UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title

Essays on the Economics of Education

Permalink
https://escholarship.org/uc/item/33f04700

Author

Sohn, Hosung

Publication Date

2013

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Essays on the Economics of Education

By

Hosung Sohn

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Public Policy

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Steven Raphael, Chair

Professor Jesse Rothstein

Professor Michael Anderson

Spring, 2013

Abstract

Essays on the Economics of Education

by

Hosung Sohn

Doctor of Philosophy in Public Policy

University of California, Berkeley

Professor Steven Raphael, Chair

This dissertation addresses three questions in the economics of education. Chapter 1 analyzes whether segregating students by gender is beneficial for students' academic achievement. Students or parents often choose peer groups by selecting school types, assuming that peers are important determinants of one's academic achievement. Among the various types of peer effects, this study addresses whether segregating students from the opposite sex is beneficial for one's academic performance by making use of the variation created by randomly assigning students to either same- or mixed-sex high schools. By using seven years of administrative data on scores in college entrance exams, I find that both male and female students benefit by being in same-sex schools. Moreover, the quantile regression analysis reveals that the effect is greater for students located at the middle quantile of the distribution of test scores. I conducted a sensitivity analysis by using a different type of test that students take, and the results are robust.

In Chapter 2, unlike estimating the effect of conventional incentive mechanisms in which good schools are rewarded and bad schools are punished, I estimate the impact of "rewarding" poor-performing schools on students' academic achievement. Because of the simple discontinuous eligibility that determines the provision of categorical school funding to underachieving schools, I use regression discontinuity designs to causally estimate the treatment effect. The results of the analysis reveal that students' academic performance in poor-performing schools improved significantly (7 to 10 percentile points) after the treatment. Moreover, the ratio of underachieving students decreased in schools that received funding (5 to 10 percentage points), relative to those that did not receive funding.

Finally, in Chapter 3, I explore whether grouping students by ability benefits students. Local education agencies often engage in educational reforms with limited resources aimed at improving the academic achievement of students. One of the low cost methods that the agency frequently employs is the use of ability tracking. In this chapter, by making use of the randomized social experiment conducted in Seoul, I provide causal estimates of the effect of ability tracking on students' achievement, using administrative data on students' test scores. Based on the results, I find that, on average, tracking promotes achievement of not only high-achieving students, but also of low-achieving students. Moreover, the magnitude of the treatment effect is similar across various quantiles of the distribution of students' performance. Therefore, contrary to the view that tracking may be detrimental to the learning of low-achieving students, tracking may not worsen inequality in students' achievement.

This dissertation is dedicated to:

my family, Eunjoo Cho and Jimin Sohn;

my parents, Bokjo Sohn and Younghee Choi; and

my parents-in-law, Dukhyung Cho and Hanchoon Hwang.

I sincerely appreciate their support and patience during my graduate studies.

Contents

List of Figures

List of Tables

Acknowledgements

First of all, I would like to thank my academic advisor, Steven Raphael, for providing me with the opportunity to pursue a Ph.D. at the University of California, Berkeley. I also thank him for his guidance throughout my entire graduate studies. His suggestions changed my life.

Next, I would like to express appreciation to my two other dissertation committee members, Jesse Rothstein and Michael Anderson. I have learned tremendously from their lectures, lecture notes, reading lists, problem sets, and office hours. There lectures were among the best I have taken during my school days.

In addition, I am grateful to Eugene Smolensky, Aaron Chalfin, Candace Hamilton, Sarah Tahamont, Natalie Ahn, Roberto Hernandez, and Layda Negrete for their comments and suggestions during the seminar.

Furthermore, I have benefited from discussions with Wonyoung Park, Kwangbin Bae, Hyungjo Hur, Joonwoo Hong, Eunsup Jang, Namho Kwon, Hyungah Kim, Dongsook Han, and Jaehee Choi. I have also benefited from Yoonchee Kim's excellent English editing. I thank their interests, time, and patience.

Finally, I thank Honggyu Hwang, a former Deputy Director of the Ministry of Education, Science, and Technology (MEST), public officials at the Seoul Metropolitan Office of Education, and EduData Service System for allowing me to access all the necessary administrative data. Without their help, this dissertation would not have been possible.

1 Distributional Impact of Gender Segregation on Student Performance: Evidence from Randomized Block Designs

1.1 Introduction

Across all ages throughout the world, education is considered to be the most compelling means to attain upward social mobility and economic wealth. As a consequence, parents put much consideration into the choice of educational alternatives such as teacher quality insomuch that their decisions affect their child's academic achievement. Among many alternatives, parents carefully choose their child's peer group believing that peer group is an important determinant of one's academic and also non-academic achievement.

There are many kinds of peer effects, and most researchers in the economics literature have focused on analyzing the ability peer effect.[1] Epple and Romano (2011) and Sacerdote (2011) also provide surveys of many of these peer effect studies. Recently, researchers have focused on estimating the causal estimate of gender peer effects. The rationale for estimating the gender peer effect is that social interactions between boys and girls often have critical effects on academic performance of both genders. For example, students may feel less distracted when the opposite sex is not present in their class or at school, and accordingly, this helps students focus more on their academics. Moreover, it is possible that the level of fatigue that teachers experience may increase when teachers have to face both genders.

As demonstrated in Manski (1993), estimating the gender peer effect is difficult because, in general, a peer group is an endogenous consequence of individual choice. To overcome the endogeneity issue, one must control for all the observed confounding variables. Even if one succeeds in controlling for all of the observed characteristics of students, however, there still exists unobservable attributes that determine the selection of peer groups and unobservable factors that affect students' performance, and these will introduce bias to estimates of peer effect.

[1] Some of the evidence of ability peer effects appear in Sacerdote (2001), Hanushek et al. (2003), Zimmerman (2003), Arcidiacono and Nicholson (2005), Foster (2006), Ding and Lehrer (2007), Carrell, Fullerton and West (2009), and Ammermueller and Pischke (2009).

The most convincing method to tackle the endogeneity problem is to randomize the gender peer group. Fortunately, middle school students in Seoul, the capital city of South Korea, are randomly assigned to either single- or mixed-sex high schools upon their graduation. Hence, in this study, I exploit the random variation in gender composition created by Seoul's random assignment policy to analyze the effect of gender segregation on students' academic achievement.

I estimate both the mean effects and the distributional effects of gender segregation using the conventional linear regression method and the quantile regression method. The results of the analysis show that, on average, both boys and girls in boys-only and girls-only schools scored 2 to 3 percentile points higher in reading, English, and basic mathematics sections of college entrance exams compared with boys and girls in coeducational schools. Furthermore, quantile regressions show heterogeneity in the estimated treatment effect. Gender segregation is highly favorable for students in middle quantiles of the distribution of percentile ranks. The estimated treatment effect for these students ranges from 4 to 6 percentile points. However, the magnitude of the treatment effect is inconsequential for students in the very bottom and the very top quantiles of the distribution.

1.2 Literature Review

There are various theoretical arguments for and against single- or mixed-sex education. These arguments can be classified into academic and non-academic theories. Regarding the former, some argue that male students are often distracted by the presence of female students, and vice versa. As to non-academic standpoints, people oppose single-sex education and contend that in single-sex schools, peer-bullying is prevalent. Since the aim of this paper is to estimate the effect of gender segregation on students' *academic* achievement, however, I will focus only on studies relevant to academic arguments.[2]

The earliest study is conducted by Coleman (1961). In a two-year study, he finds that the coeducational high school setting is detrimental to student's academic achievement because students are likely to focus more on non-academic issues such as dates. Building upon Coleman's (1961) work, researchers, especially psychologists and sociologists, have examined the single-sex school effect (e.g. Schneider and Coutts, 1982; Lee and Bryk, 1986; Marsh, 1991; Harker, 2000). Mael (1998) provides an excellent literature review of most of the work performed before 2000. Moreover, Sax's (2006) book on the effectiveness of segregating gender raises multifarious scientific evidence in favor of offering same-sex schools. The findings from these research, however, is inconclusive as to whether single-sex education is better or not.

Even though many of the earlier works have presented convincing arguments, all used methods such as ordinary least squares (OLS) and analysis of variance. These methods are limited in addressing problems of selection bias in studies of observational data (e.g., High School and Beyond Survey), and unforgiving consequences of not addressing the selection bias in an empirical analysis are well recognized and are not reiterated here.

Starting with Hoxby (2000), however, the economics literature have focused on the issue of endogeneity when estimating the gender peer effect and used various identification strategies to tackle the bias incurred by potential endogeneity. Hoxby (2000) and Lavy and Schlosser (2011), for example, exploit exogenous sources of variations generated in gender composition within the classroom and the school.

[2] Besides, there is insufficient data to analyze the effect of gender segregation on "non-academic" perspectives.

Using administrative data from the Texas Schools Project and the Israel Ministry of Education, both studies find that both male and female students benefit from an increase in the ratio of female students. Hoxby (2000) notes that the channels of the effect may be reduction in classroom disruptions and an improvement in intra-student relationship, and these channels have been confirmed in Lavy and Schlosser's (2011) study.

Even though the above two studies make use of credible idiosyncratic variations of gender compositions, more compelling identification of gender peer effects can be obtained by using randomized data. Using a widely recognized randomized data set, Tennessee's Project STAR, Whitmore (2005) makes use of random variations in classroom sex compositions. In this data set, a mean ratio of female students in a classroom is 49% with a within-school standard deviation of 11 percentage points. Using these random variations, Whitmore (2005) reports that an increase in the proportion of female students has a positive effect on boys as well as girls from kindergarten to second grade. For students in third grade, however, she finds that boys' test scores tend to decrease when they were in a classroom with a higher fraction of girls.

Rather than focusing on the class- or school-level gender peer effects, Lu and Anderson (2011) estimate gender peer effects within a sub-classroom. In China, middle school students are randomly assigned to classroom seats. Accordingly, Lu and Anderson (2011) exploit random variations of neighboring students that are generated within the subgroups of classrooms. The results show that both boys and girls benefit from having female deskmates. On the other hand, while an increased share of girls have positive spillover effects on girls, they do not find that boys are affected by a higher share of girls.

Oosterbeek and van Ewijk (2010) estimate gender peer effects in a post-secondary setting. They conducted an experiment using first-year college students in economics and business classes by manipulating the ratio of female students in workgroups. From the experiment, they observe that boys' dropout ratios are delayed when the ratio increased. Contrarily, boys underperformed on mathematically intensive courses when they were in a workgroup with a higher percentage of girls. Nonetheless, Oosterbeek and van Ewijk (2010) do not find statistically significant gender peer effects.

Although all of the above studies find positive spillover effects of girls' ratios, in general, one cannot generalize the results of these studies to the question of whether "segregating" boys from girls (and vice versa) are beneficial for students' academic achievement compared to coeducational environments. Since the amendments of Title IX (prompted by the No Child Left Behind Act) have allowed school districts to offer single-sex classes or schools, however, estimating the effect of gender segregation is of great importance to current policy practice (Cable and Spradlin, 2008). In addition, as some studies find that boys' scores tend to decrease as they are surrounded by more girls, it is desirable to investigate whether boys will benefit from being in boys-only classes or schools.

In fact, Jackson (2012) estimated the effect of attending single-sex schools using the data from Trinidad and Tobago by exploiting the rule-based assignments used to assign students to either same- or mixed-sex schools. By simulating the rule used by the Ministry of Education, he extracts exogenous variations in school attendance and finds little evidence of single-sex school advantages. Moreover, two studies in a recent psychology literature conducted experiments to test the effectiveness of single-sex schooling. First, Inzlicht and Ben-Zeev (2000) experimented using 70 to 90 undergraduate students at Brown University to test a hypothesis that female students experience a so-called "stereotype threat" when they were outnumbered by male students, and thereby create a performance deficit of female students. The result of the experiment shows that female students' performance in math tests are raised under the same-sex setting. Second, using 401 students in eighth grade, Kessels and Hannover (2008) tested whether girls' self-concept of ability with respect to physics is promoted under girls-only classes. Their results confirm that this is, indeed, the case. Finally, Eisenkopf et al.'s (2011) study also estimated the effect of same-sex education by making use of a couple of hundred female students who are randomly assigned to either mixed- or same-sex classes within a particular high school in Switzerland.[3] According to their analysis, math scores of female students were improved when these students were assigned to female-only classes.

Even though the three studies mentioned above used randomized experiments to estimate the effect of gender segregation, there exists several limitations. To begin with, the results are restricted to the impacts of female-only classes and we cannot infer from these studies whether male-only schooling is beneficial for male

[3] The high school used in their study is catered specifically to students who intend to attend a college of education upon graduation.

students. Second, since their studies rely on a small number of observations, the external validity of those results is limited. Besides, all of the research have been conducted using particular sets of students, and accordingly, it is somewhat hard to generalize their results to students in general. Next, to analyze the effect of gender segregation on students' performance at a secondary or a higher education level, gender segregation implemented at the school level or at the grade level is a more pertinent way of measuring the treatment since students at these levels spend less amount of time in their classes (Lavy and Schlosser, 2011). Lastly, it is natural to think that the effect of gender segregation is heterogeneous. That is, the treatment may have differential effects on students' academic achievement depending on students' performance, and one cannot capture "distributional" impact of gender segregation on students' performance from these studies.

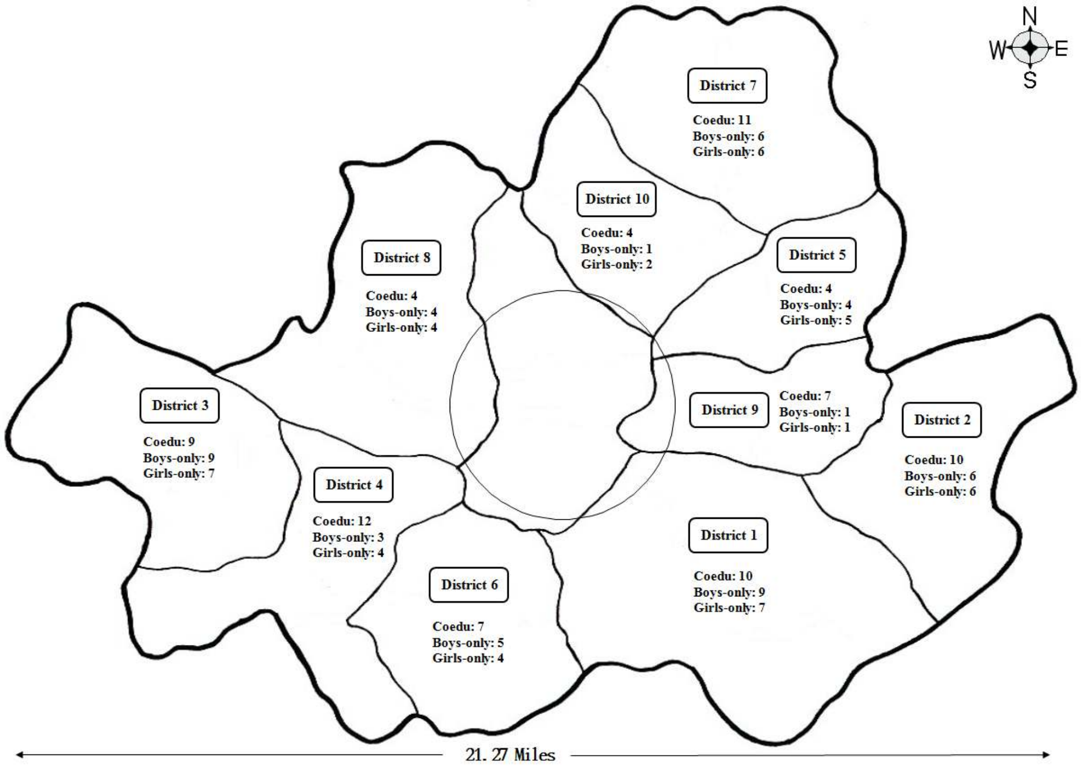
1.3 Random Assignment Mechanism

The average age at which a child enters the education system in South Korea is six. The school year for all grades begins at the beginning of March. Currently, six years of elementary school and three years of middle school are compulsory for all children in South Korea. Upon completing nine years of compulsory education, students enter three-year high school. In Korea, there are mainly three types of high schools. Special purpose high schools serve students who intend to major in arts, music, or physical education. These schools also specialize in science or foreign languages. The second type is vocational high schools. Students who graduate from vocational high schools normally enter the job market right after their graduation or go on to two-year vocational colleges. The third type is called the general high school and the majority of middle school graduates enter high schools of this type. Students are *not* randomly assigned to the first two types of high schools. Hence, in this paper, I will use data pertaining to students attending general high schools. I also note that random assignment of students was conducted in Seoul until 2009. From 2010, the Office of Education changed its policy and students are no longer randomly assigned to general high schools.

An assignment to the general high school is conducted as follows.[4] First, when students graduate from middle school, the Office of Education assigns each student a high school district based on their residence. In Figure 1.1, I show a number of school districts in Seoul for 2009. I also show a number of coeducational, boys-only, and girls-only schools in each district, and as can be seen from the figure, three kinds of schools exist in every school district. Next, within the assigned school district, each student is classified into one of three blocks based on their middle school graduate standing percentile rank. I show, using Figure 1.2, how students are randomly assigned to high schools based on these three blocks. The first block includes students whose middle school graduate standing percentile rank is between 0.001 to 9.999 (upper-ranked). The second block contains students with the graduate standing percentile rank within 10.000 to 49.999 (middle-ranked). Lastly, students with percentile rank 50.000 or above (lower-ranked) are placed in the third block. Next, the Office of Education randomly assigns students in each block to schools using a computer-assisted lottery system. (See Figure 1.2 for the case with four

[4] All information are based on the Office of Education's annual official documents.

Figure 1.1: School Districts in Seoul (2009)

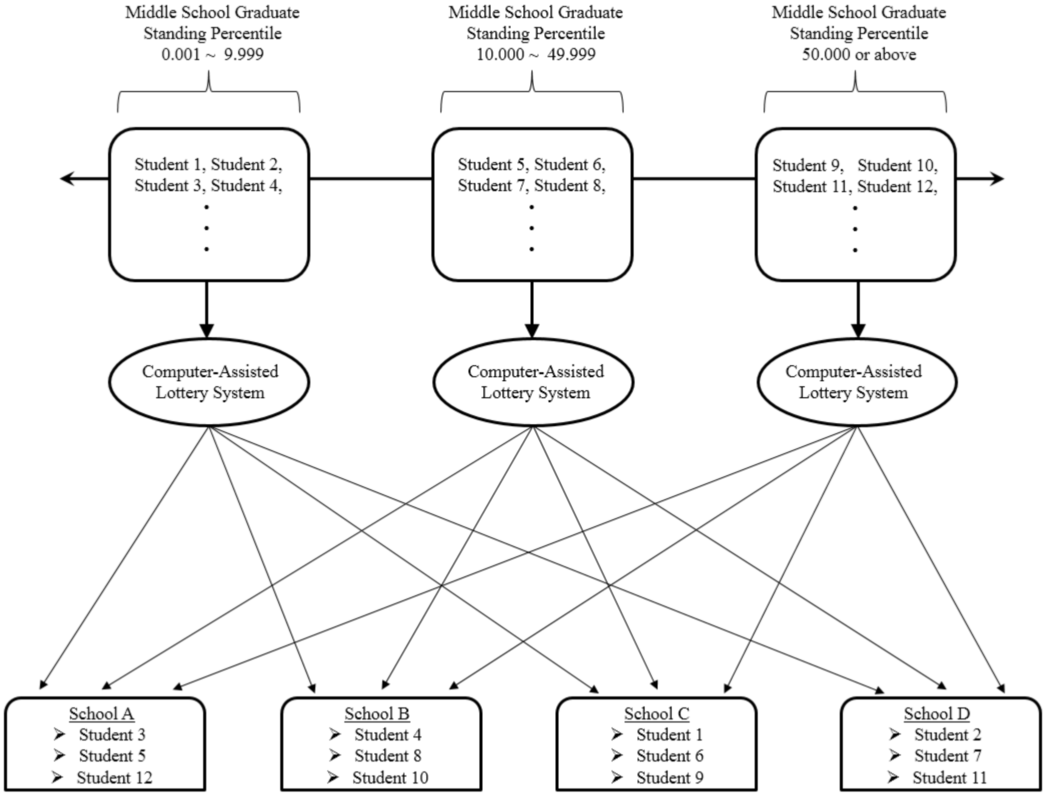


schools in a school district).

Hence, the Office of Education uses randomized block designs to assign students to high schools. By implementing the randomized block design-type assignment, the Office of Education ascertains that within the school district, the ratio of upper-, middle-, and lower-ranked students is equally distributed among schools. Using the middle school graduate standing percentile rank for dividing blocks is favorable for equalizing the average performance level of students across high schools because in Korea, almost all students are randomly assigned to middle schools after completing elementary school, and as a result, the academic level of students in middle schools are highly homogeneous across schools.

This kind of random assignment in Seoul has been adopted since 1974 and reasons for adopting the random assignment policy are twofold. First, prior to 1974, parents in Korea suffered from paying the high costs of private education for their children. At that time, admissions to high schools were determined by the high school entrance exam. Accordingly, in order to admit their children to prestigious

Figure 1.2: Random Assignment Mechanism



high schools in Seoul, parents desperately relied on private tutoring so that their children may earn high scores in the entrance exam. Second, as a result of soaring competition among middle school students, a formal middle school education system was no longer viable as teachers in middle school focused on preparing their students solely for the entrance exam rather than "educating" their children. To resolve these problems within the education system, the Korean government decided to change the high school admission system from the test-based system to a no-test-based system. Furthermore, because the above problems were more serious in the capital city, the Seoul Metropolitan Office of Education complemented the no-test-based system with random assignment of students to high schools within the school district (Ministry of Education, 1998).[5]

[5] Compared to Seoul, other cities allow students to list several schools.

1.4 Data

To draw a causal estimate of gender segregation on student performance, I use several sources of data. For student-level data, I use administrative records of the College Scholastic Aptitude Test (CSAT) and the National Assessment of Educational Achievement (NAEA) maintained by the Ministry of Education, Science and Technology in Korea. For school-level data, I make use of the data that are publicly available at the School Information Website (SIW)[6]. Furthermore, I use school-level information that I personally obtained from both the Ministry of Education and the Seoul Metropolitan Office of Education. Lastly, I use the Statistical Yearbook of Seoul Education.

1.4.1 CSAT and NAEA

The CSAT is similar to the SAT in the U.S. Unlike the SAT, however, the CSAT is a high-stakes test that most students in Korea can take only once at the end of their *third year* of high school. For example, students who entered high school in 2009 took the CSAT at the end of 2011. The CSAT held in 2011 is called CSAT *2012*, rather than CSAT 2011. Although figures vary by year and university, CSAT scores typically determine 50 to 100% of college admissions. Because of its importance in determining students' futures, and since the test begins at 8:00 A.M., every firm and government office delays commuting hour to 9:00 A.M. to obviate rush hours so that students are not late for the exam. Moreover, middle school students and students in the first and second grade of high school are given the day off because both middle schools and high schools are used as testing centers during the exam day. The test runs from 8:00 A.M. to 6:00 P.M., and students are, in general, tested on five sections; reading, English, mathematics, social studies, and science studies. The CSAT has been conducted since 1993, and with an exception of the test held in 1993, it is conducted only once per year.

The NAEA, on the other hand, is analogous to the National Assessment of Educational Progress in the U.S. The purpose of the NAEA is to assess whether each student in all levels of education is keeping up with the curriculum. The test has been conducted annually since 2008. The NAEA data is available from

[6] www.schoolinfo.go.kr. The website is operated by the Ministry of Education, Science and Technology.

2009, and in 2009, first year high school students were tested on basic knowledge in reading, mathematics, science, social studies, and English. In 2010, the same students (who became second grade high school students) were again tested on reading, mathematics, and English. From 2010 onwards, second year high school students are tested on these three subjects.

Administrative records of these two tests have not been made public until 2009 due to concerns that the results might reveal educational gaps among schools and promote a sense of incongruity. However, after several years of administrative litigations filed by the members of National Assembly, researchers, and parents, the Supreme Court ruled in favor of the disclosure and ordered the government to publicize the test scores solely for the purpose of scientific research. As a consequence, the government has allowed researchers to apply for the data starting from 2010.

In order to obtain the data, one has to first submit research plans to the Ministry of Education, and by following necessary administrative procedures, I received the data for CSAT from 2002 to 2004 and CSAT from 2009 to 2012. For the NAEA, I obtained the data for 2009 and 2010. In Appendix (Section 1.9.1), I present brief explanations on administrative procedures required for obtaining the data.

There are four reasons why I am using these particular sets of data. First, note that for the school-level variables, the CSAT contains only the name of the city in which the school is located and the name of the schools. On the other hand, the student-level variables include gender, diploma types, majors, scores on each subject, and names of the test districts in which the students took the test. As a result, I do not have data on school types. Because of the presence of school names, however, I was able to impute school types by consulting the Statistical Yearbook of Seoul Education. For instance, when imputing the school types for schools in CSAT 2002, I make use of the Statistical Yearbook of 1999 because students who took CSAT 2002 have been randomly assigned to high schools in 1999.[7] Hence, I have to use the Yearbook that is three years before the CSAT year. The problem with using the information on school types in the Yearbook, however, is that prior to Yearbook 1999, some schools' types are miscoded.[8] Using miscoded school types creates serious threat to the validity of the analysis. Fortunately, starting from the

[7] Note that CSAT 2002 is held at the end of 2001.

[8] For example, I verified that in the Yearbook of 1996, school types of five schools are miscoded. Note, furthermore, that this does not necessarily imply that other schools are correctly specified.

Yearbook of 1999, two variables are reported in the Yearbook; school types and the ratio of female students. As a consequence, I was able to double check the type of schools by looking at the ratio of female students.[9] Therefore, I do not use data prior to CSAT 2002.

Secondly, test formats are identical for CSAT 2002 to 2004. During CSAT 2005 to 2008, however, the format was changed and for CSAT 2008, only the discrete rank has been reported for each subject. Accordingly, I did not apply for CSAT 2005 to CSAT 2008. Thirdly, rather than using consecutive CSATs, I decided to use the CSATs that were taken several years after CSAT 2004. Accordingly, since the format of the CSAT 2009 to 2012 is the same, I applied for the second CSAT dataset for these periods.

Finally, I am using NAEA data for testing the sensitivity of the estimates obtained from CSAT datasets. The advantage of using NAEA data is that since NAEA 2009 assessed students in the first grade and NAEA 2010 tested the same students who went on to second grade, I am able to observe how the impact of gender segregation changes by the amount of time students are segregated.[10]

To investigate the treatment effect, I generated two variables and merged them to the datasets. The first variable is the school type of each school, and since the way I created the variable is mentioned above, I do not repeat here. The second variable is the school district indicator for each school. The CSAT data has information on the test districts of each student. For most students, test districts are equivalent to school districts. However, some students take the CSAT in districts outside of their school districts.[11] Moreover, even if the test districts match their school districts, one cannot use test districts for school districts because the name of the test district in CSAT data corresponds to the school in CSAT year minus one. To reiterate, since students are randomly assigned three years before the CSAT year, it is essential to use the school district three years before the CSAT year because some schools may

[9] A ratio of female students in mixed-sex schools should be greater than 0. Contrarily, the ratio should be equal to 0 for boys-only schools and 100% for girls-only schools. Using this strategy, I verified that the school type of only one school is miscoded in the Yearbook between 1999 to 2009.

[10] NAEA 2009 was held in October of 2009 and at the time of this assessment, students had spent about 8 months in either single- or mixed-sex schools. On the other hand, NAEA 2010 was held in July of 2010. Hence by the time NAEA 2010 was taken, students had spent about 17 months in either single- or mixed-sex schools.

