample is used in the analysis much of the variation in characteristics between ability-grouped and ungrouped classrooms are, in fact, found to be attributable to between-school differences in those characteristics rather than between-classroom differences. Furthermore, the following analyses show that it is rather school factors than classroom factors that are important in predicting whether or not classrooms use ability grouping (this is further discussed below).

These findings suggest two important points. First, the results of the analyses of classroom ability grouping using "mixed schools" are not generalizable to classrooms in schools where ability-grouping is a school-wide practice. In particular, ungrouped classrooms in "mixed schools" are not comparable to classrooms in ungrouped schools. Second, differences/similarities in school characteristics across the three types of schools may be important factors to consider when we interpret the results of the classroom-level analyses in comparison with the results of the school-level analyses of ability grouping (this is further discussed in Chapter 6).

With regard to the generalizability of the classroom-level analyses, the population of the "mixed schools" cannot be clearly defined because "mixed schools" in the ECLS-K data do not necessarily represent "mixed schools" among the U.S. schools. However, because the analytic sample includes a substantive number of "mixed schools" and it comes from the data with a nationally representative sample of students, the results can be roughly generalizable to a larger population of the U.S. schools that have both ability-grouped and ungrouped classrooms within the same schools.

Predictors of classrooms adopting ability grouping. The next analyses use the sample of "mixed schools" to examine bivariate associations between classroom characteristics and ability grouping status. It is noted that classroom-level covariates have limitations partly because many of the measures are aggregated student characteristics by classrooms and sampled students do not represent students in classrooms, but those in schools. Also, the estimation of the aggregate characteristics of classrooms may be problematic because student sample sizes are relatively small per classroom. The mean sample size per classroom is 3.7 students. In addition, 16 percent of all classrooms in "mixed schools" have only one sampled student. The same student-level covariates used in the school-level analyses are aggregated by classrooms in the classroom-level analyses. Other measures come from the ECLS-K teacher questionnaires on his/her classrooms.

When bivariate relationships are examined between classroom-level covariates and classroom ability grouping status, many covariates are found to have no significant relationships. The results in Appendix C show that only 16 covariates and 2 sets of dummy variables have significant bivariate associations with classroom ability grouping.

However, interestingly, preliminary analyses have also found that when the whole sample is included in the same analysis, 64 classroom-level covariates are found to have bivariate associations with classroom ability grouping status. This suggests that although ability-grouped classrooms differ from ungrouped classrooms in many characteristics, such differences, in fact, reflect differences in the characteristics of the schools where these classrooms are located. In other words, much of the differences between ungrouped classrooms and ability-grouped classrooms are attributable to the fact that ungrouped schools have very different characteristics than ability-grouped schools or "mixed" schools. Indeed, when only classrooms in the "mixed school" are examined, ability-grouped and ungrouped classrooms have relatively similar characteristics.

Appendix C shows descriptive statistics on classroom covariates by classroom ability grouping status among "mixed schools". Appendix D shows bivariate associations between these covariates and classroom ability grouping status. The first column in Appendix D displays z-statistics with robust standard errors. Compared to the results of the school-level analyses (see Appendix A), Appendix D shows that much fewer covariates are associated with ability grouping status in the classroom-level analyses. However among these covariates, the classroom-level results are somewhat consistent to the school-level results. For example, ability-grouped classrooms are generally more heterogeneous in student cognitive and behavioral characteristics than classrooms without ability grouping. Also, ability-grouped classrooms have more students with low literacy skills in the beginning of the school year, while the average cognitive skills and behavior do not have significant associations with classroom ability grouping status.

While both classroom- and school-level analyses suggest that cognitive and behavioral heterogeneities have stronger associations with ability grouping status than the average cognitive skills and behavior, it is noted that in the classroom-level analyses, unlike the school-level analyses, neither IRT scores nor Academic Rating Sales are associated with classroom ability-grouping status.

Appendix D also shows that classrooms with ability grouping have a greater percent of Hispanic students, students from non-English speaking homes, and LEP students than classrooms without ability grouping. In addition, parents of students in ability-grouped classrooms have lower expectations on child readiness for kindergarten and they are less likely to participate in educational and cultural activities than parents of students in ungrouped classrooms. These findings are consistent to the findings of school-wide ability grouping.

However, differences between classroom- and school-level results are found in that the average income is higher for ability-grouped classrooms than ungrouped classrooms. Also, teachers in ability-grouped classrooms have higher educational levels than teachers in ungrouped classrooms.

It is important to note that the above comparisons between ability-grouped and ungrouped classrooms do not take into account the fact that classrooms are nested in schools and schools may differ in various characteristics. It is possible that some of the observed differences between ability-grouped and ungrouped classrooms may be explained by between-school differences in these characteristics. In order to examine this claim, fixed effect models are used to estimate the average within-school differences between ability-grouped and ungrouped classrooms.

The results of the fixed effect models (Column 2 in Appendix D) show that when school differences are taken into account, some classroom covariates are no longer statistically different between ability-grouped and ungrouped classrooms. These characteristics include the percent of Hispanic students, the percent of students from non-English speaking homes, and teacher's educational levels. These results indicate that schools do not segregate classrooms within the same school with regard to these demographic and teacher characteristics. However, ability-grouped classrooms are found to have greater heterogeneities in student cognitive skills and behavior than ungrouped classrooms even after controlling for all school characteristics. This provides some support for a claim that teachers use ability grouping in response to heterogeneity in the student body.

Estimation of the propensity score of a classroom to adopt ability grouping. The strategy to select covariates to estimate a classroom propensity of adopting ability grouping replicates the strategy used in the school level analysis. First, a stepwise logistic regression is performed to select initial covariates to be included in the propensity model. Second, the propensity scores are estimated using these covariates. Third, classrooms are subdivided into strata using two stratification methods—quintile stratification and interval-based stratification. Then, the balance is checked for all covariates within each stratum. If the balanced is not met for covariates that are not included in the initial propensity model, those variables are added and the propensity scores are re-estimated. This process is repeated until a satisfactory balance is achieved for all covariates within each stratum. Also, in the classroom propensity models, school-level error terms are set as zero because preliminary analyses suggest that the error variance does not vary across schools.

The final propensity model includes 10 covariates and two sets of dummy variables (Appendix D). Of those variables, two variables are included as measures of classroom cognitive characteristics, which are the percent of students who read when school began and the standard deviation of teacher ratings on child math skills in spring kindergarten. Three variables are included as measures of classroom behavioral characteristics. They are the standard deviation of teacher ratings on child self control in fall kindergarten, the standard deviation of teacher ratings on child interpersonal relationship in spring kindergarten, and the mean parent ratings on child loneliness. Two classroom demographic variables include the standard deviation of family income and the percent of Hispanic students. Three aggregate measures—the mean level of importance for children to draw before kindergarten, the frequency of parents talking about nature to children, and the percent of students who visited museum in the past month—are constructed using items from parent questionnaires in fall kindergarten. Two sets of dummy variables are the percent of LEP students and teacher's educational levels (see Appendix D for dummy coding).

By following Rosenbaum (1986), missing predictors are imputed by using the means of ability-grouped classrooms and ungrouped classrooms accordingly. Dummy variables are constructed to indicate missing cases and these dummy variables are included in the propensity models when appropriate. Of covariates included in the final propensity model, one covariates and two sets of dummy variables have missing cases of less than 5 percent, five covariates have missing cases between 5 and 10 percent, and four covariates have missing cases over 15 percent.

Table 5.10 displays the descriptive statistics of the estimated propensity scores and their logit transformations by classroom ability grouping status. The mean propensity scores for ability-grouped classrooms and ungrouped classrooms are, respectively, .65 and .58. The ranges of the estimated propensity scores are from .24 to .97 for ability-grouped classrooms and from .14 to .87 for ungrouped classrooms. The means of the logit of the estimated propensity scores for ability-grouped and ungrouped classrooms are, respectively, .66 and .32. They range from -1.03 to 2.76 for ability-grouped classrooms and from -1.66 to 1.82 for ungrouped classrooms.

Table 5.10. Estimated Propensity Scores and Logit of Estimated Propensity Scores by Classroom Ability Grouping Status

	Ability-Grouped Classrooms			*Ungrouped Classrooms*		
	Mean	SD	Range	Mean	SD	Range
Prop scores	.65	.12	.24-.97	.58	.13	.14-.87
Logit prop scores	.66	.54	-1.16-3.60	.32	.56	-1.8-1.86

Figure 5.3 and Figure 5.4 present the distributions of the estimated propensity scores and the logit of the estimated propensity scores, respectively. These two figures show that compared to the results of school-wide ability grouping (see Figure 5.1 and Figure 5.2), ability-grouped and ungrouped classrooms have very similar propensity score distributions. For both classrooms, the estimated propensity scores are approximately normally distributed. These results suggest that characteristics of ability-grouped classrooms and ungrouped classrooms are relatively similar in

"mixed schools". The fact that considerable variations exist between ability-grouped schools and ungrouped schools rather than between ability-grouped classrooms and ungrouped classrooms suggests the importance of between-school stratification processes rather than within-school processes.

Figure 5.3. Distribution of Estimated Propensity Scores by Classroom Ability Grouping Status (Treated = Ability-grouped classrooms. Untreated = Ungrouped classrooms)

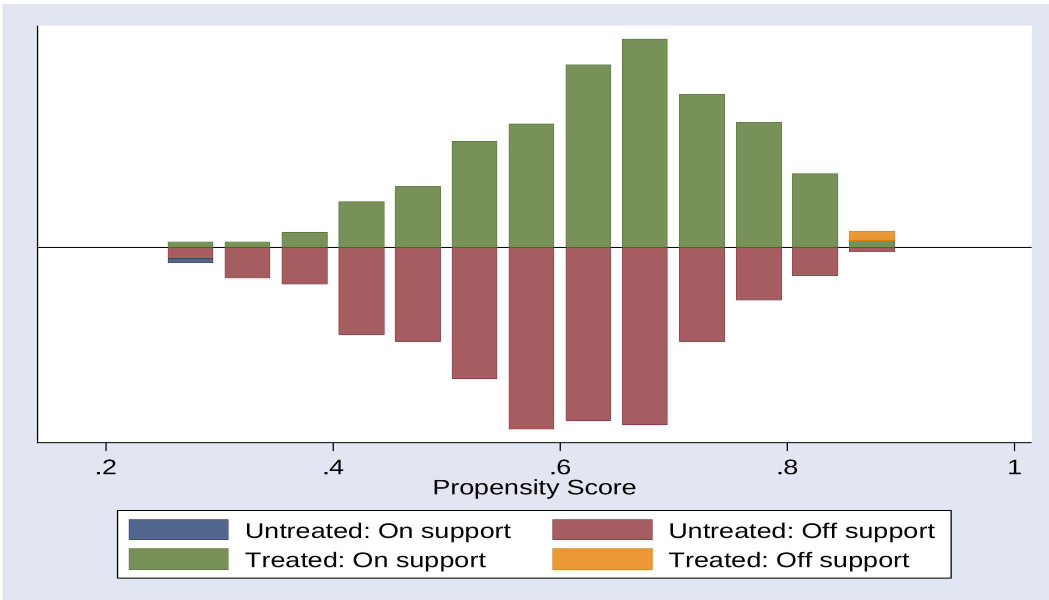
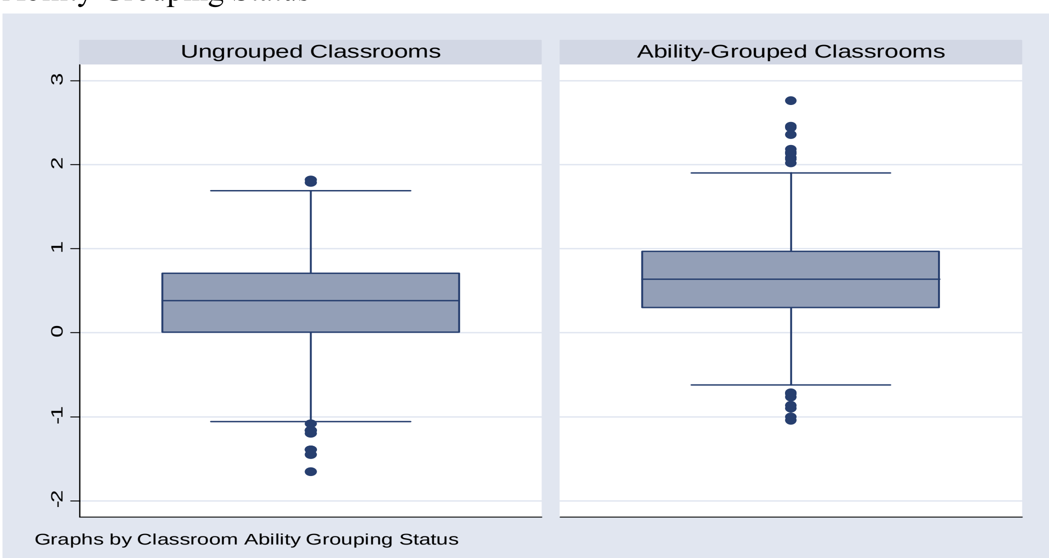


Figure 5.4. Distribution of the Logit of Estimated Propensity Scores by Classroom Ability Grouping Status



Results of propensity score stratifications. This section presents the results of the two propensity score stratifications—quintile stratification and interval stratification. Table 5.11.1 presents a stratification using the propensity score quintiles. This creates five strata with equal sample sizes. Table 5.11.2 presents a stratification based on the propensity score intervals. In both tables higher strata indicate greater propensities of classrooms using ability grouping.

The strategy for the interval-based stratification is the same as the strategy used in the school-level analyses of ability grouping. First, the sample is stratified into five strata of equal propensity score intervals. Then, the mean estimated propensity scores are compared between ability-grouped and ungrouped classrooms within each stratum. When significant differences are found in a given stratum, that stratum is further divided into half.

The resulting interval-based stratification is composed of seven strata. Table 5.11.2 shows that the stratum with the highest propensity scores (stratum seven) contains 39 students in 11 ability-grouped classrooms, and these classrooms do not have their comparison groups in the same stratum. Similarly, the stratum with the lowest propensity scores (stratum one) has 10 students in 5 ungrouped classrooms and this stratum has no ability-grouped classrooms. In the subsequent analyses, the top two strata (stratum six and stratum seven) and the bottom two strata (stratum one and stratum two) are combined. This did not change the substantive results of the subsequent analyses.

It is also noted that the results of the two propensity score stratifications in the classroom-level analyses differ considerably from those in the school-level analyses. For example, recall that in the school-level analyses most ability-grouped schools are observed in the highest stratum while ungrouped schools are relatively evenly distributed across the range of the propensity score distributions (see Figure 5.1). As a result, two stratification strategies have produced very different stratifications in terms of the sample sizes and the range of propensity scores that each stratum contains.

Compared to these school-level results, the results of the classroom-level analyses show that the estimated propensity scores of ability-grouped and ungrouped classrooms are both normally distributed with a good overlap (see Figure 5.3). Consequently, the two stratification strategies have produced relatively similar distributions in terms of the sample sizes and propensity score ranges regardless of how stratification is defined (see Table 5.11.1 and Table 5.11.2).

