The Korean Economic Review
Volume 36, Number 2, Summer 2020, 319-352.
DOI : 10.22841/kerdoi.2020.36.2.002

Impact of the Simultaneous Use of the Stigmatization and Categorical School Funding Policy on the Test and Post-Secondary Outcomes of Lower-Achieving Students*

Seunghoon Han** · Hosung Sohn***

This study analyzes the impact of one type of school-based accountability system—the simultaneous use of stigmatization and categorical school funding—on test scores and post-secondary outcomes. We conduct randomization inference in the context of regression discontinuity design by exploiting the discontinuous rule used in the accountability system in South Korea. The results show that the joint use of stigmatization and funding leads to a statistically and practically significant increase in test scores (7, 6, and 5 percentile points for reading, math, and English, respectively). Subgroup analyses by urban or rural areas show that the policy leads to a practically but not statistically significant increase in the share of students taking college entrance exams (4 percentage points) and a practically and statistically significant increase in the share of students matriculating into four-year colleges (9 percentage points) only for schools located in rural areas. We do not find any practically and statistically significant increase in the post-secondary outcomes for schools located in urban areas.

JEL Classification: I21, I28, L15
Keywords: Categorical School Funding, School-based Accountability System, Stigmatization

I. Introduction

Countries across the world operate school-based accountability systems to

Received: Feb. 13, 2019. Revised: June 17, 2019. Accepted: July 26, 2019.

* This Research was supported by the Chung-Ang University Research Grants in 2017.

** First Author, Assistant Professor, College of Social Sciences, School of Public Service, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06794, Korea. Email: sehan@cau.ac.kr. Phone: +82-2-820-5704. Fax: +82-2-826-5442

*** Corresponding Author, Assistant Professor, College of Social Sciences, School of Public Service, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06794, Korea. Email: hsohn@cau.ac.kr. Phone: +82-2-820-5123. Fax: +82-2-826-5442

promote student achievement. While many types of school-based accountability systems exist, most of these systems have been adopted on the premise that monitoring the efforts of school administrators and teachers is difficult (i.e., the principal-agent problem). The information that policymakers can draw from accountability systems, such as how schools perform relative to some common standardized metrics that policymakers set, can be used to overcome this problem.

Researchers have extensively analyzed the efficacy of accountability systems on student achievement, especially in the US and the UK contexts. Figlio and Ladd (2015) conducted an extensive literature review regarding school accountability systems. They found that, in general, the pressure induced by the system raises short-term student achievement (especially in math) and long-term outcomes, such as graduation, college attainment, and earnings. School accountability systems seem to have a positive impact on student outcomes, but the impact differs according to how the systems are designed. Thus, from a policy perspective, determining the relative effectiveness of different designs and how the effects differ by design is crucial. A close examination of previous literature indicates that school-based accountability systems generally operate in two different forms (high-stakes and low-stakes).

High-stakes accountability systems use negative incentives by identifying low-performing schools—using, for example, information provided in nationwide educational achievement assessments—and stigmatizing them as failing schools. The impact on the behavior of teachers and administrators who work at "failing" schools may be significant for several reasons. First, schools classified as failing may face local and community pressure. Many studies have confirmed that parents and communities respond to such information (e.g., Black, 1999; Figlio and Lucas, 2004; Hastings and Weinstein, 2008; Figlio and Kenny, 2009; Hart and Figlio, 2015). Second, stigmatization may influence the so-called identity utility of principals or teachers (Akerlof and Kranton, 2005). Third, studies in psychology and economics have shown that people tend to be last-place averse (Kuziemko et al., 2014), meaning teachers and administrators in high-stakes accountability contexts may put increased effort into avoiding being categorized as failing. Existing research has found evidence that punishment is effective in influencing the behavior of educators (e.g., Diamond and Spillane, 2004; Hanushek and Raymond, 2005; Jacob, 2005; Burgess et al., 2005; Lemke, Hoerandner, and McMahon, 2006; West and Peterson, 2006; Krieg, 2008; Reback, 2008; Sims, 2008; Neal and Schanzenbach, 2010; Rockoff and Turner, 2010; Dee and Jacob, 2011; Chakrabarti, 2013a; Chakrabarti, 2013b; Chakrabarti, 2014). However, it often leads to unintended strategic behavior, such as cheating (Jacob and Levitt, 2003; Jacob, 2005).

Low-stakes accountability systems use positive incentives by providing financial support for low-performing schools. High-stakes accountability systems assume that the main reason for low performance is that school efforts are not sufficient for

promoting student achievement. Therefore, teachers and administrators are viewed as "culprits" for not meeting the standards set by policymakers. However, poor-performing schools are failing because of deficiencies in school resources, not because teachers and administrators are not doing their jobs. If this insufficiency is the case, then merely stigmatizing or punishing schools may not effectively boost student achievement. Therefore, this scheme may have a significant positive impact on underperforming schools, especially when such schools lack the capacity and resources to respond in ways that are consistent with policy goals. Note, however, that such schemes often produce less significant effects than other schemes (e.g., Bacolod, DiNardo, and Jacobson, 2012).

The Korean accountability system is unique in that it utilizes both schemes (i.e., financial support and stigmatization). Given that these schemes have advantages and disadvantages, the Korean system may produce significant positive impacts on underperforming schools for two reasons. First, low-performing schools located in rural areas often lack financial resources but have more autonomy than schools in urban areas because they experience less stigmatization-related pressure. Second, low-performing schools located in urban areas do not lack financial resources but often face high levels of competition from other schools. In such circumstances, an accountability system operated based on both schemes may help raise the achievement of students in low-performing schools.

The Korean system may generate different impacts depending on the school type (urban/rural). For example, given that urban schools experience high levels of competition, the system may have only short-term impacts, such as improving student test scores. In contrast, the system may generate both short- and long-term impacts for schools located in rural areas because these schools may not experience such high levels of competition. Therefore, analyzing the accountability system in Korea should help determine whether the aforementioned arguments are empirically supported. From a policy perspective, analyzing the intended and unintended consequences of this policy should have the additional benefit of enabling educational administrators to adopt specific systems that suit their purposes.

This study adds to the literature on the efficacy of school-based accountability systems by analyzing the causal impact of the simultaneous use of stigmatization and school funding.[1] To achieve this goal, we examine the school accountability

[1] Borba (2003) analyzed California's school accountability system, whose main intervention is to use external evaluators to assist low-performing schools, and Roy and Kochan (2012) studied the effect of the Alabama School Assistance Team, which helps a low performing school—systems that both resemble the Korean approach; however these systems do not involve the provision of additional funding. Moreover, the studies assessed principals' perceptions regarding these external evaluators, not student achievement, and the analyses relied on case studies. Meanwhile, Woo et al. (2015) also examined the school accountability system in Korea, but they analyzed students in middle schools.

system adopted in Korea to isolate the causal impact of providing financial support and punishment. Starting from the school year of 2009, the Ministry of Education implemented a countrywide program to make schools accountable for student achievement. The Ministry conducts annual nationwide assessments of student achievement and identifies low-performing schools on the basis of the outcomes of the assessments. The Ministry stigmatizes and provides funding to low-performing schools. Hence, the accountability system provides a favorable setting for examining the impact of an accountability model that uses both schemes.

This study contributes to existing literature in two ways. First, given that no previous studies have quantitatively analyzed the causal impact of the stigmatization and school funding policies on students in disadvantaged schools (i.e., vocational schools), the findings of this study shed light on whether this policy effectively promotes the achievement of disadvantaged students. Many countries have started to move toward more supportive and less punitive accountability systems than the existing methods (e.g., the Renewal School program recently implemented by New York City Public Schools). Thus, the results of this study may have important policy implications. In addition, whereas many studies have focused on analyzing school accountability systems in the US and other western countries, few studies have focused on Asian countries. By analyzing the Korean system, this study may contribute to the existing research by expanding the evidence regarding the efficacy of school accountability systems. Second, relatively few studies have analyzed the long-term outcomes of school accountability systems.[2] This study thus contributes to the literature by examining the effect of the Korean system on post-secondary outcomes, such as college matriculation.

This study exploits the discrete nature of the assignment mechanism used in the school accountability system and uses a regression discontinuity (RD) design to estimate causally the impact of the stigmatization and school funding policy on student achievement. Using administrative records of student- and school-level data, we find that students in schools that are stigmatized and receive school funding perform significantly better than those in similar schools that are not stigmatized and do not receive school funding. More specifically, the estimated effects of the policy on reading, math, and English are approximately 7, 6, and 5 percentile points, respectively. We also find that the share of underachieving students in reading and math has decreased by 10 and 5 percentage points, respectively. Moreover, the share of students who classify as "average or above" in reading has increased by 5 percentage points, and the share in math and English has increased by

[2] In analyzing the Florida accountability system, Chiang (2009) and Rouse et al. (2013) found that the positive effects of the system, in fact, persist for several years. Meanwhile, examining the Texas accountability system, Deming et al. (2016) found that students in low-performing schools are more likely to have attended college than those in better-performing schools. The study also found that these students have higher earnings at age 25.

approximately 5 and 1 percentage points, though these last two estimates are statistically insignificant. Lastly, the subgroup analysis by urban or rural areas shows a positive impact on the likelihood of taking the college entrance exams for schools located in rural areas, though the estimates are imprecisely estimated. Moreover, we find a practically and statistically significant increase in the share of students matriculating into a four-year college, again, for schools located in rural areas.

The remainder of this paper is organized as follows. In Section II, we provide an institutional background on the Korean school accountability system. Section III discusses the data and our empirical strategy, and Section IV presents the result of the tests of the identifying assumptions of the RD design. In Section V, we present the estimation results, which we then discuss in Section VI. Section VII concludes the paper.