[11] I verified this fact by matching the test districts of students and school districts of schools after imputing the school district for each school.

have moved to other school districts during the three-year period. Hence using the test districts as the school districts is inappropriate because the name of the test districts in the CSAT data might not equal to the name of the school districts. Fortunately, although the Yearbook does not have information on school districts, it contains an address of each school and by using the address, I was able to impute school districts for each school as school districts are determined by the address of each school.[12]

In order to draw a "causal" inference, I also make step-by-step restrictions to the initial sample for each of the CSAT and NAEA datasets. This is because students are randomly assigned to a partial set of schools that are located in Seoul only. In Appendix (Section 1.9.2), I provide, in detail, a step-by-step explanation on sample restrictions.

1.4.2 SIW and Other Data

Another set of data that I use for testing the validity of the randomization is administrative records of school-level data stored in the SIW. The website was made in 2008 as part of the "Act on Special Cases Concerning the Disclosure of Information by Education-Related Institutions," and it includes a rich set of information on school-level data such as the state of students, teachers, school activities, school conditions, and budget and account for every school in Korea. For the most part, the data covers 2007 to 2011. In this study, I collected data, by grade level, on the number of students in each school, number of students who transferred to other schools, number of students who dropped out from high school, and the ratio of students that received free lunch.

Moreover, I obtained information from the Office of Education on the number of students who were supported by the government or third parties during their *first year* of high school (from 2006 to 2009).[13] These include students who are supported with tuition reductions or in the form of fellowships. In order to qualify for financial support, students should be from low-income families and/or be protected by the

[12] For example, for CSAT 2002, I imputed school districts for each school by retrieving the address of each school specified in the Yearbook of 1999 and matching it with the corresponding school districts.

[13] It is important to use the number of *first-year* students because students are randomly assigned to high schools during the first year.

law such as “National Basic Living Security Act.” Furthermore, for each school, the Office of Education further provided me with data on the ratio of students in each of the blocks mentioned in Section 1.3 for year 2009.[14]

Finally, I resort to the Statistical Yearbook of Seoul Education, and these Yearbooks contain school-level information such as addresses of schools, school types, number of classes by grade, number of students by grade, number of female students by grade, and number of teachers.

[14] I would like to express my gratitude to the public officials at the Office of Education for providing me with the necessary data.

1.5 Statistical Framework

Let T_i be a binary variable equal to 1 if student i is assigned to a same-sex school and 0 if assigned to a mixed-sex school. I denote the test scores (Y_i) for student i in same-sex schools as $Y_i(1)$, and $Y_i(0)$ for student i in mixed-sex schools. In a canonical Rubin causal model framework under the "Fundamental Problem of Causal Inference" (Rubin, 1974; Holland, 1986), the average treatment effect on the treated, $\hat{\tau}$, is

$$\begin{aligned}\hat{\tau} &= E\left[Y_i|T_i=1\right]-E\left[Y_i|T_i=0\right]\\ &= E\left[Y_i(1)|T_i=1\right]-E\left[Y_i(0)|T_i=1\right]+E\left[Y_i(0)|T_i=1\right]-E\left[Y_i(0)|T_i=0\right], \quad (1)\end{aligned}$$

where the last two terms in Equation (1) constitute a selection bias. The bias term is the difference in the expected values of "$Y_i(0)$" between those in same-sex schools and those in mixed-sex schools. To give an example how this bias may affect the treatment effect, suppose an educationally motivated parents may choose to make their children attend same-sex schools believing that the coeducational high school setting would harm their children's academic achievement. Then it is likely that those who attend single-sex schools have higher values of $Y_i(0)$, and it would make a direction of the bias to be positive and the resulting estimate would overstate the treatment effect. The point is that in order to obtain a reliable estimate of the treatment effect, this endogeneity issue should be embodied in the econometric framework.

Because of the random assignment of students to either same- or mixed-sex high schools, however, T_i is independent of potential outcomes, and accordingly, the bias term in Equation (1) fades away and we can estimate the treatment effect using a standard regression framework. Thus, for each CSAT and NAEA year, I estimate the following two specifications:

$$y_{isd}=\begin{cases}\alpha+\beta_1 B_{sd}+\gamma_d+\varepsilon_{isd}, & \text{if student } i \text{ is a male student}\\ \alpha+\beta_1 G_{sd}+\gamma_d+\delta_{isd}, & \text{if student } i \text{ is a female student,}\end{cases}$$

where y_{isd} is the percentile rank in CSAT or NAEA for student i in school s and in school district d. Since the randomization has been conducted within the school district, I include school district fixed effects γ_d where $d \in \{1, 2, ..., 10\}$.[15] Moreover, note that I am comparing boys in boys-only schools with boys in mixed-sex schools, and girls in girls-only schools with girls in mixed-sex schools. Thus, the treatment indicator B_{sd} is equal to 1 if male students attend boys-only schools and 0 if male students attend mixed-sex schools. On the other hand, G_{sd} is equal to 1 if female students attend girls-only schools and 0 if female students attend mixed-sex schools. Lastly, ε_{isd} and δ_{isd} are the error terms.

I first run the regression separately for each year to observe the sensitivity of the treatment effect. After that, I pool seven years of CSAT data and run the following regression:

$$y_{isdc} = \begin{cases} \alpha + \beta_1 B_{sdc} + \xi_c \times \gamma_d + \varepsilon_{isdc}, & \text{if student } i \text{ is a male student} \\ \alpha + \beta_1 G_{sdc} + \xi_c \times \gamma_d + \delta_{isdc}, & \text{if student } i \text{ is a female student,} \end{cases}$$

where subscript c indicates CSAT year. In the regression, I include district-by-year fixed effects by interacting the school district fixed effect γ_d with CSAT year dummies ξ_c where $c \in \{2002, 2003, 2004, 2009, 2010, 2011, 2012\}$. This produces 70 district-by-year fixed effects in the pooled regression.

Note that the standard linear regression is summarizing the average relationship between the dependent variable and a treatment based on the conditional mean function. This provides only a partial view of the relationship between the outcome variable y and the regressor T (i.e., mean-effects of T on y). Estimating the mean impact, however, may miss the heterogeneous effect that the treatment has on students. In the context of this paper, the effect of gender segregation may be different for high-, middle-, or low-performing students. Therefore, I estimate heterogeneous treatment effects by estimating the effect of treatment across the distributions of students' percentile ranks in CSAT. To estimate the effect, I implement the quantile regression method first developed by Koenker and Bassett (1978) and run the following using the pooled data:

$$y_{isdc}(q) = \begin{cases} \alpha + \beta_1(q) B_{sdc} + \xi_c \times \gamma_d + \varepsilon_{isdc}, & \text{if student } i \text{ is a male student} \\ \alpha + \beta_1(q) G_{sdc} + \xi_c \times \gamma_d + \delta_{isdc}, & \text{if student } i \text{ is a female student,} \end{cases}$$

[15] For CSAT 2002 to 2004, $d \in \{1, 2, ..., 9\}$.

where $q \in (0,1)$ denotes a quantile, and $y_{isdc}(q)$ refers to the outcome in q-th quantile. Hence, by running the quantile regression, we can retrieve different values of the treatment effect β_1 by choosing the different values of quantile q. Note, however, that in the current setting, the assignment to the single-sex school is exogenously determined, conditional on school districts (i.e., the selection on observables assumption). In this instance, the classical quantile regression estimator proposed by Koenker and Bassett (1978) cannot be used. Instead, I make use of an estimator suggested by Firpo (2007).[16]

Another statistical issue should be addressed in this study. As pointed out by Moulton (1986), failing to account for the group structure of the data would understate standard errors. In the current setting, since students in the same school are likely to be correlated for they are subject to same academic environment such as in teacher quality, it is necessary to cluster standard errors by schools. Hence, I report robust standard errors clustered by school names.

Relative to existing research, this study carries several advantages. First, in Seoul, most middle school students are randomly assigned to high schools after they graduate from middle school. Fortunately, the share of boys-only, girls-only, and mixed-sex high schools in Seoul is the same. Because of the variations in the number of school types, the setting in Seoul is suitable for estimating the effect of gender segregation on students' academic performance. Note that since boys are randomly assigned to either boys-only or coeducational schools, we would also be able to learn whether boys benefit from being isolated from female students.

Second, this study uses 7 years of CSAT data and 2 years of NAEA data, and although the number of observations varies depending on which test scores are analyzed, the final pooled data used in the estimation process comprise more than 200,000 observations. As a consequence of this large-scale experiment, the study has sufficient statistical power and the results of the study, I believe, possess a high degree of external validity. Third, high schools in Seoul are remarkably homogeneous in terms of school-level variables purported to affect students' academic achievement. For example, teacher salaries are all controlled by the government, and the salary level is solely determined by a pay step based on the length of one's service. Hence, teacher quality is highly homogeneous across schools. Furthermore,

[16] For the estimation, I benefited from the Stata command developed by Frölich and Melly (2010).

every school in Seoul faces the same governmental regulations and all students learn similar curriculums. Accordingly, other than school types, school-level characteristics are similar, in important aspects, among school types.

Fourth, many of the prior studies pay little attention to attrition. As pointed out in many of the economics of education literature (e.g. Krueger, 1999), it is necessary to check whether the data suffer from sample attrition, and I show that the attrition problem is minimal in this paper. In addition, experimental studies in previous research lack consideration on non-conforming behaviors, and I present, in this paper, that the study does not suffer from these behaviors. Finally, using the quantile regression method, I estimate heterogeneous treatment effects based on the percentile ranking of students' performance. The quantile regression estimates will give a picture of how gender segregation affects students' academic achievement differently, and compare to only estimating the mean effects, this paper gives more policy implications.

1.6 Validity of Randomization

In order to extract a valid identification from the randomized experiment, several conditions should be confirmed before analyzing the data. In this section, I address the following three conditions: A) balance in predetermined covariates, B) non-conforming behaviors, and C) attrition problems. Since CSAT and NAEA datasets consist of few student-level variables, I make use of administrative records on school-level variables which can be retrieved from the SIW. Because some of the school-level variables such as the ratio of high school dropouts are derived from the student-level data, I can use these school-level data to check the validity of the randomization. In testing the validity of the randomization, I also make use of the information that I personally obtained from the Ministry of Education and the Office of Education.

1.6.1 Balance in Predetermined Covariates

Suppose students in same-sex schools performed higher than those in mixed-sex schools. Note that, although students are randomly assigned to high schools within a school district by a computer-assisted lottery system, students with backgrounds that are favorable to one's academic achievement may have been assigned to, say, a same-sex school by chance variations. If this is the case, one cannot conclude that same-sex schooling is beneficial for students' academic achievement even though students are randomly assigned to high schools. Hence, in this subsection, I test balance in predetermined covariates that are purported to positively affect one's academic achievement.

According to the "Education Production" function,

$$A_{it} = f(A_{it-1}, F_{it}, P_t, S_{it}, T_{it}, \alpha_i, \xi_{it}),$$

factors that affect one's academic performance (A_{it}) are students' prior academic achievement (A_{it-1}), family background (F_{it}), peer (P_t), school resources (S_t), teacher quality (T_t), innate ability (α_i), and noise terms (ξ_{it}). To test the balance in "observable" covariates among aforementioned parameters, and because of the data availability, I first use a ratio of students in each block (in Figure 1.2) to test whether A_{it-1} is equally balanced across schools. Moreover, I use a ratio of students who are financially supported by the government (via tuition reductions or fellowships) and students in free-lunch status to test for the balance in F_{it}. These variables are good proxies for students' family background because the eligibility for receiving above support is determined by family income. For S_t, I use class size as well as student-teacher ratio. With respect to T_t, I argue that average teacher quality of schools in Seoul is homogeneous across schools because the salary level for all teachers in Korea is set by the Ministry of Education and is entirely determined by the pay-step based on the length of service.

In Table 1.1, I present tests of within-district balance in the ratio of students in each block. The Office of Education no longer holds the data prior to 2009, and consequently, I was only able to obtain the data for the academic year 2009. Since this study compares boys in boys-only schools with boys in mixed-sex schools, it is necessary to use the ratio where the numerator is the number of boys in each block and the denominator is the number of total boys in each school. Likewise, we need to use the corresponding ratio for girls. Fortunately, I was able to receive the data on the number of students in each block by gender, and accordingly, I can precisely analyze whether the ratio of high-, middle-, and low-achieving students are equally balanced by gender. In order to analyze the balance in the ratio, I calculated the ratio of upper-, middle-, and lower-ranked boys by dividing the number of boys in each category by the total number of boys in each school, and similar for girls. In Table 1.1, I run a regression of each ratio on a dummy variable indicating the school type conditional on school districts. In Panel A, I compare the ratios in boys-only schools with that of mixed-sex schools. As can be expected from the randomized block design-type assignment procedures described in Figure 1.2, the ratio of upper-, middle-, and lower-ranked students are almost identical between boys-only schools and coeducational schools. Panel B presents the corresponding results for girls. Again, the difference in the ratio of students in each category is almost zero. Therefore, I conclude that differences do not exist in students' previous academic achievement among school types.

I underscore the importance of checking the baseline covariates in a randomized control trial by presenting the ratio of students in each block for high schools in which students are not randomly assigned to high schools (i.e., schools located in the circled area in Figure 1.1). In Panel C of Table 1.1, I present the analysis that compare the ratio of girls-only schools with that of mixed-sex schools. Because students "apply" for the schools in this circled area, the ratio of students in each category differs

Table 1.1: Tests of Within-District Balance in Baseline Graduate Standing Percentile

Dependent Variable (Ratio)	School Type	2009	S.E.
	A. Boys in Boys-Only School vs Boys in Mixed-Sex Schools		
Upper-ranked students	Boys-only school (1 = yes)	−0.001	(0.000)
($\Psi < 10$)	Constant (= Ratio in Mixed-sex)	0.064	(0.000)
	Number of schools	126	
Middle-ranked students	Boys-only school (1 = yes)	0.000	(0.001)
($10 \leq \Psi < 50$)	Constant (= Ratio in Mixed-sex)	0.414	(0.000)
	Number of schools	126	
Lower-ranked students	Boys-only school (1 = yes)	0.000	(0.002)
($\Psi \geq 50$)	Constant (= Ratio in Mixed-sex)	0.521	(0.001)
	Number of schools	126	
	B. Girls in Girls-Only School vs Girls in Mixed-Sex Schools		
Upper-ranked students	Girls-only school (1 = yes)	−0.002	(0.000)
($\Psi < 10$)	Constant (= Ratio in Mixed-sex)	0.135	(0.000)
	Number of schools	123	
Middle-ranked students	Girls-only school (1 = yes)	0.000	(0.000)
($10 \leq \Psi < 50$)	Constant (= Ratio in Mixed-sex)	0.501	(0.000)
	Number of schools	123	
Lower-ranked students	Girls-only school (1 = yes)	0.002	(0.001)
($\Psi \geq 50$)	Constant (= Ratio in Mixed-sex)	0.363	(0.000)
	Number of schools	123	
	C. Non-Randomly Assigned High Schools		
Upper-ranked students	Girls-only school (1 = yes)	−0.059	(0.023)
($\Psi < 10$)	Constant (= Ratio in Mixed-sex)	0.188	(0.019)
	Number of schools	19	
Middle-ranked students	Girls-only school (1 = yes)	−0.073	(0.032)
($10 \leq \Psi < 50$)	Constant (= Ratio in Mixed-sex)	0.542	(0.028)
	Number of schools	19	
Lower-ranked students	Girls-only school (1 = yes)	0.132	(0.049)
($\Psi \geq 50$)	Constant (= Ratio in Mixed-sex)	0.269	(0.042)
	Number of schools	19	

Note: I estimate all coefficients from running a regression of each ratio on a dummy variable indicating the school type conditional on school districts. Standard errors in parentheses. Ψ denotes middle school graduate standing percentile rank. Smaller the number, the higher the ranking.

significantly between girls-only schools and mixed-sex schools. For example, the ratio of lower-ranked students in girls-only schools is 0.132 higher compared with coeducational schools. Hence, this highlights that when students are not randomly assigned, then baseline characteristics may differ significantly among schools.

In Table 1.2, I test balance in baseline covariates for F_{it}. The data used for checking the balance are 2007, 2008, and 2009. Data prior to 2007 were not available from both the SIW and the Office of Education. First variable, "Financially supported", corresponds to the ratio of students who are financially supported by the government in the form of tuition waiver or fellowships. I use first-year high school students for calculating the ratio of financially supported students in each school. As can be seen from the table, none of the coefficients show significance with respect to school-type dummies. For the overall test of significance, I run a joint-test of the hypothesis that the school-type indicators jointly had no effect. P-values from the F-test are 0.39, 0.91, and 0.66 implying that none of the hypotheses can be rejected at the statistically significant level.

Next variable used for checking the balance is the ratio of students receiving free lunch. Although it is necessary to use the number of "first-year" high school students in calculating the ratio, the data are only available for the total number of students in each school, and accordingly, I calculated this ratio by the number of students receiving free lunch divided by the total number of students in each school. Note that the data for the ratio of students receiving free lunch were not available for 2007. Again, the ratio is not statistically different among same-sex and mixed-sex schools.

Note that in order to precisely test the balance in F_{it}, however, it is necessary to use the ratios calculated separately for boys and girls. However, the ratios of the above two variables are not separately available by gender. Thus, I admit that the analysis presented here have some limitations.

To test the balance in school resources, I use class size and the student-teacher ratio. Although some coefficients are statistically significant at the 5% level, the magnitude of the coefficient is negligible. For example, in 2007, the average class size in coeducational schools is 37.971 and the size differs by less than one student for same-sex schools (all coefficients are less than one student). Accordingly, class size is almost identical among school types. On the other hand, in 2007, the average number of students per teacher in coeducational schools is 16.498. As with class

Table 1.2: Tests of Within-District Balance in Baseline Covariates

		Year		
Dependent Variable	School Type	2007	2008	2009
Financially supported	Boys-only school (1 = yes)	0.004	−0.000	−0.012
		(0.010)	(0.010)	(0.014)
	Girls-only school (1 = yes)	−0.010	0.003	−0.001
		(0.010)	(0.010)	(0.015)
	Constant (= coeducational)	0.100	0.117	0.163
		(0.006)	(0.006)	(0.009)
	F-test	0.391	0.910	0.660
	Number of schools	162	164	166
Receiving free lunch	Boys-only school (1 = yes)		−0.010	0.002
			(0.006)	(0.010)
	Girls-only school (1 = yes)		−0.002	0.000
			(0.007)	(0.010)
	Constant (= coeducational)		0.078	0.099
			(0.004)	(0.006)
	F-test		0.317	0.962
	Number of schools		163	168
Average class size	Boys-only school (1 = yes)	−0.066	0.218	0.590
		(0.186)	(0.229)	(0.248)
	Girls-only school (1 = yes)	0.240	−0.318	−0.808
		(0.190)	(0.233)	(0.252)
	Constant (= coeducational)	37.971	37.932	37.055
		(0.118)	(0.145)	(0.154)
	F-test	0.280	0.098	0.000
	Number of schools	164	166	172
Student-teacher ratio	Boys-only school (1 = yes)	0.660	0.654	1.029
		(0.238)	(0.219)	(0.242)
	Girls-only school (1 = yes)	0.705	0.548	0.732
		(0.243)	(0.222)	(0.246)
	Constant (= coeducational)	16.498	16.996	17.026
		(0.151)	(0.139)	(0.150)
	F-test	0.004	0.005	0.000
	Number of schools	164	166	172

Note: I estimate all coefficients from running a regression of each dependent variable on a dummy variable indicating the school type conditional on school districts. Standard errors in parentheses. Numbers in F-test are the p-values retrieved from joint-test of the hypothesis that the school-type indicators jointly had no effect. There are two to six missing values for the “Financially supported”, and one and four missing values for “Receiving free lunch”.

size, the difference in the student-teacher ratio between same- and mixed-sex schools is less than one student indicating that the student-teacher ratio is not different among school types. Indeed, the balance in the class size and the student-teacher ratio certainly holds because these two variables are strictly enforced by the Office of Education every year. It is obvious that if these are not balanced across schools, parents would not abide by the random policy adopted by the Office.

All in all, estimates in Table 1.2 show that school resources and family background are equally balanced across schools.

1.6.2 Non-Conforming Behavior

In a randomized control trial, if subjects do not conform to their initially assigned treatment, it creates a threat to the validity of the analysis. In this case, the validity of the analysis can be revitalized by using the initial assignment as an instrumental variable (IV), and the IV regression would be estimating the intent to treat effect (e.g. Howell et al., 2002). Unfortunately, I lack data on students' initial assignments. However, based on several facts below, I argue that this study does not suffer from non-conforming behaviors.

There are several reasons why parents in Korea have very few incentives to not conform to their children's initially assigned high schools. First, the Office of Education does not accept the request for reassignments. The assignment is a once-for-all procedure in Korea. This is clearly stated in the webpage of the Seoul Metropolitan Office of Education.[17] Second, by Article 84 (8) of the Education Act, if a student does not enroll in the high school that one has been assigned to, then that student would not be assigned another school in that academic year. Third, Article 86 of the Education Act lists the case in which the reassignment procedures can be conducted, and according to this article, parents cannot request a reassignment just because they do not like the initially assigned high school. Fourth, Article 89 (2) of the Education Act states that when all family members of a student has moved to another school district *after the initial assignment has taken place*, then the parents can request a reassignment. Even if this is the case, however, parents *cannot* choose the school in the district to which they have moved. That is, the student has to go through another computer-assisted lottery assignment procedure and is randomly

[17] english.sen.go.kr.

Table 1.3: Tests of Within-District Differences in Transferred Students

	During First Year			During Second Year		
Explanatory Variable	2007	2008	2009	2007	2008	2009
Boys-only school (1 = yes)	−0.002	−0.000	−0.002	0.000	0.000	0.000
	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Girls-only school (1 = yes)	−0.005	−0.003	−0.005	0.001	−0.003	−0.003
	(0.003)	(0.002)	(0.002)	(0.001)	0.001	(0.001)
Constant (= coeducational)	0.037	0.040	0.039	0.012	0.015	0.016
	(0.001)	(0.001)	(0.001)	(0.001)	0.000	(0.001)
School district fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of school districts	10	10	10	10	10	10
Number of schools	164	166	166	162	164	164
F-test	0.227	0.281	0.143	0.588	0.013	0.146

Note: I estimate all coefficients from running a regression of a ratio of transferred students (during first or second year of high school) in each school on dummy variables indicating the school type conditional on school districts. Standard errors in parentheses. For the last row, I performed an F-test and tested the hypothesis that the school-type indicators jointly had no effect. The numbers in the last row are p-values for joint F-tests. For "During First Year", there are six missing values. For "During Second Year", there are two to eight missing values.

assigned a high school in the new school district. Accordingly, educationally motivated parents have few incentives to move to another school district just to get a "random" reassignment opportunity.

Lastly, one might argue that some parents may engage in lobbying the officials at the Office of Education before the lottery assignment takes place. However, in order to prevent corruptive behaviors during this procedure, the lottery assignment is conducted by each school district's High School Admission Lottery Management Committee which consists of parents, principals, assistant principals of high schools, and commissioners of local education offices. Hence, it is hard to imagine that parents would lobby the administrators and engage in manipulation of lottery assignment.

Table 1.3 shows tests of within-district differences among boys-only, girls-only, and mixed-sex schools.[18] I first run an OLS by regressing a ratio of transferred students on dummies indicating school types, and since the randomization is conducted within the school districts, all regressions include school district fixed effects. During these periods, all coefficients for school type dummies are around zero and most of the coefficients are statistically insignificant implying that the ratio of transferred students is similar across school types. Moreover, although not presented in the ta-

[18] Data retrieved from the SIW.

ble, the ratio of transferred students during the third year of high school is close to 0 for all school types. As an overall test, I conducted an F-test of the null hypothesis that an assignment to boys-only, girls-only, and mixed-sex schools has no effect on the ratio of transferred students during the first year or second year of high school. I report p-values in the last row, and with an exception of 2008 (during second year), none of the school type indicators exhibit a statistically significant relationship with respect to the ratio of transferred students. Even though the p-value is significant in 2008 for the second year, I argue that since the coefficients for boys- and girls-only dummies are 0.000 and -0.003, respectively, there are few differences in the ratio between school types.

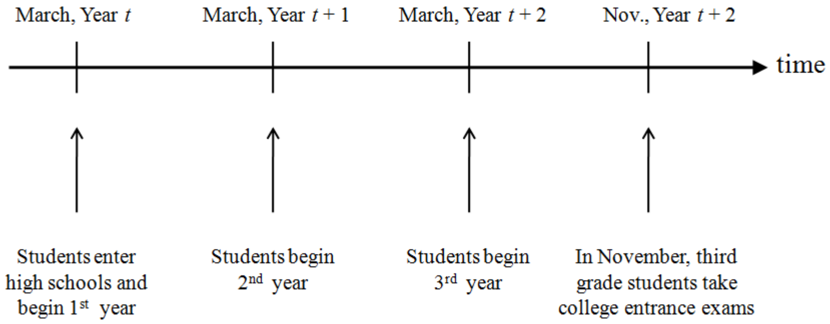
These numbers do not necessarily "prove" the fact that parents are conforming to their initially assigned high schools. However, note that on average, the ratio of transferred students during the first year of high school is around 3.5%. Besides, during the second and third year, the ratio is around 1.5% and 0%, respectively. Hence given the small ratio of transferred students across school types coupled with the reasons mentioned above, I argue that parents do conform to their initially assigned high schools.

1.6.3 Attrition

In Korea, students spend three years in high school. In Figure 1.3, I demonstrate a time frame of high school periods. Once they are randomly assigned to high schools, students begin their first year in March. In the following March, students start their second year of high school. After spending one more year, the third year begins again in March. Then during November of the third year, students take their college entrance exams.

Hence, students take the exam 30 months after entering high school. As a consequence, there is a probability that some students quit high school during this time frame. If students who were originally assigned to, say, coeducational high schools quitted the school had lower exam scores, on average, compared to those who were originally assigned to boys-only or girls-only high schools who also quitted the school, then the estimated treatment effect will be biased downwards. I illustrate an attrition bias mathematically and conditions in which one can bypass the problem of attrition, and further argue that the magnitude of the bias is negligible in this

Figure 1.3: Time Frame During High School Periods



study.

Let S and M denote an assignment to same- and mixed-sex schools, respectively. Note that student i is assigned to a treatment $T_i \in \{S, M\}$, and let $Y_i(S)$ and $Y_i(M)$ indicate test scores of student i assigned to same- and mixed-sex schools, respectively. The treatment effect we want to estimate is $\bar{\tau} = E\left[Y_i(S)\right] - E\left[Y_i(M)\right]$. Now, some students might drop out from high school and not take exams. As a consequence, for student $i \in \{S, M\}$, we observe

$$D_i = \begin{cases} 1, & \text{if observed (not attrit)} \\ 0, & \text{if missing (attrit).} \end{cases}$$

Furthermore, for student $i \in \{S, M\}$, we potentially have information on either of the following:

$$Y_i(S) = \begin{cases} Y_i^{obs}(S), & \text{if observed} \\ Y_i^{miss}(S), & \text{if missing} \end{cases} \qquad \text{or} \qquad Y_i(M) = \begin{cases} Y_i^{obs}(M), & \text{if observed} \\ Y_i^{miss}(M), & \text{if missing.} \end{cases}$$

In the presence of sample attrition, rather than estimating the $\bar{\tau} = E\left[Y_i(S)\right] - E\left[Y_i(M)\right]$, we end up estimating $\bar{\tau}^* = E[Y_i^{obs}(S)] - E[Y_i^{obs}(M)]$. Observe that for student $i \in S$,

$$E[Y_i(S)] = E\left[D_i|T_i = S\right] \times E[Y_i^{obs}(S)] + (1 - E\left[D_i|T_i = S\right]) \times E[Y_i^{miss}(S)]$$

$$\Rightarrow E[Y_i^{obs}(S)] = E[Y_i(S)] + \left(\frac{1 - E\left[D_i|T_i = S\right]}{E\left[D_i|T_i = S\right]}\right)\Big(E[Y_i(S)] - E[Y_i^{miss}(S)]\Big). \quad (2)$$

In the same manner, for student $i \in M$,

$$E[Y_i^{obs}(M)] = E[Y_i(M)] + \left(\frac{1 - E\left[D_i|T_i = M\right]}{E\left[D_i|T_i = M\right]}\right)\Big(E[Y_i(M)] - E[Y_i^{miss}(M)]\Big). \quad (3)$$

Then by Equation (2) and Equation (3), we have the following:

$$\begin{aligned}\bar{\tau}^* &= E[Y_i^{obs}(S)] - E[Y_i^{miss}(M)] \\ &= E[Y_i(S)] - E[Y_i(M)] + \underbrace{\left(\frac{1 - E\left[D_i|T_i = S\right]}{E\left[D_i|T_i = S\right]}\right)\Big(E[Y_i(S)] - E[Y_i^{miss}(S)]\Big)}_{\phi_S} \\ &\qquad - \underbrace{\left(\frac{1 - E\left[D_i|T_i = M\right]}{E\left[D_i|T_i = M\right]}\right)\Big(E[Y_i(M)] - E[Y_i^{miss}(M)]\Big)}_{\phi_M}. \quad (4)\end{aligned}$$

Thus, the last two terms (ϕ_S and ϕ_M) in Equation (4) correspond to the bias created by the attrition in single- and mixed-sex schools, respectively. From the equation, there are three cases in which the bias terms disappear. First case is when $E\left[D_i|T_i = M\right]$ or $E\left[D_i|T_i = S\right]$ is close to 1 (i.e., there are few attrition). In this case, the two fraction terms approach zero and the bias terms vanish. Second, when the observation is missing at random (i.e., $E[Y_i(S)] - E[Y_i^{miss}(S)] = 0$ and $E[Y_i(M)] - E[Y_i^{miss}(M)] = 0$), then again both bias collapse to zero. Finally, when there is a pattern in attrition but similar in expectation between single- and mixed-sex schools, then ϕ_S and ϕ_M that constitute the bias will jointly cancel out leaving no bias.