After classrooms are subclassified into strata, statistical tests are conducted to see the balance on covariates between ability-grouped classrooms and ungrouped classrooms. The results in Appendix E show that the overall balance (i.e., the average within-stratum difference) is met for all covariates at the probability level of .05 in both stratification methods. Also, within each stratum, all covariates are balanced at the probability level of .05 in both stratification methods.

Table 5.11.1. Propensity Score Stratification by Quintiles

	Total Prop. Scores		*Ability-Grouped Classrooms*		*Ungrouped Classrooms*	
Stratum	Mean	Range	N Classrooms	N Students	N Classrooms	N Students
5	.79	.72-.97	211	623	59	197
4	.69	.66-.73	194	691	76	298
3	.63	.60-.66	164	651	106	439
2	.56	.51.-60	144	519	126	498
1	.43	.14-.51	125	364	146	447
		Total	838	2,848	513	1,879
		Prop score mean	.65	NA	.58	NA
		SD	.12		.13	

Table 5.11.2. Propensity Score Stratification based on Propensity Score Intervals

	Total Prop. Scores		*Ability-Grouped Classrooms*		*Ungrouped Classrooms*	
Stratum	Mean	Range	N Classrooms	N Students	N Classrooms	N Students
7*	.93	.90-.97	11	39	0	0
6*	.83	.80-.90	78	206	14	46
5	.74	.70-.80	207	693	67	233
4	.65	.60-.70	270	1,021	159	654
3	.52	.40-.60	247	826	220	792
2*	.34	.22-.40	25	63	48	144
1*	.18	.14-.20	0	0	5	10
		Total	838	2,848	513	1,879
		Prop score mean	.65	NA	.58	NA
		SD	.12		.13	

*the bottom two strata and top two strata are combined in the subsequent analyses

Comparison between students in ability-grouped classrooms and ungrouped classrooms: descriptive results. The next analysis compares first-grade reading achievement between students in ability-grouped classrooms and those in ungrouped classrooms by propensity strata. Descriptive statistics are shown in Table 5.12.1 and Table 5.12.2. These two tables present rough estimates of the effects of classroom ability grouping on first-grade reading achievement.

Table 5.12.1. Descriptive Statistics on Mean Reading Scores in First Grade by Classroom Ability Grouping Status and Propensity Strata (Quintile Stratification)

	Ability-Grouped Classrooms			*Ungrouped Classroom*			
Stratum	N student	Mean	SD	N student	Mean	SD	Mean diff
5	623	56.51	13.42	197	57.30	13.27	-.79
4	691	54.85	12.32	298	56.25	13.39	-1.4
3	651	55.65	13.05	439	55.08	13.60	.57
2	519	56.66	12.65	498	56.75	12.92	-.09
1	364	56.77	12.00	447	56.53	13.19	.24
Total	2,848	55.97	12.90	1,879	56.29	13.27	-.32

Table 5.12.2. Descriptive Statistics on Mean Reading Scores in First Grade by Classroom Ability Grouping Status and Propensity Strata (Interval Stratification)

	Ability-Grouped Classrooms			*Ungrouped Classroom*			
Stratum	N student	Mean	SD	N student	Mean	SD	Mean diff
5	245	57.45	13.65	46	55.83	12.53	1.62
4	693	55.61	12.75	233	56.88	13.71	-1.27
3	1,021	55.27	12.86	654	55.58	13.46	-.31
2	826	56.48	12.81	792	56.82	13.15	-.34
1	63	58.40	12.51	154	55.76	12.49	2.64
Total	2,848	55.97	12.90	1,879	56.29	13.27	-.32

The results in Table 5.12.1 and Table 5.12.2 show that, overall, the average first-grade reading scores of students in ability grouped classrooms are 55.97 and those of students in ungrouped classrooms are 56.29. The difference between the two groups is -.32 points. Within-stratum comparisons show that in the quintile stratification (Table 5.12.1), the average first-grade reading scores for students in ability-grouped classrooms and students in ungrouped classrooms are, respectively, 56.77 and 56.53 in stratum one, 56.66 and 56.75 in stratum two, 55.65 and 55.08 in stratum three, 54.85 and 56.25 in stratum four, and 56.51 and 57.30 in stratum five. The results of the interval-based stratification (Table 5.12.1) show that the average first-grade reading scores for students in ability-grouped classrooms and those in ungrouped classrooms are, respectively, 58.40 and 55.76 in stratum one, 56.48 and 56.82 in stratum two, 55.27 and 55.58 in stratum three, 55.61 and 56.88 in stratum four, and 57.45 and 55.83 in stratum five.

These descriptive results suggest that the first-grade achievement for students in ability-grouped classrooms does not seem to be different from that for students in ungrouped classrooms in most strata whether stratification is defined by the propensity score quintiles or intervals. If there is any effect of ability grouping on student achievement, it may be found for students in classrooms that are least likely to practice ability grouping (i.e., the stratum one classrooms in the interval stratification in Table 5.11.2). In these classrooms, students in ability-grouped classrooms have higher reading achievement by 2.64 than their counterparts in ungrouped classrooms. The next section uses HLM models to estimate ability grouping effects on first-grade reading achievement in "mixed schools".

Model based estimates of ability grouping effects on first-grade reading achievement. To reiterate the following research questions are addressed in the classroom-level analyses. 1) Does reading achievement differ between students who are grouped by ability and those who are not ability-grouped? 2) Does reading ability grouping have differential effects by students' initial abilities? If so, do differential effects contribute to increasing achievement gaps between high and low achievers? 3) Do the effects of ability grouping vary by classrooms? 4) Do the effects of ability grouping vary by students' initial abilities *and* classrooms?

To address the above four research questions, the following four analyses use HLM models discussed in Chapter 4. The first analysis estimates the average ability grouping effect for the overall population of "mixed schools" (Equation 11), the second analysis estimates the average ability grouping effect by student initial ability levels (Equation 12), the third analysis estimates the average ability grouping effect by propensity score strata (Equation 13), and the fourth analysis estimates the average ability grouping effect by student initial ability levels and propensity score strata (Equation 14).

The average treatment effects of ability grouping for the overall population of "mixed schools". The first analysis uses an HLM model in Equation 11 in Chapter 4 to estimate the average effect of ability grouping on first-grade reading achievement for the overall population of "mixed schools". The results of classroom-level analyses are similar to those of school-level analyses (Table 5.13). No significant differences are found in the overall first-grade reading achievement between students in ability grouped classrooms and those in ungrouped classrooms in both stratification methods. The average causal effect is -.36 (SE=.33) for the quintile stratification and -.42 (SE=.33) for the interval-based stratification. These findings provide little evidence that ability grouping practices lead to overall higher achievement.

Table 5.13. Model-based Estimation of The Average Effects of Classroom Ability Grouping on First-Grade Reading Achievement

Stratification by Quintiles		*Stratification by Propensity Intervals*	
Coefficient	SE	Coefficient	SE
-.36	.33	-.42	.33

Significant level: ***p<.001, ** p<.01, * p<.05, + p<.10

Differential effects of ability grouping by initial ability levels. The next analysis examines whether the effect of ability grouping varies by student initial ability levels. First, the sample is divided into three initial ability levels—high, middle, and low achievers—based on the spring-kindergarten reading scores. Dummy variables are constructed to indicate student initial ability levels and the analysis compares reading achievement between students in ability-grouped classrooms and students in ungrouped classrooms at each ability level (Equation 12).

Also preliminary analyses suggest that first-grade reading scores do not vary by classrooms for students with high initial abilities. In other words, first-grade reading achievement for high initial ability students is similar across classrooms. This suggests no effect of ability grouping for these students. In the subsequent analyses, the slope for high initial ability students is set as fixed. The effects of ability grouping are estimated only for students with low and middle initial abilities (Table 5.14).

Table 5.14. Model-based Estimation of the Average Effects of Classroom Ability Grouping on First-Grade Reading Achievement by Initial Ability Levels

	Stratification by Quintiles		*Stratification by Propensity Intervals*	
	Low achievers	Middle achievers	Low achievers	Middle achievers
Coeff.	-.59	-.12	- .62	-.22
SE	.57	.50	.57	.48

Significant level: ***p<.001, ** p<.01, * p<.05, + p<.10

The results in Table 5.14 show that in both stratifications ability grouping has no significant effects on overall first-grade reading achievement at any ability level. The average ability grouping effects for students with low and middle initial abilities are, respectively, -.59 and -.12 (p>.10) in the quintile stratification. The comparative figures in the propensity interval stratification are - .62 for low initial ability students and -.22 for middle initial ability students (p>.10). These findings do not support for a claim that ability grouping leads to higher achievement for students at all initial ability levels. Similarly, there is little evidence to suggest that ability grouping only benefits students with high initial abilities, while it leads to lower achievement for low initial ability students.

Heterogeneity in the effects of ability grouping by classrooms contexts. The next analysis examines whether ability grouping effects vary by classroom contexts. This is done by estimating the stratum-specific effect of ability grouping on first-grade reading achievement (Equation 13).

Recall the results of school-level analyses, which suggest that ability grouping leads to higher student achievement for schools that are less likely to practice ability grouping and it leads to lower student achievement for schools that are more likely to practice ability grouping. In contrast to these findings, the results of classroom-level analyses show little evidence for differential effects of ability grouping on student achievement by classroom characteristics (Table 5.15). The results of the quintile stratification and interval stratification both show no significant effects of ability grouping on first-grade reading achievement for students in any propensity score stratum.

The results of the quintile stratification show that the effects of ability grouping on reading achievement are -.69 for stratum one, -.74 for stratum two, .57 for stratum three, -.45 for stratum four, and -.75 for stratum five (p>.1). The ability-grouping effects for the interval-based stratification are -.03 for stratum one, -.18 for stratum two, -.06 for stratum three, -.76 for stratum four, and .05 for stratum five (p>.1). Thus, for schools that have both ability-grouped and ungrouped classrooms, the effect of ability grouping does not vary by the context of classrooms.

Table 5.15. Model-based Estimation of the Average Effects of Classroom Ability Grouping on First-Grade Reading Achievement by Propensity Score Strata

	Stratification by Quintiles			*Stratification by Propensity Intervals*		
Stratum	Prop. Score	Coefficient	SE	Prop. Score	Coefficient	SE
5	.72-.97	-.75	.88	.80-.97	.05	2.8
4	.66-.73	-.45	.73	.70-.80	-.76	1.18
3	.60-.66	.57	.68	.60-.70	-.06	.79
2	.51.-60	-.74	.68	.40-.60	-.18	.76
1	.14-.51	-.69	.74	.14-.40	-.03	2.63

Significant level: ***p<.001, ** p<.01, * p<.05, + p<.10

Heterogeneity in the effects of ability grouping by student initial ability levels and classrooms contexts. The final analysis examines whether the effects of ability grouping on first-grade achievement vary by student initial ability levels and the characteristics of classrooms. This is done by estimating the stratum specific effects of ability grouping by student initial ability levels (Equation 14). Because preliminary analyses suggest that first-grade reading scores do not vary by classrooms for high initial ability students, the ability grouping effects are set as zero for these students.

The results of this analysis are presented in Table 5.16. It is shown that in either stratification method ability grouping has no significant effect on first-grade reading achievement at any initial ability levels in any strata. However, the results of the two stratification methods seem to suggest that ability grouping may lead to lower achievement for low initial ability students in classrooms that are least likely to practice ability grouping (i.e., classrooms in stratum one). For example, the quintile stratification shows that in stratum one first-grade reading achievement is 1.84 lower for low initial ability students in ability-grouped classrooms than it is for their ungrouped counterparts. Similarly, the interval-based stratification shows that ability grouping leads to lower achievement by 1.56 points for low initial ability students in stratum one classrooms. Neither of these results, however, yields statistical significance.

Table 5.16. Model-based Estimation of the Average Effects of Classroom Ability Grouping on First-Grade Reading Achievement by Initial Ability Levels and Propensity Strata

	Stratification by Quintiles			*Stratification by Propensity Intervals*		
Stratum	Prop. Scores	Low Coeff.	Middle Coeff.	Prop. Scores	Low Coeff.	Middle Coeff.
5	72-.97	-1.71	-.96	.80-.97	-.67	.05
		(1.38)	(1.48)		(3.32)	(2.80)
4	.66-.73	-1.03	.15	.70-.80	-.75	-.77
		(1.25)	(1.12)		(1.36)	(1.19)
3	.60-.66	.92	.10	.60-.70	-.19	-.06
		(1.11)	(1.06)		(.92)	(.79)
2	.51.-60	-.24	-.84	.40-.60	-.95	-.18
		(1.26)	(.93)		(.95)	(.76)
1	.14-.51	-1.84	.87	.14-.40	-1.56	-.03
		(1.41)	(1.14)		(2.96)	(2.63)

Significant level: ***p<.001, ** p<.01, * p<.05, + p<.10
Standard errors are in parenthesis

It is also noted that the magnitude of ability grouping effects in the stratum one classrooms in the quintile and interval stratifications (1.84 and 1.56, respectively) is as large as the magnitude of the school-wide ability grouping effects that have yielded statistical significance. A lack of statistical significance in the classroom results may be due to small sample sizes of students in classrooms. As noted earlier, the average sample size per classroom is 3.7. Also, when the sample is divided into three ability levels, it may create many empty cells or cells that are too sparse to efficiently estimate the effect of classroom ability grouping on student achievement by initial ability levels. Thus, the conclusions drawn from the class-level analyses are not warranted.

With regard to a question of achievement inequalities, the results of the classroom-level analyses provide little evidence that the practice of ability grouping increases, or reduces, achievement inequalities when "mixed schools" are examined. In contrast to the school-level analyses, which showed differential effects of ability grouping by student initial abilities and school characteristics, the classroom-level analyses show little effects of ability grouping for students at any initial ability levels in any types of classrooms. Chapter 6 discusses factors that may explain why school-level analyses and classroom-level analyses have produced different findings. In particular, discussions focus on the importance of school contexts rather than classroom contexts as both determinants and consequences of ability grouping and how school contexts may affect student learning experiences through ability grouping.

Issues on Kindergarten Ability Grouping

It is noted that some kindergartens use ability grouping and some may argue that the causal effect of first-grade ability grouping may be confounded with the effect of kindergarten ability grouping. Table 5.17 presents kindergarten ability grouping status by first-grade ability grouping status. Of 451 ability grouped schools in first grade, 132 schools use school-wide ability grouping in kindergarten, 147 schools do not use kindergarten ability grouping, and 169 schools are "mixed" kindergartens. Of 142 ungrouped first-grade schools, 99 schools have no ability grouping in kindergarten, 19 schools use school-wide ability grouping in kindergarten, and 18 schools are "mixed" kindergartens. Of 306 "mixed" first-grade schools, 45 schools use school-wide ability grouping in kindergarten, 123 schools do not use kindergarten ability grouping, and 137 schools are "mixed" kindergartens.