II. Institutional Background

The education system in Korea has three levels: elementary (six years), middle (three years), and high (three years) school. High school has three divisions: general, special purpose, and vocational. Since 2008, the school-based accountability system has been implemented at all educational levels. The accountability system in Korea is similar to the No Child Left Behind Act of 2001 in the US. The system was designed to identify schools that do not meet the standards set by the Ministry of Education.

To identify such schools, the Ministry of Education conducts an annual countrywide assessment of the academic achievement of students. The assessment is based on the National Assessment of Educational Achievement (NAEA). Every student in Korea is tested using this assessment, and the Ministry identifies low-performing schools on the basis of the NAEA results.

The purpose of the NAEA is to examine the overall level of academic achievement of every student in every school and to determine how many students are not meeting the basic academic standards. The Ministry of Education administers the test annually, testing students on three to five subjects depending on their educational levels.

The Ministry of Education identifies low-performing schools on the basis of the NAEA results as follows. After the test, each student receives a grade from one of the four achievement categories: 1) proficient, 2) average, 3) basic, and 4) underachieving. Next, the Ministry of Education calculates the proportion of students classified as underachieving—for each subject—within each school and then calculates the mean of the share. Then, it designates schools with a mean share of underachieving students above a certain cutoff as schools in need of achievement

improvement (SINAI).[3] After being identified as low-performing, the government provides categorical school funding that must be used to promote student achievement. For elementary, middle, and both general and special purpose high schools, the amount of funding depends on school size, as measured by the number of students in each school. For vocational high schools, the amount is determined by the average number of underachieving students per school. For instance, vocational high schools who have fewer than 100 underachieving students on average receive approximately $30,000, while schools with 100 to 200 underachieving students receive approximately $50,000. The mean per student spending in 2008 was reported to be approximately $7,000. Thus, the magnitude of funding that the SINAI-designated schools receive from the accountability policy is approximately 5% of the annual per-pupil spending.

The SINAI designation carries no consequences other than stigmatization and school funding. For example, the SINAI-designated schools are not punished for not being able to promote student achievement in subsequent years. To publicize and share the best practices for other schools to follow, the government identifies certain schools that employ the best practices among those that succeed in increasing student achievement in subsequent years; however, the schools the government identifies do not receive any positive incentives such as bonuses or additional funding.

Thus, the accountability part of the program is mostly driven by the stigmatization part. The stigma effect with respect to vocational schools is quite strong in the context of the Korean educational system. Given that students self-select vocational schools in Korea and that many different kinds of schools are available for students and parents to choose from in given localities, principals and teachers are sensitive to the relative rankings of their schools. While vocational high schools in Korea are constrained by limits on the government side, such as in increasing their capacity to accommodate additional students, they are likely to be sensitive to their rankings because these rankings may affect the pool of students to whom they can send out admission letters. While the accountability dimension of the program is likely to affect the efforts of teachers, it is less likely to affect students directly. Students do not experience any disadvantages as a result of being designated as "underachieving." This designation may affect their motivation, as students typically tend to be last-place averse (Kuziemko et al., 2014). However, given that the test results do not affect the likelihood of students graduating from high school, let alone attending college, we argue that the effect of the accountability dimension of the program on students is weak. Overall, we believe that any resulting improvements in student outcomes are primarily driven by the

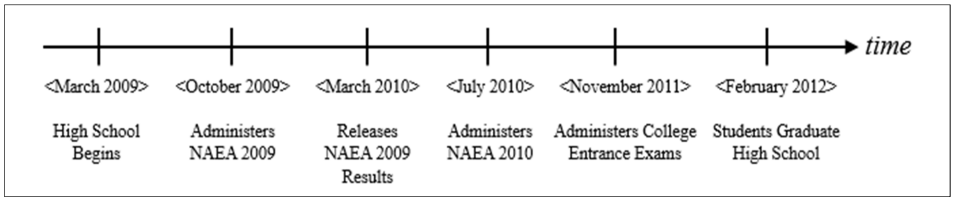
[3] The cutoffs used for the designations are as follows: elementary school (5%), middle school (20%), general and special purpose high school (20%), and vocational high school (40%).

incentives schools receive.

The Korean setting is favorable for examining the causal impact of the stigmatization and school funding policy for several reasons. First, the treatment is determined solely by a single variable (i.e., the share of underachieving students in a school) over which schools have imprecise control. This setting enables us to estimate the causal impact of the treatment using an RD design. Second, the RD setting facilitates the isolation of treatment effects that are driven mainly by stigmatization and school funding because it provides the opportunity to compare the two groups with similar observable and unobservable characteristics but different treatment status. In Section IV, we describe how we tested for the identifying assumptions of an RD design so that we could consider the estimated treatment effects as reflections of the treatment.

Figure 1 presents the timeline of this study. The students in the sample entered vocational high school in March 2009. In October of the same year, NAEA exams were administered for "first-year" vocational high school students (i.e., students who entered in March 2009). In March 2010, the NAEA results were disclosed, and on the basis of these test results, schools were designated as SINAI. In July 2010, the same students—who became second-year students—took the second NAEA exam. In this study, we use these two sets of test results. In November 2011, the same students became third-year students, and these students took the college entrance exam if they intended to matriculate into post-secondary schools. In short, we focus on a high school cohort who entered high school in 2009.

[Figure 1] Timeline of the Study



III. Data and Empirical Strategy

3.1. Data

In this study, we use the administrative records of student-level test scores on the NAEA 2009, NAEA 2010, and the College Scholastic Ability Test (CSAT) held in November 2011.[4] These data are not publicly available. At the end of 2010, however,

[4] Data for 2008, the first year of the NAEA administration, 2011, and after 2011 are no longer available from the government. The government disclosed population data only for 2009 and 2010.

the Supreme Court ruled in favor of disclosing the test scores, and the Ministry of Education began disclosing them to researchers in 2011.[5] Following a series of data application steps, we obtain the NAEA data for 2009 and 2010, and CSAT data for 2011. The data include test scores of every student in Korea who has taken the test.[6]

The NAEA data contain information on student characteristics and test scores, and the CSAT data include information mostly on test-related outcomes. Although the NAEA datasets contain information on all the test scores and information on gender and some family background, they are limited in the sense that they do not contain certain important student-level variables—such as family income—that can be used for testing the identifying assumptions of an RD design. Note, however, that using student-level data for the analysis in this study is problematic for conducting statistical inference (see the Empirical Strategy subsection). To test whether the identifying assumptions of the RD design hold, we use school-level data. We obtain this data from the EduData Service System (EDSS).[7] One of the advantages of using the school-level data retained by EDSS is that the data contain some important student-, teacher-, and school-level variables that we can use for covariate balancing tests and post-secondary outcome analyses. We use the CSAT data to analyze the effect of the accountability system on the share of students taking the college entrance exam. The Ministry of Education has provided us student-level population data for the CSAT 2012 (i.e., the CSAT held in 2011). Given that the dataset contains specific high school names, we are able to calculate the share of students taking the college entrance exam for all vocational high schools considered in the analysis.

To analyze the impact of the simultaneous use of the stigmatization and school funding policy, we use the test scores of students in vocational high schools. Every middle school graduate who intends to proceed to high school receives a graduate standing percentile rank. On the basis of the ranking, only approximately the top 65% can enter general high schools, and the rest (35%) enter vocational high schools. The majority of students in general high schools matriculate into four-year colleges, whereas students in vocational high schools matriculate into two-year colleges or enter the labor market after graduation. Hence, vocational high school students can be considered lower-achieving students in Korea. We reason that analyzing such students enables us to examine whether the policy promotes the achievement of

[5] To obtain the datasets, researchers must submit a research proposal to the Ministry. Then, the Ministry forms a committee consisting of members from inside and outside the Ministry. These committee members examine the feasibility of the research and decide whether the test scores should be disclosed to the researcher.

[6] Note, however, that the Ministry of Education no longer discloses population data. Instead, it discloses a random sample starting from 2014.

[7] The EDSS website operates a formal application system that researchers can use to request school-level data. The authors can assist researchers who are interested in obtaining such data (www.schoolinfo.go.kr).

disadvantaged students.

One of the challenges of analyzing the treatment effect using the NAEA datasets is that the datasets do not contain school names. To implement an RD design properly, the name of the school must be identified so that the school-level data can be merged into the NAEA datasets. Therefore, to identify the school names in the NAEA data, we use the government-run school information website (SIW).[8] On this website, the descriptive statistics of the results of the NAEA for each school are posted with the school names. These descriptive statistics have been calculated using the same data that we have obtained from the Ministry of Education. The descriptive statistics include information on the proportions of students in each of the achievement categories mentioned in Section II (up to the first decimal place) and the number of test-takers for each school. These statistics have been calculated for each subject (the data include three subjects). Using the data used in this study, we calculate the same descriptive statistics presented on the SIW, successfully matching each number with a 100% matching rate. Thus, we are able to identify the names of the schools in the NAEA data.[9]

After completing this matching, we conduct a series of sample restrictions. The original sample consists of 125,850 students in 536 vocational high schools for NAEA 2009 and 123,196 students in 540 vocational high schools for NAEA 2010. In Step 1, we exclude one school in NAEA 2009 whose name we cannot identify from the SIW. The share of underachieving students in this unmatched school is 0.172. Thus, it is designated as SINAI. In Step 2, we drop the schools that are not in both datasets. Specifically, one school in the 2009 dataset does not show up in the 2010 dataset. The share of underachieving students in this school is 0.240. Thus, this school is not a SINAI-designated school. Meanwhile, six schools in the 2010 dataset do not show up in the 2009 dataset. These schools may have been closed or newly established during this two-year period, or they may have taken only one NAEA exam. The resulting final sample consists of 125,803 students in 534 schools for NAEA 2009 and 121,943 students in 534 schools for NAEA 2010.