As the first case implies, we can ignore attrition problems when the attrition rate is close to zero. In the current setting, however, the attrition rate is not close to zero. To test the second case, one can check the baseline covariates of students who are assigned to single- and mixed-sex schools and determine whether $E[Y_i(S)] - E[Y_i^{miss}(S)] = 0$ or $E[Y_i(M)] - E[Y_i^{miss}(M)] = 0$. However, since I lack data on

Table 1.4: Mean Ratio of High School Dropouts by School Type

		School Grade			
Peer Students	School Type	First Year	Second Year	Third Year	Total
AY 2007	Mixed-sex school	0.025	0.013	0.004	0.042
		(0.015)	(0.006)	(0.004)	(0.018)
	Boys-only school	0.027	0.016	0.004	0.047
		(0.017)	(0.008)	(0.003)	(0.020)
	Girls-only school	0.019	0.010	0.003	0.032
		(0.014)	(0.005)	(0.003)	(0.014)
AY 2008	Mixed-sex school	0.020	0.017	0.004	0.041
		(0.009)	(0.008)	(0.003)	(0.014)
	Boys-only school	0.028	0.017	0.003	0.048
		(0.013)	(0.004)	(0.002)	(0.016)
	Girls-only school	0.020	0.013	0.003	0.036
		(0.015)	(0.005)	(0.003)	(0.016)
AY 2009	Mixed-sex school	0.027	0.018	0.004	0.049
		(0.009)	(0.007)	(0.004)	(0.015)
	Boys-only school	0.032	0.017	0.003	0.052
		(0.012)	(0.007)	(0.003)	(0.015)
	Girls-only school	0.027	0.015	0.002	0.044
		(0.010)	(0.006)	(0.002)	(0.044)

Note: AY 2007 in the first column corresponds to student peer groups who entered high school in academic year 2007. Numbers in Second Year and Third Year columns correspond to the mean ratio of dropouts for second grade and third grade students who entered high school in AY 2007. Standard deviations in parentheses. Ratios in the last column have been calculated by summing the ratios over the three-year period.

student-level covariates, I cannot test whether this holds.

On the other hand, there is a school-level data on a ratio of students who dropped out from high school, and by calculating the mean of the ratio by school types, I can use these as proxy variables for the expectation terms in ϕ_S and ϕ_M. In Table 1.4, I show an average ratio of high school dropouts by school types and by school grades. To interpret the numbers in the table, AY 2007 indicates students who entered high school in 2007, and on average, 2.5% of students in mixed-sex schools quitted high school during the first year. When these students became second grade, another 1.3% of students in mixed-sex schools dropped out, on average. Lastly, 0.4% of students quitted during their third year of high school. Hence, the mean ratio of dropouts during the high school period is 4.2% for students in mixed-sex schools who entered high school in 2007.

Therefore, based on the last column, an average ratio of high school dropouts in mixed-sex schools for three waves is 4.4%. Likewise, corresponding dropout ratios for boys- and girls-only schools are 4.9% and 3.7%, respectively. Using these three ratios, I can estimate two fraction terms in ϕ_S and ϕ_M. If we were to compare between boys-only schools versus coeducational schools, $E\left[D_i|T_i = S\right] = 0.951$ and accordingly, the estimate for the fraction term in ϕ_S is 0.052. Contrarily, when comparing with girls-only schools, $E\left[D_i|T_i = S\right] = 0.963$ and the corresponding estimate is 0.038. For coeducational schools, $E\left[D_i|T_i = M\right] = 0.956$ and as a result, the estimate for $\left(1 - E\left[D_i|T_i = M\right]\right)/\left(E\left[D_i|T_i = M\right]\right)$ in ϕ_M is 0.043.

As a consequence, the magnitude of the attrition bias is increased by the factor of either 0.052 or 0.038 for same-sex schools and 0.043 for mixed-sex schools. According to Equation (4), then, the size of the bias depends on the two terms, $(E[Y_i(S)] - E[Y_i^{miss}(S)])$ and $(E[Y_i(M)] - E[Y_i^{miss}(M)])$ multiplied by above factors. Note that if these two values are similar, then the size of the attrition bias is minimal because the difference in the attrition multiplier between the treatment and the control groups is only 0.009 or 0.005, in absolute terms. In this study, I contend that the two values are similar in expectation because there are few reason to believe that students who quit from mixed-sex schools are significantly different, academically, from those who quit from same-sex schools. Therefore, I conclude that the attrition problems are negligible in the sample.

1.7 Estimation Results

1.7.1 Mean Effects

The first thing I estimate is the mean effect of gender segregation on students' academic performance. The dependent variables used in the estimation are percentile ranks in reading, English, and mathematics. For mathematics, there are two types. In Korea, high school students are divided into one of three majors; liberal arts, natural sciences, or athletics and arts. For the analysis using math scores, I do not use students in the athletics and arts major because they did not take math tests in CSAT 2009 to 2012. On the other hand, the format of the math tests are different for the liberal arts students and the natural sciences students. Students in liberal arts are tested on basic mathematics whereas students in natural sciences take more advanced mathematics. Hence, I created math percentile ranks within each major and conducted a separate analysis for students in liberal arts and students in natural science. Note, however, that since every student takes exactly the same test for reading and English, I created percentile ranks and ran the analysis using all observations for these two subjects.

Table 1.5 to Table 1.6 report mean effects. In Panel A of Table 1.5, I ran a regression of percentile ranks in reading test on a dummy variable indicating the school type. Since students are randomly assigned to high schools within the school districts, all regressions are conducted with a set of school district fixed effects. For the analysis, I estimated the coefficients for each CSAT separately, and then ran a regression using pooled sample. For the analysis, I compared boys in boys-only schools with boys in mixed-sex schools, and girls in girls-only schools with girls in mixed-sex schools. For Panel B of Table 1.5, everything is the same except that the dependent variable is percentile ranks in English tests. Note that for dependent variables, I used percentile ranks calculated within the sample used in the analysis. The estimates are similar in magnitude when used with percentile ranks calculated at the national level.

For boys, CSAT-by-year regressions report that, on average, boys in boys-only schools performed better than those in mixed-sex schools. The magnitude of the mean effects varies by CSAT year but every estimate shows boys in same-sex schools have higher percentile ranks in reading. Turning to the effect of gender segregation

Table 1.5: Mean Effects of Gender Segregation (Reading and English)

	Boys-Only vs Boys in Coedu			Girls-Only vs Girls in Coedu		
CSAT	Boys-Only	S.E.	Sample	Girls-Only	S.E.	Sample
A. Dependent Variable: Percentile Ranks in Reading Test						
2002	0.970	(1.013)	42,461	1.478	(1.052)	33,967
2003	1.523*	(0.830)	38,385	2.498**	(1.004)	30,995
2004	1.634*	(0.900)	39,519	2.540**	(1.034)	33,469
2009	2.517***	(0.904)	33,694	2.402**	(0.942)	31,394
2010	2.815***	(0.844)	39,028	2.589***	(0.810)	36,422
2011	2.578***	(0.783)	39,108	1.998**	(0.875)	36,026
2012	2.121**	(0.847)	40,188	2.648***	(0.903)	35,275
Pooled	2.039***	(0.331)	272,383	2.314***	(0.353)	237,548
B. Dependent Variable: Percentile Ranks in English Test						
2002	1.882	(1.260)	42,386	2.017	(1.481)	33,940
2003	2.322**	(1.079)	38,299	2.857**	(1.416)	30,961
2004	1.720	(1.133)	39,436	2.738**	(1.321)	33,439
2009	2.879**	(1.214)	33,473	3.170***	(1.197)	31,251
2010	4.167***	(1.108)	38,778	2.830**	(1.152)	36,307
2011	3.611***	(1.136)	38,754	2.906**	(1.288)	35,840
2012	2.924***	(1.113)	40,188	2.975**	(1.235)	35,275
Pooled	2.816***	(0.435)	271,314	2.799***	(0.487)	237,013

Note: Standard errors in parentheses are clustered at the school level. A number of clusters range from about 90 to 120 depending on the CSAT year. All regressions include school district fixed effects (9 or 10 school districts per year). For the pooled sample, there are 769 clusters for boys sample and 734 clusters for girls sample, and 70 school district fixed effects for both samples. * significant at the 10%; ** significant at the 5%; *** significant at the 1%, all under clustered standard errors.

for girls, the estimates are also favorable for female students in same-sex schools. The mean effects of being in a girls-only school range from 1.47 percentile points to 2.64 percentile points. According to the pooled regression, when students attend same-sex schools, reading percentile ranks increase by 2.03 and 2.31 for boys and girls, respectively.

In the table, I report standard errors clustered by school names. For both panels, all coefficients are statistically significant at the 1% level under the conventional standard errors. However, when evaluated under the clustered standard errors, some coefficients (CSAT 2002) are insignificant even at 10% level. This highlights the importance of using clustered standard errors when the errors are correlated. Even if used with a large sample size (e.g., 272,383 for boys and 237,548 for girls), the clustered standard errors are roughly three times higher than the conventional standard errors. Thus, it reveals that conducting a statistical inference based on the conventional standard errors will yield misleading conclusions.

The effects of gender segregation on one's percentile rank in English tests are presented in Panel B of Table 1.5. For both genders, the size of the mean effects increased compared with the results obtained from reading tests. In all CSAT years, every corresponding coefficient is larger for the counterparts in Panel A of Table 1.5. The estimates range from 1.88 to 4.16 percentile points for boys and 2.01 to 3.17 for girls when assessed with percentile ranks calculated within the sample. For boys, with exceptions of CSAT 2002 and 2004, all coefficients are statistically significant either at the 5% or 1% under the clustered standard errors. For girls, all the estimates except CSAT 2002 are statistically significant either at the 5% or 1%. The mean effects of gender segregation using pooled sample show that boys in boys-only schools earn 2.81 percentile points higher than boys in coeducational schools. On the other hand, girls in girls-only schools earn 2.79 percentile points higher than girls in coeducational schools.

Two points should be noted from Table 1.5. First, within the same subject, the extent of mean estimates are highly homogeneous across CSAT years. That is, all coefficients show that students in same-sex schools scored better than those in mixed-sex schools, and the magnitude of the mean effects does not vary much across years. These facts imply that, in terms of academic performance, same-sex setting is beneficial for boys as well as girls. Second, the mean effects for English tests are greater than that of reading tests. This is probably because a large proportion of English education is conducted within schools and the curriculum and content of the subject matter is relatively well-defined and sequenced compared to reading education, and as a result, the effect of gender segregation is larger for English tests.

The argument that the gender peer effect will be higher for the well-sequenced subjects that are mostly taught at schools is also confirmed in Panel A of Table 1.6. I estimate the mean effects of gender segregation using students in liberal arts who took basic mathematics. Compared with the results in Panel A of Table 1.5, the

Table 1.6: Mean Effects of Gender Segregation (Basic Math and Advanced Math)

	Boys-Only vs Boys in Coedu			Girls-Only vs Girls in Coedu		
CSAT	Boys-Only	S.E.	Sample	Girls-Only	S.E.	Sample
A. Dependent Variable: Percentile Ranks in Basic Math Test						
2002	2.819***	(0.984)	17,442	1.018	(1.057)	22,196
2003	4.049***	(1.131)	15,582	3.036**	(1.178)	19,943
2004	3.891***	(1.208)	16,924	3.139**	(1.341)	20,987
2009	2.659***	(1.019)	22,202	3.276***	(0.969)	23,702
2010	3.341***	(0.970)	25,657	2.530***	(0.958)	28,297
2011	2.216***	(0.794)	25,546	2.051*	(1.112)	27,715
2012	3.005***	(0.772)	25,670	2.584***	(0.904)	27,239
Pooled	3.049***	(0.365)	149,023	2.505***	(0.402)	170,079
B. Dependent Variable: Percentile Ranks in Advanced Math Test						
2002	−0.401	(1.376)	18,983	3.222**	(1.595)	6,034
2003	−0.610	(1.129)	17,300	2.341	(1.501)	5,775
2004	−0.411	(1.265)	16,614	1.852	(1.442)	6,427
2009	2.334*	(1.400)	9,769	0.403	(1.708)	4,553
2010	0.243	(1.253)	11,463	1.684	(1.708)	5,116
2011	1.023	(1.240)	11,571	1.105	(1.756)	5,084
2012	2.434	(1.375)	12,664	0.696	(1.512)	5,348
Pooled	0.511	(0.495)	98,364	1.619***	(0.607)	38,337

Note: Standard errors in parentheses are clustered by school names. A number of clusters range from about 90 to 120 depending on the CSAT year. All regressions include school district fixed effects (9 or 10 school districts per year). For the pooled sample in basic math tests, there are 769 clusters for boys sample and 734 clusters for girls sample. For the advanced math tests, there are 769 clusters for boys and 729 clusters for girls. And 70 school district fixed effects for both samples. * significant at the 10%; ** significant at the 5%; *** significant at the 1%, all under clustered standard errors.

magnitude of the estimates, again, is larger. In addition, for boys in particular, most coefficients are larger than the counterparts estimated with English tests. The magnitude of the effect is roughly equal to 3 percentile points for boys and 2.5 percentile points for girls. For boys, all of the coefficients are statistically significant at the 1% level. Contrarily, for girls, the level of statistical significance are slightly less than that of boys' estimates. At any rate, I find that placing a male or a female student at a same-sex school increases his/her percentile points in basic math tests.

The advantage that students in same-sex schools experience with respect to basic math tests disappears, however, when estimated with scores in advanced mathematics tests (see Panel B of Table 1.6). For example, for boys, only the estimates obtained from CSAT 2009 and 2012 sample show that boys in boys-only schools gained higher percentile points. Besides, some coefficients are negative and most of the estimates are not statistically significant. Even the coefficient obtained from a pooled sample is statistically insignificant. The results are the same for girls. Although most of the estimates demonstrate that girls in girls-only schools received higher percentile points in the test, few of them are statistically significant. Note, however, that when analyzed using the pooled sample, female students in female-only schools scored about 1.6 percentile points higher than those in mixed-sex schools, and the estimate is statistically significant at the 1% level.

Note that students who took basic mathematics tests are different from those who took advanced mathematics tests. In general, students in Korea who take advanced math tests as a consequence of being in a natural sciences major tend to perform better academically than students who choose to be in liberal arts and possess stronger mathematical abilities. Accordingly, ambiguous effects reported in Panel B of Table 1.6 point to one important fact that gender segregation can play on students' academic performance. That is, it is likely that gender segregation has heterogeneous treatment effects and the extent of the effects may vary depending on the academic performance of students. I address this issue in the following subsection.

1.7.2 Distributional Effects

I estimate heterogeneous treatment effects using the quantile regression. The estimates corresponding to four quantiles (20%, 40%, 60%, and 80%) are presented in Table 1.7 together with standard errors and sample size. Note that the quantile regression has been conducted with pooled sample rather than separately estimating the heterogeneous treatment effect for each CSAT year. Three intriguing points can be seen from the table. First, Table 1.7 clearly shows that depending on the distribution of one's percentile rank, gender segregation has differential impact on students' academic performance. Second, in all of the subjects, male students in

Table 1.7: Distributional Effects of Gender Segregation (Pooled CSAT)

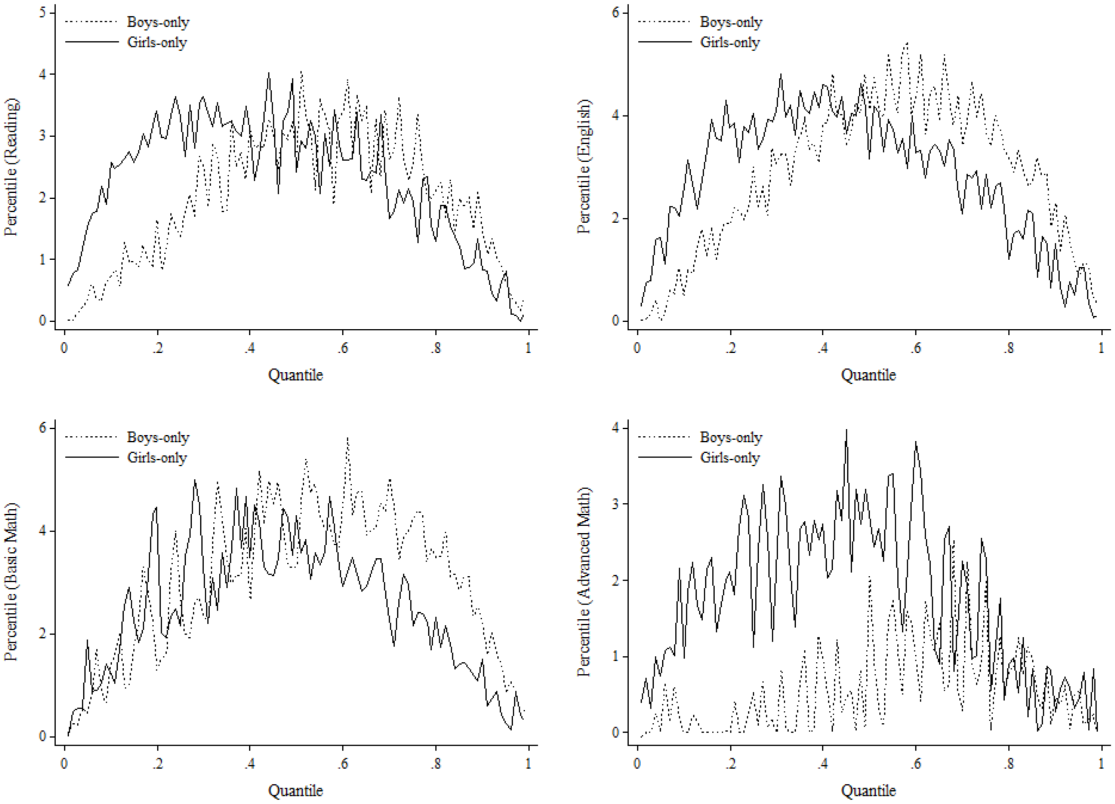
		Boys			Girls		
Subject	Quantile	Same-Sex	S.E.	Sample	Same-Sex	S.E.	Sample
Reading	0.2	1.638	(0.142)	272,383	3.399	(0.204)	237,548
	0.4	3.111	(0.204)		3.134	(0.231)	
	0.6	3.382	(0.235)		2.606	(0.214)	
	0.8	2.134	(0.210)		1.295	(0.174)	
English	0.2	1.883	(0.146)	271,314	3.768	(0.205)	237,013
	0.4	3.839	(0.212)		4.588	(0.227)	
	0.6	4.375	(0.239)		3.276	(0.221)	
	0.8	3.173	(0.203)		1.204	(0.178)	
Basic Math	0.2	1.299	(0.218)	149,023	4.449	(0.243)	170,079
	0.4	2.658	(0.278)		3.476	(0.286)	
	0.6	4.558	(0.303)		2.926	(0.276)	
	0.8	3.485	(0.271)		2.327	(0.210)	
Adv. Math	0.2	0.000	(0.292)	98,364	2.114	(0.469)	38,337
	0.4	0.944	(0.396)		2.737	(0.532)	
	0.6	0.904	(0.409)		3.824	(0.545)	
	0.8	0.826	(0.334)		0.898	(0.472)	

Note: Standard errors adjusted for heteroskedasticity. The quantile regression has been conducted with 70 school district fixed effects. For boys, I am comparing boys in boys-only schools vs boys in mixed-sex schools. For girls, I am comparing girls in girls-only schools vs girls in mixed-sex schools.

40 and 60 percentiles experience the highest gain from gender segregation. For instance, for basic math tests, students in 60 percentiles who attended boys-only schools earned 4.55 percentile points higher than boys in coeducational schools. Third, for advanced math tests, although small in magnitude, the estimated coefficients for quantiles other than 20 quantile yield statistically significant coefficients, which we do not see in the mean effects.

I show more complete pictures of heterogeneous treatment effects in Figure 1.4 by estimating the quantile treatment effects for quantile values $q \in \{0.01, 0.02, ..., 0.99\}$. The upper left figure in Figure 1.4 plots the coefficient $\hat{\beta}_1(q)$ estimated using reading percentile ranks. The dotted line corresponds to the estimates for boys and the line corresponds to that of girls. The overall shape of the curve is inverted U-shape implying that gender segregation is favorable for students in middle quantiles. Note that for reading tests, even the female students in quantiles between 0.1 to 0.3 also benefit from attending female-only schools where as for male students, the case is

Figure 1.4: Distributional Effects of Gender Segregation by Subject



not true.

The upper right figure in Figure 1.4 plots the quantile treatment effects for English. Likewise, the curve is inverted U-shaped and it indicates the fact that students in middle quantiles benefit most from being isolated from the opposite sex. Observe that for boys in 58 quantile, the effect of gender segregation even goes up to 5.43 percentile points which is roughly three times higher than the effect estimated in 20 quantile.

The lower left figure corresponds to the distributional effects for basic math tests, and again, the estimates yield a bell-shaped curve. For basic math tests, the difference between the minimum estimated effect and the maximum estimated effect is approximately 6 percentile points for boys and 5 percentile points for girls. Finally, the lower right panel plots the estimates obtained from advanced math

Table 1.8: Sensitivity Analysis Using Cohorts (Students who Entered in 2009)

Subject	Test Name	Same-Sex	S.E.	Sample
A. Boys in Boys-Only Schools vs Boys in Mixed-Sex Schools				
Reading	NAEA 2009	2.614***	(0.978)	41,884
	NAEA 2010	2.990***	(0.924)	40,986
	CSAT 2012	2.121**	(0.847)	40,188
English	NAEA 2009	3.017**	(1.329)	41,864
	NAEA 2010	3.474***	(1.230)	40,895
	CSAT 2012	2.924***	(1.113)	40,188
Mathematics	NAEA 2009	3.086***	(1.072)	41,882
	NAEA 2010	3.650***	(0.979)	40,943
	CSAT 2012	3.005***	(0.772)	25,670
B. Girls in Girls-Only Schools vs Girls in Mixed-Sex Schools				
Reading	NAEA 2009	1.291	(0.922)	36,171
	NAEA 2010	2.556***	(0.833)	35,632
	CSAT 2012	2.648***	(0.903)	35,275
English	NAEA 2009	1.503	(1.320)	36,163
	NAEA 2010	2.229*	(1.233)	35,602
	CSAT 2012	2.975**	(1.235)	35,275
Mathematics	NAEA 2009	2.238**	(1.022)	36,173
	NAEA 2010	2.417**	(0.954)	35,621
	CSAT 2012	2.584***	(0.402)	27,239

Note: For Mathematics in CSAT 2012, I used students who took basic mathematics to estimate the treatment effect. Standard errors in parentheses are clustered by school names. A number of clusters are approximately 120. All regressions include school district fixed effects (10 school districts per year). Note that first year high school students took NAEA 2009, and the same students took NAEA 2010 as second year high school students. Finally, at the end of the third year, these same students took CSAT 2012. * significant at the 10%; ** significant at the 5%; *** significant at the 1%, all under clustered standard errors

tests. With regard to advanced math tests, I find that, in general, placing students in single-sex schools do not improve one's percentile rank. Although female students in quantiles between 0.2 to 0.6 experience gains from gender segregation, the estimated coefficients are highly volatile probably because of a small sample size, and I argue that one should be cautious as to interpreting the results.

In a nutshell, Figure 1.4 demonstrates the fact that the effect of gender segregation is heterogeneous and that the students in the middle quantiles benefit most from attending same-sex schools. On the other hand, for students located in the very lower and upper distribution of quantiles, the magnitude of the impact of gender segregation is small.

1.7.3 Sensitivity Analysis and Dynamic Treatment Effects

In this subsection, I explore the sensitivity of the results to changes in the type of test that students take. I use NAEA exams that students take during their first and second year, and I use CSAT that students take at the end of their third year of high school. In Table 1.8, I present estimates obtained from running a regression of percentile rank in NAEA exams on a dummy indicating school types. As with the analysis based on CSATs, the effect of being in a same-sex schools is favorable for students' academic achievement. Moreover, the magnitude of the estimated effect is very similar to the ones obtained from the CSAT sample.

Another finding is noteworthy in Table 1.8. Students who took NAEA 2009 as first grade high school students also took NAEA 2010 as second grade high school students. Furthermore, these students took CSAT 2012 as third grade high school students. As a consequence, the amount of time that students are exposed to the treatment is different between these three periods, and it provides an opportunity to observe dynamic treatment effects of gender segregation. Provided that staying in single-sex setting is propitious as to one's academic performance, I should observe the effect of gender segregation increasing between this three-year period. The argument is reasonable because as students are more exposed to a helpful academic environment, their academic performance should improve more compared to the case in which students are exposed to the treatment for a shorter period of time. Interestingly, for all subjects, the effects of being in boys-only schools have increased over two-year periods. Contrarily, for girls, the estimates continually increased over the three-year period. These imply that the impact of gender segregation increases as the time one has been exposed to the treatment increases.

1.8 Conclusion

Consider male or female students who move from coeducational schools to same-sex schools. The findings from this study imply that, on average, students' academic performance increase by more than 2 percentile points when segregated from the opposite sex. Furthermore, this study highlights the fact that estimating the mean treatment effect leaves out much information and gives an incomplete picture of the effect of gender segregation. A more complete picture of the effect has been demonstrated by estimating the quantile treatment effects, and as a result of the quantile regression strategy, I find that students in the middle quantiles benefit the most from being segregated from the opposite sex. On the other hand, gender segregation has little impact on students located at the very bottom and the very top quantiles of the distribution of percentile ranks. This strongly suggests that estimating the quantile treatment effect is essential for evaluating the effect of educational interventions. Lastly, the sensitivity analysis indicates that the higher the exposure to the treatment, the higher the benefits they experience from being in the treatment group.

I do not argue that this study is impeccable. However, by leveraging the benefit of large-scale randomized social experiments conducted in Seoul coupled with careful analysis of the validity of the randomization in terms of three dimensions, I contend that the study suffers less from the lack of internal and external validity of the analysis and was successful in obtaining a "causal" estimate of the gender segregation on students' academic performance. Accordingly, because Title IX of the Education Amendments of 1972 enabled school districts to provide same-sex schools, I believe that the result of the analysis clearly has policy implications to both the educational administrators as well as parents in the United States.