Table 5.17. Kindergarten Ability Grouping Status by First-Grade Ability Grouping Status

Kindergarten	*First-Grade Ability Grouping*			
Ability grouping	Grouped	Ungrouped	Mix	Total
Grouped	132	19	45	196
Ungrouped	147	99	123	369
Mix	169	18	137	324
Unknown	3	6	1	10
Total	451	142	306	899

In the school-level analyses of ability grouping, kindergarten ability grouping status was not included as a covariate to predict the first-grade ability grouping status of the school. This is partly because if ability grouping is indeed a school decision, schools may decide the policies of ability grouping in kindergarten and first grade simultaneously, and the implementation of kindergarten ability grouping may not necessarily precede the implementation of first-grade ability grouping. The same factors (e.g., heterogeneities in student characteristics) may predict whether kindergartens use, or do not use, ability grouping. In such cases, it may not be appropriate to use kindergarten ability-grouping status as a predictor of first-grade ability grouping.

To examine the extent to which kindergarten ability grouping may bias the estimates of first-grade ability grouping effects, additional analyses are conducted to see the balance on kindergarten ability grouping status between ability-grouped and ungrouped schools within each stratum in both stratification methods. The results (not shown) suggest that in both stratification methods the distributions of kindergarten ability grouping status are balanced at the probability level of .05 in all strata but stratum one in the interval stratification. In the classroom level analyses, the distributions of student ability-grouping status in kindergarten are balanced in all strata in both stratifications.

Also, in order to examine whether kindergarten ability grouping status may bias the effects of school-wide ability grouping, all models are re-estimated by including the kindergarten ability grouping status of schools as an additional covariate in HLM models. Similarly, in the classroom-level analysis, the kindergarten ability grouping status of students is included as a covariate in HLM models. The results (not shown) suggest that the inclusion of kindergarten ability grouping status of schools or students does not alter the general findings of this study in both school- and classroom-level analyses. Thus, there is little evidence for the "carryover effect" of kindergarten ability grouping on first-grade achievement.

CHAPTER 6

ACADEMIC CONSEQUENCES OF WITHIN-CLASS ABILITY GROUPING

Ability grouping is a widely used and often controversial practice in American schools. Ability grouping begins as early as kindergarten, and during early school years teachers typically use within-classroom ability grouping. While much of previous research has examined between-classroom ability grouping, or tracking, in secondary schools, this dissertation has specifically focused on the academic consequences of within-classroom ability grouping on reading achievement in first grade.

This dissertation addresses four research questions: 1) Does reading achievement differ between students who are grouped by ability and those who are not ability-grouped? 2) Does reading ability grouping have differential effects by students' initial abilities? If so, do differential effects contribute to increasing achievement gaps between high and low achievers? 3) Do the effects of ability grouping vary by schools or classrooms? 4) Do the effects of ability grouping vary by students' initial abilities *and* schools or classrooms?

Unlike many other studies on ability grouping, this dissertation is unique in three important ways. First, most research on ability grouping that uses observational data examines the relationship between ability grouping placement and student achievement. The research on ability group placement addresses a question as to how students achieve if they are placed in different ability group levels (e.g. high, middle, and low ability group levels). However, these conventional ability grouping studies, and sociological studies in particular, have not examined a critical policy question of "detracking" or ungrouping. That is, "what does achievement look like when students are ability grouped compared to when these students are not ability grouped?" This question is specifically addressed in this dissertation. Also, placement research has several methodological limitations, which are discussed in Chapter 4.

Second, while there are some experimental or quasi-experimental studies that have addressed questions of ability grouping practices, these studies often have weak external validity due to small sample sizes. In addition, few previous studies have examined the consequences of reading ability grouping practices in the early years of elementary school. This study uses the ECLS-K data with a nationally representative sample of kindergarteners and applies propensity score approaches to address questions of ability grouping practices in first grade.

Third, few previous studies have examined how school or classroom contexts interact with ability grouping practices to produce student achievement outcomes. Even though organizational perspectives suggest important relationships among larger school organizations, classrooms, and ability grouping in the production of knowledge, previous studies of ability grouping have often overlooked the contexts of schools or classrooms in understanding how ability grouping practices produce particular student achievement. This study addresses this critical gap in the previous literature. As discussed below, the findings of this dissertation indeed provide a complex picture of how schools produce student learning.

As discussed in Chapter 1, two possible factors may explain why the effects of ability grouping vary by classroom or school contexts. First, the same policy may produce different outcomes because of the differences in school or classroom contexts. For example, the practice of ability grouping may be similar across schools/classrooms; however, due to the differences in their student ability compositions, ability grouping may produce differential effects by schools/classrooms. Second, the consequences of ability grouping may differ by schools or classrooms because they implement the policy differently. For example, because schools/classrooms differ in student compositions or teacher experiences, schools/classrooms may practice ability grouping differently. Ability group numbers, group size, and ability group compositions may differ across classrooms and schools, which may produce differential effects of ability grouping on student achievement by schools/classrooms. These two possible explanations are further explored below.

Three types of schools are identified in the analyses. The first type includes schools where all classrooms within a school use ability grouping (i.e., ability-grouped schools). The second type includes those schools where no classrooms use ability grouping (i.e., ungrouped schools). The third school type has classrooms with and without ability grouping in the same schools (i.e., "mixed schools").

To examine the above four research questions, two research designs were proposed. The first research design looks at ability-grouping as a school-wide practice. The analytic sample of this design includes ability-grouped schools and ungrouped schools. The second research design views ability grouping as a classroom-level policy, assuming that individual teachers decide whether or not to use ability grouping. The analytic sample of the second research design consists of only schools that have both ability-grouped and ungrouped classrooms (i.e., "mixed schools"). In both research designs, it is assumed that ability grouping is a school or classroom response to the characteristics of students who typically enroll in these schools or classrooms.

Importance of School Contexts as Determinants of Ability Grouping Practices

The findings of this dissertation suggest the importance of school characteristics, not only as factors determining whether schools use ability grouping, but also as factors influencing how ability grouping produces student learning. The ECLS-K data provide no information as to whether ability grouping is a school or classroom policy, and some may argue that even when all classrooms, or no classrooms within a school, use ability grouping, it is individual teachers who decide whether or not to use ability grouping. However, the results of this study suggest that policies of ability grouping—particularly policies for not using ability grouping—are school-level decisions when all classrooms, or no classrooms, use ability grouping in the same schools[22].

First, one of the most important determinants of school ability grouping status is whether schools are public or private. Over 90 percent of public schools have classrooms that use ability grouping. Of these public schools, 51 percent are ability-grouped schools and 40 percent are "mixed schools," while less than 10 percent of all public schools are ungrouped schools. In contrast, 47 percent of all private schools are ability-grouped schools and 14 percent are "mixed schools, while 40 percent of private schools do not use ability grouping.

Second, school characteristics differ among ability-grouped, ungrouped, and "mixed" schools. In particular, schools that do not use ability grouping have particularly distinctive characteristics compared to ability-grouped schools and "mixed schools. In addition to being private schools, these schools are smaller and more likely to have various admission policies than ability-grouped and "mixed schools." These admission policies can be viewed as a mechanism to control the student body of the entrants. In fact, ungrouped-schools have higher mean achievement and more homogenous cognitive and behavioral characteristics. In addition, these schools have higher mean SES, fewer minority students, fewer language minority students, and fewer students who failed the Oral Language Development Scales than ability-grouped and "mixed schools."

[22] Whether schools have formal ability grouping policies is an empirical question. It is also possible that schools have informal policies, such as traditions and consensus among teachers regarding the use of ability grouping.

As for the determinants of school ability grouping status, the findings suggest that heterogeneities in student cognitive skills and the proportion of low achieving students are more important factors than the school average of student cognitive characteristics. More importantly, it is teachers' perceptions of student cognitive skills rather than their objective measures that are associated with whether schools adopt ability grouping. Furthermore, heterogeneities in student behavior, but not the school average of student behavior, influence school ability grouping status.

Third, the importance of school characteristics is illuminated by the fact that much of the dissimilarities between ability-grouped classrooms and ungrouped classrooms are, in fact, attributable to the dissimilarities in the characteristics of schools in which these classrooms are situated. For example, when school contexts are ignored, the ECLS-K data shows that ability-grouped and ungrouped classrooms have very different characteristics. However, this study shows that when the whole sample is used in the analysis, a likelihood of classroom ability grouping is largely determined by school characteristics rather than classroom characteristics. Also, when we compare characteristics between ability-grouped and ungrouped classrooms within the same school using only "mixed schools", many classroom differences, particularly those in student demographic characteristics and mean achievement levels, disappear, although differences in some measures of cognitive and behavioral heterogeneities remain significant.

The findings of this dissertation provide support for organizational perspectives, which suggest that the practice of ability grouping is an organizational response to the diversity of the student body. These perspectives suggest that it is not the average student achievement, but the heterogeneity and the proportion of low achieving students that determine the use of ability grouping. While Bar and Dreeben (1983) discuss little about how heterogeneities in student behavior affect the formation of ability grouping, this study highlights the importance of both cognitive and behavioral heterogeneities in determining the use of ability grouping. Overall, this dissertation illuminates how larger organizational characteristics of schools affect ability grouping practices.

Consequences of Ability Grouping on Student Achievement and Achievement Inequality: School Contexts Matter

This dissertation illuminates complex relationships among school contexts, ability grouping practices, and student achievement. In particular, the findings of this study highlight the significance of school contexts, rather than classroom contexts, in influencing how ability grouping affects student learning.

Pedagogical rationales for ability grouping and social reproduction perspectives provide two seemingly opposing views on how ability grouping affects student achievement. Pedagogical rationales suggest that all students benefit from ability grouping because content is taught at the difficulty level and pace that is commensurate with student ability. In contrast, social reproduction perspectives suggest that ability grouping is a key organizational practice that contributes to reproducing the existing academic and social inequalities. As discussed below, this dissertation provides partial support for both perspectives.

Most sociological research suggests that ability grouping increases achievement inequalities because students who are placed in higher ability groups gain more than students in lower groups. These results are often interpreted as evidence to suggest that ability grouping only benefits high ability students at the expense of low ability students. However, when first-grade within-class ability grouping is examined the results show little evidence to support this claim[23].

Both school- and classroom-level analyses find that, overall, ability grouping has little effect on average student achievement. In addition, overall, the average first-grade achievement does not differ by initial ability levels (i.e., low, middle, and high initial ability levels) between students in ability-grouped schools and their counterparts in ungrouped schools. These results indicate that ability grouping has little effect on the overall patterns of the existing achievement inequalities.

However, the school-level analyses suggest that the effects of ability grouping on the average achievement depend on the characteristics of schools that students attend. Also, the findings show that the ability grouping effects for low, middle, and high initial ability students vary by school contexts.

As discussed above, schools with more advantageous characteristics are less likely to practice ability grouping. However, when these schools do use ability grouping, all students are found to benefit from such practices. The results also suggest that students with low initial ability benefit the most from ability grouping practices. This indicates that in these schools ability grouping may reduce achievement inequalities because low initial ability students gain more than high ability students. These findings support pedagogical rationales for ability grouping.

In contrast, the results suggest that schools with more disadvantageous characteristics are more likely to use ability grouping. When these school use ability grouping, students have lower achievement than they would if these schools had not used ability grouping. In these schools, ability-grouped students with low initial ability have particularly lower achievement than their

[23] However, this is not to say that ability grouping, or tracking, in secondary schools has little effects on achievement inequalities.

counterparts in ungrouped schools. This indicates that in these schools ability grouping may increase achievement inequalities because this practice is more likely to have adverse consequences on students with low initial ability compared to high initial ability students. This provides support for social reproduction perspectives.

Unlike the school-level analyses, the classroom-level analyses show little evidence that the effects of ability grouping on the average student achievement vary by classroom contexts. Neither do they find that the ability grouping effects for low, middle, and high initial ability students vary by classroom contexts.

The findings of this dissertation lead to the following questions: 1) Why do schools that are not likely to practice ability grouping produce higher achievement, especially for students with low initial abilities, when they do adopt ability grouping? 2) In schools which typically use ability grouping, why does this practice lead to lower achievement, especially for low initial ability students? 3) Why does ability grouping have little effect on student achievement in schools that have both ability grouped and ungrouped classrooms?

Before discussing possible explanations for these questions, it should be noted that in many schools ability grouping *does not* have an effect on the average achievement or achievement by the initial ability levels of students. The fact that the effects of ability grouping are not uniform across schools suggests that an important question is not ability grouping per se, that is, whether classrooms or schools use or do not use ability grouping. Rather it is important to ask how ability grouping is practiced in classrooms in certain school contexts and how it affects instruction and student learning experiences. School context may very well be a neglected piece for understanding the puzzle of why ability grouping apparently benefits lower ability students in schools with certain characteristics while disadvantaging lower ability students in schools with different characteristics.

I argue that a key to understanding how ability grouping produces student learning lies in the context of schools and how teachers use ability grouping in classrooms within a certain school context. In other words, the school contexts shape the context of classrooms, which, in turn, affects the practice of ability grouping in classrooms. All three of these elements may be integrally related, and therefore, all should be considered in studies of ability grouping. The interrelationships among school contexts, classroom contexts, and ability grouping practices imply that ability grouping may not be implemented similarly across classrooms and schools. In other words, practices of ability grouping may vary by classroom and school contexts. Thus, a possible explanation for variable effects of ability grouping by school contexts may be that schools, or more specifically classrooms within these schools, practice ability grouping differently[24].

Although this dissertation does not directly address how schools/classrooms differ in the practice of ability grouping, this relationship can be addressed theoretically. This dissertation integrates two theoretical perspectives—organizational perspectives and structural perspectives—in understanding how school and classroom characteristics affect the practice of ability grouping, and how this, in turn, influences student achievement. First, organizational perspectives imply that the practice of ability grouping differs by classrooms because of the differences in classroom compositions. Bar and Dreeben (1983) argue that cognitive

[24] However, it should be noted that even if schools indeed practice ability grouping differently because of the differences in their school contexts and this produces differential ability grouping effects, an alternative explanation—ability grouping practices are similar across schools, but they produce differential effects because of school contextual differences—cannot be empirically tested unless we observe schools that differ in school contexts but similar in ability grouping practices. In other word, these two explanations are not mutually exclusive--even if the former explanation is true, this cannot exclude a possibility of the latter explanation.

heterogeneities and the proportion of students with low ability in classrooms are the major factors that influence teachers' decisions on the structure of ability grouping, such as the size, number, and discreetness. These characteristics of ability grouping partly determine instructional activities given to students, and this, in turn, affects what students learn.

Because classrooms are situated in schools, classroom contexts reflect school contexts. For example, as discussed above differences in classroom characteristics between ability-grouped and ungrouped classrooms are mainly due to differences in characteristics of schools in which these classrooms are located. Thus, it is reasonable to assume that classroom compositions, such as cognitive heterogeneities and the proportion of low achieving students, mirror the compositions of the schools[25]. The findings of this study show that schools that are not likely to use ability grouping are more homogeneous and have fewer low ability students than other schools. This indicates that classrooms in these schools are also more homogenous and have fewer low ability students. Similarly, schools that are likely to use ability grouping are found to be more heterogeneous and have a larger number of low ability students. Classrooms in these schools would also reflect such student characteristics.