Table 1 presents descriptive statistics for the final sample, separately by the treatment and control groups. As the table shows, 5,219 students in 35 schools are treated, and 120,584 students in 499 schools are not treated. The NAEA 2009 results show that the difference in the mean percentile rank between the treated and control group is huge for every subject. The differences in the means are more than 16 percentile ranks, and the differences are all statistically significant. Moreover,

---

[8] www.schoolinfo.go.kr.

[9] Although analyzing the effect of the policy not only on vocational high school students but also on general high school students are desirable for increasing external validity, the data for general high school students are not used in this study because the names of general high schools cannot be matched at this point. The information is not available for general high schools, at this point, on this website.

most of the student-, teacher-, and school-level characteristics are different between the two groups.

[Table 1] Descriptive statistics by treatment or control groups

	NAEA year				
	2009			2010	
Variable	Treated	Control	*t*-test	Treated	Control
Mean reading percentile rank	21.897	47.942	26.045	32.605	48.139
	(5.060)	(16.479)	[0.000]	(7.838)	(17.453)
Mean math percentile rank	29.451	45.851	16.398	38.858	45.474
	(3.642)	(13.184)	[0.000]	(13.670)	(13.277)
Mean English percentile rank	22.087	47.465	25.378	32.164	48.105
	(3.578)	(16.536)	[0.000]	(11.184)	(17.337)
Mean total percentile rank	19.739	49.012	29.272	32.332	48.758
	(4.156)	(18.051)	[0.000]	(11.381)	(18.710)
Share of test takers	0.841	0.930	0.088	0.867	0.912
	(0.075)	(0.062)	[0.000]	(0.115)	(0.096)
% living with both parents	0.628	0.711	0.083	—	—
	(0.124)	(0.095)	[0.000]		
% receiving government subsidy	0.209	0.153	−0.055	—	—
	(0.088)	(0.086)	[0.000]		
% receiving private tutoring	0.147	0.186	0.038	—	—
	(0.115)	(0.119)	[0.064]		
Class size	25.440	29.117	3.676	—	—
	(5.291)	(6.059)	[0.001]		
Student-teacher ratio	11.364	14.465	3.101	—	—
	(2.470)	(4.496)	[0.000]		
% newly hired teachers	0.067	0.049	−0.018	—	—
	(0.099)	(0.062)	[0.112]		
% part-time teachers	0.073	0.103	0.030	—	—
	(0.054)	(0.104)	[0.091]		
Number of schools	35	499		35	499
Number of students	5,219	120,584		4,885	117,058

Notes: The numbers reported in the "*t*-test" column are the difference in the means for the corresponding variable. Standard deviations in parentheses and *p*-values calculated from the two-sample *t*-test with equal variances are presented in brackets. The number of schools used in the *t*-test varies because of missing values (ranges from 520 to 530).

Interestingly, the gaps in the achievements are reduced after the treatment. For example, the difference of 26 percentile points before the treatment in reading reduces to 15 percentile points. In the total achievement, the difference is reduced by 16 percentile points, an approximately 55% reduction from the NAEA 2009 results. While the reductions in the differences indicate treatment effects, the reductions may not be due to the policy effect, as the baseline characteristics between the treated and control group are different. Therefore, the revealed

differences in baseline characteristics call for some identification strategies to isolate the effect that is driven by the policy. This study uses an RD design to isolate the policy effect.

3.2. Empirical Strategy

The SINAI-designation and the provision of school funding (T_s) to school s is determined by the following simple discontinuous rule:

$$T_s = 1\{\bar{X}_s \geq 0.4\},$$

where $\bar{X}_s$ is the average share of underachieving students in school s. Every school is treated when the average share of underachieving students in a school is equal to or greater than 40% (i.e., the sharp RD setting). Under this setting, the average causal effect of the treatment τ_i in this study is obtained by estimating the following conditional expectation functions:

$$\tau_i = \lim_{x \downarrow 0.4} E[Y_{is} \mid \bar{X}_s = x] - \lim_{x \uparrow 0.4} E[Y_{is} \mid \bar{X}_s = x], \quad (1)$$

where Y_{is} denotes one of the outcome variables.

Cattaneo, Titiunik, and Vazquez-Bare (2017) demonstrated that two approaches can be used to estimate the conditional expectation function in Equation (1). One approach is a continuity-based approach that uses parametric or nonparametric local polynomial regressions (e.g., local linear regressions) to estimate the function. This approach relies on extrapolation and asymptotic approximations of the expectation functions using the sample around the threshold of a running variable. The other approach is based on the idea of local randomization. In this approach, observations close to the cutoff that determines the treatment are considered as randomly assigned to the treatment and control groups.

In this study, we adopt a locally randomized approach to implement an RD analysis for two reasons. First, schools have no control over the assignment variable (see Section IV). Moreover, whether a school ends up being placed on the left or right of the cutoff point (i.e., 40%) can be thought of "as good as" random. When the assignment variable can be considered as good as random, the setting can be analyzed as a locally randomized experiment (Lee, 2008; Lee and Lemieux, 2010). The second reason is related to conducting statistical inference. As a matter of course, calculating correct standard errors is of great importance in any empirical work; both coefficient estimates and standard errors are critical components of statistical inference (Cameron and Miller, 2015). Adopting a continuity-based approach and conducting the inference based on such an approach poses a huge

challenge in the context of this study.

The treatment (i.e., stigmatization and funding) in this study varies at the school level, and errors are likely to be clustered at this level. To account for the clustering of errors, the conventional cluster-robust standard errors can be estimated. However, the theoretical development of a variance estimator in the context of an RD design is at a nascent stage. Although Calonico et al. (2019) developed a cluster-robust variance estimator in an RD design, their proposed variance estimator is conditional on the number of clusters being infinite. Typically, however, the number of clusters is likely to be small because identification in an RD setting amounts to comparing observations around the threshold. In this study, the number of schools is small because the identification involves comparing students and schools near the threshold. Thus, the assumption required for using the variance estimator is not met. Many studies have found that using the conventional cluster-robust variance estimator under the presence of a small number of clusters significantly underestimates standard errors (e.g., Donald and Lang, 2007; Cameron, Gelbach, and Miller, 2008; Conley and Tabler, 2011; MacKinnon and Webb, 2016). Even though these studies have proposed many other variance estimators that solve for issues related to the small number of clusters, such methods have not been formalized and developed for use in RD settings.

To eliminate the clustering issues and conduct proper statistical inference, we collapse the data at the school level. Collapsing the data has two advantages. First, the resulting data has no dependence problems. Second, many school-level data are available for use in the analysis, such as for testing the RD assumptions. Note, however, that conducting inferences using school-level data is problematic because the number of schools within a small window—where the treatment can be thought of as good as random—will be very small. In addition, applying the conventional inferential approach that relies on large-sample approximations will not yield standard errors that are of correct size.

Cattaneo, Frandsen, and Titiunik (2015) developed an inferential method in the RD setting where the number of observations near the threshold is very small. They proposed using a finite-sample Fisherian approach to conduct inferences. The idea behind this proposed method is to conduct randomization tests (or permutation tests). Specifically, given the null hypothesis that no discontinuity exists in outcomes at the cutoff point, we permute the observations located at the left and right of the cutoff within the chosen bandwidth, with each observation always keeping its outcome values unchanged. For each permutation, we compute the test statistic. After deriving all the test statistics for each permutation, we construct the permutation distribution of the statistics and find the p-value by locating the original true test statistic on the permutation distribution. Standard errors based on the Fisherian approach are finite-sample exact and are immune to clustering issues. This approach allows researchers to conduct inferences without relying on

asymptotic approximations.

In this study, we adopt a randomized-based approach to estimate the treatment effect and conduct the randomization inference developed by Cattaneo, Frandsen, and Titiunik (2015). For the purpose of robustness check, estimation results based on the continuity-based approach, which makes use of student-level data, are presented in Appendix (Table A.1). The estimated treatment effects (i.e., coefficient estimates) are similar in magnitude and direction. However, the statistical significance is quite different compared with those obtained from the randomized-based approach. All coefficient estimates for outcome variables are statistically significant under the continuity-based approach, which are expected given the arguments provided by the previous literature (e.g., clustering, small and unbalanced number of clusters).

IV. Tests of Identifying Assumptions

4.1. Manipulation of an Assignment Variable

One of the identifying assumptions in an RD design is that individuals have imprecise control over the assignment variable. The assumption in the context of this study is that schools cannot manipulate their share of underachieving students. We test for the assumption using the density test proposed by McCrary (2008), which tests the null hypothesis of the continuity of the density of the assignment variable as it crosses the eligibility cutoff.