1.9 Chapter 1: Appendix

1.9.1 Administrative Procedures

Once the Ministry of Education receives research plans, an external committee reviews the appropriateness and feasibility of the research plan. Once the committee approves the plan, the Ministry of Education organizes screening committees, in this case, with inside members, and decide upon whether to approve the data disclosure. Once the final approval has been made, researchers submit a written pledge regarding the data usage followed by several administrative procedures. After all the necessary process has been made, researchers visit the Ministry to retrieve the data stored on a compact disk.

1.9.2 Sample Restrictions

First, I exclude schools that are not located in Seoul because the random assignment policy is only adopted in Seoul during these periods. Second, within Seoul, the randomization has only been conducted on students in general high schools. Accordingly, I discard students in special purpose high schools and vocational high schools. Third, among general high schools in Seoul, some schools located at the center of the city (circled area in Figure 1.1) are excluded from the school pools used for randomly assigning students. The Office of Education allows students who reside in Seoul to list three to five schools located in this area and students are assigned to the school of their choice. The number of schools in the circled area varies by year, and I obtained the list of schools in this area, by year, from the Office of Education and sample restrictions are conducted based on the information given by the Office of Education. This policy has been adopted by the government since 1996 to palliate the resentment of parents who were deprived of their right of school choices for their children. Anyhow, since students in these schools are not randomly assigned, I do not use students in these schools.

Fourth, some general high schools are operated both at the daytime and at night. These schools consist of general high school students as well as vocational high school students. Since I cannot determine whether the student in CSAT or NAEA dataset is a general high school student or a vocational high school student, I drop students in these schools. Fifth, there are cases in which some single-sex

Table 1.9: Sample Restrictions

	Number of Students and Schools						
Dataset	Initial	Step 1	Step 2	Step 3	Step 4	Step 5	Final
CSAT 2002	717,975	191,518	162,256	138,156	125,837	122,653	76,431
	(2,011)	(295)	(181)	(149)	(138)	(134)	(133)
CSAT 2003	654,765	176,335	149,299	126,754	117,337	115,458	69,380
	(1,990)	(296)	(183)	(150)	(141)	(139)	(138)
CSAT 2004	641,807	174,176	147,281	126,681	123,141	119,961	72,992
	(1,967)	(297)	(183)	(153)	(149)	(145)	(144)
CSAT 2009	558,259	146,046	115,140	96,993	96,374	96,374	65,104
	(2,115)	(322)	(200)	(163)	(162)	(162)	(162)
CSAT 2010	637,063	159,774	126,497	106,772	106,081	106,081	75,474
	(2,156)	(325)	(203)	(165)	(164)	(164)	(164)
CSAT 2011	667,949	163,603	137,766	115,602	114,911	114,911	75,158
	(2,225)	(328)	(206)	(167)	(166)	(166)	(166)
CSAT 2012	691,985	164,777	139,514	117,349	116,705	116,705	75,463
	(2,169)	(330)	(212)	(173)	(172)	(172)	(172)
NAEA 2009	643,106	118,312	94,176	78,792	78,322	78,322	78,322
	(2,201)	(306)	(212)	(173)	(172)	(172)	(172)
NAEA 2010	631,362	116,077	92,648	77,717	77,256	77,256	77,256
	(2,199)	(306)	(212)	(173)	(172)	(172)	(172)

Note: Step 1 excludes schools not located in Seoul. Step 2 excludes special purpose and vocational high schools. Step 3 excludes schools that do not admit students via lottery system. Step 4 excludes schools that operate both at the daytime and at night. Step 5 excludes schools that modified their school types during the three-year period prior to CSAT year. In the final step, I exclude students who already graduated from the high school at the time of the test. Also, for CSAT 2002 to 2004, I exclude one district, where only one girls-only school is present, because there are no other control groups to compare with. The number of high schools in parentheses.

schools change to mixed-sex schools since the randomization. I remove these schools from the sample.[19] Sixth, I eliminate students who have graduated from high school because their treatment status are contaminated by being graduates. Finally, for CSAT 2002 to 2004, I exclude one school district because only girls-only school is present in this time period and there are no other schools to compare with. The series of the aforementioned sample restrictions and the resulting sample size are summarized in Table 1.9.

[19] For example, note that students who took CSAT 2002 entered high school in 1999. Accordingly, I do not have to drop schools which changed their school types four years before the CSAT year (i.e., in 1998) because students who took CSAT 2002 are not affected by the modification done in 1998, and hence the three-year period.

2 The Effect of “Rewarding” Poor-Performing Schools Under the School Accountability System: Evidence from Regression Discontinuity Designs

2.1 Introduction

Governments around the world conduct various educational interventions to enhance the performance of their schools. Among those is the school accountability system that evaluates school performance based on the academic achievement of its students. Until recently, most school accountability systems in the United States operated under two incentive mechanisms: 1) providing rewards to schools that outperformed the standards set by the states, or 2) imposing penalties on those who failed to meet the standards.[20] Existing research findings, however, is inconclusive regarding the effectiveness of such school accountability systems.

Furthermore, as previously documented by many researchers, the conventional incentive mechanisms adopted in the school accountability system generated several unintended side effects. One that has received much attention is the problem of strategic behavior committed by schools (i.e., cheating). Also, some argue that the system makes teachers and school administrators more likely to concentrate on teaching the subjects that students are tested on and put less emphasis on non-tested subjects. Recently, some argue that award-winning schools are not using their rewards in productive ways, and as a result, the awards are less effective in improving the subsequent performance of these schools. Hence, previous studies highlight that under these strategic behaviors committed by schools, it is difficult to causally estimate the effect of the conventional incentive mechanisms.

On the other hand, in 2010, the U.S. Department of Education introduced the federal grant program named the School Improvement Grants (SIGs) under the Elementary and Secondary Education Act of 1965. This grant adopts a slightly different incentive mechanism. Rather than punishing schools that do not meet the standards or rewarding schools that do, it “rewards” grants to poor-performing schools. The program has been administered since 2010 and studies on the effect

[20] According to Kane and Staiger’s (2002) compilation, 18 states provide financial rewards and 20 states impose sanctions.

of the SIGs are yet to be done. Accordingly, the SIGs provide a rare opportunity to analyze the effect of rewarding poor-performing schools. However, the SIGs are not randomly distributed to the local educational agencies and it is hard to elicit a "causal" effect of the program. Besides, the SIGs are operated under several intervention models, and under these models, the provision of grants is determined in a complex manner. Hence, it is difficult for one to come up with credible quasi-experimental estimates of the effect of the SIGs.

In order to estimate the effect of rewarding underachieving schools under the school accountability system, I make use of the school accountability system in South Korea in which poor-performing schools are rewarded with a grant based on the simple eligibility cutoff. To be more specific, schools in South Korea receive categorical funds from the government when its ratio of underachieving students exceeds a certain cutoff in a nationwide assessment of student achievement. Accordingly, by using the regression discontinuity design (RDD), the case in South Korea allows me to obtain the causal effect of providing awards to poor-performing schools.

To analyze the effect of providing awards to underperforming schools, I use students' test scores in the National Assessment of Educational Achievement (NAEA) in 2009 and 2010. In order to properly execute the RD estimator, I carefully tested the validity of the RDD by checking the possible manipulation of an assignment variable as well as the continuities in predetermined covariates. Based on several measures, I argue that the validity of the RDD is met, and accordingly, I use local linear regressions (LLR) to estimate the RD estimates.

Based on the analysis, I find that students in schools that received the grant performed significantly better than those in schools that did not receive the grant. While the magnitude of the effect varies depending on the subjects, students in funded schools, on average, scored 10 percentile points higher based on the total scores. Furthermore, I find that the ratio of underachieving students decreased in funded schools by more than 5 percentage points compared to that of non-funded schools.

2.2 Review of Previous Literature

In this section, I briefly present the results of previous literature that analyzed the effect of the school accountability system and the methodological obstacles to eliciting causal effects of the accountability system. In addition, I address some of the limitations of the conventional school accountability system raised by many researchers and how this study suffers less from those limitations.

Since the adoption of the school accountability system under Title I of the Elementary and Secondary Education Act of 1965 (reauthorized by the No Child Left Behind Act in 2002), many researchers have studied the impact of the system on students' academic achievement.[21] Numerous studies on the effect of the school accountability system have been conducted, thus I do not present reviews of every related study here. Rather, I resort to the comprehensive literature reviews presented in Figlio and Ladd (2008), and Figlio and Loeb (2011). According to these studies, the results are mixed. Moreover, even in the studies that find significant gains in students' test scores, the estimated results are less consistent. For instance, some studies find that the improvement is observed in math but not in reading scores.

Reasons for the ambiguous results and less consistent results presented in the previous literature are easily identifiable. To begin with, as noted in Figlio and Loeb (2011), estimating the effect of the school accountability system is difficult because such a system is conducted simultaneously with other states or federal educational reforms. In this case, it is not an easy task for one to disentangle the effect of the school accountability system from that of other reform measures. In addition, some studies use cross-state trends in academic achievement, and therefore, it is difficult to develop counterfactuals to credibly estimate the causal effect of the accountability system.

Apart from methodological challenges in eliciting the causal effect of the accountability system, some researchers claim that the incentive mechanisms adopted in the accountability system generate unintended consequences. For example, Jacob and Levitt (2003) convincingly show that under high-powered incentive mechanisms,

[21] Some of the studies published in academic journals are Richards and Sheu (1992), Ladd (1999), Smith and Mickelson (2000), Carnoy and Loeb (2002), Hanushek and Raymond (2005), Figlio and Rouse (2006), Chiang (2009), Neal and Schanzenbach (2010), Rockoff and Turner (2010), and Dee and Jacob (2011).

schools or teachers engage in cheating by manipulating the answer sheets of students. Moreover, Figlio and Winicki (2005) find that schools faced with sanctions try to promote students' academic performance by increasing the calorie intake in lunch menus. Jacob (2005) and Cullen and Reback (2006) also find that teachers respond strategically to the accountability system by increasing special education placements as well as manipulating the test-taking pool in advance of the test time. Likewise, Rouse et al. (2007) uncover that gains in the test scores under the school accountability system have been driven by the strategic behavior of schools such as changing school policies and practices. The aforementioned strategic behaviors are well-documented in the widely recognized theorems of principal-agent models. For instance, suppose an agent is faced with an incentive program in which the agent will be rewarded or sanctioned based on various dimensions. Suppose, however, that the principal cannot observe or can only partially observe the dimensions. In this setting, it is likely that the agent will focus only on the verifiable dimensions to comply with the incentives (Holmstrom and Milgrom, 1991). At any rate, given that school administrators and teachers engage in undesirable behavior a priori, it is difficult for one to conclude that the improved school performance is, indeed, due to the accountability system.

On the other hand, Bacolod, DiNardo and Jacobson (2012) examine the ex post facto behavior of schools under the school accountability system. They study whether schools that won financial rewards use the funds productively. According to their study, they do not find evidence that award-winning schools are using the rewards in favor of students' academic achievement, and moreover, they find that the academic achievement of students in the winning schools did not improve in a subsequent assessment. This is the limitation of the incentive mechanism of rewarding outperforming schools. Since outperforming schools face fewer incentives to exert efforts in the subsequent tests, they are less likely to use the rewards in an efficient manner.

Lastly, the conventional incentive mechanisms adopted in the accountability system possess another limitation. Note that one of the rationales behind the adoption of school accountability systems is the principal-agent model. In this model, school teachers and administrators are viewed as "culprits" for schools not meeting the standards set by the local educational agencies. However, it might be the case that poor-performing schools are failing due to a deficiency in school resources rather than because administrators and teachers are not doing their jobs. If this is the case, local educational agencies may mistakenly be penalizing the underperforming schools rather than providing support, and as a result, the agencies may not be able to identify schools that are truly in need of help.

Although not impeccable, the nature of the school accountability system in South Korea has several advantages in addressing the limitations presented above. First, the provision of school funding to poor-performing schools is determined by a simple eligibility rule, and accordingly, the institutional setting in South Korea, I believe, is favorable for estimating the causal effect of the rewards. Secondly, previous studies show that schools engage in strategic behavior and one might argue that the case is no different in this study. For example, in the context of this study, schools may engage in strategic behaviors to receive funding. As can be seen in the subsequent section, however, I show that this is close to impossible. Thirdly, since schools in South Korea receive "categorical" funding, the school accountability system suffers less from the point raised by Bacolod, DiNardo and Jacobson (2012) because schools are mandated to use the rewards in productive ways. For example, schools that receive the reward have to report the usage of their acquired funding. Besides, an annual evaluation on the assessment of students' academic performance is performed and therefore, schools have few incentives to use the reward in unproductive ways. Finally, given that the purpose of the school accountability system is to promote underperforming schools, and since schools that do not perform well receive funding, the nature of the accountability system in South Korea is more appropriate for meeting the goal of the school accountability system.

2.3 Institutional Background

South Korea implemented a nationwide program in 2008 that is similar to the U.S.'s No Child Left Behind Act of 2001. The program is intended to identify schools that are not meeting the standards set by the Ministry of Education. In order to identify these schools, the Ministry conducts a countrywide assessment of students' academic achievement at all educational levels.[22] The assessment is conducted by using the test results of the NAEA (a test every student in South Korea must take) which is analogous to the National Assessment of Educational Progress (NAEP) in the U.S.

The purpose of the assessment is to estimate students' overall level of academic achievement in each school and to calculate the number of students who do not meet the basic academic standards. The test is administered annually, and students are tested on three to five subjects depending on their educational level. For example, students in middle schools are tested on five subjects; reading, social studies, mathematics, science studies, and English.

After the test, the Ministry of Education collects all the tests and grades them. Then, each student is given a grade in one of the four categories: 1) excellent achievement, 2) normal achievement, 3) elementary achievement, or 4) less than elementary achievement. Next, for each school, the government calculates the ratio of students who received *less than elementary achievement* (hereafter, underachieving). Then, when the ratio in any school is above a certain cutoff, that school is designated as the school in need of achievement improvement (SINAI), and the government provides a categorical school funding to these SINAI-designated schools.[23] The amount of funding depends on the school size as measured by the number of students in each school (for elementary, middle, and general high schools) or the number of underachieving students (for vocational high schools). For instance, vocational high schools whose number of underachieving students is less than 100 receive about $30,000. Likewise, schools with 100 to 200 underachieving students receive about $50,000.

[22] There are three educational levels in South Korea; elementary (six years), middle (three years), and high school (three years). Also, note that high schools are divided into general high schools and vocational high schools.

[23] The cutoff for the ratio is as follows: elementary schools (5%), middle schools (20%), general high schools (20%), and vocational high schools (40%).

2.4 Data and Methods

2.4.1 Data

In this study, I use the administrative records of student-level test scores in NAEA 2009 and NAEA 2010. The data is not publicly available. At the end of 2010, however, the Supreme Court ruled in favor of disclosing the test scores, and the Ministry of Education has started to disclose the test scores to researchers since 2011. To obtain the datasets, researchers have to submit a research proposal to the Ministry, and the Ministry forms a committee consisting of members from inside and also outside the Ministry. These committee members then examine the feasibility of the research and decide upon disclosing the test scores to the researcher. Following the series of application steps, I obtained the NAEA data for 2009 and 2010. The dataset includes test scores of students as well as some school-level characteristics together with some answers to survey questionnaires conducted upon principals and students.

To analyze the effect of rewarding poor-performing schools, I make use of students in vocational high schools. Reasons for using students in vocational high schools are twofold. First, note that this study uses students' test scores in the NAEA for 2009 and 2010.[24] Now, in order to analyze whether students benefit from school funding, it is essential that students who take the test in these two time periods should be the same. Unfortunately, students in elementary and middle school are different for these two periods. In 2009, elementary school students in the sixth grade and middle school students in the third grade took the test. Likewise, in 2010, students in the same grades took the test. As a consequence, the students are not the same for these two time periods. Contrarily, in 2009, high school students who took the NAEA (both in general and vocational high schools) in their first grade took the NAEA again as second grade students in 2010. Hence, high school students are suitable for analyzing the effects of school funding.

Secondly, I do not use students in general high school students. Unfortunately, the data provided by the government does not contain the "name" of each school. To properly implement the RDD, however, it is necessary for one to identify the name

[24] Although the NAEA has been conducted since 2008, the census data for 2008 is not available. The government only provides 5% of total observations. On the other hand, for 2009 and 2010, government provides census data.

of the school so that one can merge with other datasets that contain school-level covariates and test the validity of the RDD. Hence, to identify the school name, I used the government-run website.[25] In this website, the descriptive statistics of the results of the NAEA for each school are uploaded with school names. These descriptive statistics are calculated using the same data used in this study. The descriptive statistics include the ratio of students in each categorical grade (up to first decimal place), the ratio of test-takers (up to first decimal place), and the number of test-takers. This information is calculated for each subject. Using the data used in this study, I calculated the same descriptive statistics that are presented in the SIW, and was able to match each statistic with a 100% matching rate. Consequently, I was able to identify the names of the schools in the dataset. Now, the number of vocational high schools in 2009 is 536, and since the SIW does not allow one to download the information uploaded on the website, I imputed the names of each school in the datasets manually by comparing descriptive statistics posted on the website (by clicking each school) with the ones that I calculated myself.[26] On the other hand, there are more than 1,500 high schools in 2009, and as a consequence, I do not have sufficient time and resources to match the names of each general high school. This is the reason I was able to use only the data of students in vocational high schools.

Table 2.1 presents a series of sample restrictions conducted for analyzing the effect of rewarding, as well as the resulting descriptive statistics. Panel A describes the sample restrictions. The original sample consists of both general high schools and vocational high schools. In Step 1, I exclude general high schools, and in NAEA 2009, there are 125,850 students in 536 vocational high schools. In Step 2, I exclude one school in NAEA 2009 whose name could not be identified from the SIW. The ratio of underachieving students in this unmatched school is 0.172, and this school is not entitled to be designated as SINAI. In Step 3, I drop schools that were not in both datasets. Specifically, in NAEA 2009, there is one school that is not present in NAEA 2010. The ratio of underachieving students in this school is 0.240, and again, this school is not entitled to be designated as SINAI. Contrarily, in NAEA

[25] School Information Website (SIW); www.schoolinfo.go.kr.

[26] I could not match the name of one school because this school was not listed on the SIW. There were 536 vocational high schools in the NAEA dataset, but 535 vocational high schools on the SIW.

Table 2.1: Sample Restrictions and Descriptive Statistics

Panel A

Original Sample		Step 1		Step 2		Step 3	
2009	2010	2009	2010	2009	2010	2009	2010
643,106	631,362	125,850	123,196	125,821	123,196	125,803	121,943
(2,191)	(2,196)	(536)	(540)	(535)	(540)	(534)	(534)

Panel B

	NAEA Year	
Variable	2009	2010
Ratio of female students	0.436	0.438
Ratio of test-takers	0.965	0.950
Ratio of reading test-takers among test-takers	0.987	0.994
Ratio of math test-takers among test-takers	0.991	0.994
Ratio of English test-takers among test-takers	0.995	0.998
No. of schools above 40% cutoff in NAEA 2009	35	—
No. of students in SINAI-designated schools	5,219	4,885

Note: In Panel A, numbers in parentheses indicate the number of schools. Numbers not in parentheses are number of students. The original sample column includes general and vocational high schools. In Step 1, I exclude general high schools. In Step 2, I exclude one school whose school name is not matched. In Step 3, I drop schools that are not present in both NAEA datasets. In Panel B, the 40% cutoff indicates the ratio of underachieving students in each school for NAEA 2009. When this ratio is equal to or greater than 40%, the school is designated as SINAI and this school receives categorical school funding.

2010, there are six schools that are not present in NAEA 2009. The reason for the presence of these schools may be that the schools have been closed or newly established during this two-year period, or these schools might have taken only one NAEA exam. Anyhow, I exclude these schools from the analysis.

Panel B lists some of the descriptive statistics that are estimated from the resulting sample restrictions. The ratio of female students is approximately 44%. Moreover, the ratios of test-takers are more than 95%. On the other hand, for those who came for the test, there were instances in which some students do not take one or two subjects among the three subjects. For each subject, I calculated a ratio of partial test-takers among those who are present on the test date. The ratio is almost 100%. As previously mentioned, vocational high schools are designated as SINAI and receive rewards when the ratio of underachieving students is equal to or greater than 40%. Using the result from NAEA 2009, the government designated

35 schools as SINAI.[27] Indeed, in the dataset, a number of schools whose ratio is equal to or greater than 40% is 35, and there are a total of 5,219 (2009) and 4,885 students (2010) in these schools.

2.4.2 Methods

In this study, for student i, we have two potential outcomes; $Y_i(0)$ for potential untreated outcome and $Y_i(1)$ for potential treated outcome. Then we observe the following outcome:

$$Y_i = (1 - T_i)Y_i(0) + T_iY_i(1),$$

where $T_i \in \{0, 1\}$ denotes the binary indicator for the treatment.

Now, because of the simple discontinuous rule used for assigning the treatment, this study uses the RDD to estimate the effect of school funding on students' performance. To illustrate the eligibility, let T_i be the indicator for whether students attend SINAI-designated schools. Then,

$$T_i = \mathbb{1}\{X_{is} \geq 0.4\},$$

where X_{is} is the ratio of underachieving students in school s in which student i attends. Note that depending on the probability of receiving the treatment, the RDD is classified into the sharp RDD or the fuzzy RDD, and since the probability of receiving the treatment is equal to one when $X_{is} \geq 0.4$, this study makes use of the sharp RDD.

To estimate the causal effect of the treatment under the sharp RDD setting, it must be the case that the conditional expectation functions of potential outcomes are smooth functions of the running variable X_{is}. There is no reason to believe that this assumption does not hold in this study because it is unlikely that the potential outcomes (i.e., students' academic performance) is not smooth in X (i.e., ratios of underachieving students in schools). Provided that the assumption holds, the average causal effect of the treatment (τ_i) in this study is

$$\tau_i = \lim_{x \downarrow 0.4} E[Y_i|X_{is} = x] - \lim_{x \uparrow 0.4} E[Y_i|X_{is} = x], \tag{5}$$

[27] Retrieved from the official government report of the Ministry of Education, Science, and Technology.

where Y_i denotes the percentile rank of student i.

In order to estimate the conditional expectation functions in Equation (5), one needs to estimate the conditional expectation functions near the boundary. That is, we need to estimate two conditional expectation functions: left and right of the cutoff ($X_{is} = 0.4$). Many researchers propose using the LLR (e.g., Imbens and Lemieux, 2008). Moreover, according to Hahn, Todd and van der Klaauw (2001), the LLR nonparametically provides a consistent estimator for the treatment effect in the context of an RDD. Hence, I use the LLR to estimate the conditional expectation functions in Equation (5). In the context of this study, the LLR estimator is based on minimizing the following problems:

$$\min_{\alpha_l,\, \beta_l} \sum_{i:\ 0.4-h<X_{is}<0.4} [Y_i - \alpha_l - \beta_l(X_{is} - 0.4)]^2 K\left(\frac{X_{is} - x}{h}\right)$$

and

$$\min_{\alpha_r,\, \beta_r} \sum_{i:\ 0.4\leq X_{is}<0.4+h} [Y_i - \alpha_r - \beta_r(X_{is} - 0.4)]^2 K\left(\frac{X_{is} - x}{h}\right)$$

where α_l, β_l, α_r, and β_r correspond to the intercepts and the slope coefficients of the LLR estimator at the left and the right of the 0.4 threshold, respectively. As can be seen from the minimization problems, one has to make discreet choices on two aspects; the kernel, $K(\frac{X_{is}-x}{h})$, and the bandwidth, h. As noted in previous literature, the choice of the kernel has little impact on the RD estimate. However, Fan and Gijbels (1996) show that the triangle kernel is optimal for estimating the LLR at the boundary. Following their suggestions, I use the triangle kernel.[28] On the other hand, the choice of the bandwidth is very important in the RDD. Recently, Imbens and Kalyanaraman (2012) developed an asymptotically optimal bandwidth choice. I use the proposed bandwidth given by Imbens and Kalyanaraman's (2012) algorithm, and by using this bandwidth as a benchmark, I present the sensitivity of the RD estimate based on other bandwidth choices.

[28] In this study, the choice of the kernel has little impact on the RD estimates.

2.5 RDD Validity Check

2.5.1 Manipulation of the Running Variable

The use of the RDD for estimating the causal effect of the treatment requires that individuals have imprecise control over the running variable (Lee and Lemieux, 2010). In the context of this study, this requires that schools are not able to precisely manipulate the ratio of underachieving students. To formerly test whether the school is gaming the ratio, I use the density test proposed by McCrary (2008) that tests the null hypothesis of continuity of the density of the running variable as it crosses the eligibility cutoff.

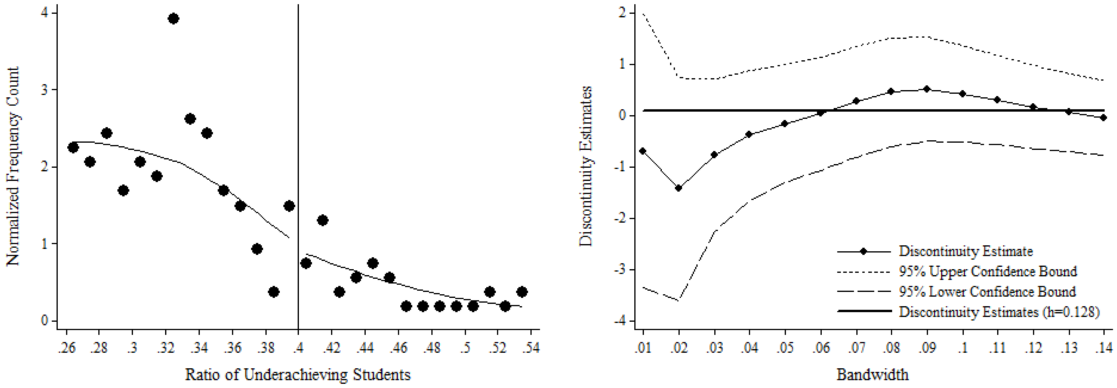
The left graph in Figure 2.1 corresponds to the density test proposed by McCrary (2008). The histogram has been created using the binwidth (0.01) suggested by McCrary (2008). For the bandwidth, I use 0.14 because almost no school has an underachieving ratio greater than 0.54.[29] As can be seen from the histogram, there seems to be no sign of manipulation at the eligibility cutoff. The formal test proposed by McCrary (2008) suggests using the bandwidth of 0.128 when estimating the discontinuity estimate. Based on this bandwidth, the discontinuity estimate at the 40% cutoff is 0.076 with a standard error of 0.393 implying that the discontinuity estimate is not statistically significant.

As presented in the simulation results in McCrary (2008), the discontinuity estimate is sensitive to the choice of bandwidth. Accordingly, McCrary (2008) recommends using the discontinuity estimate from the proposed bandwidth as a benchmark and conducting a sensitivity test of the estimate using varying bandwidths. Following this suggestion, I present in the second graph of Figure 2.1, the discontinuity estimates derived from different bandwidths together with 95% confidence intervals. In all of the discontinuity estimates, none of the estimates were statistically significant indicating that no matter which bandwidth one chooses, there seems to be a continuity of the density of the running variable across the cutoff.

However, McCrary (2008) further notes that the density test can still fail to detect the manipulation of the assignment variable when the number of schools manipulating the ratio of underachieving students to go up is offset by the number of schools manipulating the ratio to go down. Hence, it is desirable to examine,

[29] For the sake of consistency, I use the same bandwidth and the same binwidth throughout the paper when presenting the density.