School and classroom ability compositions may be a particularly important factor to understand why ability grouping practices are more likely to affect students with lower initial ability than students with higher initial ability. One possible reason is that 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, if classrooms have very few low ability students, it is likely that the low ability students in these classrooms are ability grouped with students with higher ability levels than themselves. In contrast, if classrooms have

---

[25] This is, of course, unless schools segregate students within schools, which is not a common practice in early elementary school, although it is common in middle schools and quite prevalent in secondary schools.

a large number of low ability students, teachers many decide to group low ability students with other low ability students.

This suggests that in schools that are unlikely to practice ability grouping, when they do use ability grouping low ability students are likely to be grouped with higher ability students because of the school characteristics discussed above. In such a case, the level of instruction would be more challenging than instruction given in an ability group with many low achieving students. Thus, it is in this context where ability grouping may be particularly beneficial for lower ability students[26].

In comparison, in schools that are more likely to use ability grouping, students are more likely to be ability grouped with other low ability students because these classrooms have many low ability students. In such a group, the quality of instruction may be lower than it is in a group where higher ability students are at presence.

Previous studies indeed suggest that the academic performance of low ability students is lower when they are in homogeneous ability groups than when they are in heterogeneous groups (Lou et al. 1996). Group compositions may influence student achievement not only because of the differences in instructional activities between these groups, but also because of social psychological processes and peer influence (Eder 1981; Page 1991; Rist 1970). For example, when teachers group many low achieving students in a same group, it may lead to more negative stereotyping partly because it create more salient group distinction between this group and other groups than when these low ability students are grouped with higher ability students. Also, grouping low achieving students may create disruptive and inattentive behavior. This study finds that the teachers' perceptions of student behavior and cognitive skills, rather than more objective measures of cognitive skills, are important factors in predicting whether schools use ability

[26] This is not testable using the ECLS-K data because of a lack of information on group ability compositions.

grouping. The results also indicate that in schools with greater perceived cognitive and behavioral heterogeneities ability grouping leads to lower achievement for students with low initial abilities. This suggests that the self-fulfilling prophecy may be at work.

Organizational perspectives, however, fail to explain why the effects of ability grouping depend on school or classroom contexts. Structural perspectives help us understand this question by describing the relationship between ability grouping structures (e.g., size, number, and discreteness) and student achievement. Structural perspectives argue that classrooms differ in ability grouping structures and such differences produce variations in the effects of ability grouping on student achievement (Sorensen 1970).

As is discussed in Chapter 2, one structural component of within-class ability grouping is selectivity—the degree of group homogeneity and group distinction in cognitive characteristics (Sorensen 1970). Sorensen (1970) argues that more selective ability grouping indicates greater differentiation in student ability and instruction by group levels. Thus, classrooms with selective ability grouping may have greater achievement inequalities than classrooms with unselective grouping. Selective grouping may also reinforce the self-fulfilling prophesy through adverse social-psychological effects, particularly on lower achievers, because it makes more salient distinctions between groups, which may lead to greater stigmatization[27].

Sorensen's structural analyses specifically focus on secondary school ability grouping or tracking. However, the concept of selectivity can be applied to the study of within-class ability grouping. For example, the number of ability groups in classrooms can be considered as one defining factor of selectivity in within-class ability grouping. In the ECLS-K data, the number of ability groups varies from two to five and above, and Nomi (2006) finds that the classrooms with

[27] For example, some studies find that low ability students in homogenous ability groups have lower achievement than low ability students in heterogeneous small groups (See, Lue et al. 1996).

more ability groups create greater distinctions in cognitive characteristics by group levels than the classrooms with fewer ability groups.

Structural perspectives, however, do not explain why some classrooms have more selective ability groups than others and what factors determines the number of groups in classrooms. Organizational perspectives may provide a link between the two. For example, using the ECLS-K data, Nomi (2006) finds that classrooms with more ability groups tend to be more heterogeneous and have more Hispanic students and students with low literacy skills than classrooms with fewer ability groups. She also finds that a teacher's perception of student ability is a more important factor in determining the number of ability groups than more objective measures of student cognitive skills. When the effect of ability group numbers on student achievement is examined, Nomi (2006) finds that classrooms with more ability groups lead to lower average achievement than classrooms with fewer ability groups. In addition, classrooms with more ability groups are found to have greater achievement inequalities, and this is because having more ability groups lowers achievement for students with low initial ability.

These findings indicate that because schools that are less likely to use ability grouping are more homogenous and have fewer low ability students, their ability grouping may be less selective. In contrast, because schools that are more likely to use ability grouping are more heterogeneous and have more students with low literacy skills, classrooms in these schools may have more selective ability groups. In other words, classrooms in these two types of schools—schools that are less likely to use ability grouping and those that are more likely to use ability grouping—may differ in ability group numbers, and this may explain why the former type of schools lead to higher student achievement, while the latter type of schools lead to lower student achievement.

An additional analysis (results not shown) indeed suggests that schools that are more likely to use ability grouping have more groups numbers than schools that are less likely to use ability grouping. For example, in ability-grouped schools that are most unlikely to use ability grouping (i.e., stratum one schools in interval stratification), 62 percent of classrooms use three groups or less. In contrast, in ability-grouped schools that are more likely to use ability grouping, most classrooms use four groups and above. This suggests that typical ability-grouped schools (i.e., disadvantageous schools) tend to use selective ability grouping while atypical ability-grouped schools (i.e., advantageous schools) have unselective ability grouping.

Differences in ability grouping numbers (i.e., selectivity) are also important because they affect peer group formations in classrooms. In a recent study of peer effects on student achievement, Hoxby (2006) shows that peer effects vary by student own ability levels as well as the achievement distribution of his/her classrooms. In general, Hoxby (2006) suggests that while being with students high ability students is beneficial to all students, individual students can also benefit from being with peers who have similar abilities, or being with lower ability students than themselves for students with initial high abilities, as long as classrooms do not create bimodal or multimodal ability distributions. In other words, low achieving students can benefit from being with other low ability students as well as being with students with higher abilities than themselves. Also, high initial ability students benefit from being with other high students as well as those with lower initial ability students than themselves. However, creating bimodal or multimodal ability distributions by changing the ability compositions of peers, whether by increasing high ability peers or low ability peers, are not beneficial to any student.

This implies that in schools with high mean student abilities and more homogeneous ability distributions, grouping students by their abilities can be beneficial to all students. This is because the group number in classrooms in these schools tends to be small (two or three groups) and students ability distributions are relatively homogenous. Thus, ability grouping is unlikely to create bimodal or multimodal ability distributions. In contrast, schools with lower mean abilities and more heterogeneous ability distributions may lead to lower student achievement because classrooms in these schools tend to have more ability groups, which can artificially create multimodal ability distributions.

In addition, the fact that ability grouping has no effects on student achievement in "mixed schools" may highlight the importance of school contexts. For example, there is little evidence to suggest that any ability-grouped classrooms in "mixed" schools produce higher student achievement than ungrouped classrooms in "mixed" schools. This may be because characteristics of these "mixed schools" are different from characteristics of schools that are shown to produce higher achievement by using ability grouping.

"Mixed schools" are found to have the most disadvantageous characteristics of all, and their schools characteristics resemble more to characteristics of ability-grouped schools than ungrouped schools. This may make us wonder why ability grouping does not lead to lower achievement in "mixed schools" just like many ability-grouped schools that have produced lower student achievement. An answer to this question may lie in what goes on in schools that do not use ability grouping even though their school characteristics suggest otherwise.

In these schools—schools that are more likely to use ability grouping, but do not use ability grouping—the decision to *not* use ability grouping may be indeed a school decision. Such a decision may be made based on the beliefs that ability grouping is not beneficial to students. These ungrouped schools may be organizing instructions more effectively to accommodate diversity in the student body at school levels. For example, in these schools, instead of using ability grouping, teachers may use various instructional strategies, or these schools may create more homogenous classrooms (i.e., ability grouping by classrooms) and adjust instruction accordingly. Teachers in these schools may have greater consensus on instructional strategies or spend more time on professional development for instruction to deal with a diverse student population without using ability grouping. These factors may explain not only why students in ungrouped schools have higher achievement than their counterparts in ability-grouped schools among schools that are more likely to use ability grouping, but also why ability-grouping in "mixed schools" have negligible effects on students achievement. However, it is also noted that because of the data limitations for the classroom-level analysis, the results of the classroom-level analyses are tentative.

Direction for Future Research

This dissertation examined the causal effects of ability grouping, but it does not examine the causes of these effects. Similarly, this study only addresses how the effects of ability grouping vary by school contexts, but does not addresses why its effects vary by school contexts. In other words, this study tells us little about what really goes on within schools and how it differs by school contexts. Thus, in future research, it is important to investigate qualitative differences within schools by school contexts to understand why ability grouping produces higher or lower achievement in some school contexts and why this practice has little effects in other school contexts.

First, future research may investigate differences in teacher's and administrator's approach to ability grouping among schools with ability grouping policy, schools with non-ability grouping policies, and schools where ability grouping decisions rest on individual teachers. Because positive effects are found in schools that are unlikely to practice ability grouping and negative effects are found in schools that are likely to use ability grouping, it may be important to investigate why schools have made a decision of using or not using ability grouping and what the consequences of such decisions are on classroom organization and instruction. It would provide us with a greater understanding of the consequences of ability grouping policies if future research examines how different schools implement ability grouping, or non-grouping, policies.

Second, as discussed earlier, differential effects of ability grouping may be explained by compositional factors of schools, how classrooms are organized in schools, how instruction is organized in classrooms, and what instructional activities are provided to students. This study suggests the importance of examining interrelationships among these factors to understand how ability grouping affects student learning. As organizational and structural perspectives might suggest, school contexts shape the organization of classrooms, which affect how teachers organize instruction and what instruction they provide to students. This, in turn, shapes what students learn in classrooms. These relationships can be further examined in future research.

For example, researchers might want to examine differences in group sizes and ability group compositions across classrooms that differ in student compositions and how they affect student achievement. The fact that schools use ability grouping do not mean that teachers instruct students in the same way. It is likely that schools within which the achievement of low ability students is raised may be engaging in fundamentally different instructional practices. Qualitative studies can also provide detailed accounts of differences in how teachers use ability grouping and what instructional activities they provide to students by different school contexts.

Third, this study finds that teachers' perceptions of student ability are more important determinants of the use of ability grouping than standardized test scores. Also, teachers' perceptions are found to be more significant than standardized measures in affecting the number of ability groups (Nomi 2006). Both qualitative and quantitative studies may help us understand, for example, how school compositions shape teachers' perceptions of student ability and how this is translated into the organization of instruction.

Fourth, it may be equally important to examine what teachers do in classrooms when they do not use ability grouping. This study does not address variations in ungrouped schools by school contexts. However, based on the results of this study, it is reasonable to assume that classroom organizations and instructional activities also vary by school contexts for schools that do not use ability grouping. Cohen (1997) discusses different instructional strategies that may be effective in heterogeneous classroom environments. Indeed, the results of this study show higher student achievement when schools do not use ability grouping, and classrooms in these schools are more likely to be diverse in student composition. Overall, to understand why the effects of ability grouping vary by school contexts, it is important to investigate what is occurring in both ability-grouped classrooms and ungrouped classrooms in various school contexts.

Lastly, this study examines the consequences of ability grouping only in first grade. However, because previous research suggests that early literacy skills have long term consequences on later academic outcomes (Entwisle and Alexander 1989, 1993; Cunningham and Stanovich 1997; Farkas and Beron 2004; Werner and Smith 1992; Walker et al. 1994), future research needs to address the long term consequences of ability grouping in early school years.

Implications for Ability Grouping Policy

Several policy implications are discussed here. First, this study suggests the importance of acknowledging heterogeneities in the consequences of ability grouping policies. Policy discussions on ability grouping often surround the question of whether schools or classrooms should or should not use ability grouping to improve student academic achievement. While ability-grouping policies can be implemented at district, school, or classroom levels, this dissertation suggests that the consequences of such policies may be context specific. Thus, one should not assume that such policies will help or hurt certain students without a consideration of the larger school composition and context. For example, ability-grouping can promote student learning in schools that have more favorable school characteristics, such as higher average achievement, higher SES, homogenous, and a small number of low ability students. However, in schools with more socially and economically disadvantaged students this practice is likely to lead lower student achievement.

The importance of school contexts is also highlighted by controversies over “detracking” in secondary schools. For example, Oakes (1994) takes a position that tracking is fundamentally discriminatory and socially unjust. She presents examples of how detracked schools can promote student learning and positive academic and social climates in schools (Oakes and Lipton 1992). However, Hallinan (1994) contends that the success of “detracking” are due to positive normative and political climate in those schools, which has provided institutional support in an effort to detrack schools. Thus, detracking does not necessarily produce positive outcomes in all schools. Hallinan (1997) argues that schools can also implement tracking more effectively to promote student learning. Although within-class ability grouping in elementary schools differ from ability grouping, or tracking, in secondary schools, similar points can be made for the policy of within-class ability grouping.

Second, ability grouping policies are most consequential for low ability students. Low ability students typically attend schools that use ability grouping, and current ability grouping policies have particularly adverse consequences for these students. While opponents of "detracking" often argue that not using ability grouping may lead to lower achievement for high ability students, this study suggests little evidence for this claim in schools with more disadvantaged characteristics. Thus, in these schools, a policy of "detracking" or "ungrouping" may lead to a more equitable organizational practice and higher achievement for low ability students without affecting achievement of high ability students. In contrast, the evidence of this dissertation suggests that low ability students may benefit the most when they attend schools with more advantaged characteristics; however, ironically these schools are least likely to practice ability grouping. In these schools, ability grouping can be encouraged as a means to promote achievement for low ability students as well as higher ability students. However, it should be stressed that how schools implement ability grouping, or ungrouping, policies are more important questions than whether schools implement such policies.

Third, although it is not directly examined in this dissertation, the issue of ability grouping structure is also an important policy issue for low ability students. This is because previous research finds that when classrooms use ability grouping, teachers are likely to use more groups when they have a larger percent of Hispanic students and students with low literacy skills and when they perceive greater heterogeneity in student literacy skills. This, in turn, results in lowering achievement for students with low initial ability (Nomi, 2006).

These negative consequences possibly occur because students in each group receive less time for supervised work from teachers and spend more time working independently when teachers use more groups. In earlier years of schooling, this may be particularly consequential for low ability students because of their young age and unfamiliarity to formal schooling. Classrooms with more ability groups may have teacher's aids; however, they may not be as effective as classroom teachers in instructing students. Or creating more salient distinctions between groups may reinforce the self-fulfilling prophecy. Because these classrooms tend to have a larger proportion of students with limited literacy skills as they are evaluated by teachers, instructional material in low ability groups in such classrooms may be not as challenging as that in low ability groups in classrooms with smaller number of groups. These factors are important to consider for low ability students when schools implement ability grouping policies.