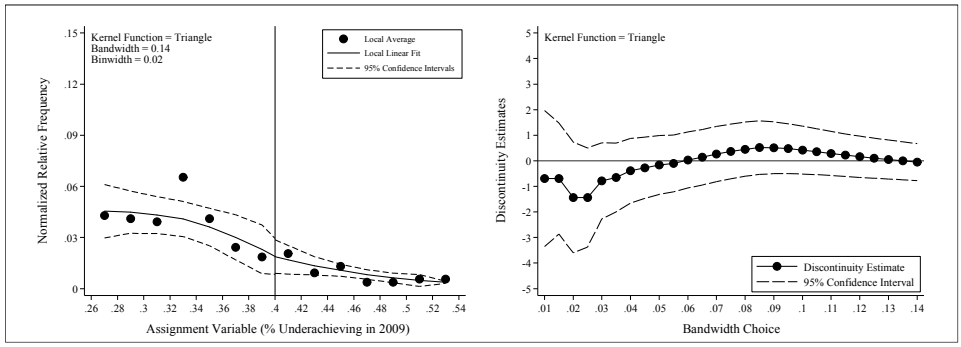
The left panel in Figure 2 depicts the density of the share of underachieving students in equally spaced bins with separate fitted lines left and right of the cutoff based on local linear regression (the dotted lines correspond to the 95% confidence interval). For the sake of consistency, we use the same bandwidth and binwidth, and the local linear fit throughout the paper when presenting the density of variables.[10] As can be seen from the histogram, no statistically significant discontinuity is observed in the density of the assignment variable at the eligibility cutoff (i.e., 0.4). Moreover, the 95% confidence intervals overlap at the cutoff.

Testing for the sensitivity of a discontinuity estimate to the choice of a bandwidth is necessary in any RD application. The right panel in Figure 2 shows the results from our sensitivity test. The graph presents varying discontinuity estimates by bandwidth choice with the corresponding 95% confidence interval. As the graph shows, none of the discontinuity estimates are statistically significant at the 5% level. Moreover, the 95% confidence interval encompasses the zero horizontal line.

---

[10] The local linear fit is based on the triangle kernel function with a bandwidth and binwidth equal to 0.14 and 0.02. Note that the fit is insensitive to the choice of the kernel function and bandwidth.

[Figure 2] Tests of Manipulation Using McCrary's Density Tests



Although the results of the density test show no sign of discontinuity at the threshold, the density test may still fail to detect the manipulation of the assignment variable when the number of schools manipulating the share of underachieving students to increase is offset by the number of schools manipulating the share to decrease. McCrary (2008) emphasized that a researcher should not only conduct a formal test of discontinuity at the cutoff point but also examine the institutional background regarding how the assignment variable and treatment are determined.

Schools in Korea face two incentives under the school accountability system. First, schools do not want to be stigmatized as low-performing schools and may manipulate their shares of underachieving students to avoid this designation. Second, schools—especially those that are underperforming and have insufficient resources—may prefer to receive school funding, even at the risk of stigmatization. Therefore, the setting is susceptible to the case mentioned above.

Nevertheless, schools cannot game the assignment variable for several reasons. First, schools cannot manipulate the scores of students because schools do not grade the tests. Once the tests are completed, schools must send the answer sheets to the Office of Education in each city immediately. Afterward, the Office of Education sends the sheets to the central government, which then grades all the tests. Second, answers to the tests are not disclosed until the tests are over. Third, even if we assume that schools are able to manipulate the test scores, gaming the share of underachieving students is virtually impossible because the eligibility cutoff is announced by the central government long after the tests are conducted. Lastly, the score range that determines the achievement category of each student is disclosed when the students receive their scores. Thus, school administrators or teachers gaming the scores of their students, let alone manipulating their placement to be placed at approximately the 40% cutoff, is highly unlikely.

[Table 2] Randomization inference for tests of balance in baseline covariates

	Bandwidth choice					
Outcome variable	$h = 1.5$	$h = 2.0$	$h = 2.5$	$h = 3.0$	$h = 3.5$	$h = 4.0$
Reading percentile rank	0.719	−0.166	0.229	0.509	0.693	−0.837
	(0.641)	(0.894)	(0.855)	(0.669)	(0.543)	(0.520)
Math percentile rank	0.378	−0.324	−1.158	−0.856	−0.741	−0.308
	(0.763)	(0.786)	(0.393)	(0.524)	(0.506)	(0.736)
English percentile rank	−0.817	−1.369	−0.737	−0.608	−1.228	−1.601**
	(0.480)	(0.192)	(0.404)	(0.497)	(0.160)	(0.038)
Total percentile rank	−0.507	−1.165	−0.777	−0.517	−0.608	−1.507*
	(0.595)	(0.176)	(0.288)	(0.475)	(0.390)	(0.053)
% living with both parents	0.017	0.006	0.004	0.010	−0.024	−0.014
	(0.608)	(0.816)	(0.873)	(0.707)	(0.588)	(0.681)
% receiving government subsidy	−0.005	−0.006	−0.006	−0.002	0.004	0.002
	(0.880)	(0.813)	(0.795)	(0.931)	(0.832)	(0.923)
% receiving private tutoring	−0.007	−0.005	−0.014	−0.014	0.037	0.029
	(0.762)	(0.798)	(0.423)	(0.348)	(0.434)	(0.499)
Class size	2.560	1.712	0.321	−0.140	0.076	0.258
	(0.220)	(0.353)	(0.842)	(0.935)	(0.964)	(0.884)
Student-teacher ratio	−0.762	−0.289	−1.101	−1.420	−1.396	−1.190
	(0.584)	(0.820)	(0.278)	(0.177)	(0.130)	(0.141)
% newly hired teachers	−0.031	−0.002	−0.015	−0.006	−0.009	−0.018
	(0.277)	(0.951)	(0.563)	(0.805)	(0.745)	(0.451)
% part-time teachers	−0.046	−0.030	−0.025	−0.031	−0.022	−0.026
	(0.401)	(0.529)	(0.473)	(0.357)	(0.477)	(0.308)
No. of students within h	2,538	3,553	4,288	4,828	5,526	6,743
No. of schools within h	16	21	26	28	33	39

Notes: Point estimates correspond to the difference in means between the left and right of the cutoff under the constant polynomial specification. p-values—calculated from the randomization inference in the regression discontinuity design proposed by Cattaneo, Frandsen, and Titiunik (2015)—are presented in parentheses. h denotes the size of bandwidth (in the percentage format). The uniform kernel function is used for the estimation (the results rarely change when the triangle kernel function is used instead). The randomization inference is conducted with 1,000 permutations. *, **, and *** indicate statistical significance at the 10, 5, and 1 percent level, respectively.

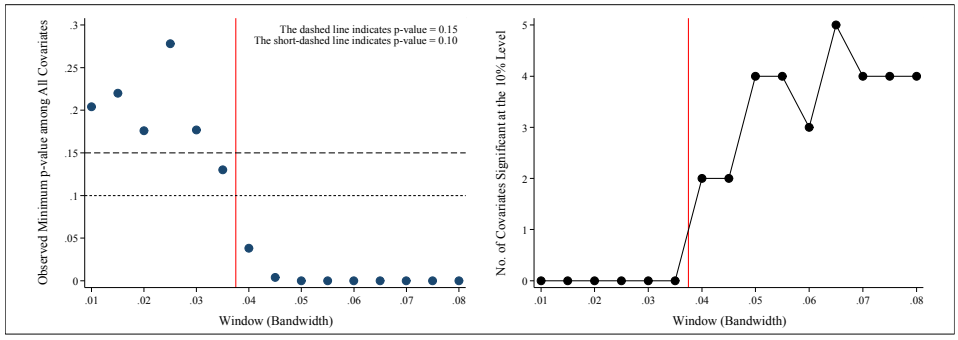
4.2. Balance in Baseline Covariates

If schools have no control over the assignment variable and the notion of the locally randomized setting is pervasive, predetermined covariates must be balanced across the cutoff. Given that this study applies the Fisherian randomization-based framework for RD analysis, the first step is to determine the window (or bandwidth) in which the locally randomized assumption is plausible on the basis of balance in baseline covariates (Cattaneo, Titiunik, and Vazquez-Bare, 2017). The proposed window selection procedure treats each baseline covariate as an outcome variable

and tests the null hypothesis of no effect in the window of increasing width. Therefore, the chosen window is the largest window such that the smallest *p*-value among all the tests is above a chosen cutoff (Cattaneo, Frandsen, and Titiunik, 2015).

Our choice of window is based on the 11 predetermined covariates shown in Table 2. The left panel in Figure 3 plots the minimum *p*-value obtained from the procedure as a function of window choice.[11] We start with a window of 0.015 and increase it by 0.005 at each step. As the figure shows, none of the test results produce a *p*-value less than 0.1 until the window length equals 0.035. When the chosen window length is equal to or greater than 0.04, the estimated minimum *p*-value is below the conventional 5% level. The right panel in Figure 3 depicts the number of baseline covariates that are statistically significant. None of the variables are statistically significant when the choice of window is less than 0.04. Starting from 0.04, however, some variables turn out to be statistically significant starting from 0.04. Thus, the local randomization setting is plausible within the 0.035 window.