Figure 2.1: A Density Test and Discontinuity Estimates Under Varying Bandwidths



Note: The left figure corresponds to the density test. The histogram has been created using the binwidth of 0.01 and the bandwidth of 0.14. The line has been fitted with the local polynomial regression with the first-order polynomial and the triangle kernel. The right figure plots the discontinuity estimates estimated from varying bandwidths. In the right figure, h corresponds to the bandwidth, and 0.128 is a proposed bandwidth from the density test (McCrary, 2008). The resulting discontinuity estimate from the proposed bandwidth is 0.076 with a standard error of 0.393.

more in detail, whether the precise gaming of the assignment variable is indeed less likely in the current setting. In order to conclude whether gaming is indeed close to impossible, I present two more arguments.

First, schools face little possibility of manipulating the test scores of students. The reason behind this argument is that schools do not grade their students' tests. Once the test is completed, they are sent to the Office of Education, and the Office of Education further sends the tests to the central government, who then grades all the tests. Moreover, all of the answers to the tests are not disclosed until the tests are over. As a consequence, it is highly unlikely that school administrators or teachers engage in manipulating the scores of their students.

Second, even if we assume that schools can manipulate the test scores, it is virtually impossible for them to game the ratio of underachieving students. This is because the eligibility cutoff (i.e., 40%) was announced by the central government much after the test date. Thus, it is impossible for schools to manipulate the ratio of underachieving students to be placed around the 40% cutoff.

Therefore, the density test as well as the institutional background behind the school accountability system indicate that the current study does not suffer from sorting of schools with respect to the running variable.

2.5.2 Continuity in Baseline Covariates

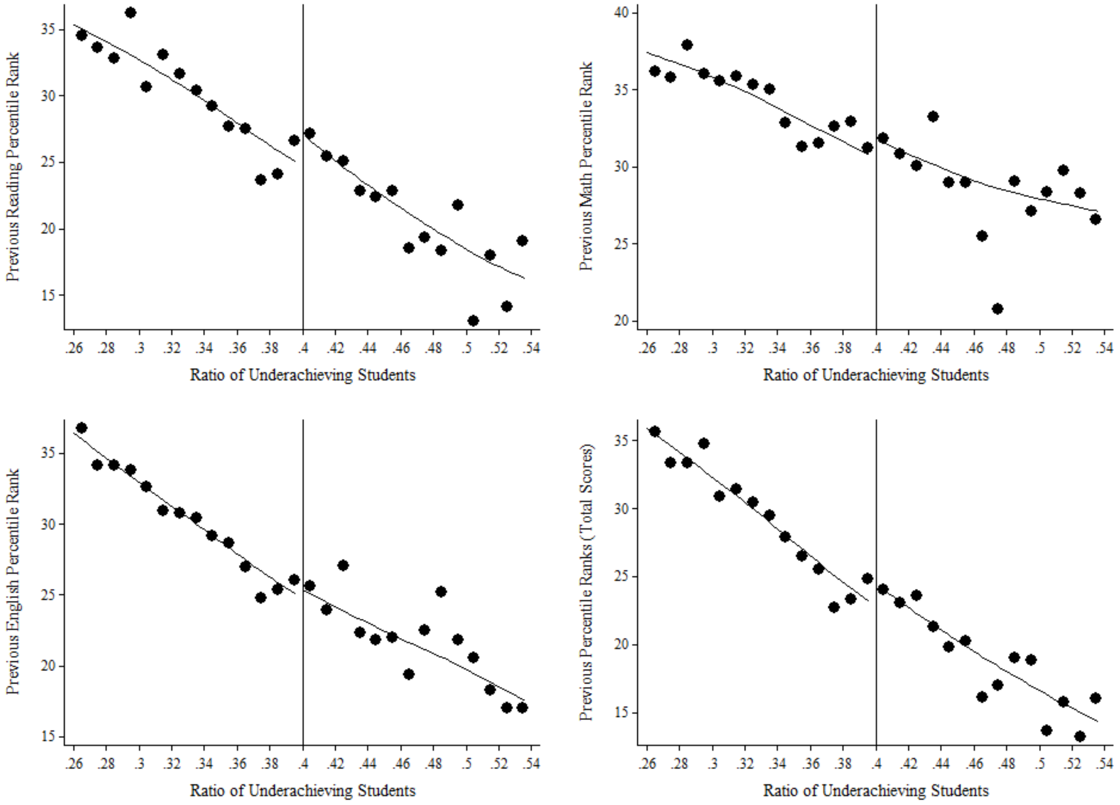
The second validity check that one should examine is whether the predetermined covariates display discontinuities at the eligibility cutoff. Since the RDD is based on the notion that the provision of the treatment is locally randomized at the cutoff, plotting the baseline characteristics against the assignment variable allows us to examine whether the characteristics are balanced across the cutoff. The rationale behind this idea is that if the provision of school funding is locally randomized, the baseline covariates should not show discontinuities in the threshold.

In order to test the continuities, I first use the average school performance measured by students' average percentile ranks. Since the average academic achievement of students in NAEA 2009 is determined prior to the realization of the running variable, the average percentile ranks for each subject should not display discontinuities at the threshold. Furthermore, I use class size, a student-to-teacher ratio, and a ratio of students living in poor households. These variables are likely to be highly correlated with the academic performance of schools, and since these covariates are determined a priori, these baseline covariates should not show discontinuities at the threshold if the local randomization is valid.

In Figure 2.2, the histogram has been plotted with a bandwidth of 0.14 and a binwidth of 0.01. The line fit is conducted using the LLR with a triangle kernel. The first two figures in the first row correspond to the average percentile ranks in reading and math tests in NAEA 2009. As can be seen from the figure, average performance on reading and math seem to display a smooth relationship at the 40% cutoff. To give a more formal analysis of the discontinuity estimate, I present RD estimates obtained from varying bandwidths in Table 2.2. For the reading percentile rank, one estimate, obtained from bandwidth 0.12, is statistically significant at the 5% level. However, other estimates derived from other bandwidths are not statistically significant. On the other hand, for math tests, none of the RD estimates display significant discontinuities at the threshold.

The two figures in the second row of Figure 2.2 show distributions of students' percentile ranks in English test and the percentile ranks calculated from scores in

Figure 2.2: Continuities in the Baseline Achievement

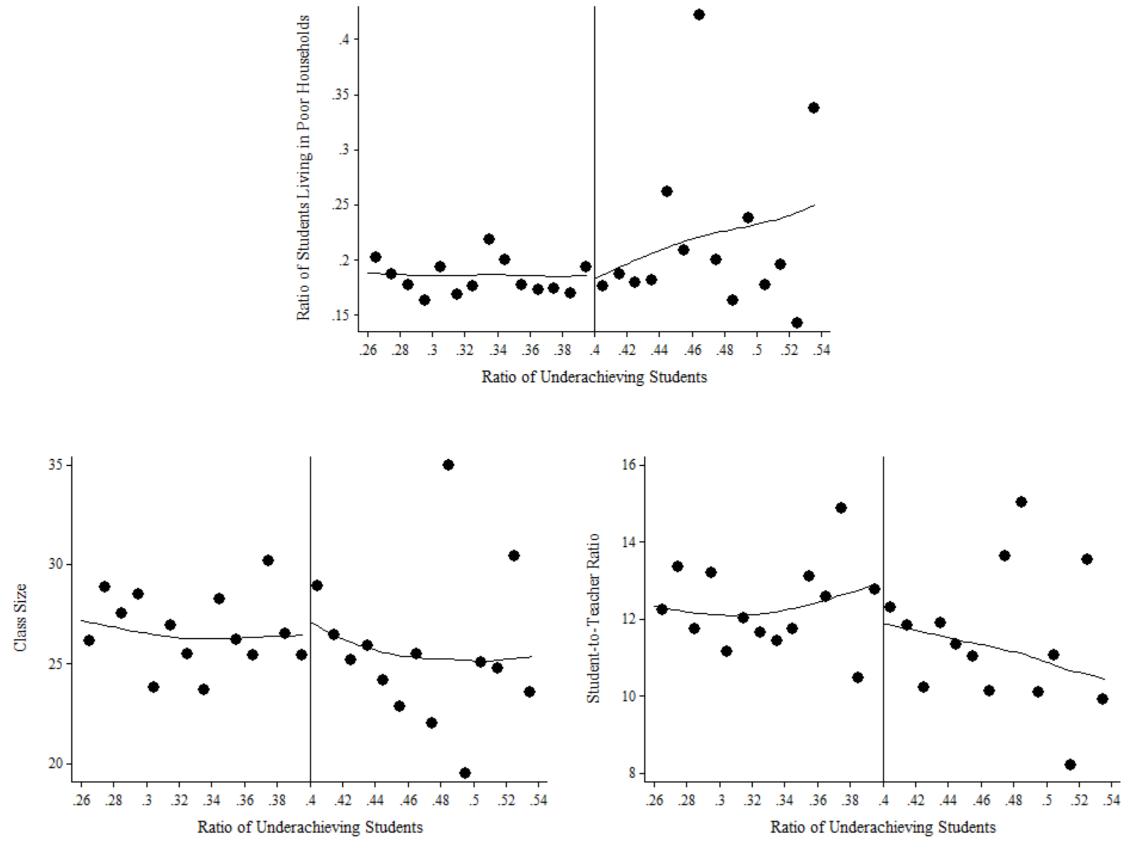


Note: The density has been plotted with the following: bandwidth of 0.14, binwidth of 0.01. The line fit has been conducted with the LLR using triangle kernel.

all three subjects (reading, mathematics, and English). As with the case of the other two subjects, the distributions of school performance in English display a smooth downward relationship with respect to the assignment variable. Moreover, in Table 2.2, none of the RD estimates are statistically significant implying that students' prior test scores in English are equally balanced across the eligibility cutoff. For the total scores, on the other hand, although the density displays a smooth relationship with respect to the assignment variable, some of the RD estimates turn out to be statistically significant. However, since the estimates are sensitive to the choice of the bandwidth, it is less likely that there is in fact a discontinuity.

In the first row of Figure 2.3, I present distributions of family background of students measured as a ratio of students living in poor households. The ratio of students living in poor households indicates the ratio of students who are living in

Figure 2.3: Continuities in the Baseline Covariates



Note: The density has been plotted with the following: bandwidth of 0.14, binwidth of 0.01. The line fit has been conducted with the LLR using triangle kernel.

a household that receives governmental subsidies under the National Basic Living Security Act. This variable is a good proxy for denoting students' family background. As can be expected from the locally randomized scenario, the distribution of the ratio displays no discontinuity at the threshold. Besides, as can be seen from Table 2.2, none of the discontinuity estimates are statistically significant.

Finally, in the last row of Figure 2.3, I plot density of class size and student-to-teacher ratios to examine the continuity in the distributions of baseline school-level characteristics. As can be seen from the figure, there does not seem to be discernible discontinuities at the 40% threshold. Indeed, the formal RD estimates presented in Table 2.2 show that the discontinuity estimates are not statistically significant under

Table 2.2: RD Estimates of the Baseline Covariates

	Bandwidth			
Variable	0.03	0.06	0.09	0.12
Reading Percentile Rank	−0.009	2.052	2.963	2.966**
	(2.675)	(1.754)	(1.525)	(1.377)
Mathematics Percentile Rank	1.959	0.494	1.691	1.717
	(2.017)	(1.365)	(1.239)	(1.156)
English Percentile Rank	0.597	1.018	1.209	0.816
	(2.695)	(1.542)	(1.281)	(1.104)
Percentile Ranks Using Three Subjects	−1.213	0.868	1.950**	1.978**
	(1.403)	(1.054)	(0.973)	(0.883)
Ratio of Students in Poor Households	−0.019	−0.019	−0.011	−0.007
	(0.048)	(0.034)	(0.029)	(0.026)
Class Size	5.385	2.877	1.410	0.652
	(3.041)	(2.173)	(1.915)	(1.746)
Student-to-Teacher Ratio	0.241	−0.740	−1.365	−1.394
	(2.086)	(1.492)	(1.246)	(1.057)

Note: RD estimates estimated from the LLR with a triangle kernel. Standard errors in parentheses.

any bandwidths. Moreover, the magnitude of the RD estimate is quite small. For example, the difference in class size is roughly three students, and the number of students per teacher is approximately one student. The homogeneity in class size and student-to-teacher ratios across schools is presumable because in South Korea, these two factors are strongly controlled by the Ministry of Education. That is, to maintain homogeneity of school-level characteristics, the Ministry of Education determines the number of students that each vocational school can accept. As a consequence, it is less likely that there will be differences in the school-level covariates across schools.[30]

All in all, the density of the baseline covariates and the formal RD estimates imply that these predetermined characteristics are equally balanced across the 40% threshold. Consequently, I argue that the validity of using the RDD to analyze the effect rewarding poor-performing schools is assured in this study.

[30] This is one of the main advantages of using the data of South Korea to analyze the causal effect of rewarding poor-performing schools.

2.6 Results

To draw a causal effect of providing positive incentives on poor-performing schools, I use students' percentile ranks in NAEA 2010 exams as the outcome variable. Students are tested on three subjects; reading, mathematics, and English. Accordingly, I estimate the effect on these three subjects. Furthermore, to test the overall impact, I add students' test scores in these three subjects and estimate an RD estimate based on the total scores. Lastly, I examine whether the treatment reduces the ratio of underachieving students in tracked schools.

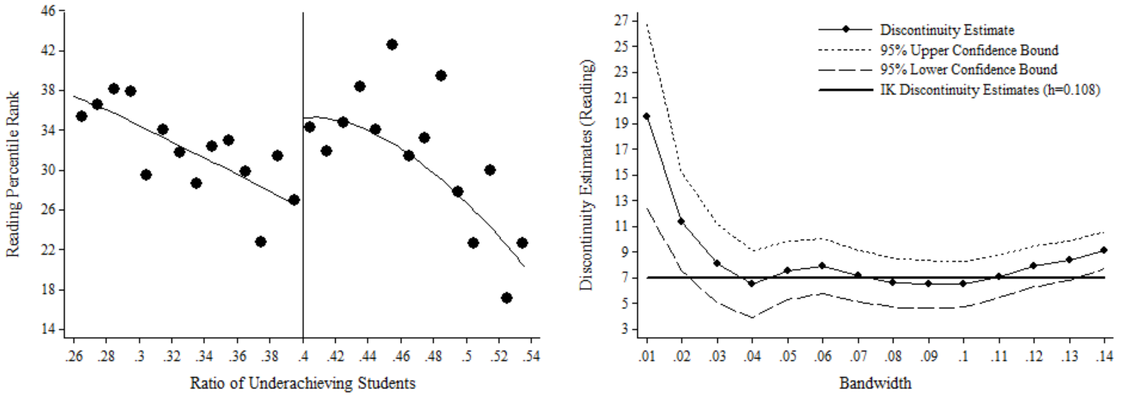
To examine the treatment effect, I first show the density of the discontinuities. When illustrating the density, I use a bandwidth of 0.14 with a binwidth of 0.01.[31] For the RD estimation, I use the LLR estimators. As noted in Section 2.4, I first estimate an RD estimator based on the optimal bandwidth given by Imbens and Kalyanaraman (2012, hereafter IK) and present the sensitivity of the estimate using varying bandwidths.

At the left column of Figure 2.4, I present the density of the percentile ranks in reading by the ratio of underachieving students in NAEA 2009. On the right column, I give RD estimates obtained from varying bandwidths to test the sensitivity of the RD estimate. In the graph, I provide 95% confidence intervals. Furthermore, I superimpose the RD estimate (horizontal line) obtained from the optimal bandwidth proposed by IK.

As can be seen from the figure, there is a clear visual break at the eligibility cutoff. The optimal bandwidth given by IK is 0.108, and based on this bandwidth, the RD estimate is 6.939 percentile points with a standard error of 0.846 implying that students in SINAI designated schools performed better compared to those who were not in SINAI-designated schools. On the right side, I present the sensitivity of the RD estimate using varying bandwidths. As it turns out, the RD estimates are very stable for the bandwidth greater than 0.02. In Table 2.3, I give specific RD estimates obtained from various bandwidths together with the standard errors and the resulting sample size used in the estimation process for each bandwidth. From Table 2.3, we can see that all the RD estimates are highly significant, and it

[31] The choice of the bandwidth and the binwidth rarely changes the graphical presentations of the data points. Hence, I use the same bandwidth and the binwidth used previously for the sake of consistency.

Figure 2.4: Density by the Running Variable and RD Estimates (Reading)



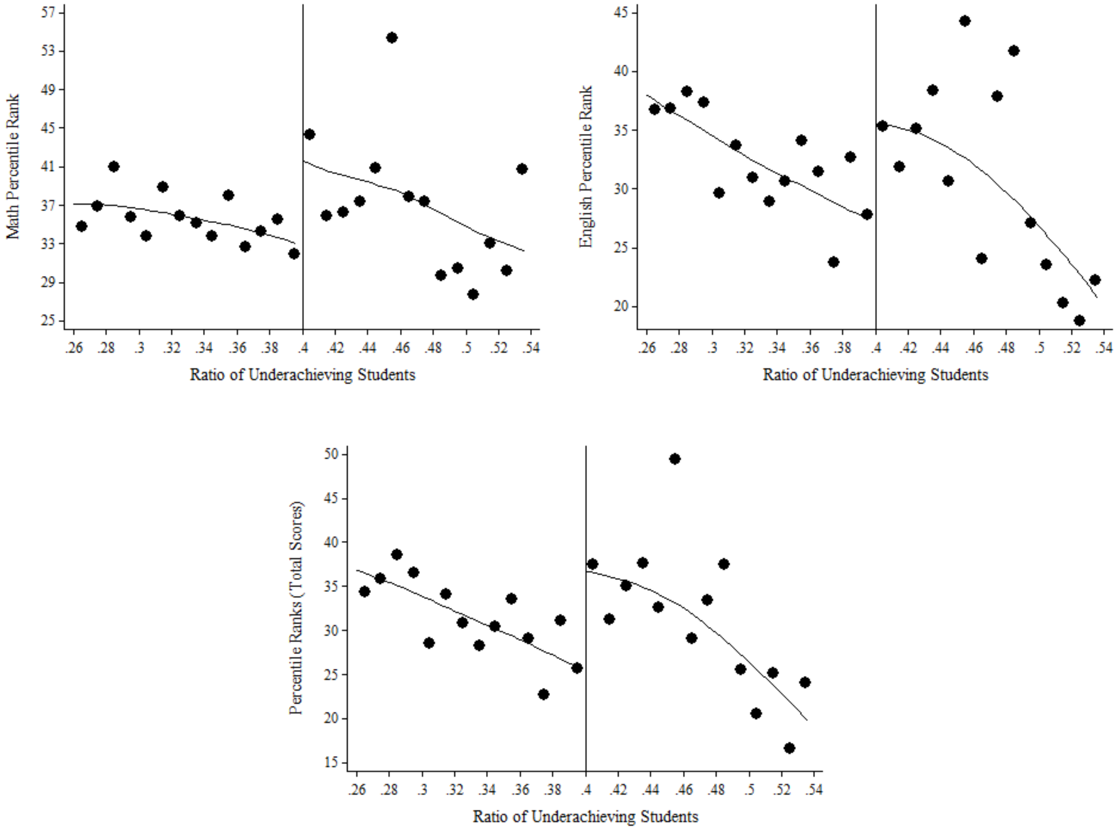
Note: The left column presents the density of the percentile ranks depicted with a bandwidth of 0.14 and a binwidth of 0.01. The right column presents the RD estimates derived from varying bandwidths. IK denotes the RD estimates based on the proposed bandwidth (h) from Imbens and Kalyanaraman (2012).

is reasonable to conclude that students' reading achievement in SINAI-designated schools improved significantly compared to those not in SINAI-designated schools.

Two figures in the first row of Figure 2.5 correspond to the density of math and English percentile ranks by the running variable. Compared to the histogram for reading percentile ranks, the data points at the right side of the cutoff in the histogram of math percentile ranks are quite noisy. Regardless, there does seem to be an even larger treatment effect on math scores. Indeed, the RD estimate based on IK's optimal bandwidth (0.083) is 7.905 with a standard error of 1.105. Since some of the data points just to the right of the threshold are quite noisy, the RD estimates are fairly sensitive when the choice of the bandwidth is less than 0.06 (see the appendix in Section 2.9). However, since all the estimates in Table 2.3 are statistically significant, I argue that students in SINAI-designated schools clearly benefited with respect to mathematics.

For English, on the other hand, the density is quite similar to that of reading tests. There is a jump in the density of the students' achievement at the 0.4 cutoff, and the RD estimate using IK's proposed bandwidth (0.077) is 6.072 and the corresponding standard error is 1.022 percentile points. As expected from the density of English percentile ranks, the estimated RD estimate is very close to that of reading tests. Furthermore, the pattern of the sensitivity of RD estimates is similar to that

Figure 2.5: Density by the Running Variable (Math, English, and Total Scores)

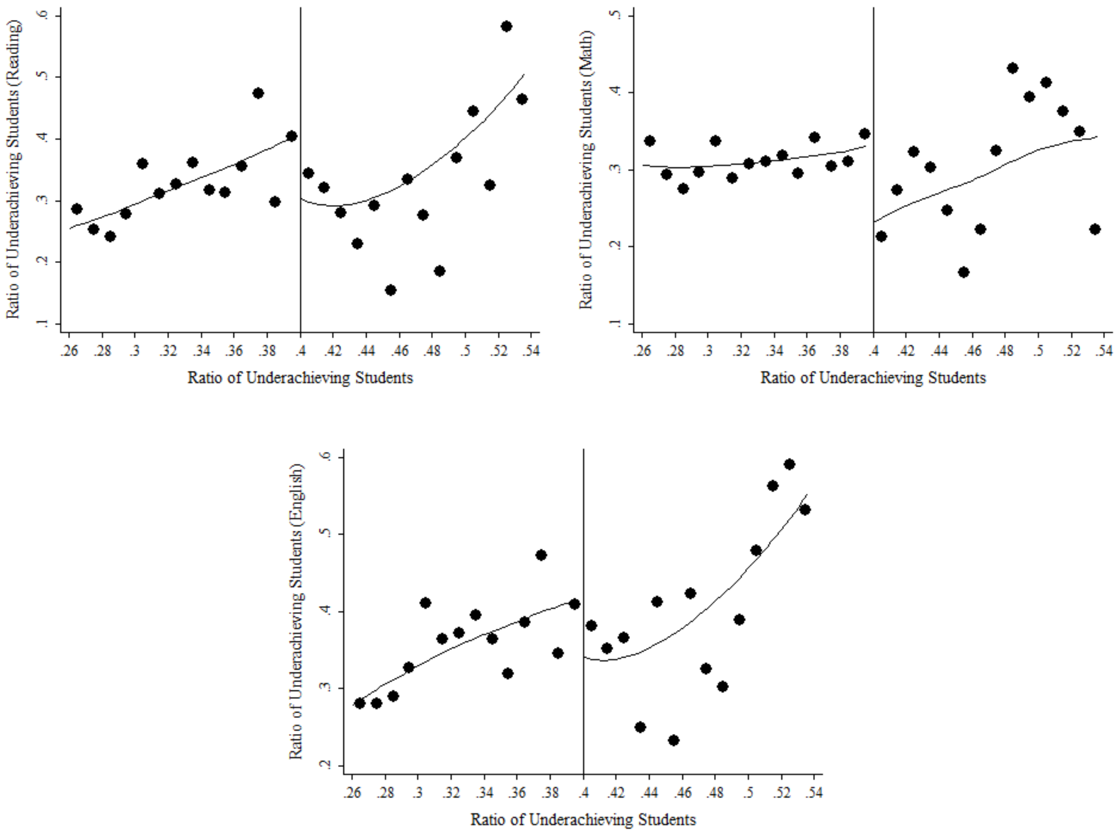


Note: The density has been plotted with the following: bandwidth of 0.14, binwidth of 0.01. The line fit has been conducted with the LLR using triangle kernel.

of reading tests (see the appendix in Section 2.9).

To estimate the overall treatment effect, I combined students' test scores in all three subjects and calculated the percentile ranks based on total scores. The bottom figure in Figure 2.5 corresponds to the density plot of total scores. On the left side of the threshold, the data points show a smooth downward relationship with respect to the running variable. There is a clear jump at the 40% cutoff and after that, it again shows a downward slope. This relationship is reasonable because the very end of the data point (near 0.5) includes students of the worst performing schools, and it is likely that school administrators and teachers have a hard time promoting their test scores in a very short amount of time compared with students in schools near the threshold. Because of the smooth relationship between total scores and

Figure 2.6: Density by the Running Variable (Ratio of Underachieving Students)



Note: The density has been plotted with the following: bandwidth of 0.14, binwidth of 0.01. The line fit has been conducted with the LLR using triangle kernel.

the assignment variable observed above, the optimal bandwidth proposed by IK is 0.143. If one uses a bandwidth of 0.143, most of the observations on the right side of the cutoff will be used in the estimation process. According to Table 2.3, when one uses a bandwidth of 0.14, the sample used in the estimation process is approximately 30,000 students. Contrarily, if one uses a bandwidth of 0.01, one ends up using only about 3,500 students. In any case, the estimated discontinuity based on this bandwidth is 11.418 with a standard error of 0.723 implying that students in SINAI-designated schools improved significantly compared with those in non-SINAI-designated schools.

Note that the purpose of designating schools as SINAI and providing categorical school funding is to encourage schools to engage in helping disadvantageous students

Table 2.3: RD Estimates Under Varying Bandwidths

	Bandwidth							
Variable	0.02	0.04	0.06	0.08	0.10	0.12	0.14	IK
Reading	11.326	6.489	7.922	6.622	6.483	7.883	9.138	0.108
	(1.943)	(1.320)	(1.085)	(0.974)	(0.897)	(0.809)	(0.737)	—
	[3,452]	[6,412]	[11,254]	[16,691]	[20,516]	[24,813]	[30,184]	—
Math	23.789	14.068	17.072	8.061	7.816	8.524	8.790	0.083
	(2.306)	(1.531)	(1.243)	(1.121)	(1.031)	(0.920)	(0.833)	—
	[3,472]	[6,450]	[11,350]	[16,852]	[20,707]	[25,024]	[30,415]	—
English	14.273	6.841	7.912	5.949	5.578	6.924	8.292	0.077
	(2.033)	(1.388)	(1.130)	(1.020)	(0.941)	(0.843)	(0.764)	—
	[3,473]	[6,458]	[11,369]	[16,880]	[20,743]	[25,065]	[30,436]	—
Total Scores	20.593	11.698	11.854	9.063	8.632	10.042	11.265	0.143
	(2.005)	(1.379)	(1.119)	(1.007)	(0.927)	(0.827)	(0.747)	—
	[3,443]	[6,399]	[11,234]	[16,664]	[20,480]	[24,773]	[30,107]	—
δ Reading	−0.159	−0.048	−0.078	−0.062	−0.060	−0.081	−0.104	0.094
	(0.038)	(0.025)	(0.021)	(0.019)	(0.017)	(0.015)	(0.014)	—
	[3,452]	[6,412]	[11,254]	[16,691]	[20,516]	[24,813]	[30,184]	—
δ Math	−0.204	−0.146	−0.120	−0.096	−0.095	−0.102	−0.101	0.136
	(0.035)	(0.023)	(0.019)	(0.017)	(0.016)	(0.014)	(0.013)	—
	[3,472]	[6,450]	[11,350]	[16,852]	[20,707]	[25,024]	[30,415]	—
δ English	−0.093	−0.005	−0.050	−0.035	−0.032	−0.050	−0.075	0.086
	(0.037)	(0.026)	(0.021)	(0.019)	(0.017)	(0.016)	(0.014)	—
	[3,473]	[6,458]	[11,369]	[16,880]	[20,743]	[25,065]	[30,436]	—

Note: IK denotes the optimal bandwidth proposed by Imbens and Kalyanaraman (2012). δ indicates the ratio of underachieving students. Standard errors in parentheses and the number of observations in brackets.

in their schools. Accordingly, as a final exercise, I estimate whether the treatment was successful in reducing the ratio of underachieving students. To estimate the treatment effect, I use the following dependent variable:

$$D_i = \begin{cases} 1, & \text{if student } i \text{ received an ``underachieving'' grade} \\ 0, & \text{otherwise} \end{cases} .$$

Contrary to the density of students' percentile ranks, it is likely that the data points on the left of the threshold show an upward trend followed by a drop at the cutoff. After that, the density, again, is likely to display an upward relationship with respect to the running variable. Figure 2.6 presents the results based on the above dependent variable, and as can be seen from the figure, the density clearly follows the predicted path. The first two figures in the first row correspond to the ratio of underachieving students in reading and math tests. The discontinuity estimate is -0.059 and -0.103, respectively, suggesting that on average, the ratio of underachieving students in reading and math dropped by 5.9 and 10.3 percentage points in SINAI-designated schools compared to that of non-SINAI-designated schools. The figure in the second row pertains to the ratio of underachieving students in English. The corresponding RD estimates is -0.033 based on the optimal bandwidth proposed by IK. As with the analysis of students' percentile ranks, I present the sensitivity of the RD estimates in the in Section 2.9. All the corresponding graphs show that the ratio of underachieving students in SINAI-designated schools decreased under various bandwidth choices.