Third, it is important to note that as the number of language minority students is expected to increase in the future due to immigration, ability grouping policy questions need to address how the organization of instruction influences student learning for English learners. In particular, Hispanic immigrants are the fastest growing population in the United States, and they have a greater proportion of young children than any other race/ethnic group. Hispanic students are more likely to attend schools with ability grouping and their classrooms are more likely to have a larger number of groups than other types of ability-grouped classrooms. Also, they begin schooling with much lower literacy skills than White students or Asian immigrants. Thus, the current ability grouping practices are likely to have adverse effects for Hispanic students.

Lastly, if cognitive and behavioral heterogeneities in schools set constraints on organizing instruction for teachers, an alternative policy may be early interventions to improve literacy skills for low ability students. This may reduce cognitive and behavioral heterogeneities in classrooms. However, the effect of such policies may be minimal unless it changes the perceptions of teachers on student ability and behavior.

Conclusions

Although ability grouping policies are among the most controversial issues in educational research, the use of ability grouping has been supported by many, including teachers, school administrators, and parents, as a means of organizing instruction more effectively when students come from diverse academic, social, and cultural backgrounds.

Brown v. Board of Education (1954) and the civil rights movement raised public consciousness regarding equalities in educational opportunities for all students regardless of their racial backgrounds. The Coleman report (1966), which examined the extent of segregation and educational inequalities in American schools, showed that much of achievement variations lie within schools rather than between schools. The publication of this report led many sociologists of education to examine within-school processes of achievement inequalities. In this context, ability grouping was regarded as a key organizational practice in maintaining the existing achievement inequalities because it differentially allocates opportunities-to-learn in school.

Ability grouping is one of the most researched issues in education. While some studies have found that ability grouping produces higher student achievement for all students (Lou et al. 1996; Slavin 1987), most studies have suggested that this practice leads to greater achievement inequalities because it benefits only high ability students and adversely affects low ability students (for example, see Oakes 1985; Rist 1970; Rosenbaum 1970). This study presents a complex picture of how within-class ability grouping produces student learning. It is important to emphasize that the effects of education policies are not uniform—the same policy can be effective in some schools, but not others. For ability grouping policies, the differential effects may be attributable to the fact that schools implement ability grouping policies differently because of the differences in school contexts.

Variability in the effects of ability grouping is not surprising if one considers the diversity across American schools. For example, schools differ in student compositions, including SES, race, and ability compositions, and these schools may practice ability grouping, or non grouping, differently. For example, group size and compositions may differ depending on school characteristics. Prior ability grouping research has, however, paid little attention to the importance of school contexts in affecting how ability grouping produces certain achievement outcomes. This study shows that all students benefit from ability grouping in schools with the most favorable characteristics, while this practice leads to lower achievement for low and middle ability students in schools with more disadvantaged characteristics. These findings, however, may not extend to secondary school ability grouping because secondary schools practice ability grouping differently in many important ways. It would be interesting to examine how school contexts influence ability grouping and student achievement in secondary schools. Lucas and Berends (2002), for example, suggest the sociodemographic composition of schools plays a part in maintaining de fact tracking, which may have important consequences for students' future academic success.

This dissertation suggests that ability grouping is most likely to impact the achievement of low ability students both positively and negatively, depending on school context. From education policy perspectives, this is promising because it suggests a possibility of raising achievement for *all* low ability students in early stages of their educational paths through effective education policies. More importantly, this study suggests that the achievement of low ability students can be raised without lowering the achievement of higher ability students. This is important because in policy discussions on ability grouping and "detracking", improving the achievement of some students are often thought to lower the achievement of others.

It is, however, not known from this dissertation the mechanisms through which ability grouping, or ungrouping, can improve student achievement. Schools can purse a policy of ability grouping or no grouping to improve student achievement, and future educational research should address how such policies impact student learning experiences.

References

Alexander, Karl and Martha Cook. 1982. "Curricula and Coursework: A Surprise Ending to a Familiar Story." *American Sociological Review* 47:626-640.

Alexander, Karl, Martha Cook, and Edward L. McDill. 1978. "Curriculum Tracking and Educational Stratification: Some Further Evidence." *American Sociological Review* 43: 47-66.

Alexander, Karl, Doris R. Entwisle, and Carrie Horsey. 1997. "From First Grade Forward: Early Foundation of High School Dropout." *Sociology of Education* 70:87-107.

Alexander, Karl L., Doris R. Entwisle, and Linda S. Olson. 2001. "Schools, Achievement, and Inequality. A Seasonal Perspective." *Educational Evaluation and Policy Analysis* 23:171-191.

Alexander, Karl L. and Edward L. McDill. 1976. "Selection and Allocation within Schools: Some Causes and Consequences of Curriculum Placement." *American Sociological Review* 1976:963-980.

Argys, Laura M., Daniel I. Rees, and Dominic J. Brewer. 1996. "Detracking America's Schools: Equity at Zero Costs?" *Journal of Policy Analysis and Management* 15:623-645.

Baker, David P. and David Stevenson. 1986. "Mother's Strategies for Children's School Achievement: Managing the Transition to High School." *Sociology of Education* 59:156-166.

Ball, Stephen. 1981. *Beachside Comprehensive: A Case-Study of Secondary Schooling*. Cambridge, England: Cambridge University Press.

Barr, Rebecca and Robert Dreeben. 1983. *How Schools Work*. Chicago: University of Chicago Press.

Becker, Sascha O. and Andrea Ichino. 2002. "Estimation of Average Treatment Effects Based on Propensity Scores." *The Stata Journal* 2:358-377.

Betts, Julian R. and Jamie L. Shkolnik. 2000. "The Effects of Ability Grouping on Student Achievement and Resource Allocation in Secondary Schools." *Economics of Education Review* 19:1-15.

Bourdieu, Pierre. 1973. "Cultural Reproduction and Social Reproduction." in *Knowledge, Education, and Cultural Change*, edited by R. Brown. London: Tavistock.

Bowles, Samuel and Herbert Gintis. 1976. *Schooling in Capitalist America*. New York: Basic Books.

Brophy, Jere E. and Thomas L. Good. 1974. *Teacher-Student Relationships: Causes and Consequences*. New York: Holt, Rinehart and Winston.

Cochran, William G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrika* 24:205-213.

Cohen, Elizabeth G. 1994. *Designing groupwork: Strategies for Heterogeneous Classrooms* (2nd ed.). New York: Teachers College Press.

------. 1997a. "Equity in Heterogeneous Classrooms: A Challenge for Teachers and Sociologists." Pp. 3-14 in *Working for Equity in Heterogeneous Classrooms*, edited by E. G. Cohen and R. A. Lotan. New York: Teachers College Press.

------. 1997b. "Understanding Status Problems: Sources and Consequences." in *Working for Equity in Heterogeneous Classrooms*, edited by E. G. Cohen and R. A. Lotan. New York: Teachers College Press.

Coleman, James S., Ernest Q. Cambell, Carol J. Hobson, James M. McPartland, Alexander M. Mood, Frederick Weinfeld, and York Robert L. 1966. "Equality of Educational Opportunity." Washington: Government Printing Office.

Cox, David R. 1958. *Planning of Experiment*. New York: Willy.

Cunningham, Ann E. and Keith.E. Stanovich. 1997. "Early Reading Acquisition and its Relation to Reading Experience and Ability 10 Years Later." *Developmental Psychology* 33:934-945.

D'Agostino, Ralph B. Jr. 1998. "Tutorial in Biostatistics: Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-randomized Control Group." *Statistics in Medicine* 17:2265-2281.

Dauber, Susan L., Karl L. Alexander, and Doris R. Entwisle. 1996. "Tracking and Transitions through the Middle Grades: Channeling Educational Trajectories." *Sociology of Education* 69:290-307.

Dehejia, Rajeev H. and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94:1053-1062.

------. 2002. "Propensity Score Matching Methods for Non-Experimental Causal Studies." *Review of Economic and Statistics* 84:151-161.

Deil-Amen, Regina and James E. Rosenbaum. 2002. "The Unintended Consequences of Stigma-Free Remediation." *Sociology of Education* 75:249-268.

Drake, Christiana. 1993. "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect." *Biometrika* 49:1231-1236.

Dreeben, Robert and Adam Gamoran. 1986. "Race, Instruction, and Learning." *American Sociological Review* 51:660-669.

Dynarski, Susan M. 2003. "Does Aid Matter? Measuring the Effect of Student Aid on College Attendance and Completion." *The American Economic Review* 3:279-288.

Eder, Donna. 1981. "Ability Grouping as a Self-Fulfilling Prophecy: A Micro-analysis of Teacher-Student Interaction." *Sociology of Education* 54:151-162.

Entwisle, Doris R. and Karl Alexander. 1989. "Early Schooling as a 'Critical Period' Phenomenon." Pp. 27-55 in *Sociology of Education and Socialization*, vol. 8, edited by K. Namboodiri and R. G. Corwin. Greenwich, CT: JAI Press.

------. 1993. "Entry into School: The Beginning School Transition and Educational Stratification in the United States." *Annual Review of Sociology* 19:401-423.

Entwisle, Doris R., Karl Alexander, and Linda S. Olson. 1997. *Children, Schools, and Inequality*. Boulder, CO: Westview Press.

Farkas, George. 1996. *Human Capital or Cultural Capital? : Ethnicity and Poverty Groups in an Urban School District*. New York: A. de Gruyter.

Farkas, George and Kurt Beron. 2004. "The Detailed Age Trajectory of Oral Vocabulary Knowledge: Differences by Class and Race." *Social Science Research* 33:464-497.

Finley, Marilee K. 1984. "Teachers and Tracking in a Comprehensive High School." *Sociology of Education* 57:233-243.

Gamoran, Adam. 1986. "Instructional and Institutional Effects of Ability Grouping." *Sociology of Education* 59:185-198.

------. 1987. "The Stratification of High School Learning Opportunities." *Sociology of Education* 60:135-155.

------. 1992. "The Variable Effects of High School Tracking." *American Sociological Review*

57:812-828.
Gamoran, Adam and Robert D. Mare. 1989. "Secondary School Tracking and Educational Inequality: Compensation, Reinforcement, or Neutrality?" *American Journal of Sociology* 94:1146-1183.
Gamoran, Adam, Martin Nystrand, Mark Berends, and Paul C. LePore. 1995. "An Organizational Analysis of the Effects of Ability Grouping." *American Research Journal* 32:678-715.
Gamoran, Adam and Matthew G. Weinstein. 1989. "Differentiation and Opportunity in Restructured Schools." *American Journal of Education* 106:385-415.
Gaskell, Jane. 1985. "Course Enrollment in the High School: the Prospective of Working Class Families." *Sociology of Education* 58:48-59.
Goldberg, Miriam L., Harry A. Passow, and Joseph Justman. 1966. *The Effects of Ability Grouping*. New York: Teacher College, Columbia University.
Goodlad, John. 1984. *A Place Called School*. New York: McGraw-Hill.
Haller, Emil J. 1985. "Pupil Race and Elementary School Ability Grouping: Are Teachers Biased Against Black Children?" *American Educational Research Journal* 22:465-483.
Haller, Emil J. and Sharon A. Davis. 1980. "Does Socioeconomic Status Bias the Assignment of Elementary School Students to Reading Groups?" *American Educational Research Journal* 17:409-418.
Hallinan, Maureen T. 1984. "Effects of Race on Assignment to Ability Groups." in *The Social Context of Instruction: Group Organization and Group Processes,* edited by L. P. Peterson, L. C. Wilkinson, and M. T. Hallinan. New York: Academic Press.
------. 1991. "School Differences in Tracking Structures and Track Assignments." *Journal of Research on Adolescence* 1:251-275.
------. 1994a. "School Differences in Tracking Effects on Achievement." *Social Forces* 72:799-820.
------. 1994b. "Tracking: From Theory to Practice." *Sociology of Education* 67:79-97.
------. 1996. "Track Mobility in Secondary School." *Social Forces* 74:983-1002.
Hallinan, Maureen T. and Warren N. Kubitschek. 1999. "Curriculum Differentiation and High School Achievement." *Social Psychology of Education* 3:41-62.
Hallinan, Maureen T. and Aage B. Sorensen. 1983. "The Formation and Stability of Instructional Groups." *American Sociological Review* 48:838-851.
Hart, Betty and Todd Risley. 1995. *Meaningful Differences in Everyday Experience of Young American Children.* Baltimore, Maryland: Paul Brookes Publishing Co.
Hauser, Robert, William Sewell, and Duane Alwin. 1976. "High School Effects on Achievement." Pp. 309-341 in *Schooling and Achievement in American Society,* edited by W. Sewell, R. Hauser, and D. Featherman. New York: Academic Press.
Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64:605-654.
------. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65:261-294.
Heyns, Barbara. 1974. "Social Selection and Stratification Within Schools." *American Journal of Sociology* 79:1434-1451.
Hoffer, Thomas B. 1992. "Middle School Ability Grouping and Student Achievement in Science and Mathematics." *Educational Evaluation and Policy Analysis* 14:205-227.

Holland, Paul W. 1986. "Statistics and Causal Inference (with discussion)." *Journal of the American Statistical Association* 81:945-970.

Hong, Guanglei and Stephen W. Raudenbush. 2003. "Causal Inference for Multi-level Observational Data with Application to Kindergarten Retention Study." Pp. 1849-1865 in *2003 Proceedings of the American Statistical Association*. Social Science Section, Alexandria, VA: American Statistical Association.

------. 2005. "Effects of Kindergarten Retention Policy on Children's Cognitive Growth in Reading and Mathematics." *Educational Evaluation and Policy Analysis* 27:205-4224.

Hoxby, Caroline M. 2006. "Taking Race out of the Equation: School Reassignment and the Structure of Peer Effects." in *Unpublished Manuscript*.

Jalan, Jyotsna and Martin Ravallion. 2003. "Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching." *Journal of Business and Economic Statistics* 21:19-30.

Jencks, Christopher, Marshall Smith, Henry Acland, Mary Jo Bane, David Cohen, Herbert Giontis, Barbara Heyns, and Stephan Michaelson. 1972. *Inequality: A Reassessment of the Effects of Family and Schooling in America.* New York: Basic Books.

Kellam, Sheppard G. 1993. "The Social Adaptation of Children in Classrooms: A Measure of Family Childrearing Effectiveness." in *Exploring Family Relations with Other Social Context*, edited by R. D. Parke and S. G. Kellam. Hillsdale, NJ: Erlbaum.

Kerckhoff, Alan C. 1986. "Effects of Ability Grouping in British Secondary School." *American Sociological Review* 51:842-858.

Kerckhoff, Alan C. and Elizabeth Glennie. 1999. "The Matthew Effect in American Education. "*Research in Sociology of Education and Socialization* 12:35-66.

Kozol, Jonathan. 1991. *Savage Inequalities*. New York: Crown Publishers.

Kulik, James A. and Chen-Lin C. Kulik. 1982. "Effects of Ability Grouping on Secondary School Students: A Meta-Analysis of Evaluation Findings." *American Educational Research Journal* 19:415-428.

------. 1984. "Effects of Accelerated Instruction on Students." *Review of Educational Research* 54:409-426.

Lareau, Annette. 1989. *Home Advantage: Social Class and Parental Intervention in Elementary Education.* Philadelphia: Falmer Press.