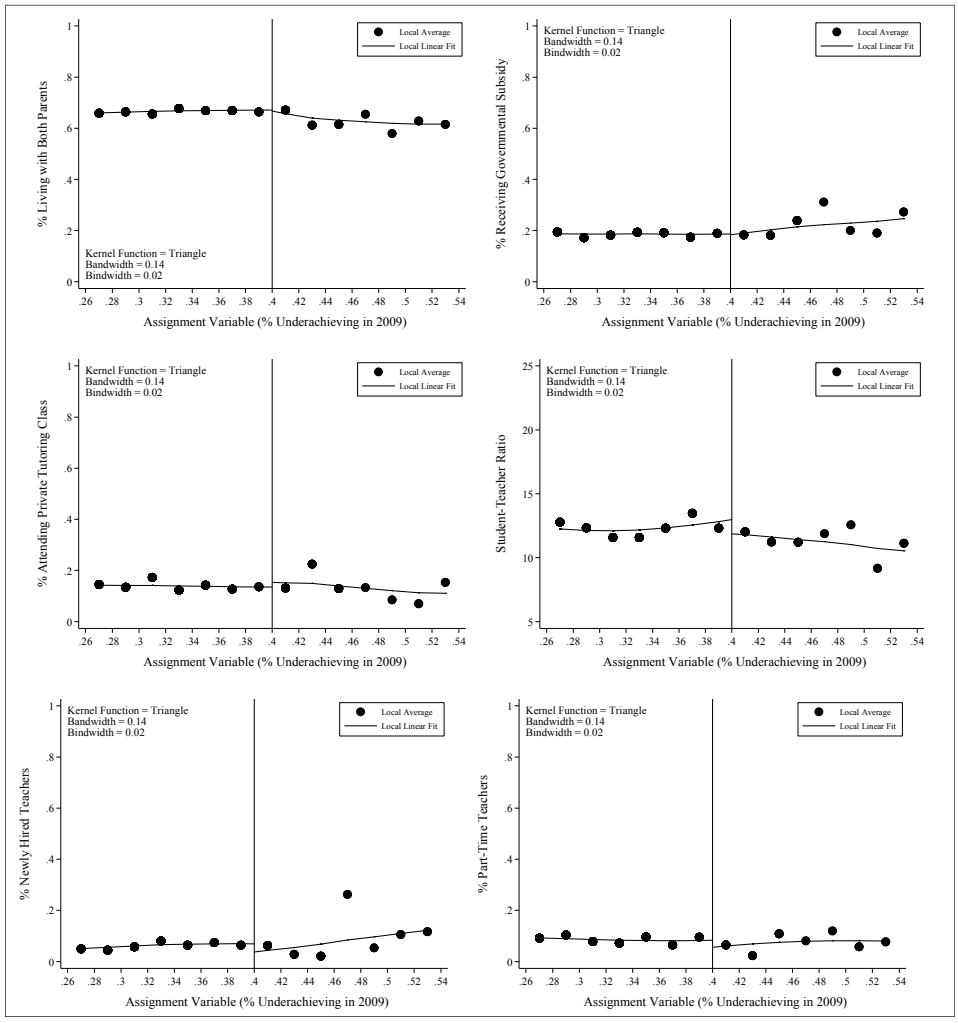
[Figure 3] Minimum *p*-values and the No. of Statistically Significant Covariates



For the sake of transparency, we provide the densities of the baseline covariates in Figure 4. We do not juxtapose the confidence intervals because the intervals based on the conventional inferential method are not used to conduct the statistical inference. Instead, we conduct randomization inference. As all the panels in Figure 4 indicate, the densities are smooth across the assignment variable, and discontinuities do not seem to be present at the 40% cutoff. For example, the share of students living with both parents is approximately 70%, and the share is not different across the assignment variable. The share of students who receive private tutoring is approximately 18%, and the share is stable across the assignment variable.

[11] The test statistic used in the procedure is the difference in the means statistic. The results rarely change if the Kolmogorov-Smirnov statistic is used.

[Figure 4] Continuity in Baseline Student, Teacher, and School Characteristics



While the figures show no discernible differences in the baseline characteristics across the assignment variable, we provide formal test results based on randomization inference in Table 2 to show the balance in predetermined covariates. The table presents point estimates of the difference in means between the left and right of the cutoff and the sensitivity of the estimates to the choice of window selection. The first four variables indicate baseline achievements. As the point estimates show, the difference is very small. Most of the point estimates are less than one percentile point. The next three variables are proxy for family income, and the estimated difference is trivial (0 to 2 percentage points). Class size and the student-teacher ratio are also similar. The difference is approximately zero to two students—a finding we expect because the central government puts considerable

effort into homogenizing school-level characteristics. Finally, the difference in teacher quality, based on the share of newly hired and part-time teachers, appear to be minimal. As can be predicted from the window selection procedure results in Figure 3, none of the baseline covariates are statistically significant when the choice of window is less than 0.04. Moreover, the magnitude of the differences is small. When the bandwidth choice is 0.04, the estimated difference for the two variables (i.e., English and Total) become statistically significant, though the magnitude is small (approximately 1.5 percentile points). The randomization test results lend support to the local randomization assumption when the comparison is based on schools within a window length of 0.035.

4.3. Strategic Behavior, Attrition, and Mean Reversion

In any policy analysis, estimated effects may not reflect policy effects if individuals behave strategically. In the context of school accountability literature, many studies have found that schools often engage in strategic behavior, such as test score and test-taking pool manipulation (e.g., Jacob and Levitt, 2003; Figlio and Winicki, 2005; Jacob, 2005; Cullen and Reback, 2006; Figlio, 2006; Rouse et al., 2013).

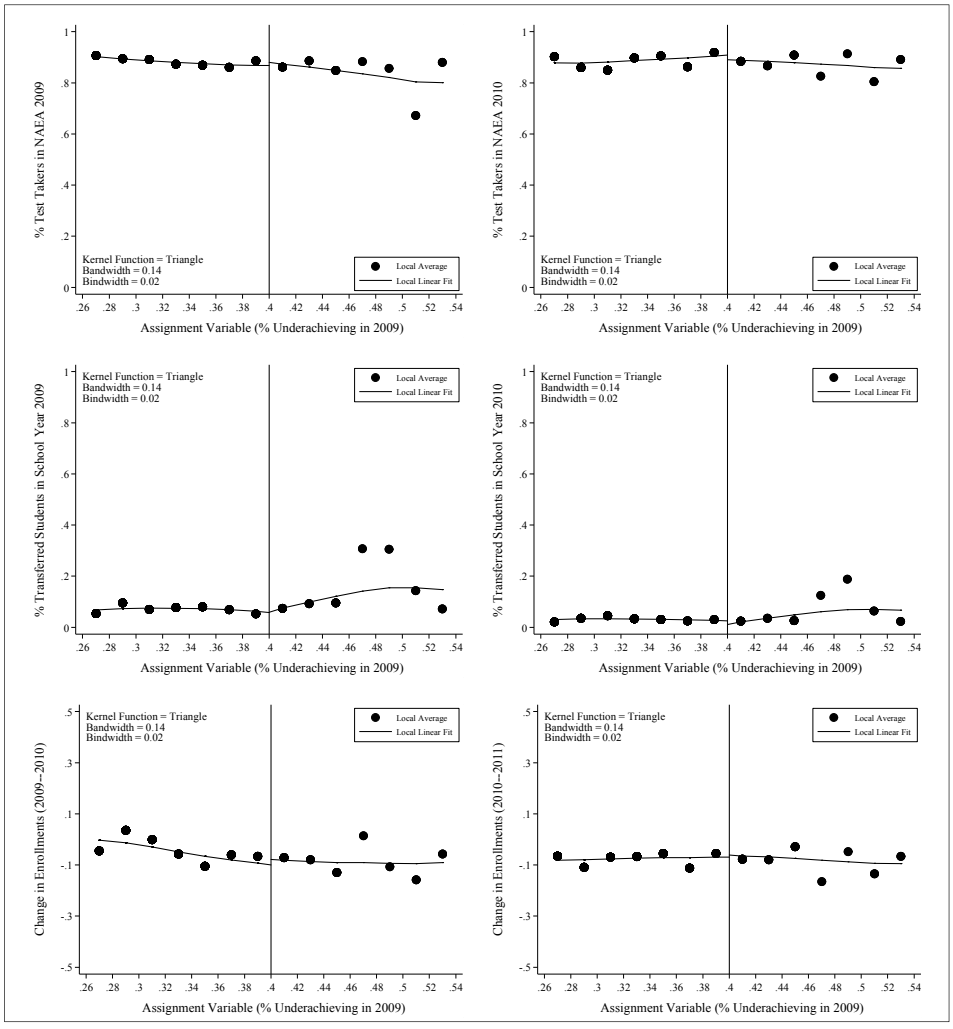
In the context of this study, manipulating test takers and test scores is infeasible because the personnel from the Office of Education inspects the NAEA administration. Furthermore, every student in Korea must take the test. Nevertheless, we test for the balance in the share of test takers, and the results are presented in Figure 5 and Table 3. The top two panels in Figure 5 display the densities of the share of test takers in NAEA 2009 and 2010. As the panels show, the share is approximately 0.9 with no signs of discontinuity at the 40% cutoff. Some schools at approximately 50% cutoff have a share of test takers at 0.75, but these schools are not part of the analysis sample used in this study. We use a window length of 0.035 for the randomization inference. The first two rows of Table 3 show the discontinuity estimates. As the table shows, none of the differences are statistically significant, and the estimated differences are all small (i.e., from -0.002 to -0.035).

Meanwhile, effect estimates may also be biased when significant differences are observed between the treatment and control groups in the selection in and out of the sample after the treatment (e.g., noncompliance and attrition). To test for this condition, we examine the differences in the share of student transfers and enrollment changes. The two middle panels in Figure 5 show the proportions of transferred students in 2009 and 2010. In Korea, students rarely transfer to other schools. Consequently, the proportions are small—at approximately 0.05 for 2009 (i.e., the first year of high school) and 0.02 for 2010 (i.e., the second year of high school). Moreover, the densities are smooth across the assignment variable and are

continuous at the threshold. Some outliers are present within 0.46 and 0.48. However, they are not part of the analysis sample. Table 3 presents the formal estimation results. In 2009, the estimated differences are in the range of 0.015 and 0.034, and all of the estimates are statistically insignificant. The differences are even smaller for 2010—at −0.013 or below.

The last two panels in Figure 5 display the densities of enrollment changes between two periods: 2009–2010 and 2010–2011. Again, the densities are smooth and continuous across the assignment variable with the percent change being close to −0.1 for most schools. The estimated differences in the percent change between the treated and untreated schools are less than one percentage point, with the exception of the estimate under the bandwidth choice of 0.015 (i.e., −0.047).