In sum, it turns out that students clearly benefited from being in SINAI-designated schools. While the RD estimates vary depending on the subjects, the estimates are quite stable as to the choice of bandwidths. Hence, I conclude that students' academic achievement as well as the ratio of underachieving students, on average, improved significantly in NAEA 2010 as a result of rewarding poor-performing schools.

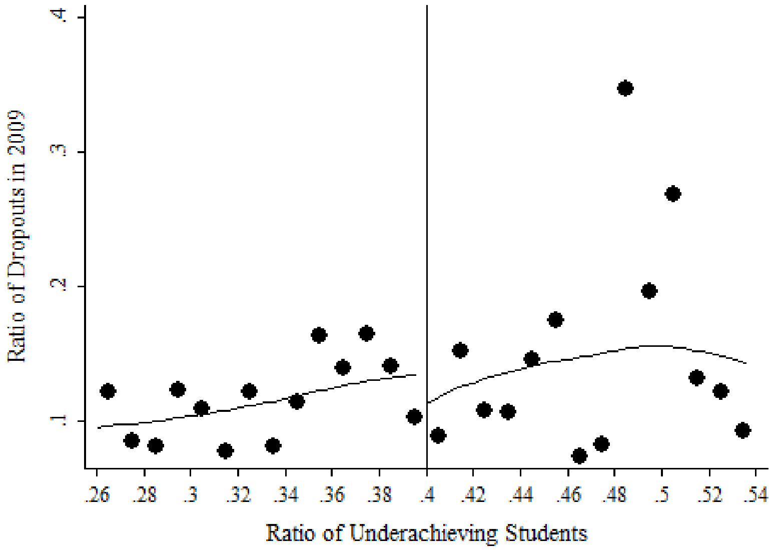
2.7 Attrition and Strategic Behaviors

In this section, I address two issues that may challenge the causality of the RD estimates; attrition and potential strategic behaviors committed by school administrators and teachers. When properly executed, the RDD is a convincing tool for estimating the causal effect of the treatment. However, when one uses a data that is longitudinal, one has to pay attention on the problem of attrition. Researchers may ignore the problem of attrition under three scenarios. The first scenario is when the attrition is close to zero. The second case is when the observation is missing at random. In this case, the attrition, on average, does not affect the estimated treatment effect. The third case is when the attrition is not random but similar in expectation between the treatment and the control group.

In the context of this study, the attrition problem exists because vocational high school students may have dropped out from the period between NAEA 2009 and NAEA 2010. For instance, if students who were originally present in non-SINAI-designated schools dropped out from the school had higher test scores, on average, compared to those who were initially present in SINAI-designated schools who also dropped out, then the treatment effect estimated above is biased upward. In Figure 2.7, I plot the RD-type histogram of a ratio of high school dropouts by the assignment variable. High school completion rate is extremely high in Korea, and as can be expected from the high completion rate, the ratio of dropouts is around 10 to 15% even in the worst performing schools.

At any rate, since the attrition rate is not close to zero, we cannot ignore the problem of attrition. To test whether attrition is random, one can use the baseline test scores of students in the treated group and the untreated group. Unfortunately, students' IDs are not linked between these two time periods, and as a consequence, I cannot test whether this is true. However, I contend that, on average, the pattern of attrition is similar, in expectation, between the treatment and the control group. First, it is hard to believe that students who dropped out from non-SINAI designated schools are academically different from students who dropped out from SINAI-designated schools, on average. Second, if the treatment had an impact on students' behaviors, then there must be some discontinuities in the distribution of dropouts. As can be seen from Figure 2.7, however, the density of the ratio of dropouts is quite smooth across the threshold (except for schools near 50%). The

Figure 2.7: Ratio of High School Dropouts by the Running Variable



Note: The density has been plotted with the following: bandwidth of 0.14, binwidth of 0.01, and the line fit has been conducted with a LLR using triangle kernel.

formal RD estimator gives a discontinuity estimate of −0.021 with a standard error of 0.030.[32] Hence, the RD estimate is statistically insignificant.

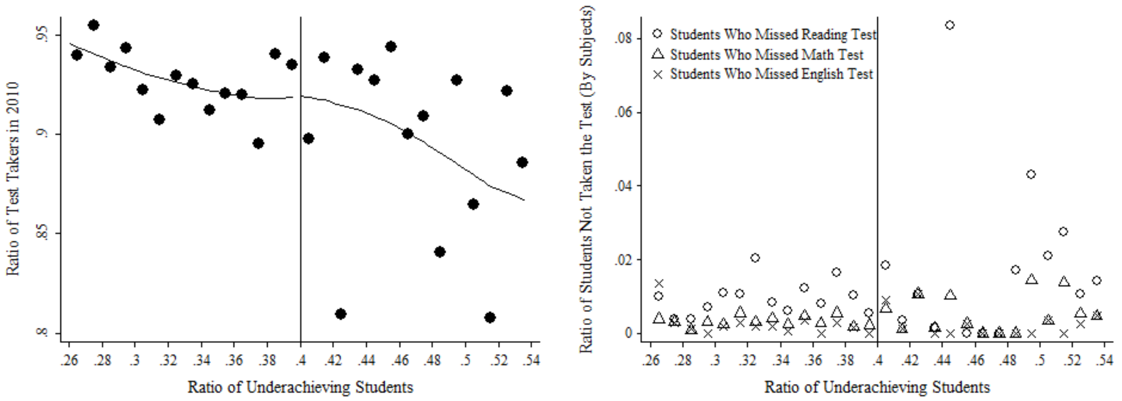
Although the aforementioned arguments do not "prove" that academic performance of students in non-SINAI-designated schools who dropped out is similar, in expectation, with those in SINAI-designated school, I argue that it is highly likely that, on average, they are similar.

Another issue to address in this study is when school administrators and teachers engage in strategic behaviors before, during, or after treatment periods. I have presented some arguments in Section 2.5 that these are unlikely in pre-treatment periods. Thus I show in this section, some data to check two possible strategic behaviors that school administrators or teachers may engage in during or after treatment periods.

Suppose school A has received fund in 2009 because the ratio of underachieving students in school A has exceeded the 40% cutoff. Since the usage of the fund will

[32] The RD estimator has been conducted via the LLR with a bandwidth of 0.1 and a triangle kernel. In addition, every discontinuity estimate is statistically insignificant under any bandwidth choices.

Figure 2.8: Ratio of Test-Takers and Ratio of Students who Partially Missed Tests



Note: The left column presents the density of the ratio of test-takers in NAEA 2010 depicted with a bandwidth of 0.14 and a binwidth of 0.01. The data is retrieved from the SIW. The right column presents the scatterplots of the ratio of students who missed the test for each subject.

be monitored and audited by the government based on the results in the NAEA in 2010, this school has an incentive to strategically behave in NAEA 2010. Two possible behaviors can be identified using the current dataset. First, school A may discourage students to come to school on the test date. If this is the case, then it is likely that we observe a spike in the density of the ratio of test-takers near the threshold.

In the left panel of Figure 2.8, I present the density of the ratio of test-takers in NAEA 2010 plotted against the running variable. The data points do not display any discernible spikes near the threshold, and most of the ratios of test-takers are approximately above 90%. In fact, the LLR estimate of the discontinuity is -0.016 with a standard error of 0.017.[33] Thus, I contend that schools do not engage in this kind of strategic behaviors.

The other behavior may be discouraging students from taking certain subjects. To give an example, suppose an academically poor student shows up on the test date. To improve overall school performance, school A might engage in preventing these students from taking certain subjects. In the right panel of Figure 2.8, in order to test whether this is true, I plot the density of the ratio of students who missed

[33] The RD estimate has been obtained from the bandwidth of 0.1 and a triangle kernel. The discontinuity estimate and the statistical significance rarely change when derived from other bandwidth choices.

the test by subject. As shown on the graph, there are few students who missed any of the three subjects. Most ratios are less than 2%. Hence, I argue that schools do not engage in this type of behavior.

Therefore, given the above arguments, I believe that this study does not suffer from the possible bias incurred by attrition and strategic behaviors.

2.8 Conclusion

To test whether "rewarding" poor-performing schools is beneficial for students' academic achievement, this study makes use of South Korea's unique quasi-experimental setting in which the provision of categorical school funding is determined via simple eligibility cutoff. Because of the discontinuous nature of the eligibility cutoff, I use the sharp RDD to estimate the effect of the treatment.

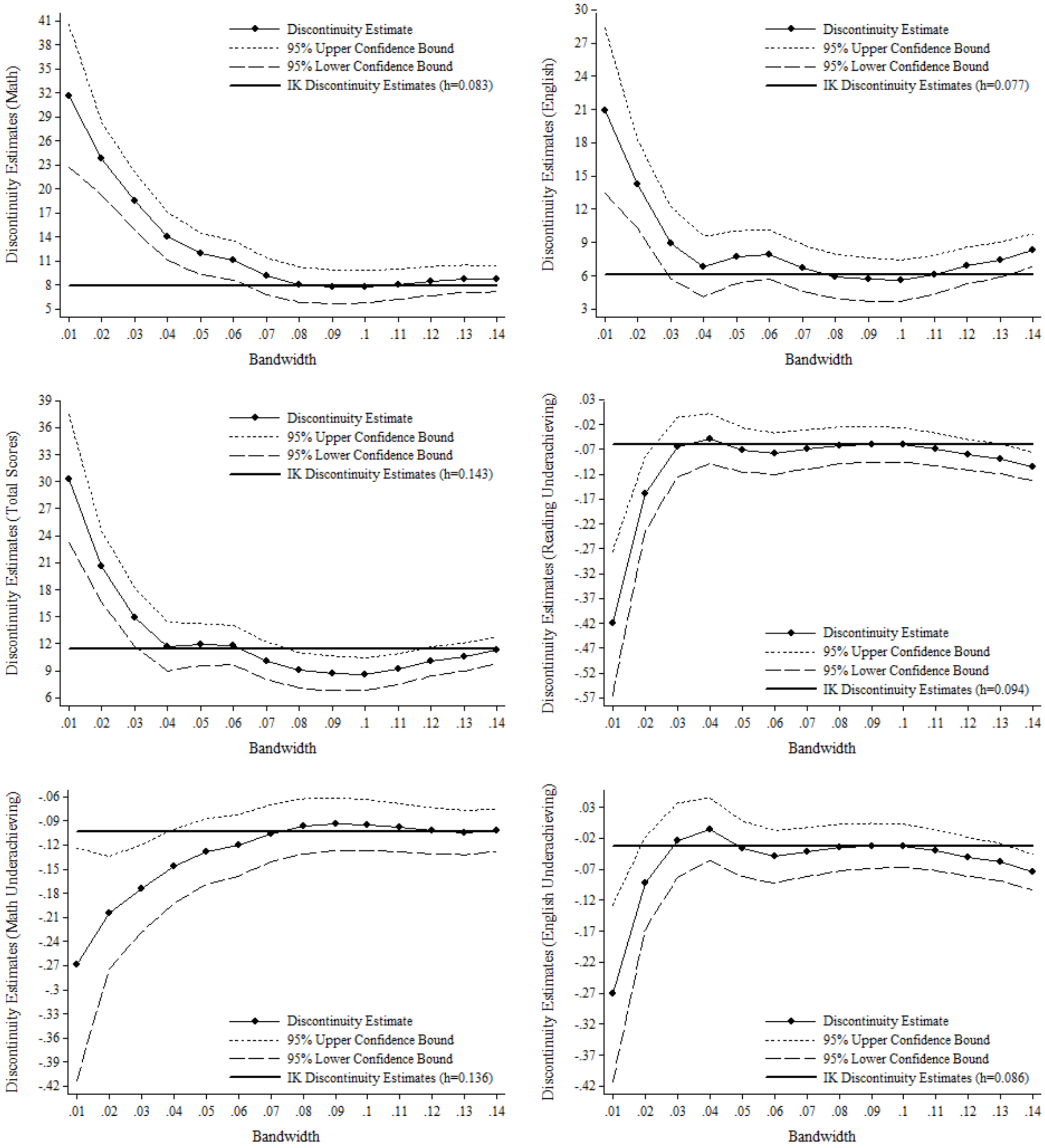
By carefully checking the validity of the RDD and executing the RD estimator, I find that rewarding underachieving schools significantly improved the overall academic performance of students. On average, students in SINAI-designated schools scored 7 to 10 percentile points higher in every subject than those in non-SINAI-designated schools. Furthermore, the ratio of underachieving students decreased by 5 to 10 percentage points in SINAI-designated schools compared to non-SINAI-designated schools.

One limitation of this study is that since student IDs are not linked between NAEA 2009 and NAEA 2010, I was unable to detect whether the sample is missing at random. However, since the attrition rate is around 10%, and due to the nature of the institutional background together with a unique incentive mechanism adopted in South Korea, I believe this study suffers less from problems of attrition or strategic behaviors committed by school administrators and teachers. Consequently, I assert that the estimated treatment effect is true and accurate. That is, I conclude that rewarding poor-performing schools "causally" promoted academic achievement of students.

As previously mentioned, the United States currently operates a federal grant program named SIGs under the Elementary and Secondary Education Act of 1965. I believe that the results presented in this study is in favor of the effect of the SIGs on school performance and provides helpful policy implications for effectively designing the school accountability system.

2.9 Chapter 2: Appendix

Figure 2.9: Sensitivity of RD Estimates



Note: Figures illustrate RD estimates derived from varying bandwidths. IK denotes the RD estimates based on the proposed bandwidth (h) from Imbens and Kalyanaraman (2012).

3 Experimental Estimates of the Distributional Effects of Ability Tracking on Students' Achievement

3.1 Introduction

Academic tracking (also known as ability grouping) is one of the widely used low-cost educational interventions that are aimed at promoting students' academic achievement. Academic tracking is defined in two ways. In Europe, for example, tracking refers to a bilateral educational system in which students are divided into either a general or a vocational educational system. In the U.S., however, academic tracking usually refers to the practice of grouping students within the school or classroom based on students' test scores or prior academic achievement. In the literature, the former is normally referred to as tracking whereas the latter is usually termed as ability tracking. Since this study deals with "ability tracking," I note that tracking in this study refers to ability tracking.

The rationales behind practicing ability tracking are mainly twofold. The first rationale is the peer effect. Provided that students learn better when they are surrounded by "good" students, it is likely that they benefit more from tracking. The other rationale is the possibility of an efficient use of school resources such as teachers. For instance, it is probable that teachers' level of fatigue decreases when they teach a class with a small variance of students' achievement because teachers can tailor their class materials with more ease compared to the situation in which they face students of different academic levels.

The effect of ability tracking is not unanimously agreed upon by reseachers. If students benefit from high-achieving students, then students who are in the low-achieving track will not benefit from tracking. As a consequence, it may worsen the inequality of academic achievement. Moreover, being placed in the low-achieving track may induce students to feel a sense of inferiority, and this may discourage the student from actively engaging in learning. Consequently, some argue that tracking will mostly benefit high-achievers. Contrarily, some contend that both high-achievers and low-achievers will benefit from tracking. Hence, the benefits of ability tracking should be tested empirically.

However, assessing the causal effect of ability tracking is easier said than done for two main reasons. First, as is normally the case in most of the empirical studies, there may exist an endogenous selection into the school in which ability tracking is implemented. The problem of selection bias can be obviated by comparing students within the school. However, there still exists an omitted variable bias that plague the estimate of the tracking effect. The first problem can be surpassed by randomly assigning students to schools that track and schools that do not track. Or, one may randomly assign students in tracked and untracked classes within the school. Even if students are randomly assigned to schools or classes, however, disentangling the effect of tracking is still challenging because each class or school might be different in important aspects that affect students' achievement.

Second, the variable of interest (i.e., the tracking indicator), may suffer from measurement error issues. Because of the abstract nature of the definition of ability tracking, the definitions of tracking used in the previous literature often differ dramatically among schools, and accordingly, it is difficult for one to interpret and compare the estimates of previous literature.[34]

Because of the problems mentioned above, it is understandable that the existing research is not in agreement as to the effect of ability tracking on students' academic achievement. The institutional setting in Seoul, the capital city of South Korea, on the other hand, is favorable for overcoming the problems mentioned above for several reasons. First, students are randomly assigned to high schools within the school districts located in Seoul, and consequently, students do not self-select the schools that implement ability tracking. Second, since the Seoul Metropolitan Office of Education has adopted the "high school equalization" policy, schools located in Seoul are highly homogeneous in terms of their school resources (class size, pupil-to-teacher ratio, teacher quality) and the curriculums through which students learn because these are all controlled by the Office of Education. Third, this study does not suffer from the ambiguous nature of the definition of ability tracking because the way in which tracking is used is similar across schools.

Therefore, by taking advantage of the institutional setting in Seoul, I present experimental estimates of the mean effects as well as the distributional impact of ability tracking on students' performance. An empirical analysis reveals two main

[34] Figlio and Page (2002) provide an excellent illustration of the problems caused by the ambiguity of defining ability tracking.

points. First, using the conventional regression, I find that students who attend schools that implement ability tracking benefit academically compared to those who attend untracking schools. For mathematics, students in tracking schools scored higher by 5.7 percentile points than those in untracking schools. I also find that placing students into schools that implement tracking in reading raises their scores by 3.3 percentile points. Finally, although not statistically significant, the estimated treatment effect is 4.8 percentile points for English tracking. These estimated results also highlight the fact that the effect of ability tracking varies by subject.

Second, by making use of the quantile regression method, I find that both the students in the bottom quantiles as well as the students in the top quantiles benefit from attending schools that track students. To be more specific, the estimated quantile treatment effects are similar, in general, across all quantiles, implying that all students benefit from ability tracking. Hence, contrary to the established literature that claims the negative effects of tracking on disadvantageous students, ability tracking may, in fact, be beneficial for low-achieving students.

3.2 Related Literature

Previous literature on the effect of tracking can be classified into two categories. One is related to the former tracking mentioned in Section 3.1. The other category studies the effect of ability tracking. Since the current study focuses on estimating the effect of ability tracking, I present literature reviews that are related to ability tracking.[35]

Many of the previous studies that empirically analyze the effect of tracking have been conducted by sociologists and psychologists. These studies have made use of the ordinary least squares (OLS), analysis of variance, and matching methods. Slavin (1987) and Slavin (1990) provide comprehensive meta-analysis of these literature conducted prior to 1990. According to his studies, tracking does not seem to be an effective way of increasing students' achievement. As can be seen from the tables presented in the studies, many produce inconsistent results as to the effect of tracking. This is well expected because most of the studies included in Slavin (1987) and Slavin (1990) use observational data, and many studies fail to account for endogeneity issues. There are, in fact, six and seven randomized field experiments conducted at elementary and secondary schools, respectively. However, these involve a small sample size (52 to 603 students), and none of these studies clustered standard errors, which is an essential part of statistical inference nowadays. In sum, even though there are many studies, the jury is still out as to the question of whether ability tracking is desirable for students' achievement.

Beginning from the early 1990s, researchers began paying attention to endogeneity issues by making use of either the selection on observable or unobservable designs. Studies that make use of the selection on observable designs are Hoffer (1992), Argys, Rees and Brewer (1996), Betts and Shkolnik (2000), and Zimmer (2003). The key assumption underlying the selection on observable designs is that the variable of interest is as good as randomly assigned conditional on observable variables. However, in the context of tracking studies, this assumption is hardly testable. Furthermore, it is extremely difficult to justify that the assumption, indeed, holds.

[35] Some of the works that study tracking (not ability tracking) are Meghir and Palme (2005), Hanushek and Wöβmann (2006), Pischke and Manning (2006), and Pekkarinen, Uusitalo and Kerr (2009).

On the other hand, Figlio and Page (2002) and Lefgren (2004) use the instrumental variable (IV) method to causally estimate the effect of tracking, and Lavy, Silva and Felix (2009) use a fixed-effect approach to come up with a reliable estimate of the effect of tracking. All of these methods are part of the selection on unobservable designs, and once one finds exogenous variations in the treatment variable, then the estimated effect will be likely to be causal. As pointed out by Bound, Jaeger and Baker (1995), however, when the IV is weakly correlated with the variable of interest, the estimate will be biased. In Figlio and Page (2002) and Lefgren (2004), the R^2 estimated from the first-stage is quite low, and accordingly, these studies may suffer from the weak instrument issues. The problem of weak instruments may not bias the estimate if the exclusion restriction truly holds. However, it is hard to convincingly argue that the IVs used in both studies are not correlated with the outcome variables. In any case, Figlio and Page (2002) find quite a large positive effect of tracking, in particular, for low-ability students. On the other hand, Lefgren (2004) and Lavy, Silva and Felix (2009) find small effects.

In terms of solving the selection bias issue, the randomized field trial is superb. As noted previously, there are some randomized experiments conducted in the United States. However, since all of the studies use a very small number of students, the external validity is limited. Besides, none of the studies correct for within class or school correlations, and accordingly, the estimated effects in these studies may not be statistically significant. Contrarily, Duflo, Dupas and Kremer (2011) study a large-scale randomized experiment conducted in Kenya which involves about 5,800 first grade students. In the experiment, students are randomly tracked based on their initial achievement, and the estimated mean-effects reveal that students in tracked schools performed better than the those in untracked schools. In addition, they find that low-ability as well as high-ability students benefited from being tracked. Another notable point to note from this study is that they find evidence of teachers being tailoring their teaching methods. Also, they find that small changes in the peer group affect students' achievement. Accordingly, this study strongly suggests that the mechanisms through which tracking is benefiting students are via peer effects as well as teacher effects.

Although Duflo, Dupas and Kremer's (2011) study provides a compelling evidence of the positive effect of ability tracking, the result from the randomized experiment is no panacea. As pointed out by the authors, since the study is conducted in a developing country, they are cautious to extrapolate the results to developed countries. Betts (2011) further notes that since the experiment was conducted by hiring contract teachers, the external validity may be limited not only for other countries but also for Kenya itself. Lastly, one must note that, in general, ability tracking is more prevalent at the secondary education level. Hence, since the above study uses first grade elementary students, the results may not be applicable to the case of secondary school students.

To the contrary, since this study uses a large-scale randomized social experiment conducted at secondary schools in the highly urbanized city of South Korea (Seoul), results from this study may have more policy implications for educational practices in the United States. Therefore, I contribute to the existing studies by providing the experimental estimates of the effect of ability tracking on students' academic achievement.

3.3 Simple Model

The most compelling rationale behind the implementation of ability grouping is that it induces peer effects that are favorable for students' academic achievement.[36] Previous research, however, pays less attention on the theoretical mechanisms of the effect of ability tracking on students' achievement. The study by Epple, Newlon and Romano (2002) is the first to provide theoretical models on how ability tracking may affect students' achievement. According to their work, given that students benefit from being surrounded by high-achieving peers, students in the high-achieving track benefit from ability tracking. Students in the low-achieving track, however, do. As a consequence, they show that tracking interferes with the learning of disadvantageous students, thereby promoting inequality in students' academic achievement. On the other hand, Duflo, Dupas and Kremer (2011) present models that show mechanisms in which all students may benefit from tracking. In particular, they show that when the incentive mechanism targeted at teachers is properly designed, all students benefit from tracking.

In this section, by applying the model developed by Bénabou (1996) and Brunello and Checchi (2007), I present a very simple model showing that the overall achievement of students in tracking schools can be higher than that of untracking schools.[37] According to the well-documented educational production function, the academic achievement of student i is determined by the following production function:

$$A_{it} = f(A_{it-1}, P_t, F_{it}, S_t, T_t, \alpha_i, \xi_{it}),$$

where the variables in the function correspond to prior academic achievement, peers, family background, school resources, teachers, innate ability, and the error term, respectively.

In order to illustrate the effect of ability tracking on students' achievement induced by the peer effect, I assume, for the sake of simplicity, that students' achievement is solely determined by one's peers following the logic of Bénabou (1996) as

[36] Sacerdote (2011) and Epple and Romano (2011) provide excellent reviews on peer effects in education.

[37] Bénabou (1996) presents the model to show the neighborhood effects, and he notes that these include peer effects, role models, and social networks. Brunello and Checchi (2007) link the model of tracking to family background and how tracking affects human capital accumulation.

follows:

$$A_i = f(P_t) = (\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho})^{-\frac{1}{\rho}}. \tag{6}$$

Equation (6) is the CES (constant elasticity of substitution) function that captures the spillover effects generated by the mix of peers in schools. I am assuming that students belong to one of the two following types: high-capability students (c_H) or low-capability students (c_L). Also, suppose that the total number of students in a school is N, the total number of high-capability students is n_H, and the total number of low-capability students is n_L, with $n_H + n_L = N$. In Equation (6), ϕ_1 denotes the ratio of high-capability students in a school (n_H/N), and ϕ_2 denotes the ratio of low-capability students in a school (n_L/N).

Note that when $\rho < -1$, individual levels of students' achievement are complements, and A_i is convex (proof of convexity is in the appendix of Section 3.10.1). If, on the other hand, $\rho > -1$, then individual levels of students' achievement are substitutes, and A_i, in this case, would be concave (proof of concavity is in the appendix of Section 3.10.1).

Now, suppose the school is implementing ability tracking within the school under the two regimes; high-track and low-track. Here, student i's achievement is determined by

$$A_i = \begin{cases} c_H, & \text{if student } i \text{ is in the high-track} \\ c_L, & \text{if student } i \text{ is in the low-track,} \end{cases}$$

and the sum of achievement of students in the school that exercises ability tracking is given by

$$A_{track} = \sum_{i=1}^{n_H} c_H + \sum_{j=1}^{n_L} c_L$$

for $i \neq j$.

Contrarily, if the school is not practicing ability tracking and the classroom is consisted of students of heterogeneous capabilities, achievement of student i is determined by

$$A_i = (\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho})^{-\frac{1}{\rho}},$$

regardless of whether student i is c_H or c_L. In this case, the total achievement of students in the school that does not track students is given by

$$A_{untrack} = \sum_{i=1}^{n_H}(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho})^{-\frac{1}{\rho}} + \sum_{j=1}^{n_L}(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho})^{-\frac{1}{\rho}}$$

for $i \neq j$.

Whether A_{track} is bigger than $A_{untrack}$ depends on whether individual levels of students' achievement are complements or substitutes. If $\rho < -1$, students' achievement are complements, and because A_i is convex, we have

$$\begin{aligned}
A_{untrack} &= \sum_{i=1}^{n_H}(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho})^{-\frac{1}{\rho}} + \sum_{j=1}^{n_L}(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho})^{-\frac{1}{\rho}} \\
&= n_H(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho})^{-\frac{1}{\rho}} + (N - n_H)(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho})^{-\frac{1}{\rho}} \\
&\leq n_H(\phi_1 c_H + \phi_2 c_L) + (N - n_H)(\phi_1 c_H + \phi_2 c_L) \\
&= N(\phi_1 c_H + \phi_2 c_L) \\
&= N\left[\frac{n_H}{N} c_H + \left(\frac{N - n_H}{N}\right) c_L\right] \\
&= n_H c_H + n_L c_L \\
&= \sum_{i=1}^{n_H} c_H + \sum_{j=1}^{n_L} c_L \\
&= A_{track}.
\end{aligned}$$

Likewise, if $\rho > -1$, then students' achievement are substitutes, and A_i is concave. And following similar step as the above, we have the following:

$$A_{untrack} \geq A_{track}.$$

Thus, whether ability tracking is beneficial for students' achievement depends on whether the students' abilities are complements or not, and it is a matter of empirical questions. In fact, Hoxby and Weingarth (2005) present various kinds of models that hypothesize the mechanism of peer effects and they empirically find in support of the Boutique and Focus model which assume that students benefit mostly from students with similar abilities which are in favor of the complementarity of students' abilities.