Lee, Valerie E. and David T. Burkam. 2002. "Inequality at the Starting Gate: Social Background and Achievement at Kindergarten Entry." Presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

Levine, David I. and Gary Painter. 2003. "The Schooling Costs of Teenage Out-Of-Wedlock Childbearing: Analysis with a Within-School Propensity-Score-Matching Estimator." *Review of Economic and Statistics* 85:884-900.

Lou, Yiping, Philip C. Abrami, John C. Spence, Catherine Poulsen, Bette Chambers, and Sylvia d'Apollonia. 1996. "Within-Class Grouping: A Meta-Analysis." *Review of Educational Research* 66:423-458.

Loveless, Tom. 1994. "The Influence of Subject Matter on Middle School Tracking Policies." *Research in Sociology of Education and Socialization* 10:145-175.

Lucas, Samuel R. 1999. *Tracking Inequality: Stratification and Mobility in American High School.* New York: Teachers College Press.

Lucas, Samuel R. and Mark Berends. 2002. "Sociodemgraphic Diversity, Correlated Achievement, and De Fact Tracking." *Sociology of Education* 75:328-348.

Lucas, Samuel R. and Adam Gamoran. 2002. "Tracking and the Achievement Gap." Pp. 171-184 in *Bridging the Achievement Gap,* edited by J. Chubb and T. Loveless. Washington DC: Brookings Institution Press.

McDonough, Patricia M. 1997. *Choosing Colleges.* Albany: State University of New York.

McPartland, James M., J.R. Coldiron, and Jomills J. Braddock. 1987. "School Structures and Classroom Practices in Elementary, Middle and Secondary Schools." Johns Hopkins University, Center for Research on Elementary and Middle Schools, Baltimore.

Merton, Robert K. 1948. "The Self-fulfilling Prophecy." *Antioch Review* 8:193-210.

Metz, Mary. 1978. *Classrooms and Corridors: The Crisis of Authority in Desegregated Secondary Schools.* Berkeley, CA: University of California Press.

Morgan, Stephen L. and David J. Harding. 2005. "Matching Estimators of Causal Effects: From Stratification and Weighting to Practical Data Analysis Routines." in *Unpublished manuscript*.

National Center for Education Statistics. 2002. "ECLS-L Longitudinal Kindergarten-First Grade Public-Use Data Files." Washington D.C.: U.S. Department of Education.

------. 2004. "ECLS-K Third Grade Public-Use Data File." Washington D.C.: U.S. Department of Education.

Nomi, Takako. 2006. "The variable effects of within-class ability grouping: the effect of the group number on reading achievement in first grade." Presented at the annual meeting of the American Educational Research Association. San Francisco.

Oakes, Jeannie. 1985. *Keeping Track: How Schools Structure Inequality.* New Haven: Yale University Press.

------. 1994. "More than Misapplied Technology: A Normative and Political Response to Hallinan on Tracking." *Sociology of Education* 67:84-89.

Oakes, Jeannie, Adam Gamoran, and Reba N. Page. 1992. "Curriculum Differentiation: Opportunities, Consequences, and Meaning." in *Handbook of Research on Curriculum,* edited by P. W. Jackson. New York: Macmillan.

Oakes, Jeannie, Karen Hunter Quartz, Steve Ryan, and Martin Lipton. 2000. *Becoming Good American Schools: the Struggle for Civic Virtue in Education Reform.* San Francisco: Jossey-Bass Publishers.

Page, Reba N. 1991. *Lower Track Classroom: A Curricular and Cultural Perspective.* New York: Teachers College Press.

Pallas, Aaron M., Doris R. Entwisle, Karl L. Alexander, and Francis M. Stluka. 1994. "Ability-Group Effects: Instructional, Social, or Institutional?" *Sociology of Education* 67:27-46.

Raudenbush, Stephen W. and Anthony Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications, Ltd.

Reardon, Sean. 2003. "Sources of Educational Inequality: The Growth of Racial/Ethnic and Socioeconomic Test Score Gaps in Kindergarten and First Grade." Presented at the annual meeting of the American Sociological Association, Atlanta, Georgia.

Rehberg, Richard A. and Evelyn R. Rosenthal. 1978. *Class and Merit in the American High School: An Assessment of the Revisionist and Meritocratic Arguments.* New York: Longman Inc.

Riordan, Cornelius. 1997. *Equality and Achievement: an Introduction to the Sociology of Education*. New York: Longman.

Rist, Ray. 1970. "Social Class and Teacher Expectations: The Self-Fulfilling Prophecy in Ghetto Education." *Harvard Educational Review* 40:411-451.
Rosenbaum, James E. 1976. Making Inequality*: The Hidden Curriculum of High School Tracking.* New York: Wiley.
------. 1980. "Social Implications of Educational Grouping." *Review of Research in Education* 7:361-401.
------. 1984. "The Social Organization of Instructional Grouping." in *The Social Context of Instruction*, edited by L. P. Peterson, L. C. Wilkinson, and M. T. Hallinan. New York: Academic Press.
Rosenbaum, Paul R. 1986. "Dropping out of high school in the United States: An Observational Study." *Journal of Educational Statistics* 11:207-224.
------. 2002. Observational Studies. New York: Springer-Verlag New York, Inc.
Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41-55.
------. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79:516-524
------. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Incorporating the Propensity Score." *The American Statistician* 39:33-38.
------. 2002. Observational Studies. *New York*: Springer-Verlag New York, Inc.
Rosenthal, Robert and Lenore Jacobson. 1968. *Pygmalion in the Classroom*. New York: Holt, Rinehart & Winston.
Rowan, Brian and Andrew W. Miracle Jr. 1983. "Systems of Ability Grouping and the Stratification of Achievement in Elementary Schools." *Sociology of Education* 56:133-144.
Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6:34-58.
------. 1979. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 76:318-324.
------. 1986. "Which Ifs Have Causal Answers?" *Journal of the American Statistical Association* 81:961-962.
------. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100:322-331.
Schafer, E. Walter and Carol Olexa. 1971. *Tracking and Opportunity: The Locking-Out Process and Beyond.* Scarnton, PA: Chandler.
Shadish, William R., Thomas D. Cook, and Donald Thomas Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Company.
Slavin, Robert E. 1987. "Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis." *Review of Educational Research* 57:293-336.
------. 1990. "Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis." *Review of Educational Research* 60:471-499.
Smith, Jeffrey A. and Petra E. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125:305-353.
Sorensen, Aage B. 1970. "Organizational Differentiation of Students and Educational Opportunity." *Sociology of Education* 43:355-376.

Sorensen, Aage B. and Maureen T. Hallinan. 1984. "Effects of Race on Assignment to Ability Groups." in *The Social Context of Instruction: Group Organization and Group Processes,* edited by *P.* L. Peterson, L. C. Wilkinson, and M. Hallinan. New York: Academic Press.
Stevenson, David, Katherine Schiller, and Barbara Schnieder. 1994. "Sequences of Opportunities for Learning." *Sociology of Education* 67:184-198.
Tach, Laura and George Farkas. Forthcoming. "Learning-related Behaviors, Cognitive Skills, and Ability Grouping When Schooling Begins." *Social Science Research*.
Useem, Elizabeth L. 1991. "Student Selection into Course Sequences in Mathematics: The Impact of Parental Involvement and School Policies." *Journal of Research on Adolescence* 1:231-250.
Vanfossen, Beth E., James D. Jones, and Joan Z. Spade. 1987. "Curriculum Tracking and Status Maintenance." *Sociology of Education* 60:104-122.
Vaughn, V., J. F. Feldhusen, and J. W Asher. 1991. "Meta-Analysis and Review of Research on Pull-out Programs in Gifted Education." *Gifted Child Quarterly* 35:92-98
Walker, Dale, Charles Greenwood, Betty Hart, and Judith Carta. 1994. "Prediction of School Outcomes Based on Early Language Production and Socioeconomic Factors." *Child Development* 65:606-621.
Werner, Emmy and Ruth Smith. 1992. *Overcoming the Odds: High Risk Children from Birth to Adulthood*. Ithaca, New York: Cornell University Press.
West, Jerry, Kristin Denton, and Lizabeth M. Reaney. 2001. "The Kindergarten Year: Findings from the Early Childhood Longitudinal Study." National Center for Educational Statistics, Washington D.C.
Yates, Alfred. 1966. *Grouping in Education*. New York: John Wiley and Sons.
Zhao, Zhong. 2005. "Sensitivity of Propensity Score Methods." *IZA Discussion Paper series, No. 1873.* Institute for the Study of Labor, Bonn, Germany.

Appendix A. Descriptive Statistics

Variables	*Total*			*Ability-grouped schools*		*Ungrouped schools*		*Bivariate logistic regression*	
	Mean	SD	% Missing	Mean	SD	Mean	SD	Z statistics	Prop model*
Direct Cognitive Assessments									
K1 Mean math scores	19.77	4.03	6.75	19.45	3.97	20.82	4.07	-3.38	
K1 Mean general knowledge scores	22.36	4.45	6.75	21.92	4.40	23.82	4.33	-4.29	
K2 Mean reading scores	32.63	5.53	0	32.35	5.40	33.51	5.84	-2.18	X
K2 Mean math scores	28.12	4.85	0	27.72	4.80	29.39	4.82	-3.63	
K2 Mean general knowledge scores	27.41	4.75	0	26.89	4.74	29.09	4.38	-4.91	X
K2 SD reading scores	8.55	2.45	0	8.67	2.37	8.15	2.64	2.24	X
K2 SD general knowledge scores	6.08	1.32	0	6.14	1.30	5.88	1.36	2.03	
Pct Failed OLDS	.06	.14	2.50	.08	.16	.02	.07	3.79	X
Academic Rating Scale									
K1 Mean literacy	2.60	.48	9.10	2.57	.48	2.73	.44	-3.45	X
K1 Mean general knowledge	2.65	.67	15.30	2.59	.66	2.84	.69	-3.46	
K1 SD literacy	.60	.18	9.10	.62	.18	.54	.16	4.25	
K1 SD math	.56	.22	18.10	.57	.22	.51	.18	2.71	
K1 SD general knowledge	.69	.27	15.30	.71	.27	.64	.24	2.23	
K2 SD literacy	.66	.20	1.10	.68	.19	.59	.22	4.79	X
K2 SD math	.66	.21	2.30	.68	.20	.59	.23	4.42	
K2 SD general knowledge	.75	.25	1.60	.77	.24	.66	.28	4.53	
Teacher ratings on child									
K1 Mean "use complex sentence structure"	2.85	.79	7.50	2.78	.78	3.12	.76	-4.35	
K1 Pct not yet beginning on "using complex sentence structure"	.42	.27	7.50	.45	.27	.32	.25	4.93	
K1 Mean "use computer for various purposes"	2.33	.73	21.10	2.30	.71	2.46	.76	-1.97	
K1 Mean "use senses to observe/explore"	2.74	.69	11.50	2.69	.68	2.94	.71	-3.51	
K1 Pct not yet beginning on "use senses to observe/explore"	.45	.29	11.50	.47	.29	.36	.29	3.62	
K1 SD approach to learning	.59	.14	7.50	.60	.14	.55	.15	3.52	
K1 SD externalizing behavior	.57	.18	7.50	.58	.17	.54	.19	2.10	
K1 SD internalizing behavior	.45	.15	7.50	.46	.15	.41	.14	3.66	

*the last column indicates variables included in the propensity model

Appendix A (continued). Descriptive Statistics

Variables	*Total*			*Ability-grouped schools*		*Ungrouped schools*		*Bivariate logistic regression*	
	Mean	SD	% Missing	Mean	SD	Mean	SD	Z statistics	Prop model*
Teacher ratings on child (continued)									
K1 SD "use complex sentence structure"	1.01	.28	7.50	1.02	.29	.95	.24	2.51	
K1 SD "use computer for various purposes"	.63	.36	21.10	.65	.36	.56	.35	2.28	
K1 SD "use senses to observe/explore"	.80	.27	11.50	.81	.28	.74	.23	2.51	
K1 SD "solve problems involving numbers"	.82	.33	17.10	.84	.34	.74	.31	2.75	
K1 SD "use strategies to solve math problems"	.82	.33	12.20	.76	.30	.68	.29	2.70	
K2 Pct not yet beginning on "using complex sentence structure"	.16	.15	1.10	.17	.15	.12	.15	3.07	
K2 Mean "use complex sentence structure"	3.76	.60	1.10	3.72	.58	3.90	.63	-3.15	
K2 SD approach to learning	.60	.14	1.10	.61	.14	.58	.15	2.27	
K2 SD externalizing behavior	.56	.17	1.10	.57	.17	.53	.17	2.52	
K2 SD interpersonal relationship	.55	.14	1.10	.56	.14	.52	.14	3.05	
K2 SD "use complex sentence structure"	.98	.31	1.10	1.01	.29	.88	.34	4.62	
K2 SD "use strategies to read unfamiliar words"	1.08	.29	2.00	1.11	.28	.98	.30	4.56	X
K2 SD "use computer for various purposes"	.79	.35	10.80	.83	.34	.64	.35	5.04	X
K2 SD "use senses to observe/explore"	.82	.28	1.80	.85	.27	.73	.30	4.46	
K2 SD "solve problems involving numbers"	.95	.28	1.80	.97	.27	.89	.29	3.15	
K2 SD "use instrument for measuring"	.83	.30	6.10	.86	.29	.73	.29	4.21	
K2 SD "use strategies to for math problems"	.92	.27	2.00	.94	.26	.83	.28	4.32	
Parent Ratings on child									
K1 Mean approaches to learning	3.12	.16	6.75	3.11	.15	3.15	.16	-2.47	
K1 Mean social interaction	3.33	.18	6.75	3.31	.18	3.37	.17	-3.47	
K1 Mean child articulate	.36	.16	6.75	.35	.15	.40	.19	-3.60	X
K1 SD sad/lonely	.37	.09	7.30	.37	.09	.35	.09	2.64	