[Figure 5] Continuity in Test-takers, Student Transfers, and Enrollments



[Table 3] Tests of balance in test takers and student movements

	Bandwidth choice					
Outcome variable	$h = 1.5$	$h = 2.0$	$h = 2.5$	$h = 3.0$	$h = 3.5$	$h = 4.0$
% test takers in NAEA 2009	−0.017	−0.024	−0.012	−0.011	−0.003	−0.002
	(0.493)	(0.308)	(0.638)	(0.635)	(0.904)	(0.908)
% test takers in NAEA 2010	−0.012	−0.035	−0.023	−0.025	−0.008	−0.008
	(0.817)	(0.430)	(0.607)	(0.537)	(0.825)	(0.787)
% student transfers in 2009	0.026	0.022	0.034	0.028	0.015	0.018
	(0.357)	(0.302)	(0.169)	(0.213)	(0.518)	(0.335)
% student transfers in 2010	−0.013	−0.007	−0.005	−0.005	−0.002	0.000
	(0.216)	(0.440)	(0.451)	(0.465)	(0.868)	(0.997)
Change in enrollments	−0.047	−0.005	−0.009	−0.001	−0.009	−0.011
between 2009 and 2010	(0.569)	(0.944)	(0.894)	(0.984)	(0.853)	(0.782)
Change in enrollments	0.007	−0.022	−0.007	−0.008	0.005	0.009
between 2010 and 2011	(0.919)	(0.689)	(0.895)	(0.854)	(0.898)	(0.773)
No. of students within h	2,538	3,553	4,288	4,828	5,526	6,743
No. of schools within h	16	21	26	28	33	39

Notes: Point estimates correspond to the difference in means between the left and right of the cutoff under the constant polynomial specification. p-values—calculated from the randomization inference in the regression discontinuity design proposed by Cattaneo, Frandsen, and Titiunik (2015)—are presented in parentheses. h denotes the size of bandwidth (in the percentage format). The uniform kernel function is used for the estimation (the results rarely change when the triangle kernel function is used instead). The randomization inference is conducted with 1,000 permutations. *, **, and *** indicate statistical significance at the 10, 5, and 1 percent level, respectively.

Note, however, that all the estimates are statistically insignificant. Based on the randomization inference for the tests of balance in test takers and student movements, we conclude that this study does not suffer from issues induced by, for example, the strategic behavior of schools and selection in and out of the sample.

Lastly, Kane and Staiger (2002) noted that any estimated positive effects may be driven simply by mean reversion (i.e., regression to the mean) rather than the treatment itself. Mean reversion arises because schools with lower scores in a given year tend to obtain higher scores in the subsequent year for two reasons. First, a one-time event that adversely affects the test scores of students may occur (e.g., construction noise). Second, the test scores of students may have a sample variation, as each cohort that is tested is a random draw from the population. The likelihood that this study suffers from the regression to the mean issue is low for several reasons. First, although we cannot test for the first case, the second case is exceedingly unlikely because every student took the NAEA exam. Moreover, the students were the same between 2009 and 2010. Second, the effect estimates are derived from the RD setting. Chay, McEwan, and Urquiola (2005) showed that RD designs are favorable for eliminating the mean reversion problem because the

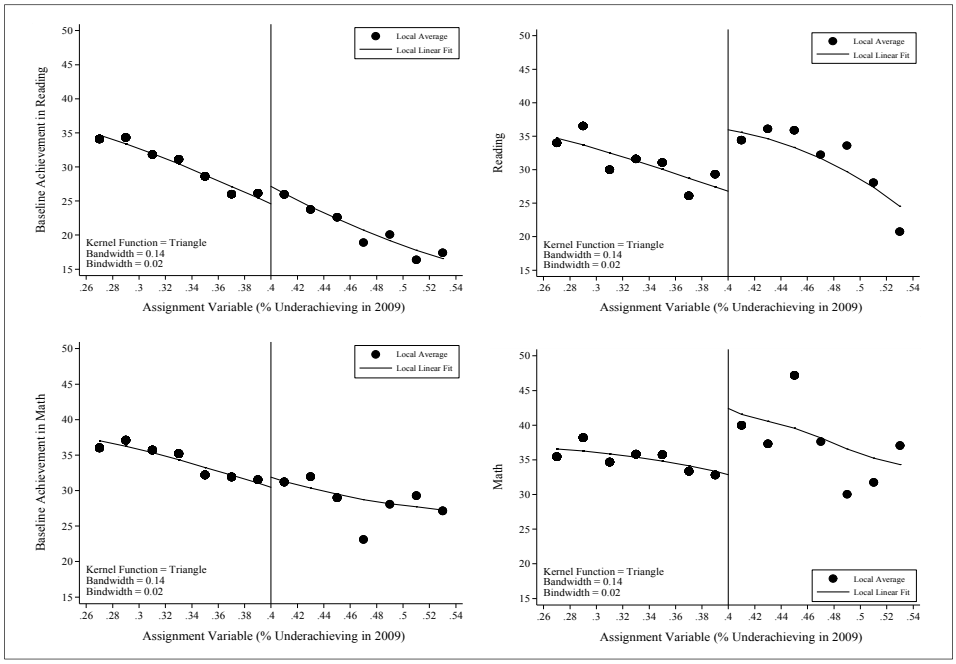
unobservable factors that affect mean reversion are likely to be similar for schools around the eligibility cutoff. Thus, the likelihood that mean reversion drives the effect estimates in this study is low.

V. Estimation Results

5.1. Effects on Test Scores

We provide results for the first set of outcome variables: test scores. Figure 6 displays the densities of the "Reading" and "Math" variables. The left panels in the figure correspond to the no-treatment period (i.e., 2009), whereas the right panels show post-treatment achievements (i.e., 2010). As the left panels show, the densities are smooth across the assignment variable with no indication of discontinuities at the cutoff. This observation coincides with the balancing test results presented in Table 2. However, as the right panels show, the achievements of most of the treated schools are higher than those that are not treated, with a discernible discontinuity at the 40% cutoff. Figure 7 presents the same information for the "English" and "Total" variables. As the two panels on the left indicate, our analysis show no discernible discontinuity in the baseline achievements at the cutoff. Meanwhile, as the two panels on the right show, discontinuities are present at the threshold.

[Figure 6] Densities of Reading and Math Achievement



[Figure 7] Densities of English and Total Achievement

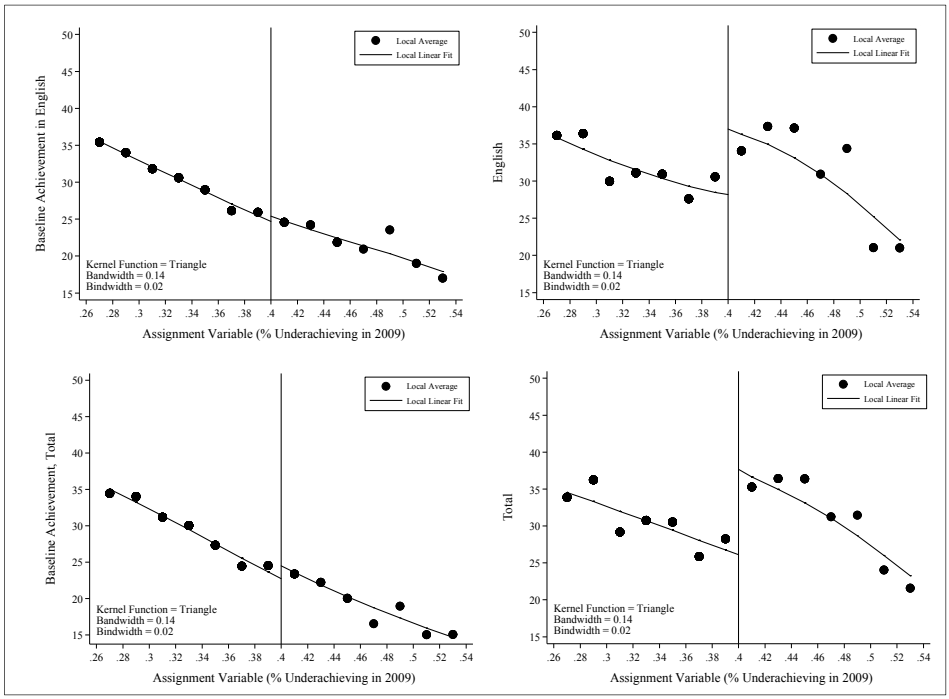


Table 4 presents estimation results for reading, math, English, and total scores. As we mentioned in the Empirical Strategy subsection, we need to collapse the data by school to conduct statistical inference correctly. We also present the estimation results based on non-collapsed data in Table A.1. Panel A in Table A.1 shows the results of the test for continuity in baseline achievement. While the discontinuity estimates vary to some extent across subjects and bandwidth choice, all the coefficients are practically and statistically insignificant with the exception of one estimate for English under the bandwidth choice of 0.02. Panel B presents the post-treatment effects. All of the effect estimates are statistically significant at the 1% level, and the magnitudes of the estimates are stable across the bandwidth choice.

Table 4 shows that the estimated treatment effect for reading ranges from 5 to 7 percentile points. For math, the effects are slightly larger—between 6 and 9 percentile points. The corresponding effect estimates for the "English" and "Total" variables are also similar to those for the "Reading" and "Math" variables. Moreover, most of the coefficient estimates are similar in magnitude to those from the student-level data analysis. Note, however, that the estimated effects are statistically less significant than those obtained from the student-level data—a finding we have expected given the issues mentioned in the Empirical Strategy subsection. The effect estimates under the bandwidth choice of 0.015 and 0.020 are statistically significant at the 10% level only for the "Math" variable. The imprecision of the

effect estimates based on the narrow window choice likely stem from the small number of schools within this window.