Accordingly, based on the hypothesis that students' abilities are complements, I present experimental evidence of the effect of ability tracking on students' achievement by using the randomized institutional setting in Seoul and determine whether ability tracking, indeed, is more efficient than untracking.

Note, however, that even if we find $A_{track} - A_{untrack} \geq 0$, it is difficult to identify whether the efficiency gain is driven by students in the high-track or those in the low-track. Hence, by conducting the quantile regression, I examine the relative contributions of students by examining the treatment effects estimated for various quantiles of achievement distributions.

3.4 Institutional Background

Students in South Korea spend six years in elementary school, three years in middle school, and another three years in high school. There are mainly two types of high schools in Korea; general high schools and vocational high schools. In Seoul, students who intend to proceed to general high schools are randomly assigned to general high schools. Contrarily, those who intend to enter vocational high schools self-select a vocational high school of their choice. Note that students in other cities are not randomly assigned to high schools. Hence, in this study, I use students who attend general high schools located only in Seoul.

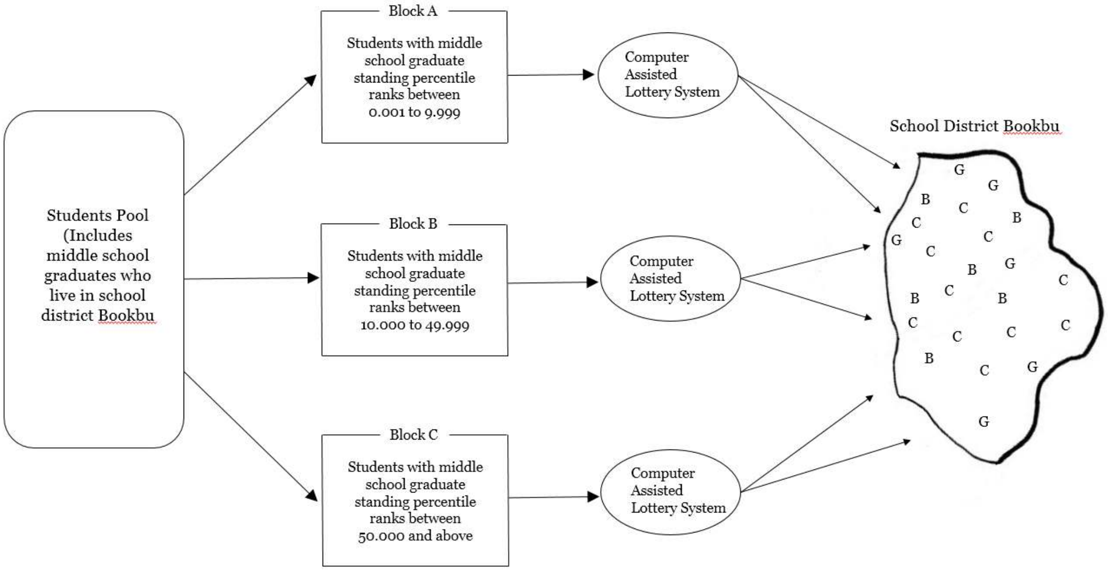
I illustrate a random assignment mechanism using Figure 3.1 (I illustrate using the "Bookbu" district). As can be seen from the figure, middle school graduates are divided into three blocks. Block A includes students whose middle school graduate standing percentile rank is between 0.001 and 9.999. These students are high-performing students. Block B consists of students whose rank is in the range of 10 to 49.999. These students are middle-performing students. Finally, Block C is made up of students whose middle school graduate standing percentile rank is 50 or above. These students are low-performing students.

Within each block, the Office of Education uses a computer-assisted lottery system to assign students to the high schools located in Bookbu district. In the district, there are a total of 23 schools. Among them, 11 schools are coeducational schools, 6 are boys-only schools, and 6 are girls-only schools. By way of randomly assigning students within the block, the Office of Education ascertains that the ratio of high-, middle-, and low-performing students are equally represented among schools in this district. The Office of Education implements this kind of assignment as part of the "high school equalization" policy.[38] In any case, because students are randomly assigned to high schools within the district, the chance of a student being assigned to a tracking or untracking school is determined randomly. As a consequence, I am able to estimate the treatment effect without the problem of a selection bias issue.

Note, however, that even if students are randomly assigned to schools, the estimated effect may still fail to capture causality if the baseline school-level covariates

[38] Using the middle school graduate standing percentile rank to equalize the level of students' achievement across schools within the district is appropriate because in Seoul, every student is randomly assigned to a middle school after their elementary education, and therefore, the performance level of middle schools within each district is highly homogeneous.

Figure 3.1: High School Randomization Mechanism


Students Pool (Includes middle school graduates who live in school district Bookbu
Block A
Students with middle school graduate standing percentile ranks between 0.001 to 9.999
Block B
Students with middle school graduate standing percentile ranks between 10.000 to 49.999
Block C
Students with middle school graduate standing percentile ranks between 50.000 and above
Computer Assisted Lottery System
Computer Assisted Lottery System
Computer Assisted Lottery System
School District Bookbu
G
G
B
C
B
C
C
G
C
B G
C
C
B B
C C
C C
B
C G
G

Note: C corresponds to coeducational schools, B stands for boys-only schools, and G indicates girls-only schools.

of tracking schools and untracking schools differ in important aspects. Due to the high school equalization policy mentioned above, however, the Office of Education makes sure that the school characteristics such as curriculums, class size, pupil-to-teacher ratios, and teacher salaries do not differ between schools. If the Office of Education does not engage in equalizing these characteristics within each district, parents or students may not abide by the random assignment policy.

All in all, the institutional setting in Seoul is favorable for analyzing the effect of ability tracking on students' achievement not only because students are randomly assigned to high schools but also because school-level characteristics are highly homogeneous across schools.

3.5 Econometric Methods

In order to provide a causal estimate of the effect of tracking, let T_s be an indicator variable equal to 1 if school s tracks students, and 0 otherwise. Since students are randomly assigned, it solves for the endogenous selection into tracked schools. Note, however, that even if students are randomly assigned to schools, one may fail to elicit a causal treatment effect if school-level characteristics are different between tracking and untracking schools in important dimensions. In Seoul, there are three types of schools; coeducational, boys-only, and girls-only schools. It is well perceived in previous literature that students are affected by the gender of their peers (e.g., Whitmore, 2005; Lavy and Schlosser, 2011). Hence, to precisely estimate the effect of tracking, I compare students within each school type. Under this setting, I run the following regression:

$$A_{isd} = \alpha + \beta_1 T_s + \beta_2 D_{isd} + \gamma_d + \delta_s + \varepsilon_{isd}, \tag{7}$$

where A_{isd} is the percentile rank of student i in school s located in school district d. D_{isd} denotes students' gender. Note that since students are randomly assigned within each school district, I include school district fixed effects (γ_d). δ_s denotes the school type fixed effects, and ε_{isd} is the stochastic error term.

Note that Equation (7) estimates the mean effects of tracking. To estimate the distributional impact of tracking, I run the quantile regression initially developed by Koenker and Bassett (1978) and Firpo (2007). Here, I run the following:

$$A_{isd}(q) = \alpha + \beta_1(q) T_s + \beta_2(q) D_{isd} + \gamma_d + \delta_s + \varepsilon_{isd},$$

where $q \in (0, 1)$ denotes the quantile, and I estimate a treatment effect $\beta_1(q)$ for each quantile q.[39]

For statistical inference, I account for the serial correlation when calculating standard errors. As demonstrated by Moulton (1986), Bertrand, Duflo and Mullainathan (2004), and Donald and Lang (2007), failing to account for the within-group dependence in calculating standard errors will easily underestimate the true standard errors. In this study, since students in the same school are likely to be

[39] In order to execute the quantile regression proposed by Firpo (2007), I use the STATA command developed by Frölich and Melly (2010).

correlated with each other, I correct for the dependence by clustering at the school level.

Note, however, that the conventional cluster-robust standard errors are asymptotically justified provided that the number of clusters goes infinity. Cameron, Gelbach and Miller (2008) point out that with a small number of clusters (5 to 30), the asymptotic tests can over-reject when used with the conventional cluster-robust standard errors. In this study, the number of schools that track students vary by subject (11 to 38 schools), and as a consequence, this study may suffer from the point made by Cameron, Gelbach and Miller (2008) due to the small number of clusters. In order to solve this issue, I estimate standard errors using the wild cluster bootstrap-t procedure proposed by Cameron, Gelbach and Miller (2008), which they find, using the Monte Carlos simulations as well as from the real data, that the procedure works very well even with a small number of clusters (as few as 6 clusters). Therefore, in the analysis to follow, I present the conventional cluster-robust standard errors as well as the p-values retrieved from the wild cluster bootstrap-t procedure.

3.6 Data and Validity Check

3.6.1 Data

This study uses administrative records of students' test scores in the National Assessment of Educational Achievement (NAEA) exam administered by the Ministry of Education in 2009. In 2009, all high school students in the first grade took this test. The purpose of the NAEA is to measure the student's overall level of achievement and to see which students do not meet the basic academic standards. Students are tested on five subjects; reading, math, English, social studies, and science. In 2010, the administrative data on the NAEA was released for the first time to those researchers who submitted a formal application to the Ministry. Once the Ministry receives the application, a judging committee is formed consisting of both outside and inside members to determine whether to disclose the data to the applicant. The NAEA dataset contains students' test scores and some student-level baseline covariates such as students' gender. It also contains answers to the survey questions administered upon students and principals.

The variable on whether schools track or do not track is also coded in the dataset. It asks whether the school tracks students in each of the five subjects. It turns out that no schools in the sample implement ability tracking in social studies and science. Furthermore, the number of schools that implement tracking varies by subject. That is, some schools track math but not English. As a consequence, the sample used in the analysis differs by subject. In the appendix (Section 3.10.2), I present a series of sample restrictions conducted for each subject.

In order to properly test the validity of the randomization, I also make use of the administrative records of school-level data posted on the government website.[40] The website contains rich sets of school- and student-level characteristics, and I use these information to test the balance in the baseline school-level covariates as well as student-level covariates.

Table 3.1 presents descriptive statistics for the sample used in the analysis. For the math subject, the number of tracking schools exceeds that of untracking schools. 29 schools track their students and 4 schools do not. There are a total of 11,992 students in tracking schools and 1,567 students in untracking schools. I also show

[40] www.schoolinfo.go.kr.

Table 3.1: Descriptive Statistics (By Sample)

	School	
Variable	Untrack	Track
A. Mathematics		
Number of schools	4	29
Number of students	1,567	11,992
Ratio of test-takers	0.972	0.983
Ratio of mathematics test-takers among test-takers	0.997	0.996
B. Reading		
Number of schools	29	9
Number of students	14,577	4,860
Ratio of test-takers	0.979	0.981
Ratio of reading test-takers among test-takers	0.996	0.997
C. English		
Number of schools	2	9
Number of students	1,127	4,307
Ratio of test-takers	0.986	0.988
Ratio of English test-takers among test-takers	1.000	0.996

the ratio of test-takers as well as the ratio of mathematics test-takers among those who have taken any of the subjects. As can be seen from the table, the ratio is almost 100%.

For reading, there are a total of 38 schools, and among the 38, 29 schools do not implement ability tracking and 9 schools do. The number of students are 14,577 and 4,860 for untracking and tracking schools, respectively. As with the math tests, the ratio of test-takers is close to 100%.

Compared to reading and mathematics, there are only 11 schools that have variations in the English tracking indicator within the school district. Among 11 schools, 2 schools do not track students and 9 schools do. The sample size for untracking schools is 1,127 and 4,307 for tracking schools, and the ratio of test-takers are similar to those of mathematics and reading.

3.6.2 Validity Check

In order to test the validity of the randomization, I present tests of balance in predetermined covariates. For student-level covariates, I use ratios of first-year high school students in the upper-, middle- and lower-rank which is determined by the middle school graduate standing percentile ranks. I also use the ratio of students in the first year of high school that receive financial aid from the government in the form of a tuition reduction or through fellowships.[41] Governmental support is provided to those whose family income is below the threshold set by the government. This ratio is a good proxy for students' family background. To complement this information, I use the ratio of students who receive lunch support from the government. These students are either from poor families or are recommended by their school. Note that in order to accurately test the balance in student-level covariates, it is necessary to use the ratio of students in the first year. The data on free lunch, however, was not available from the Office of Education as it does not keep track of the ratio of students receiving free lunch by grade. Hence, I use the ratio that includes students of all grades. I admit that the ratio does not accurately test the balance in the ratio of students receiving lunch support. However, I believe the result would not differ to a great extent even if one uses the ratio of first year high school students. For the school-level baseline characteristics, I test balance in class size as well as pupil-to-teacher ratio. The two variables, as a matter of course, correspond to the first grade.

In table 3.2, I present estimates obtained from running a regression of each variable mentioned above on a dummy variable indicating whether the school implements ability tracking. Since schools that implement ability tracking differs by subject, I run the regression separately for each subject. Moreover, since students are randomly assigned to high schools within their school district, all regressions are conducted with school district fixed effects. Finally, since students are compared within the same school type, all regressions are conditional on school types. Note that in the table, Ω denotes the middle school graduate standing percentile rank. The lower the number, the higher the rank.

Panel A corresponds to schools which track mathematics. The test reveals that

[41] I appreciate the public officials at the Office of Education for generously providing me with the data on these ratios.

Table 3.2: Tests of Balance in Baseline Student- and School-Level Covariates

Dependent Variable	Tracks (1=yes)	Standard Errors	Constant	No. of Schools
	A. Ability Tracking in Mathematics			
Students with $\Omega < 10\%$	−0.006**	0.003	0.103	33
Students with $10\% \leq \Omega < 50\%$	−0.003	0.003	0.461	33
Students with $\Omega \geq 50\%$	0.009*	0.005	0.436	33
Ratio of students receiving financial aid	−0.105*	0.056	0.253	31
Ratio of students receiving lunch support	−0.007	0.011	0.046	33
Class size	−0.165	0.666	39.267	33
Pupil-to-teacher ratio	0.100	0.880	17.778	33
	B. Ability Tracking in Reading			
Students with $\Omega < 10\%$	0.002	0.001	0.130	38
Students with $10\% \leq \Omega < 50\%$	−0.003	0.002	0.492	38
Students with $\Omega \geq 50\%$	0.000	0.002	0.378	38
Ratio of students receiving financial aid	0.012	0.027	0.104	37
Ratio of students receiving lunch support	0.002	0.004	0.018	38
Class size	−0.859**	0.343	37.758	38
Pupil-to-teacher ratio	0.032	0.473	18.795	38
	C. Ability Tracking in English			
Students with $\Omega < 10\%$	0.000	0.002	0.136	11
Students with $10\% \leq \Omega < 50\%$	−0.001	0.002	0.508	11
Students with $\Omega \geq 50\%$	0.002	0.003	0.356	11
Ratio of students receiving financial aid	−0.085	0.055	0.227	11
Ratio of students receiving lunch support	−0.004	0.017	0.052	11
Class size	−0.144	0.916	38.240	11
Pupil-to-teacher ratio	0.219	0.527	17.844	11

Note: All regressions are conditional on school district fixed effects and school types. Ω denotes a middle school graduate standing percentile rank. Ratio of students receiving financial aid include students who receive governmental support in the form of tuition reductions or fellowships. Ratio of students receiving lunch support include students in all grades (data for the ratio of first grade students is not available from the Office of Education). Coefficients for the constant corresponds to the mean of non-tracking schools. There are two missing values in Panel A, and one missing value in Panel B for the ratio of students receiving financial aid.

the ratio of upper-ranked students (i.e., $\Omega < 10\%$) is 0.6 percentage points lower for schools that track students. Although it is statically significant at the 5% level, it is not economically significant because 0.6 percentage points correspond to less than one student. The ratio of middle-ranked students is not statistically significant whereas for lower-ranked students, it is significant at the 10% level. Again, however, it is not economically significant because the estimated coefficient is only 0.9 percentage points. The ratio of students receiving financial support in tracking schools is 10.5 percentage points lower. Although the estimate is slightly significant, I argue that this small difference in the ratio has little impact on the analysis.[42] As can be expected from the fact that the government strongly engages in controlling the class size as well as student-to-teacher ratio, there is little difference between tracking and untracking schools for these two variables.

Turning to the schools which track reading, class size is the only variable that yields statistical significance. Note, however, that the estimated coefficient is -0.859. It implies that the difference in class size is less than one student among two school types, and accordingly, I conclude that class size is not different. Consequently, baseline student- and school-level covariates are equally balanced between schools which track and schools which do not track.[43] Finally, in Panel C, I test balance in baseline covariates between schools that track English and schools that do not. Compared to other subjects, none of the variables are statistically and economically significant. Hence, students' prior academic achievement, family background, and school resources are equally balanced between the two schools.

All in all, in this study, I contend that the randomization is valid and predetermined student- as well as school-level covariates are equally balanced between tracking and untracking schools.

[42] There are two missing values. For two schools, I could not obtain the data from the Office of Education.

[43] For one school, I could not obtain the data for the ratio of students receiving financial aid.

3.7 Estimation Results

In this section, I present estimation results for mean treatment effects and quantile treatment effects. All regressions are conditional on school district fixed effects and school type indicators. The dependent variable is the percentile rank that students received in NAEA 2009. Table 3.3 corresponds to the results of the mean effects of ability tracking. Panel A shows the estimation results for tracking in mathematics. With respect to math tracking, there are a total of 33 schools that have a variation in the treatment indicator within the school districts. Within these schools, there are a total of 13,559 students. The estimated mean effect is 5.743 percentile points implying that students who are in a school that tracks mathematics scored higher in the NAEA exam. In terms of the percentile rank calculated within the sample, the effect is 5.684 percentile points which is close to the percentile rank calculated at the national level. Both coefficients are statistically significant under the cluster-robust standard errors.

As noted previously, however, since the number of clusters is 33, the usual cluster-robust standard errors may easily over reject the null hypothesis. To solve this issue, I estimate standard errors using the wild cluster bootstrap-t method (Cameron, Gelbach and Miller, 2008). As can be seen from the table, statistical significance of the estimated coefficients decreased when evaluated with the wild cluster bootstrap-t method. However, the coefficient is still significant at the 10% level.

In Panel B, I estimate the mean treatment effect of tracking in reading. The sample size is bigger than that of Panel A. The sample consists of 38 schools with a total of 19,437 students. The magnitude of the tracking effect is equal to 3.328 and 3.281 percentile points implying that tracking is beneficial for students' achievement in reading. As with Panel A, I present both the cluster-robust standard errors as well as p-values estimated from the wild cluster bootstrap-t method. Because of the relatively large number of clusters (38 clusters), the statistical significance obtained from the conventional cluster-robust standard errors and the wild cluster bootstrap-t method are similar which coincides with the arguments made in the previous literature (Bertrand, Duflo and Mullainathan, 2004; Donald and Lang, 2007; Cameron, Gelbach and Miller, 2008).

Lastly, in Panel C, I present estimation results for tracking in English. In 2009, there were variations in 11 schools within the school district (5,434 students). Con-

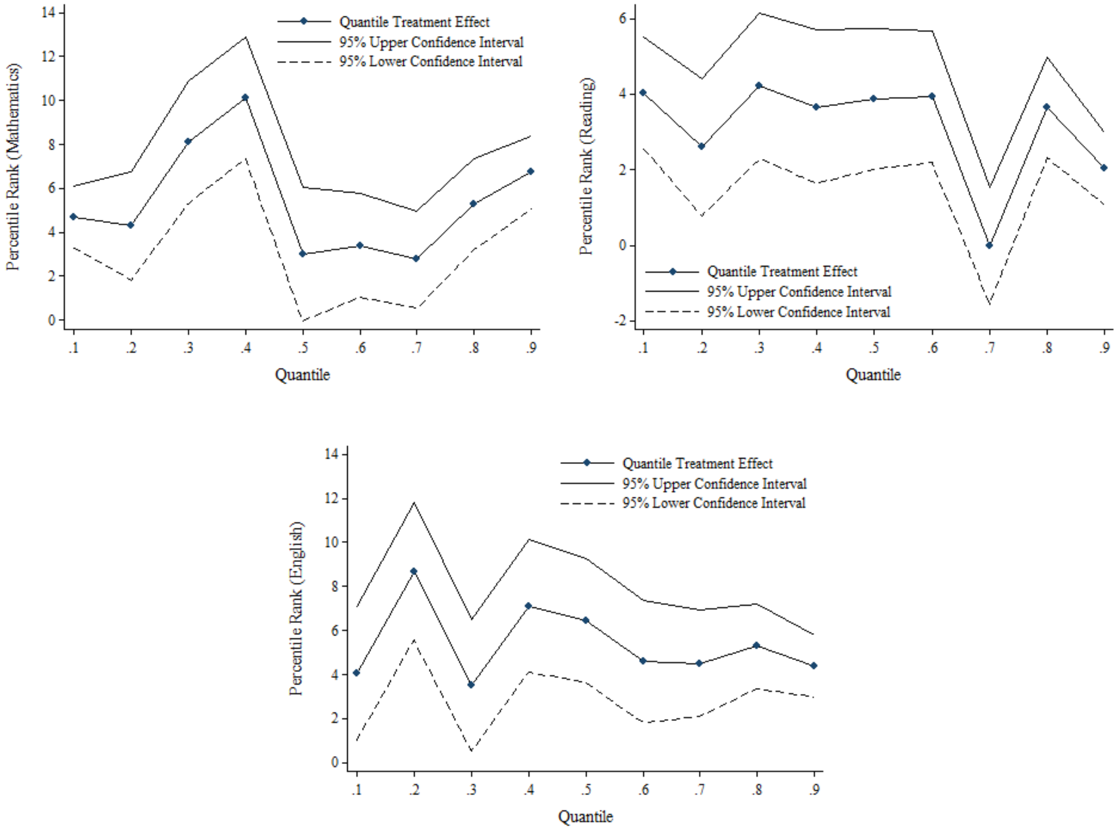
Table 3.3: Mean Effects of Ability Tracking

	Percentile	
Independent Variable	Rank I	Rank II
A. Mathematics		
Tracks students in mathematics (1=yes)	5.743	5.684
	(2.116)	(2.044)
	[0.084]	[0.072]
Female (1=yes)	2.451	2.041
	(0.608)	(0.606)
	[0.000]	[0.002]
Number of students	13,559	13,559
B. Reading		
Tracks students in reading (1=yes)	3.328	3.281
	(1.302)	(1.020)
	[0.028]	[0.020]
Female (1=yes)	10.412	10.188
	(0.968)	(0.981)
	[0.000]	[0.000]
Number of students	19,437	19,437
C. English		
Tracks students in English (1=yes)	4.805	4.971
	(2.947)	(3.227)
	[0.182]	[0.208]
Number of students	5,434	5,434

Note: All regressions are conducted conditional on school district fixed effects and school type indicators. Robust standard errors clustered at the school level are in parentheses, and *p*-values from the wild cluster bootstrap-*t* are in brackets (Null hypothesis, $\beta_{track} = 0$, has been imposed when estimating the *p*-values). For bootstraps, I use 1,000 bootstraps. There are 33 clusters for Panel A, 38 clusters for Panel B, and 11 clusters for Panel C. Percentile Rank I uses the ranking calculated at the national level. Percentile Rank II uses the ranking calculated at the sample level.

trary to Panel A and Panel B, I do not include a dummy variable for denoting female students because all the schools used in the analysis are girls-only schools. According to Table 3.3, I find that placing a female student at the school that tracks English raises her percentile rank in English exams by 4.805 at the national level.

Figure 3.2: Quantile Treatment Effects of Ability Tracking



However, the estimated effect is weakly significant. Besides, the significance level further decreases when evaluated with the wild cluster bootstrap-t method. This highlights the limitation of the usual cluster-robust standard errors when the number of clusters are relatively small. As can be seen from Panel C of Table 3.3, one may mistakenly reject the null hypothesis when evaluated with the cluster-robust standard errors.

In sum, the estimated mean treatment effects imply that students in tracking schools benefit academically. The mean effects, however, do not allow us to examine the distributional effects of ability tracking. As documented by previous studies, some argue that tracking only benefits high-achieving students and hinders academic performance of low-achieving students. In order to examine whether this argument is true, I estimate the quantile treatment effects. In Figure 3.2, I plot the estimated quantile treatment effects starting from quantile equal to 0.1 (with an increment of

0.1).

The quantile regression shows that, contrary to the above argument, students at the bottom quantiles also benefit from attending tracking schools. For example, for mathematics, the estimated treatment effect for students in quantile 0.1 is about 5 percentile points, which is close to the estimated effect for students in quantile 0.8. Furthermore, the effect is even higher (around 10 percentile points) for students in quantile 0.4. When estimated with schools that track reading, students at the bottom quantiles also benefit from tracking. In this case, the effect of ability tracking is similar across quantiles in general. Finally, I observe similar patterns of the treatment effect for English tracking.

All in all, the experimental estimates presented above demonstrate three important facts. First, ability tracking, on average, is beneficial for students' academic achievement. Second, rather than exacerbating the inequality of students' achievement, ability tracking, indeed, promotes academic achievement of students at all levels. Third, the estimated treatment effect varies by subject.

3.8 Robustness Check

In order to test the robustness of the analysis presented above, I analyze the effect of ability tracking by focusing on the homogeneous treatment indicator. Not only does the dataset contain information on whether the school tracks students, it also contains information on the number of tracking levels. That is, schools track students in two, three, or more than three levels. Now, within the school district, there are few variations in the number of schools that track students in two levels. Accordingly, I drop schools that track in two levels. Moreover, I drop schools that track students in more than three levels because I do not know the exact number of tracking levels. Hence, to test the robustness of the analysis, I use students who attend schools that track in three levels.[44]

Table 3.4 presents the estimation results obtained from the sample of schools that track students in three levels. Panel A corresponds to mathematics, and the estimated mean treatment effect is 3.882 and 3.701. That is, students who attend schools that track mathematics scored approximately 4 percentile points higher than those who attend schools that do not track mathematics. However, the coefficients are statistically insignificant under the cluster-robust standard errors. Moreover, the p-values from the wild cluster bootstrap-t method are 0.342 and 0.338.

In Panel B, I present results for reading, and I observe that ability tracking in reading raises students' academic achievement by 4.963 percentile points. Furthermore, the estimated coefficients are statistically significant under both the cluster-robust standard errors as well as the wild cluster bootstrap-t method.

I plot, in Figure 3.3, the quantile treatment effects of ability tracking. The left panel of Figure 3.3 corresponds to mathematics and the right panel corresponds to reading. As with Figure 3.2, both poor-performing students and the high-performing students benefit from tracking. Moreover, in general, the patterns of the quantile treatment effects in Figure 3.3 are roughly similar to that of Figure 3.2.

Note that the estimated treatment effects differ between Table 3.3 and Table 3.4 (approximately 2 percentile points) as well as between Figure 3.2 and Figure 3.3. These imply that the way in which the tracking system is oranized has differential impacts on students' academic performance. I cannot test whether the treatment

[44] I do not test the robustness of the analysis for English tests because there are almost no schools that track students in three levels.