*the last column indicates variables included in the propensity model

Appendix A (continued). Descriptive Statistics

Variables	*Total*			*Ability-grouped schools*		*Ungrouped schools*		*Bivariate logistic regression*	
	Mean	SD	% Missing	Mean	SD	Mean	SD	Z statistics	Prop model*
SES characteristics									
Mean SES	.04	.51	1.00	0	.51	.18	.48	-3.62	
Mean income	53,869	31,740	1.00	51,307	31,361	62,009	31,674	-3.54	
Pct mother without HS diploma	.13	.16	1.00	.14	.17	.08	.13	4.32	X
Pct mother with at least college degree	.25	.23	1.00	.23	.22	.30	.23	-2.93	
Mean mother occupational prestige	43.65	5.37	7.00	43.29	5.39	44.86	5.17	-2.41	
Pct ever on AFDC	.12	.14	7.30	.13	.15	.08	.11	3.95	X
Pct on AFDC in last 12 months	.10	.14	6.75	.11	.15	.06	.12	3.46	
Pct ever on food stamp	.18	.18	6.75	.19	.18	.13	.16	3.71	
Pct on food stamp in last 12 months	.17	.20	6.75	.18	.20	.12	.18	3.40	
Pct free lunch	.29	.33	0	.33	.33	.17	.27	4.95	X
Pct owning a computer	.56	.24	0	.55	.25	.62	.21	-3.15	
Mean number of books	74.36	32.15	6.75	71.50	32.06	83.94	30.67	-3.87	
Number of student served by Title1 (dummy)									
1 = 1-40, 0 = Otherwise	.18	.39	34.50	.17	.38	.21	.41	-.48	
1 = 41-150, 0 = Otherwise	.17	.37	34.50	.17	.38	.16	.36	.53	
1 = 150, 0 = Otherwise	.08	.28	34.50	.11	.31	.03	.17	2.19	
Demographic characteristics									
Pct 2 parent families	.77	.18	0	.76	.18	.82	.17	-3.05	
Pct mother employed part time	.22	.13	6.75	.21	.13	.26	.15	-3.46	
Pct mother's occupation not applicable	.34	.15	0	.35	.15	.29	.15	3.95	X
Pct mother not employed	.29	.16	6.75	.30	.15	.25	.16	3.07	
Mean number of moves	2.14	.48	6.75	2.18	.47	2.00	.48	3.73	X
Pct non-English speaking at home quartile									
Quartile2: 1 = Yes, 0 = Otherwise	.18	.39	0	.20	.40	.13	.34	2.44	X
Quartile3: 1 = Yes, 0 = Otherwise	.18	.39	0	.18	.38	.20	.40	.74	X
Quartile4: 1= Yes, 0 = Otherwise	.17	.37	0	.19	.40	.09	.29	3.20	X

*the last column indicates variables included in the propensity model

Appendix A (continued). Descriptive Statistics

Variables	*Total*			*Ability-grouped schools*		*Ungrouped schools*		*Bivariate logistic regression*	
	Mean	SD	% Missing	Mean	SD	Mean	SD	Z statistics	Prop model*
Demographic characteristics (continued)									
Pct LEP (dummy)									
1 = .1-10%, 0 = Otherwise	.29	.45	15.60	.32	.47	.21	.41	2.76	
1 = Over 10%, 0 = Otherwise	.13	.34	15.60	.15	.36	.08	.27	2.46	
Have LEP students: 1 = Yes, 0 = No	.45	.50	12.90	.50	.50	.31	.47	3.59	X
Pct Hispanic (dummy)									
1 = .1-5%, 0 = Otherwise	.40	.49	22.80	.43	.50	.33	.47	3.04	
1 = 5-25%, 0 = Otherwise	.21	.41	22.80	.20	.40	.25	.43	1.2	
1 = Over 25%, 0 = Otherwise	.16	.37	22.80	.19	.39	.09	.29	3.25	
Pct Hispanic students (aggregate measure)	.18	.25	0	.19	.27	.12	.18	3.07	X
Pct black students (dummy)									
1 = .1-5%, 0 = Otherwise	.42	.49	17.60	.42	.49	.43	.50	2.14	
1 = 5-10%, 0 = Otherwise	.12	.32	17.60	.11	.31	.15	.36	.86	
1 = 10-25%, 0 = Otherwise	.14	.35	17.60	.16	.36	.09	.29	2.87	
1 = Over 25%, 0 = Otherwise	.17	.38	17.60	.19	.39	.11	.31	3.14	
Pct Black students (aggregate measure)	.14	.25	0	.15	.26	.09	.21	2.40	
Pct minority students (dummy)									
1 = 10 to less than 25%, 0 = Otherwise	.19	.39	3.40	.17	.38	.25	.44	-.49	
1 = 25 to less than 50%, 0 = Otherwise	.15	.36	3.40	.16	.37	.11	.32	1.95	
1 = 50 to less than 75%, 0 = Otherwise	.12	.32	3.40	.13	.33	.09	.29	1.69	
1 = 75% or more, 0 = Otherwise	.21	.41	3.40	.23	.42	.13	.34	2.67	

*the last column indicates variables included in the propensity model

Appendix A (continued). Descriptive Statistics

Variables	*Total*			*Ability-grouped schools*		*Ungrouped schools*		*Bivariate logistic regression*	
	Mean	SD	% Missing	Mean	SD	Mean	SD	Z statistics	Prop model*
Parent expectations and activities with child									
Mean "important to be able to count"	2.34	.35	6.75	2.31	.31	2.44	.45	-3.86	
Pct yes on "attended sporting event in past month"	.44	.17	0	.43	.17	.47	.17	-2.69	
Pct yes on "do sport/exercise together everyday"	.21	.11	6.75	.22	.11	.19	.11	1.99	
Other child characteristics									
Pct yes on "child read everyday outside of school"	.39	.16	0	.41	.16	.36	.15	3.20	
Pct "pre-K parental care only"	.82	.14	0	.81	.15	.85	.12	-2.88	
Mean child health (parent rating)	.15	.19	6.75	.51	.17	.56	.17	-3.15	
Pct Head start participant	.52	.17	6.75	.16	.19	.11	.18	2.79	
School Type (dummy:1 **= Yes, 0 = Otherwise)**									
Private	.27	.45	0	.21	.41	.56	.50	-7.82	X
Non regular public school	.04	.20	0	.03	.18	.05	.22	-2.21	X
School Size/mobility									
Number of children enrolled around 10/1/1998	428.25	240.28	17.20	465.13	239.01	319.13	209.65	6.01	
Total K enrollment	64.16	43.23	14.20	71.41	44.34	42.40	30.91	6.79	
Total school enrollment (dummy)									
1 = 150-299, 0 = Otherwise	.22	.41	1.40	.19	.39	.32	.47	1.38	X
1 = 300-499, 0 = Otherwise	.26	.44	1.40	.26	.44	.26	.44	3.04	X
1 = 500-749, 0 = Otherwise	.26	.44	1.40	.30	.46	.13	.34	5.08	X
1 = Above 750, 0 = Otherwise	.14	.35	1.40	.16	.37	.07	.25	4.32	X
Number of students who left school (dummy)									
1 = 11-40, 0 = Otherwise	.30	.46	23.50	.33	.47	.24	.43	3.35	
1 = 40 or more, 0 = Otherwise	.29	.45	23.50	.33	.47	.15	.36	4.52	

*the last column indicates variables included in the propensity model

Appendix A (continued). Descriptive Statistics

Variables	*Total*			*Ability-grouped schools*		*Ungrouped schools*		*Bivariate logistic regression*	
	Mean	SD	% Missing	Mean	SD	Mean	SD	Z statistics	Prop model*
Admission policy (dummy:1 **= Yes, 0 = No)**									
Requires admission tests	.12	.33	13.20	.09	.29	.21	.41	-3.52	
Requires SAT	.14	.35	13.20	.10	.30	.28	.45	-4.89	X
Requires special aptitude	.04	.20	13.20	.03	.17	.08	.27	-2.23	
Requires interview with child	.26	.44	13.20	.19	.39	.45	.50	-5.69	
Requires recommendation letter	.14	.35	13.20	.10	.29	.28	.45	-4.97	
Requires academic records	.23	.42	13.20	.18	.38	.41	.49	-5.20	
Requires religious affiliation	.08	.28	13.20	.07	.25	.13	.34	-2.31	
Uses assessment to advise delay for school	.26	.44	12.90	.27	.45	.41	.49	-2.85	
School safety (dummy:1=Yes, 0 = No)									
Child/teacher attacked this year	.34	.48	13.50	.38	.49	.23	.43	2.95	
All visitors must sign in	.85	.36	13.50	.88	.32	.76	.43	3.38	
Hall pass required to leave class	.49	.50	13.50	.54	.50	.35	.48	3.61	
Absenteeism problem (likert scale)									
Mean “problem with teacher absenteeism”	1.91	.96	15.60	1.97	.95	1.73	.96	2.51	
Mean “problem with child absenteeism”	2.18	1.10	15.60	2.29	1.14	1.86	.91	3.87	
Teacher Characteristics									
Pct teacher Hispanic	3.26	8.00	18.30	3.84	8.79	1.66	4.91	2.63	
Pct teacher black	4.52	9.49	18.60	5.30	10.17	2.29	6.80	3.05	
Pct teacher white	86.54	23.73	18.80	84.40	25.25	92.63	17.43	-3.35	
Highest teacher base salary (dummy)									
1 = 25K-45K, 0 = Otherwise	.48	.50	17.20	.45	.50	.58	.50	1.90	
1 = 45K-60K, 0 = Otherwise	.30	.46	17.20	.34	.47	.20	.40	3.83	
1 = More than 60K	.12	.33	17.20	.14	.35	.06	.23	3.66	

*the last column indicates variables included in the propensity model

Appendix A (continued). Descriptive Statistics

Variables	*Total*			*Ability-grouped schools*		*Ungrouped schools*		*Bivariate logistic regression*	
	Mean	SD	% Missing	Mean	SD	Mean	SD	Z statistics	Prop model*
Teacher experiences									
Mean # yrs taught in first grade	7.90	5.79	.02	7.55	5.29	9.05	7.09	-2.62	
Mean # yrs taught in this school	8.55	5.80	.02	8.18	5.25	9.76	7.20	2.76	
Pct bachelor's degree	.32	.37	.02	.29	.34	.41	.44	-3.11	
Pct master's degree and above	.33	.35	.02	.35	.34	.26	.37	2.55	
Mean # of early education classes in college	.20	.29	.02	.21	.29	.15	.27	2.38	
Mean # of classes on reading method in college	3.49	1.48	.02	3.58	1.39	3.19	1.72	2.67	
Teacher agreement on the following statement (likert scale)									
Mean "important to finish tasks"	3.51	.71	4.20	3.55	.72	3.39	.66	2.34	
Mean "important to sit still"	3.69	.77	3.90	3.74	.77	3.55	.77	2.56	
Mean "help needed for parents teach child"	3.79	.94	3.70	3.84	.94	3.64	.91	2.19	
Mean "parents should give child schoolwork"	3.66	1.17	3.40	3.76	1.14	3.36	1.22	3.50	
Mean "child should get daily homework"	2.62	1.26	3.40	2.71	1.27	2.35	1.18	2.89	
Mean "child incapable of learning what I teach"	4.09	.80	12.00	4.05	.82	4.23	.71	-2.16	
Mean "staff have school spirit"	1.85	.88	12.60	1.89	.91	1.70	.73	2.13	
Mean "staff accept me as colleague"	4.46	.63	12.10	4.43	.63	4.56	.59	2.02	
Mean "paperwork interferes with teaching job"	2.91	.95	9.60	3.01	.94	2.55	.90	4.56	
Other school characteristics									
Number of computers used for instruction	52.35	35.31	20.10	57.38	36.05	38.33	28.97	5.28	
Number of FTE teacher aids	3.62	4.20	16.70	4.08	4.55	2.28	2.83	3.94	
Number of PTA meetings (dummy)									
1 = 2-3 times a year, 0 = Otherwise	.10	.30	14.00	.09	.28	.13	.34	.88	
1 = 4-6 times a year, 0 = Otherwise	.30	.46	14.00	.32	.47	.27	.45	2.55	
1 = 7 times or more, 0 = Otherwise	.52	.50	14.00	.54	.50	.47	.50	2.65	

*the last column indicates variables included in the propensity model

Appendix A (continued). Descriptive Statistics

Variables	*Total*			*Ability-grouped schools*		*Ungrouped schools*		*Bivariate logistic regression*	
	Mean	SD	% Missing	Mean	SD	Mean	SD	Z statistics	Prop model*
Other school characteristics (continued)									
Number of home visit (dummy)									
1 = 2-3 times a year, 0 = Otherwise	.07	.26	16.70	.08	.28	.04	.20	1.72	
1 = 4 times or more, 0 = Otherwise	.08	.27	16.70	.09	.29	.03	.18	2.16	
Region (dummy: 1 = Yes, 0 = Otherwise)									
Midwest	.26	.44	0	.24	.43	.31	.46	2.00	X
West	.34	.47	0	.37	.48	.25	.44	4.20	X
South	.23	.42	0	.26	.44	.15	.36	4.10	X

*the last column indicates variables included in the propensity model

Appendix B. Covariate Balance Before and After Propensity Score Stratification: Average Within-Stratum Differences between Ability-Grouped and Ungrouped Schools

Variables	*Stratification strategies*		
	No stratification	Quintiles	Propensity intervals
Direct Cognitive Assessments	Z-statistics	Z-statistics	Z-statistics
K1 Mean math scores	-3.38	1.06	.98
K1 Mean general knowledge scores	-4.29	.69	.60
K2 Mean reading scores	-2.18	.27	.44
K2 Mean math scores	-3.63	.22	.13
K2 Mean general knowledge scores	-4.91	-.08	-.21
K2 SD reading scores	2.24	.37	.28
K2 SD general knowledge scores	2.03	.95	.73
Pct Failed OLDS	3.79	.31	.64
Academic Rating Scale			
K1 Mean literacy	-3.45	.43	.67
K1 Mean general knowledge	-3.46	.10	.27
K1 SD literacy	4.25	.20	.30
K1 SD math	2.71	-.12	-.25
K1 SD general knowledge	2.23	-.40	-.58
K2 SD literacy	4.79	.18	.20
K2 SD math	4.42	-.08	-.27
K2 SD general knowledge	4.53	-.33	-.80
Teacher ratings on child			
K1 Mean “use complex sentence structure”	-4.35	.46	.50
K1 Pct not yet beginning on “using complex sentence structure”	4.93	-.07	-.26
K1 Mean “use computer for various purposes”	-1.97	.84	1.13
K1 Mean “use senses to observe/explore”	-3.51	-..44	-.24
K1 Pct not yet beginning on “use senses to observe/explore”	3.62	.40	.24
K1 SD approach to learning	3.52	-1.16	-1.11
K1 SD externalizing behavior	2.10	-.93	-1.08
K1 SD internalizing behavior	3.66	.37	.13
K1 SD “use complex sentence structure”	2.51	-1.68	-1.70
K1 SD “use computer for various purposes”	2.28	.03	-.09
K1 SD “use senses to observe/explore”	2.51	-.71	-.80
K1 SD “solve problems involving numbers”	2.75	.82	.62
K1 SD “use strategies to solve math problems”	2.70	.43	.49
K2 Pct not yet beginning on “using complex sentence structure”	3.07	-.85	-1.07
K2 Mean “use complex sentence structure”	-3.15	.59	.89
K2 SD approach to learning	2.27	-1.85	-1.89
K2 SD externalizing behavior	2.52	-.42	-.41
K2 SD interpersonal behavior	3.05	-.23	-.39
K2 SD “use complex sentence structure”	4.62	-.04	-.15
K2 SD “use strategies to read unfamiliar words”	4.56	-.23	-.01
K2 SD “use computer for various purposes”	5.04	-.32	-.33
K2 SD “use senses to observe/explore”	4.46	-.15	-.70
K2 SD “solve problems involving numbers”	3.15	-.99	-1.11
K2 SD “use senses to observe/explore”	4.46	-.15	-.70
K2 SD “use strategies to for math problems”	4.32	-.56	-.64

Appendix B (continued). Covariate Balance Before and After Propensity Score Stratification: Average Within-Stratum Differences between Ability-Grouped and Ungrouped Schools