Based on the proposed bandwidth choice of 0.035, where the local randomization assumption is plausible, the estimated treatment effects for reading, math, and English are 7.323, 6.294, and 5.714 percentile points, respectively. All these estimates are statistically significant at either the 10% or 5% levels.

[Table 4] Randomization inference for test outcomes

	Bandwidth choice					
Outcome variable	$h = 1.5$	$h = 2.0$	$h = 2.5$	$h = 3.0$	$h = 3.5$	$h = 4.0$
Reading	7.063	5.097	7.314**	7.496**	7.323**	7.433***
	(0.104)	(0.179)	(0.032)	(0.015)	(0.014)	(0.008)
Math	9.496*	7.185	6.664**	6.169*	6.294**	6.045***
	(0.091)	(0.138)	(0.038)	(0.078)	(0.048)	(0.011)
English	5.245	3.485	5.737	6.161*	5.714*	6.203**
	(0.386)	(0.487)	(0.118)	(0.085)	(0.072)	(0.029)
Total	9.677	6.998	8.936**	9.000***	8.665**	8.710***
	(0.143)	(0.226)	(0.033)	(0.010)	(0.012)	(0.000)
% underachieving	−0.084	−0.058	−0.108**	−0.104**	−0.105**	−0.103**
(reading)	(0.232)	(0.336)	(0.050)	(0.039)	(0.028)	(0.019)
% underachieving	−0.092**	−0.074**	−0.057**	−0.054*	−0.056**	−0.052**
(math)	(0.042)	(0.030)	(0.032)	(0.058)	(0.047)	(0.019)
% underachieving	−0.037	−0.029	−0.057	−0.062	−0.065	−0.078***
(English)	(0.613)	(0.593)	(0.209)	(0.181)	(0.109)	(0.026)
% average or above	0.047*	0.033	0.048**	0.050***	0.048***	0.048***
(reading)	(0.066)	(0.117)	(0.016)	(0.003)	(0.008)	(0.003)
% average or above	0.103	0.070	0.067	0.062	0.058	0.057
(math)	(0.595)	(0.862)	(0.400)	(0.398)	(0.293)	(0.155)
% average or above	0.012	0.005	0.012	0.012	0.009	0.007
(English)	(0.306)	(0.718)	(0.288)	(0.244)	(0.269)	(0.324)
No. of students within h	2,538	3,553	4,288	4,828	5,526	6,743
No. of schools within h	16	21	26	28	33	39

Notes: Point estimates correspond to the difference in means between the left and right of the cutoff under the constant polynomial specification. p-values—calculated from the randomization inference in the regression discontinuity design proposed by Cattaneo, Frandsen, and Titiunik (2015)—are presented in parentheses. h denotes the size of bandwidth (in the percentage format). The uniform kernel function is used for the estimation (the results rarely change when the triangle kernel function is used instead). The randomization inference is conducted with 1,000 permutations. *, **, and *** indicate statistical significance at the 10, 5, and 1 percent level, respectively.

5.2. Effects by Achievement

One concern raised in studies of school accountability is that educators may devote their resources to improving their signals of effectiveness rather than to

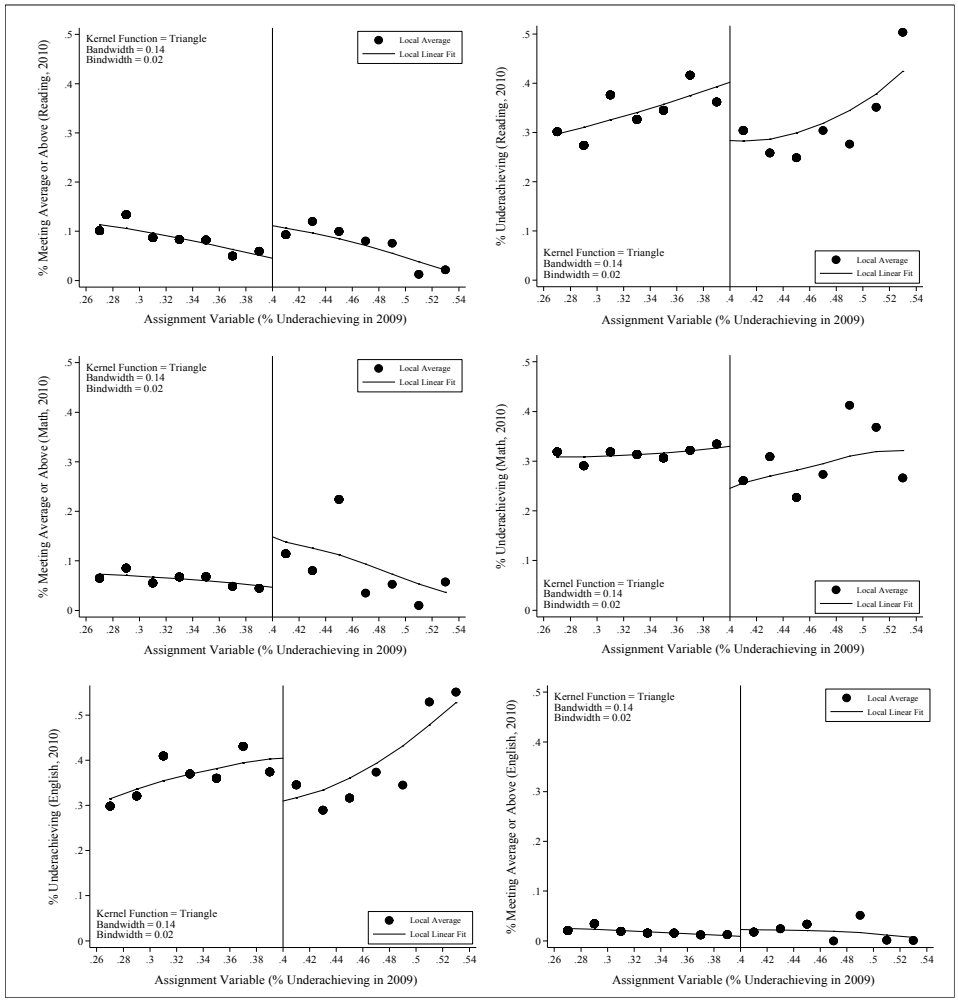
affecting the underlying achievement of students. Whether a school is stigmatized as low-performing is determined by its share of underachieving students. Accordingly, schools may focus only on promoting the achievements of underachieving students so that they can free themselves of the stigmatization. The fact that the stated policy goal of the accountability system is to promote the achievements of underperforming students makes this condition especially likely. Nevertheless, we examine whether the share of underachieving students decreases in the treated schools and whether the share of students who receive the achievement category of "average or above" increases in the treated schools.[12]

Figure 8 displays the density plots of these shares. The three panels on the right show the changes in the share of underachieving students, whereas the three panels on the left display the share of students who meet or exceed the average level. With the exception of English, our analysis shows a discontinuity in the share of students who meet or exceed the average level. In contrast, the figure shows a clear visual break at the eligibility cutoff for the share of underachieving students, indicating a decrease in this type of students in the treated schools. Table 4 shows the results of randomization inference. In all subjects, the share of underachieving students in treated schools decreases more significantly than that of untreated schools. The estimated decreases are approximately 10, 5, and 7 percentage points for reading, math, and English, respectively, and the differences are statistically significant with the exception of English.

The share of students who receive the achievement category of average or above for reading in the treated schools is approximately 5 percentage points higher than that of untreated schools, and most of the estimates are statistically significant. The shares for math and English are also higher in the treated schools (approximately 6 and 1 percentage points, respectively). However, these estimates are statistically insignificant. Although a strong conclusion cannot be drawn from these estimates, as "Math" and "English" estimates are imprecisely estimated for the % average or above in the "Math" and "English" variables, they suggest that the achievements of underachieving and other students improve in treated schools relative to untreated schools.

[12] Few vocational high school students were placed in the "proficient" achievement category. Thus, analyzing this share is not interesting.

[Figure 8] Densities of the Share of Average or Above & Underachieving Students



5.3. Effects on Post-secondary Outcome

Another concern related to the achievement effects of accountability systems is that incentives induced by the system may not promote long-term outcomes, such as graduation and college matriculation, because educators are likely to focus on short-term outcomes, such as test scores. In particular, given that schools located in urban areas face high levels of competition, the policy is more likely to be effective in the short term than in the long term in these areas, and its effects may not be long-lasting.

Note, however, that schools located in rural areas often lack financial resources but have more autonomy than schools in urban areas because of less pressure from the school accountability system. Thus, the policy may have more long-lasting

effects for schools located in rural areas. Relatively few studies have examined the long-term impacts of accountability systems. Therefore, this study contributes to the literature by conducting subgroup analyses of long-term impacts of the system on graduation, test-taking behavior, and college matriculation for urban and rural areas.

The Office of Education does not keep track of cohort graduation rates, which are accurate for examining treatment effects. As a result, we analyze the share of graduates among those who have proceeded to the third year (i.e., the final year of high school), as the SIW reports this variable. Figure 9 presents the densities of the three post-secondary outcomes for urban and rural areas. As the figure shows, we do not observe any discernible discontinuity for all the post-secondary outcomes for schools located in urban areas. For schools located in rural areas, however, we find a clear discontinuity in the share of students matriculating into a four-year college.