Table 3.4: Mean Effects of Ability Tracking (Tracks in 3 Levels)

	Percentile	
Independent Variable	Rank I	Rank II
A. Mathematics		
Tracks students in mathematics (1=yes)	3.882	3.701
	(2.696)	(2.584)
	[0.342]	[0.338]
Female (1=yes)	1.604	1.201
	(0.820)	(0.823)
	[0.046]	[0.108]
Number of students	8,109	8,109
B. Reading		
Tracks students in reading (1=yes)	4.963	4.913
	(1.587)	(1.549)
	[0.008]	[0.004]
Female (1=yes)	10.329	10.106
	(0.934)	(0.950)
	[0.000]	[0.000]
Number of students	15,589	15,589

Note: All regressions are conducted conditional on school district fixed effects and school type indicators. Robust standard errors clustered at the school level are in parentheses, and p-values from the wild cluster bootstrap-t are in brackets (Null hypothesis, $\beta_{track} = 0$, has been imposed when estimating the p-values). For bootstraps, I use 1,000 bootstraps. There are 21 clusters for Panel A, 30 clusters for Panel B, and 11 clusters for Panel C. Percentile Rank I uses the ranking calculated at the national level. Percentile Rank II uses the ranking calculated at the sample level.

effect increases as the number of tracking level increases for there are few variations in the number of schools that track in two levels or more than three levels. However, this highlights the fact that how a school organizes its tracking system matters for students' achievement.

On the other hand, the estimated results from two analyses conducted above may be biased in the presence of sample attrition. In Figure 3.4, I show the timeline of the first year of high school. Students are randomly assigned to a high school within their school districts in March 2009. In May, they take their first midterm exam. Based on their performance in this first exam, schools implement ability

Figure 3.3: Quantile Treatment Effects of Ability Tracking (Tracks in 3 Levels)

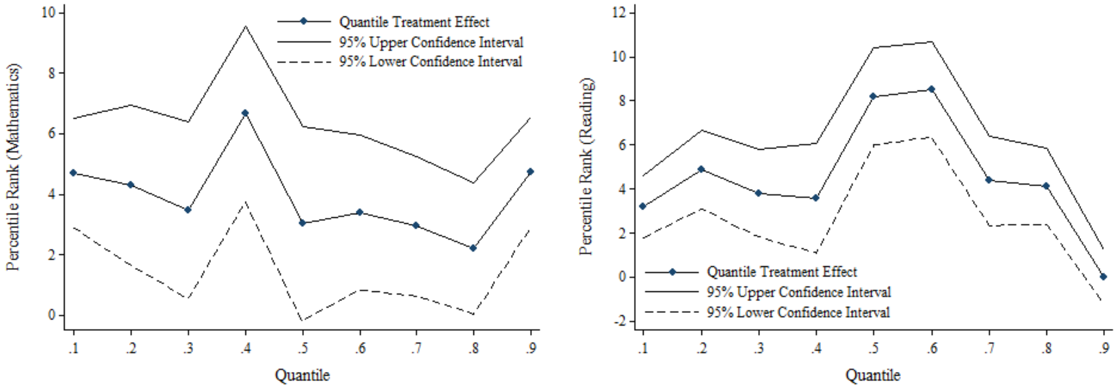
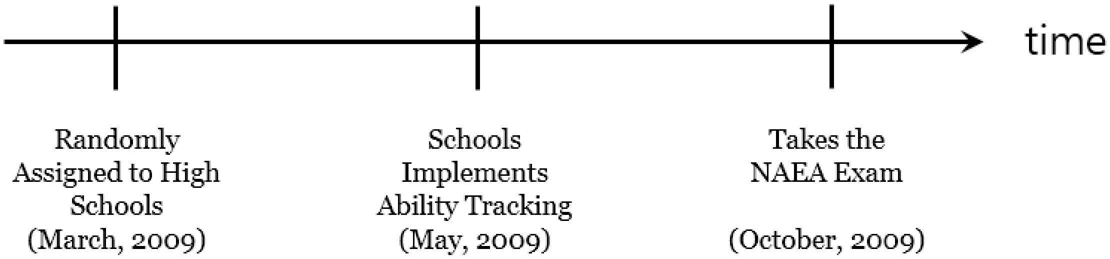


Figure 3.4: Timeline (From Random Assignment to the NAEA Exam Date)


time
Randomly Assigned to High Schools (March, 2009)
Schools Implements Ability Tracking (May, 2009)
Takes the NAEA Exam (October, 2009)

tracking. Finally, in October, 2009, all high school students take the NAEA exam on the same day. Hence, as can be seen from the timeline, there are a total of seven months in which some students may drop out from high school. If, for example, students who were originally assigned to the high schools that implement tracking had dropped out of the school had lower test scores, on average, compared to those who were initially assigned to the high schools that do not implement tracking who also quitted the school, then the estimated coefficients will be biased downward. One of the conditions in which the attrition problems do not bias the estimated treatment effects is when the pattern of attrition is similar, in expectation, between tracking schools and untracking schools. Hence, as a second method to test robustness, I test whether the attrition is problematic in this study.

I argue, using two facts, that the attrition bias is not an issue in this study. First, as can be seen from Table 3.5, the average ratio of dropouts is very small (0.027

Table 3.5: Tests of Within-District Attrition (Tracking vs Untracking Schools)

Dependent Variable	Track (1=yes)	Constant	No. of Schools
	A. Ability Tracking in Mathematics		
Ratio of dropouts	-0.002	0.027	32
	(0.005)	(0.005)	
	B. Ability Tracking in Reading		
Ratio of dropouts	0.000	0.031	37
	(0.005)	(0.009)	
	C. Ability Tracking in English		
Ratio of dropouts	-0.007	0.032	11
	(0.007)	(0.007)	

Note: All regressions are conditional on school district fixed effects and school types. In Panel A and Panel B, there is one missing value for each sample. Standard errors are in parentheses.

for mathematics, 0.031 for reading, and 0.032 for English). Moreover, although it does not “prove” that the potential outcomes of those who drop out from school are equal between students in tracking and the untracking schools, I contend that they are similar, in expectation. In Table 3.5, I run a regression of the ratio of dropouts on the dummy variable indicating whether the school tracks students to see whether the ratio of dropouts is different between tracking and untracking schools. As can be seen from the table, the estimated coefficient is close to 0, and furthermore, none of the coefficients are statistically significant.

It is likely that one’s academic achievement is closely associated with one’s likelihood of dropping out from high school. And if the pattern of attrition is not similar, one may observe some differences in the ratio of dropouts between schools that track and those that do not track. But since the difference is close to zero, I believe that, though there is a pattern in attrition, there are similar in expectation between schools that track students and schools that do not. Thus based on the two facts mentioned above, I argue that this study does not suffer from the attrition bias.

3.9 Conclusion

In this paper, I present causal estimates of the effect of ability tracking on students' academic achievement by making use of the unique random assignment mechanism conducted in Seoul. Empirical results suggest that placing a student into a high school that implements ability tracking raises his/her math scores for more than 5 percentile points compared to a student who attends a school that does not implement ability tracking. With respect to reading and English scores, tracking raises students' test scores by more than 3 and 4 percentile points, respectively. Although the estimated treatment effect for English tracking is statistically insignificant, the estimated coefficients for mathematics and reading are statistically significant. Thus, I conclude that ability tracking in general is beneficial for students' academic achievement.

Furthermore, to address the concern that tracking may hinder academic performance of low-achieving students, I estimate quantile treatment effects. Results of the quantile regression reveal that rather than worsening the inequality of students' academic performance, tracking, indeed, promotes academic achievement of students located at the bottom quantiles of the distribution. Besides, on average, I find that the magnitude of the treatment effect is as large as that of students in other quantiles.

Lastly, I find two more interesting points. First, the estimated treatment effect varies by subject. Second, the magnitude of the effect of tracking can vary depending on how schools organize and operate ability tracking. For instance, when I use a sample of schools that track students in three levels, the magnitude of the treatment effects changed. Hence, this implies that how schools operate their ability tracking matters for students' academic achievement.

From a policy perspective, the results of the current study provide one important implication. As is widely documented in previous studies, students' academic achievements are highly correlated with one's future earnings. As evidenced by the analysis above, tracking benefits poor-performing students. Accordingly, the use of ability tracking may reduce inequality in students' academic achievement, and may reduce inequality in future earnings. Thus, ability tracking may contribute to promoting income equality in the society.

3.10 Chapter 3: Appendix

3.10.1 Proof of Convexity and Concavity

To prove convexity and concavity of Equation (6), consider the following Hessian matrix;

$$D^2 f(\phi_1, \phi_2) = \begin{pmatrix} f_{\phi_1\phi_1} & f_{\phi_1\phi_2} \\ f_{\phi_2\phi_1} & f_{\phi_2\phi_2} \end{pmatrix},$$

where each entry in the matrix is the second order partial derivatives of $f(\phi_1, \phi_2)$. Note that the function is convex if and only if $f_{\phi_1\phi_1} \geq 0$, $f_{\phi_2\phi_2} \geq 0$, and $f_{\phi_1\phi_1} f_{\phi_2\phi_2} - \left(f_{\phi_1\phi_2}\right)^2 \geq 0$. Now observe that

$$\begin{aligned}
f_{\phi_1\phi_1} &= \left(-\frac{1}{\rho} - 1\right)\left(-\frac{1}{\rho}\right)\left(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho}\right)^{-\frac{1}{\rho}-2}\left(\rho^2\right)\left(\phi_1^2 c_H^{-2\rho-2}\right) \\
&\qquad + \left(-\frac{1}{\rho}\right)\left(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho}\right)^{-\frac{1}{\rho}-1}(-\rho-1)(-\rho)\left(\phi_1 c_H^{-\rho-2}\right) \\
&= (1+\rho)\left(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho}\right)^{-\frac{1}{\rho}-2}\left(\phi_1^2 c_H^{-2\rho-2}\right) \\
&\qquad - (1+\rho)\left(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho}\right)^{-\frac{1}{\rho}-1}\left(\phi_1 c_H^{-\rho-2}\right) \\
&= (1+\rho)\left(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho}\right)^{-\frac{1}{\rho}-2}\left(\phi_1^2 c_H^{-2\rho-2}\right) \\
&\qquad - (1+\rho)\left(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho}\right)^{-\frac{1}{\rho}-1}\left(\phi_1 c_H^{-\rho-2}\right)\left(c_L^{-2} c_L^2\right) \\
&= (1+\rho)\left(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho}\right)^{-\frac{1}{\rho}-2}\left(\phi_1^2 c_H^{-2\rho-2}\right) \\
&\qquad - (1+\rho)\left(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho}\right)^{-\frac{1}{\rho}-2}\left(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho}\right)\left(\phi_1 c_H^{-\rho-2}\right)\left(c_L^{-2} c_L^2\right).
\end{aligned}$$

Now, let $\xi = \left(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho}\right)^{-\frac{1}{\rho}-2}$, then

$$\begin{aligned}
f_{\phi_1\phi_1} &= (1+\rho)(\xi)\left[\left(\phi_1^2 c_H^{-2\rho-2}\right) - \left(\phi_1 c_H^{-\rho} + \phi_2 c_L^{-\rho}\right)\left(\phi_1 c_H^{-\rho-2}\right)\left(c_L^{-2} c_L^2\right)\right] \\
&= (1+\rho)(\xi)\left(\phi_1^2 c_H^{-2\rho-2} - \phi_1^2 c_H^{-2\rho-2} - \phi_1 c_H^{-\rho-2}\phi_2 c_L^{-\rho-2} c_L^2\right) \\
&= -(1+\rho)(\xi)(\phi_1\phi_2)\left(c_H c_L\right)^{-\rho-2} c_L^2.
\end{aligned}$$

Also, by taking similar steps as the above, we have

$$f_{\phi_2\phi_2} = -(1+\rho)(\xi)(\phi_1\phi_2)\left(c_H c_L\right)^{-\rho-2} c_H^2$$

and

$$f_{\phi_1\phi_2} = -(1+\rho)(\xi)(\phi_1\phi_2)\left(c_H c_L\right)^{-\rho-2} c_H c_L.$$

Then, $f_{\phi_1\phi_1} f_{\phi_2\phi_2} - \left(f_{\phi_1\phi_2}\right)^2 = 0$. Hence, $f_{\phi_1\phi_1} \geq 0$ and $f_{\phi_2\phi_2} \geq 0$ if $-(\rho+1) \geq 0 \implies \rho \leq -1$. Accordingly, $f(\phi_1, \phi_2)$ is convex if $\rho \leq -1$. On the other hand, $f_{\phi_1\phi_1} \leq 0$ and $f_{\phi_2\phi_2} \leq 0$ if $-(\rho+1) \leq 0 \implies \rho \geq -1$. Consequently, $f(\phi_1, \phi_2)$ is concave if $\rho \geq -1$. Note that when $\rho = -1$, $f(\phi_1, \phi_2)$ is both convex and concave. Therefore, when $\rho < -1$, $f(\phi_1, \phi_2)$ is always convex, and when $\rho > -1$, $f(\phi_1, \phi_2)$ is always concave.

3.10.2 Sample Restrictions

In order to causally estimate the effect of ability tracking, I make a series of sample restrictions that are presented in Table 3.6. First, the original dataset consists of 643,106 students. Next, I exclude schools that are not located in Seoul because only those in Seoul are randomly assigned to high schools. Third, I drop vocational high schools because students self-select into these schools. Fourth, I eliminate students who attend special purpose high schools because students are not randomly assigned to these high schools. Fifth, note that within Seoul, there were a total of 39 general high schools in 2009 that did not randomly admit students. Hence, I do not use students who attended these schools. Finally, I use students who are attending schools that operate only in the daytime. There is only one high school that operates both during the day and at night. This school includes general high school students as well as vocational high school students, and since the dataset does not allow one to distinguish between general and vocational high school students, I drop this school.

The resulting dataset consists of 78,322 students. Accordingly, based on this sample, I further make sample restrictions based on whether there are variations in the tracking indicator within the school district. The resulting sample size for mathematics, reading, and English is 13,559, 19,437, and 5,434 students, respectively.

Table 3.6: Series of Sample Restrictions (By Sample)

Step	Description	Resulting Sample
	A. Mathematics	
Step 1	Original dataset	643,106
Step 2	Dropping schools not located in Seoul	118,312
Step 3	Dropping vocational high schools	99,153
Step 4	Dropping special-purpose high schools	94,176
Step 5	Dropping schools in which students are not randomly assigned	78,792
Step 6	Dropping schools that operate during the day and at night	78,322
Step 7	Dropping districts with no variations in the tracking variable	13,559
	B. Reading	
Step 1	Original dataset	643,106
Step 2	Dropping schools not located in Seoul	118,312
Step 3	Dropping vocational high schools	99,153
Step 4	Dropping special-purpose high schools	94,176
Step 5	Dropping schools in which students are not randomly assigned	78,792
Step 6	Dropping schools that operate during the day and at night	78,322
Step 7	Dropping districts with no variations in the tracking variable	19,437
	C. English	
Step 1	Original dataset	643,106
Step 2	Dropping schools not located in Seoul	118,312
Step 3	Dropping vocational high schools	99,153
Step 4	Dropping special-purpose high schools	94,176
Step 5	Dropping schools in which students are not randomly assigned	78,792
Step 6	Dropping schools that operate during the day and at night	78,322
Step 7	Dropping districts with no variations in the tracking variable	5,434

References

Ammermueller, Andreas, and Jörn-Steffen Pischke. 2009. "Peer Effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study." *Journal of Labor Economics*, 27(3): 315–348.

Arcidiacono, Peter, and Sean Nicholson. 2005. "Peer Effects in Medical School." *Journal of Public Economics*, 89(2-3): 327–350.

Argys, Laura M., Daniel I. Rees, and Dominic J. Brewer. 1996. "Detracking America's Schools: Equity at Zero Cost?" *Journal of Policy Analysis and Management*, 15(4): 623–645.

Bacolod, Marigee, John DiNardo, and Mireille Jacobson. 2012. "Beyond Incentives: Do Schools Use Accountability Rewards Productively?" *Journal of Business and Economic Statistics*, 30(1): 149–163.

Bénabou, Roland. 1996. "Equity and Efficiency in Human Capital Investment: The Local Connection." *Review of Economic Studies*, 63(2): 237–264.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics*, 119(1): 249–275.

Betts, Julian R. 2011. "The Economics of Tracking in Education." In: Hanushek, Eric A., Machin, Stephen, and Woessmann, Ludger (Ed.), Handbook of the Economics of Education 3: 341–381.

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): 1–15.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak." *Journal of the American Statistical Association*, 90(430): 443–450.

Brunello, Giorgio, and Daniele Checchi. 2007. "Does School Tracking Affect Equality of Opportunity? New International Evidence." *Economic Policy*, 22(52): 781–861.

Cable, Kelly E., and Terry E. Spradlin. 2008. "Single-Sex Education in the 21st Centry." *Education Policy Brief*, 6(9): 1–12.

Cameron, Colin A., Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *The Review of Economics and Statistics*, 90(3): 414–427.

Carnoy, Martin, and Susanna Loeb. 2002. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." *Educational Evaluation and Policy Analysis*, 24(4): 305–331.

Carrell, Scott E., Richard L. Fullerton, and James E. West. 2009. "Does Your Cohort Matter? Measuring Peer Effects in College Achievement." *Journal of Labor Economics*, 27(3): 439–464.

Chiang, Hanley. 2009. "How Accountability Pressure on Failing Schools Affects Student Achievement?" *Journal of Public Economics*, 93(9-10): 1045–1057.

Coleman, James S. 1961. *The Adolescent Society: The Social Life of the Teenager and Its Impact on Education.* New York: The Free Press of Glencoe.

Cullen, Julie Berry, and Randall Reback. 2006. "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System." In: Gronberg, Timothy J. and Jansen, Dennis W. (Ed.), Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics 14: 1–34.

Dee, Thomas S., and Brian Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management*, 30(3): 418–446.

Ding, Weili, and Steven F. Lehrer. 2007. "Do Peers Affect Student Achievement in China's Secondary Schools?" *The Review of Economics and Statistics*, 89(2): 300–312.

Donald, Stephen G., and Kevin Lang. 2007. "Inference with Difference-in-Differences and Other Panel Data." *The Review of Economics and Statistics*, 89(2): 221–233.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *The American Economic Review*, 101(5): 1739–1774.

Eisenkopf, Gerald, Zohal Hessami, Urs Fischbacher, and Heinrich W. Wrsprung. 2011. "Academic Performance and Single-Sex Schooling: Evidence from a Natural Experiment in Switzerland." Unpublished Manuscript, Department of Economics, University of Konstanz.

Epple, Dennis, and Richard E. Romano. 2011. "Peer Effects in Education: A Survey of the Theory and Evidence." *Handbook of Social Economics*, 1: 1053–1163.

Epple, Dennis, Elizabeth Newlon, and Richard Romano. 2002. "Ability Tracking, School Competition, and the Distribution of Educational Benefits." *Journal of Public Economics*, 83(1): 1–48.

Fan, Jianqing, and Irene Gijbels. 1996. "Local Polynomial Modelling and Its Applications." London; New York and Melbourne: Chapman and Hall.

Figlio, David, and Susanna Loeb. 2011. "School Accountability." In: Hanushek, Eric A., Machin, Stephen, and Woessmann, Ludger (Ed.), Handbook of the Economics of Education 3: 383–421.

Figlio, David N., and Cecilia Elena Rouse. 2006. "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *Journal of Public Economics*, 90(1-2): 239–255.

Figlio, David N., and Helen F. Ladd. 2008. "School Accountability and Student Achievement." In: Ladd, Helen F. and Fiske, Edward B. (Ed.), Handbook of Research in Education Finance and Policy: 166–182.

Figlio, David N., and Joshua Winicki. 2005. "Food for Thought: The Effects of School Accountability Plans on School Nutrition." *Journal of Public Economics*, 89(2-3): 381–394.

Figlio, David N., and Marianne E. Page. 2002. "School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality?" *Journal of Urban Economics*, 51(3): 497–514.

Firpo, Sergio. 2007. "Efficient Semiparametric Estimation of Quantile Treatment Effects." *Econometrica*, 75(1): 259–276.

Foster, Gigi. 2006. "It's Not Your Peers, and It's Not Your Friends: Some Progress Toward Understanding the Educational Peer Effect Mechanism." *Journal of Public Economics*, 90(8-9): 1455–1475.

Frölich, Markus, and Blaise Melly. 2010. "Estimation of Quantile Treatment Effects with Stata." *The Stata Journal*, 10(3): 423–457.

Hahn, Jinyong, Petra Todd, and Wilbert van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica*, 69(1): 201–209.

Hanushek, Eric A., and Ludger Wö*β*mann. 2006. "Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries." *The Economic Journal*, 116(510): C63–C76.

Hanushek, Eric A., and Margaret E. Raymond. 2005. "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management*, 24(2): 297–327.

Hanushek, Eric A., John F. Kain, Jacob M. Markman, and Steven G. Rivkin. 2003. "Does Peer Ability Affect Student Achievement?" *Journal of Applied Econometrics*, 18(5): 527–544.

Harker, Richard. 2000. "Achievement, Gender and the Single-Sex/Coed Debate." *British Journal of Sociology of Education*, 21(2): 203–218.

Hoffer, Thomas B. 1992. "Middle School Ability Grouping and Student Achievement in Science and Mathematics." *Educational Evaluation and Policy Analysis*, 14(3): 205–227.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association*, 81(396): 945–960.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization*, 7(Special Issues): 24–52.

Howell, William G., Patrick J. Wolf, David E. Campbell, and Paul E. Peterson. 2002. "School Vouchers and Academic Performance: Results from Three Randomized Field Trials." *Journal of Policy Analysis and Management*, 21(2): 191–217.

Hoxby, Caroline. 2000. "Peer Effects in the Classroom: Learning from Gender and Race Variation." *National Bureau of Economic Research Working Paper Series*, No. 7867.

Hoxby, Caroline M., and Gretchen Weingarth. 2005. "Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects." Unpublished Manuscript.

Imbens, Guido, and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies*, 79(3): 933–959.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*, 142(2): 615–635.

Inzlicht, Michael, and Talia Ben-Zeev. 2000. "A Threatening Intellectual Environment: Why Females are Susceptible to Experiencing Problem-Solving Deficits in the Presence of Males." *Psychological Science*, 11(5): 365–371.

Jackson, Kirabo C. 2012. "Single-Sex Schools, Student Achievement, and Course Selection: Evidence from Rule-Based Student Assignments in Trinidad and Tobago." *Journal of Public Economics*, 96(1-2): 173–187.

Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics*, 89(5-6): 761–796.

Jacob, Brian, and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *The Quarterly Journal of Economics*, 118(3): 843–877.

Kane, Thomas J., and Douglas O. Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives*, 16(4): 91–114.

Kessels, Ursula, and Bettina Hannover. 2008. "When Being a Girl Matters Less: Accessibility of Gender-Related Self-Knowledge in Single-Sex and Coeducational Classes and Its Impact on Students' Physics-Related Self-Concept of Ability." *British Journal of Educational Psychology*, 78(2): 273–289.

Koenker, Roger, and Gilbert Jr. Bassett. 1978. "Regression Quantiles." *Econometrica*, 46(1): 33–50.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics*, 114(2): 497–532.

Ladd, Helen F. 1999. "The Dallas School Accountability and Incentive Program: An Evaluation of Its Impacts on Student Outcomes." *Economics of Education Review*, 18(1): 1–16.

Lavy, Victor, and Analia Schlosser. 2011. "Mechanisms and Impacts of Gender Peer Effects at School." *American Economic Journal: Applied Economics*, 3(2): 1–33.

Lavy, Victor, Olmo Silva, and Weinhardt Felix. 2009. "The Good, the Bad and the Average: Evidence on the Scale and Nature of Ability Peer Effects in Schools." National Bureau of Economic Research Working Paper 15600.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*, 48(2): 281–355.

Lee, Valerie E., and Anthony S. Bryk. 1986. "Effects of Single-Sex Secondary Schools on Student Achievement and Attitudes." *Journal of Educational Psychology*, 78(5): 381–395.

Lefgren, Lars. 2004. "Educational Peer Effects and the Chicago Public Schools." *Journal of Urban Economics*, 56(2): 169–191.

Lu, Fangwen, and Michael Anderson. 2011. "Peer Effects in Microenvironments: The Benefits of Homogeneous Classroom Groups." Unpublished Manuscript, Department of Agricultural and Resource Economics, University of California, Berkeley.

Mael, Fred A. 1998. "Single-Sex and Coeducational Schooling: Relationships to Socioemotional and Academic Development." *Review of Educational Research*, 68(2): 101–129.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies*, 60(3): 531–542.

Marsh, Herbert W. 1991. "Public, Catholic Single-Sex, and Catholic Coeducational High Schools: Their Effects on Achievement, Affect, and Behaviors." *American Journal of Education*, 99(3): 320–356.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics*, 142(2): 698–714.

Meghir, Costas, and Mårten Palme. 2005. "Educational Reform, Ability, and Family Background." *The American Economic Review*, 95(1): 414–424.

Ministry of Education. 1998. *50 Years of Education History: 1948-1998.* Seoul: Ministry of Education.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics*, 32(3): 385–397.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *The Review of Economics and Statistics*, 92(2): 263–283.

Oosterbeek, Hessel, and Reyn van Ewijk. 2010. "Gender Peer Effects in University: Evidence from a Randomized Experiment." Working Paper 10/24, Top Institute for Evidence Based Education Research (TIER), Maastricht University.

Pekkarinen, Tuomas, Roope Uusitalo, and Sari Kerr. 2009. "School Tracking and Intergenerational Income Mobility: Evidence from the Finnish Comprehensive School Reform." *Journal of Public Economics*, 93(7-8): 965–973.

Pischke, Jörn-Steffen, and Alan Manning. 2006. "Comprehensive Versus Selective Schooling in England in Wales: What Do We Know?" National Bureau of Economic Research Working Paper 12176.

Richards, Craig E., and Tian Ming Sheu. 1992. “The South Carolina School Incentive Reward Program: A Policy Analysis.” *Economics of Education Review*, 11(1): 71–86.

Rockoff, Jonah, and Lesley J. Turner. 2010. “Short-Run Impacts of Accountability on School Quality.” *American Economic Journal: Economic Policy*, 2(4): 119–147.

Rouse, Cecilia Elena., Jane Hannaway, Dan Goldhaber, and David Figlio. 2007. “Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure.” National Bureau of Economic Research Working Paper 13681.

Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology*, 66(5): 688–701.

Sacerdote, Bruce. 2001. “Peer Effects with Random Assignment: Results for Dartmouth Roommates.” *The Quarterly Journal of Economics*, 116(2): 681–704.

Sacerdote, Bruce. 2011. “Peer Effects in Education: How Might They Work, How Big are They and How Much Do We Know Thus Far?” *Handbook of the Economics of Education*, 3: 249–277.

Sax, Leonard. 2006. *Why Gender Matters: What Parents and Teachers Need to Know About the Emerging Science of Sex Differences.* New York: Three Rivers Press.

Schneider, Frank W., and Larry M. Coutts. 1982. “The High School Environment: A Comparison of Coeducational and Single-Sex Schools.” *Journal of Educational Psychology*, 74(6): 898–906.

Slavin, Robert E. 1987. “Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis.” *Review of Educational Research*, 57(3): 293–336.

Slavin, Robert E. 1990. “Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis.” *Review of Educational Research*, 60(3): 471–499.

Smith, Stephen Samuel, and Roslyn Arlin Mickelson. 2000. “All That Glitters is Not Gold: School Reform in Charlotte-Mecklenburg.” *Educational Evaluation and Policy Analysis*, 22(2): 101–127.

Whitmore, Diane. 2005. “Resource and Peer Impacts on Girls’ Academic Achievement: Evidence from a Randomized Experiment.” *The American Economic Review*, 95(2): 199–203.

Zimmerman, David J. 2003. "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment." *The Review of Economics and Statistics*, 85(1): 9–23.

Zimmer, Ron. 2003. "A New Twist in the Educational Tracking Debate." *Economics of Education Review*, 22(3): 307–315.