Variables	*Stratification strategies*		
	No stratification	Quintiles	Propensity intervals
Parent Ratings on child	Z-statistics	Z-statistics	Z-statistics
K1 Mean approaches to learning	-2.47	.23	.14
K1 Mean social interaction	-3.47	.15	.29
K1 Mean child articulate	-3.60	-.30	-.06
K1 SD sad/lonely	2.64	-.14	.08
SES characteristics			
Mean SES	-3.62	1.06	1.06
Mean income	-3.54	.11	.31
Pct mother without HS diploma	4.32	.05	.04
Pct mother with at least college degree	-2.93	.77	.70
Mean mother occupational prestige	-2.41	1.12	1.07
Pct ever on AFDC	3.95	-.13	-.29
Pct on AFDC in last 12 months	3.46	-.24	-.30
Pct ever on food stamp	3.71	-.47	-.55
Pct on food stamp in last 12 months	3.40	-1.08	-1.20
Pct free lunch	4.95	.05	-.07
Pct owning a computer	-3.15	1.26	1.20
Mean number of books	-3.87	.95	.96
Number of student served by Title 1			
1 = 1-40, 0 = Otherwise	-.48	1.22	1.17
1 = 41-150, 0 = Otherwise	.53	-.09	-.16
1 = 150, 0 = Otherwise	2.19	-.12	.14
Demographic characteristics			
Pct 2 parent families	-3.05	.19	.34
Pct mother employed part time	-3.46	-.19	-.42
Pct mother's occupation not applicable	3.95	.98	1.22
Pct mother not employed	3.07	1.19	.97
Mean number of moves	3.73	-.16	-.46
Pct Non-English speaking at home quartile			
Quartile2: 1 = Yes, 0 = Otherwise	2.44	.32	.42
Quartile3: 1 = Yes, 0 = Otherwise	.74	.19	.33
Quartile4: 1 = Yes, 0 = Otherwise	3.20	.53	.70
Pct LEP students (dummy)			
1 = 0.1-10%, 0 = Otherwise	2.76	0	-.02
1 = Over 10%, 0 = Otherwise	2.46	-.82	-.84
Have LEP students: 1 = Yes, 0 = Otherwise	3.59	-.09	-.16
Pct Hispanic (dummy)			
1 = 0.1-5%, 0 = Otherwise	3.04	1.70	1.76
1 = 5-25%, 0 = Otherwise	1.2	-.31	-.19
1 = over 25%, 0 = Otherwise	3.25	1.14	1.32
Pct Hispanic students (aggregate measure)	3.07	.07	.67

Appendix B (continued). Covariate Balance Before and After Propensity Score Stratification: Average Within-Stratum Differences between Ability-Grouped and Ungrouped Schools

Variables	*Stratification strategies*		
	No stratification	Quintiles	Propensity intervals
Demographic characteristics (continued)	Z-statistics	Z-statistics	Z-statistics
Pct black students (dummy)			
1 = 0.1-5%, 0 = Otherwise	2.14	1.12	1.20
1 = 5-10%, 0 = Otherwise	.86	-.43	-.45
1 = 10-25%, 0 = Otherwise	2.87	1.26	1.19
1 = over 25%, 0 = Otherwise	3.14	.72	1.09
Pct Black students (aggregate measure)	2.40	.98	.41
Pct minority students (dummy)			
1 = 10 to less than 25%, 0 = Otherwise	-.49	-1.81	-1.68
1 = 25 to less than 50%, 0 = Otherwise	1.95	-.07	.12
1 = 50 to less than 75%, 0 = Otherwise	1.69	-.38	-.17
1 = 75% or more, 0 = Otherwise	2.67	-.73	-.52
Parent expectation and activities with child			
Mean "important to be able to count"	-3.86	-1.08	-1.30
Pct yes on "attended sporting event in past month"	-2.69	-.63	-.63
Pct yes on "do sport/exercise together everyday"	1.99	-.45	-.43
Other child characteristics			
Pct yes on "child read everyday outside of school"	3.20	1.85	1.77
Pct "pre-K parental care only"	-2.88	-.21	-.36
Mean child health (parent rating)	-3.15	.53	.59
Pct Head start participant	2.79	-.68	-.81
School Type (dummy:1=Yes, 0=Otherwise)			
Private	-7.82	-.25	.19
Non regular public school	-2.21	.13	-.31
School Size/mobility			
Number of children enrolled around 10/1/1998	6.01	.96	.49
Total K enrollment	6.79	1.13	.82
Total school enrollment (dummy)			
1 = 150-299, 0 = Otherwise	1.38	.36	.10
1 = 300-499, 0 = Otherwise	3.04	.39	.05
1 = 500-749, 0 = Otherwise	5.08	.70	.30
1 = Above 750, 0 = Otherwise	4.32	.68	.41
Number of students who left school (dummy)			
1 = 11-40, 0 = Otherwise	3.35	-.20	-.31
1 = 40 or more, 0 = Otherwise	4.52	.08	.30
Admission policy (dummy:1=yes, 0=Otherwise)			
Requires admission tests	-3.52	.10	.47
Requires SAT	-4.89	-.75	-.22
Requires special aptitude	-2.23	-.88	-.63
Requires interview with child	-5.69	-.09	.34
Requires recommendation letter	-4.97	-.86	-.31
Requires academic records	-5.20	-.46	.03

Appendix B (continued). Covariate Balance Before and After Propensity Score Stratification: Average Within-Stratum Differences between Ability-Grouped and Ungrouped Schools

Variables	*Stratification strategies*		
	No stratification	Quintiles	Propensity intervals
Admission policy (continued)	Z-statistics	Z-statistics	Z-statistics
Requires religious affiliation	-2.31	.84	1.09
Uses assessment to advise delay for school	-2.85	-.82	-.81
School safety (dummy:1=Yes, 0=Otherwise)			
Child/teacher attacked this year	2.95	.97	.72
All visitors must sign in	3.38	-.01	-.29
Hall pass required to leave class	3.61	1.55	1.55
Absenteeism problem (likert scale)			
Problem with teacher absenteeism	2.51	.12	.10
Problem with child absenteeism	3.87	.72	.97
Teacher Characteristics			
Pct teacher Hispanic	2.63	1.03	1.21
Pct teacher black	3.05	.95	.89
Pct teacher white	-3.35	-1.00	-1.06
Highest teacher base salary (dummy)			
1 = 25K-45K, 0 = Otherwise	1.90	.34	.26
1 = 45K-60K, 0 = Otherwise	3.83	.06	-.10
1 = more than 60K	3.66	1.54	1.28
Teacher's experiences			
Mean # yrs taught in first grade	-2.62	-2.31	-2.26
Mean # yrs taught in this school	2.76	-2.44	-2.26
Pct bachelor's degree	-3.11	-1.52	-1.47
Pct master's degree and above	2.55	1.25	1.38
Mean # of early education classes in college	2.38	1.71	1.92
Mean # of classes on reading method in college	2.67	1.59	1.84
Teacher agreement on the following statement (likert scale)			
Mean "important to finish tasks"	2.34	1.47	1.46
Mean "important to sit still"	2.56	1.78	1.85
Mean "help needed for parents teach child"	2.19	-.01	-.05
Mean "parents should give child schoolwork"	3.50	1.72	1.58
Mean "child should get daily homework"	2.89	1.13	1.34
Mean "child incapable of learning what I teach"	-2.16	-.93	-.84
Mean "staff have school spirit"	2.13	-.05	.08
Mean "staff accept me as colleague"	2.02	-.06	0
Mean 'paperwork interferes with teaching job"	4.56	.79	.69

Appendix B (continued). Covariate Balance Before and After Propensity Score Stratification: Average Within-Stratum Differences between Ability-Grouped and Ungrouped Schools

Variables	*Stratification strategies*		
	No stratification	Quintiles	Propensity intervals
Other school characteristics	Z-statistics	Z-statistics	Z-statistics
Number of computers used for instruction	5.28	1.28	.93
Number of FTE teacher aids	3.94	.44	.35
Number of PTA meetings (dummy)			
1 = 2-3 times a year, 0 = Otherwise	.88	.08	.40
1 = 4-6 times a year, 0 = Otherwise	2.55	1.39	1.71
1 = 7 times or more, 0 = Otherwise	2.65	.52	.64
Number of home visit (dummy)			
1= 2-3 times a year, 0 = Otherwise	1.72	-.16	-.45
1= 4 times or more, 0 = Otherwise	2.16	.68	.40
Region (dummy:1=Yes, 0=Otherwise)			
Midwest	2.00	.20	.10
West	4.20	.38	.41
South	4.10	.19	.37

Appendix C. Descriptive Statistics on Classroom Characteristics

Variables	*Total*			*Ability-grouped classrooms*		*Ungrouped classrooms*	
	Mean	SD	% Missing	Mean	SD	Mean	SD
Teacher ratings on child							
K1 SD approach to learning	.57	.30	22.80	.58	.30	.54	.29
K1 SD externalizing behavior	.51	.34	23.32	.53	.35	.48	.31
K1 SD internalizing behavior	.42	.29	24.35	.44	.30	.40	.26
K1 SD self control	.49	.27	25.68	.51	.28	.46	.25
K1 SD interpersonal relationship	.51	.28	25.46	.53	.27	.49	.28
K2 SD interpersonal relationship	.53	.28	18.43	.54	.29	.50	.27
K2 SD "use strategies to solve math problems"	.86	.50	18.36	.88	.51	.81	.48
# of kids who read when school began	.55	.31	9.70	.53	.30	.57	.32
Parent ratings							
Mean "child is sad/lonely"	1.55	.27	2.66	1.53	.26	1.58	.28
Demographic characteristics							
Mean Income	46,925	36,875	1.48	48,336	41,081	44,610	28,554
SD income	26,659	36,005	17.91	28,448	41,578	23,736	24,047
Pct Hispanic	.18	.28	5.85	.19	.29	.16	.26
Mean Pct Non-English speaking at home	.17	.32	5.85	.18	.34	.15	.30
Pct LEP students (Dummy)							
1 = 25 to 50 percent, 0 = otherwise	.04	.20	0	.04	.19	.05	.23
1= over 50 percent, 0 = otherwise	.07	.26	0	.09	.28	.05	.22
Parent expectations/parent-child activities							
Important to child to draw before K	3.94	.52	7.55	3.90	.53	4.00	.50
Talk about nature to child	.09	.20	7.48	.09	.18	.11	.23
Pct visited museum in the past month	1.70	.32	2.22	1.68	.33	1.74	.30
Teacher's educational level (dummy)							
1= At least 1 yr beyond bachelor, 0=otherwise	.28	.45	3.26	.29	.45	.28	.45
1= Master and above, otherwise	.41	.49	3.26	.43	.50	.37	.48

Appendix D. Bivariate Logistic Regressions of Classroom Ability Grouping on Classroom Characteristics: Z Statistics with Robust Standard Errors and Fixed Effects

	Z stats. with robust SE	Z stats. with fixed effect	Prop model*
Teacher ratings on child	Z-statistics	Z-statistics	
K1 SD approach to learning	2.10	1.27	
K1 SD externalizing behavior	2.38	2.65	
K1 SD internalizing behavior	1.97	1.71	
K1 SD self control	2.72	3.00	X
K1 SD interpersonal relationship	2.19	2.23	
K2 SD interpersonal relationship	2.27	3.13	X
K2 SD “use strategies to solve math problems”	2.20	2.45	X
Pct of students who read when school began	-1.86	-2.02	X
Parent ratings			
Mean “child is sad/lonely”	-2.93	-2.16	X
Demographic characteristics			
Mean Income	2.17	2.26	
SD income	2.34	1.97	X
Pct Hispanic	2.47	-.44	X
Pct students from Non-English speaking homes	1.76	1.00	
Pct LEP students (Dummy)			
1 = 25 to 50 percent, 0 = otherwise	-1.55	-1.86	X
1 = over 50 percent, 0 = otherwise	2.51	1.02	X
Parent expectations/parent-child activities			
Mean “important to child to draw before K”	-3.37	-3.46	X
Mean “talk about nature to child”	-2.11	-1.68	X
Pct visited museum in the past month	-3.14	-2.25	X
Teacher’s educational level (dummy)			
1 = At least 1 yr beyond bachelor, 0=otherwise	1.72	.75	X
1 = Master and above, 0 = otherwise	2.79	1.59	X

*the last column indicates variables included in the propensity model

Appendix E. Covariate Balance after Propensity Score Stratification:
Average Within-Stratum Differences between Ability-Grouped and Ungrouped Classrooms

	Quintile Stratification	Interval Stratification
Teacher ratings on child	Z-statistics	Z-statistics
K1 SD approach to learning	1.11	1.04
K1 SD externalizing behavior	.65	.72
K1 SD internalizing behavior	1.84	2.16
K1 SD self control	-.54	-.65
K1 SD interpersonal relationship	-.10	.05
K2 SD interpersonal relationship	-.21	-.37
K2 SD "use strategies to solve math problems"	-.27	-.49
# of students who read when school began	-.14	.15
Parent ratings		
Mean "child is sad/lonely"	-.61	-.15
Demographic characteristics		
Mean Income	.56	.14
SD income	.43	-.08
Percent Hispanic	.03	-.23
Pct students from Non-English speaking homes	.57	.32
Pct LEP students (Dummy)		
1 = 25 to 50 percent, 0 = otherwise	-.59	-.19
1 = over 50 percent, 0 = otherwise	.29	-.12
Parent expectations/parent-child activities		
Important to child to draw before K	-.16	.03
Talk about nature to child	-.38	-.13
Pct visited museum in the past month	-.65	-.21
Teacher's educational level (dummy)		
1 = At least 1 yr beyond bachelor, 0 = otherwise	.33	.16
1 = Master and above, otherwise	.38	.26

Vita

Takako Nomi

Education

Doctor of Philosophy-Educational Theory and Policy with a Minor in Comparative International Education
2006 The Pennsylvania State University

Master of Arts-Sociology
2001 University of Missouri, St. Louis

Bachelor of Arts-Sociology
1998 University of Missouri, St. Louis
Sum Cum Laude

Associate of Arts-English Literature
1995 Toyo University Junior College, Tokyo, Japan

Honors and Awards

2006	The winner of the AERA Grants Program's Statistics Institute paper competition
2005-2006	American Educational Research Association Dissertation Grant
2004	Sponsored participant, American Educational Research Association Institute on Statistical Analysis for Education Policy, April, San Diego, CA.
2004	Penn State College of Education Alumni Award
2002-2004	Research Assistantships in the Educational Theory and Policy Program Penn State University
1997	The Outstanding Junior Sociology Major Award, University of Missouri-St. Louis
1997	Merit Scholarship, University of Missouri-St. Louis

Publications

Nomi, Takako and Suet-ling Pong. Forthcoming. "Self-fulfilling Prophecy." In Encyclopedia of Sociology, edited by George Ritzer. Malden, MA: Blackwell.

LeTendre, Gerald K, Roger Geertz Gonzalez, and Takako Nomi. (2006). "Feeding the Elite: The Evolution of Elite Pathways from Star High Schools to Elite Universities" *Higher Education Policy.*