[Figure 9] Discontinuity in the Post-Secondary Outcomes by Urban/Rural Area

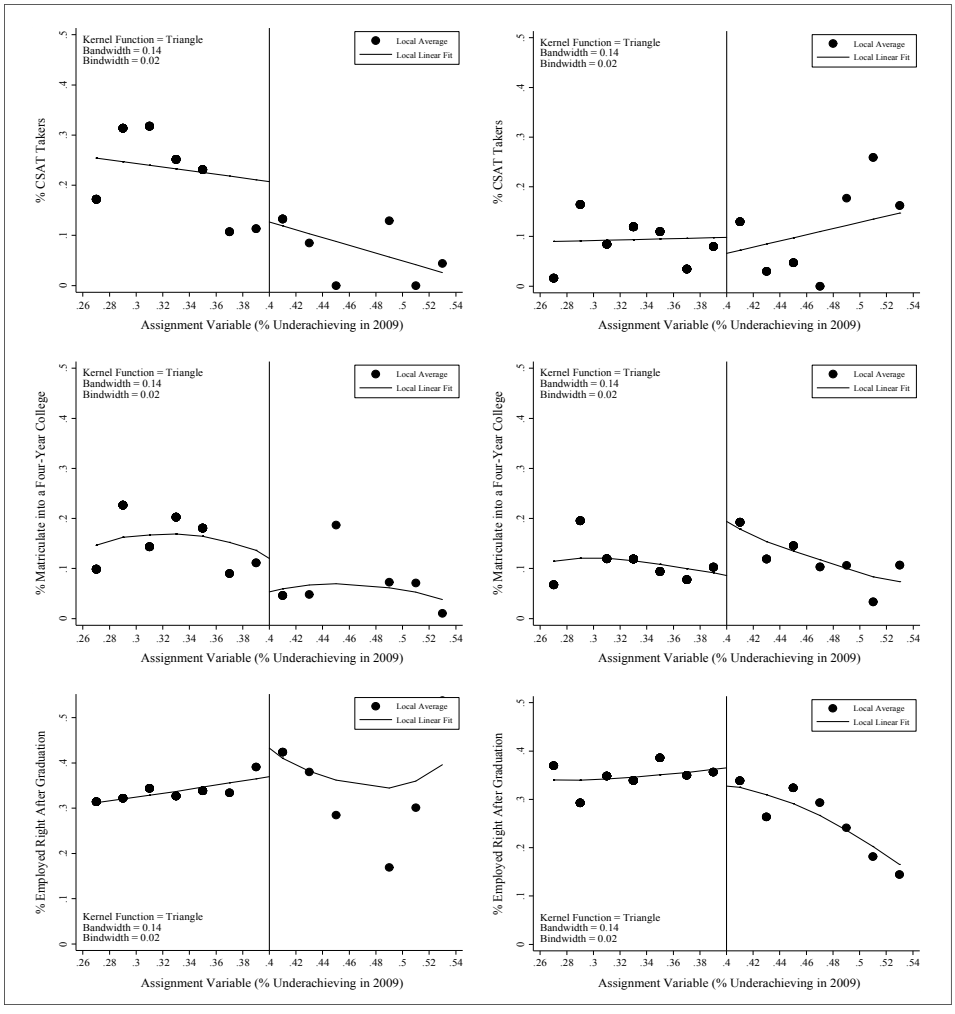


Table 5 shows the formal results. The first four rows present the results of randomization inference for the test scores. As expected, we find practically and statistically significant effects on test scores for schools located in urban and rural areas. The results imply that the policy is effective in promoting short-term outcomes, such as test scores. The next five rows show the estimated effects for post-secondary outcomes. As these results indicate, we determine that the treatment has no effect on graduation rates for both types of schools. The estimated differences are all close to zero. One reason for not seeing any effect on graduation is that almost all students who proceeded to the last year of high school have graduated. The estimated mean share of graduates based on the analysis sample used in this study is 0.982.

To analyze the effect of the policy on long-term student behavior, we use information from the 2012 CSAT data. Most colleges in Korea require students to submit CSAT scores.[13] If the treatment not only promotes test scores but also induces students to go to college, then we should see a difference in the share of students taking the CSAT between the treated and untreated schools. As Table 5 shows, we find practically significant effects on this variable only for schools located in rural areas. The difference in the share of students taking CSAT is approximately 0.05 for rural schools.[14] The estimated mean share of vocational high school students taking CSAT is 0.290. Thus, the estimated difference accounts for approximately 17% of the mean. The difference, though practically significant, is statistically insignificant.

Table 5 also shows the results of tests for the differences in three post-secondary outcomes: the shares of students matriculating into two-year colleges, the shares of students matriculating into four-year colleges, and the share of students entering the labor market right after graduation.[15] The effects on two-year college matriculation are all statistically and practically insignificant for both types of schools. Note, however, that we observe statistically and practically significant effects for students matriculating into four-year colleges only for schools located in rural areas. These findings coincide with our expectation that the policy is more effective in promoting longer-term outcomes for schools located in rural areas than for those in urban areas. Strong conclusions cannot be drawn from the randomization inference results for the share of students employed right after graduation because none of the estimates

[13] To test for the difference in the number of students taking CSAT, we obtained the CSAT 2012 data from the Ministry. The Ministry allows researchers to apply for the CSAT data. We applied for the data on CSAT 2012 because CSAT 2012 was conducted at the end of 2011, the year in which the students in our sample entered their third year of high school.

[14] We calculated the share of test takers as the number of test takers divided by the number of third-year students.

[15] The share of students entering the labor market is the proportion of third-year students who entered the labor market right after graduation.

are precisely estimated.

[Table 5] Randomization inference for the outcomes by urban or rural area

	Bandwidth choice			
Outcome variables	$h = 3.0$	$h = 3.5$	$h = 3.0$	$h = 3.5$
	Panel A: Urban area		Panel B: Rural area	
Reading score	9.231**	5.304	6.881	7.735**
	(0.026)	(0.332)	(0.133)	(0.028)
Math score	−0.474	0.994	10.242**	8.233*
	(0.973)	(0.739)	(0.038)	(0.060)
English score	6.814*	3.755	5.432	5.665
	(0.081)	(0.423)	(0.339)	(0.197)
Total score	7.530*	4.840	9.893	9.685**
	(0.054)	(0.261)	(0.108)	(0.038)
% graduates among the third year	−0.001	−0.003	−0.005	0.002
	(0.914)	(0.751)	(0.525)	(0.810)
% took the college entrance exams	0.037	0.023	0.061	0.047
	(0.490)	(0.615)	(0.334)	(0.393)
% matriculate into a two-year college	−0.096	−0.083	−0.030	−0.022
	(0.026)	(0.019)	(0.558)	(0.648)
% matriculate into a four-year college	−0.060	−0.053	0.097*	0.085*
	(0.226)	(0.202)	(0.083)	(0.068)
% employed right after graduation	0.104	0.067	−0.029	−0.011
	(0.303)	(0.567)	(0.627)	(0.842)
Number of schools within h	8	9	18	23

Notes: Seoul, Busan, Incheon, Daegu, Daejeon, Geong-gi, and Kwangju constitute urban areas. Point estimates correspond to the difference in means between the left and right of the cutoff under the constant polynomial specification. p-values—estimated from the randomization inference in the regression discontinuity design proposed by Cattaneo, Frandsen, and Titiunik (2015)—are presented in parentheses. h denotes the size of bandwidth (in the percentage format). The uniform kernel function is used for deriving the effect estimates (the results barely change when the triangle kernel function is instead used for estimation). The randomization inference is conducted with 1,000 permutations. *, **, and *** indicate statistical significance at the 10, 5, and 1 percent level, respectively.

VI. Discussion

This study is not free of limitations. First, the external validity of the findings may not be strong enough. The NAEA has been administered since 2008. This study uses the results from the 2009 and 2010 exams. Depending on the institutional context and perceived behavior of other agents, schools and students may overreact or underreact to the SINAI-designation. Additional RD analyses of

other years can certainly help promote external validity. However, the government no longer provides population data for these other years. Second, our analysis focuses on lower-achieving students. Examining whether the results observed in this study also manifest in other types of students can be valuable.

Third, the increases in student achievements in many of the subjects occurred relatively in a short period (approximately five months). Although schools may have initiated their remedial efforts prior to the announcement of the NAEA results in March 2010, the possibility that the SINAI-designated schools may have engaged in behavior such as teaching to the test cannot be precluded. We argue, however, that the improvement that has occurred not only for underachieving students but also for average or above students decreases the likelihood of schools engaging in teaching to the test behavior.[16] Finally, this study does not examine how the policy affects other cohorts (i.e., spillover effects). Given that the data for other cohorts are not available, determining whether the SINAI-designated schools have attempted to promote school-wide improvement is difficult. Answering this question also helps identify potential teaching to the test behavior.

VII. Conclusion

School-based accountability systems can be divided into two main types: low-stakes and high-stakes. This study analyzes the causal impact of the unique feature of the Korean school accountability system—the simultaneous use of stigmatization and school funding—on student test scores and post-secondary outcomes.

Using an RD design, we compare student achievements in treated schools with those in schools that are not treated but are purportedly similar in observable and unobservable baseline characteristics. Our analyses show that stigmatization and school funding have led to an increase of 7, 6, and 5 percentile points for reading, math, and English, respectively. We also find that the share of students classified as "underachieving" has declined by 10, 5, and 7 percentage points for reading, math, and English, respectively. The estimated reduction for English is, however, statistically insignificant at the conventional level. Furthermore, the share of students classified as "average performing or above" has also increased by approximately 5 percentage points for reading and 6 percentage points for math, though the math estimates are statistically insignificant.

Meanwhile, our subgroup analyses of post-secondary outcome variables for urban and rural areas show that the policy has a positive impact on the likelihood of students taking the college entrance exams (practically significant) and

[16] Some argue, however, that knowing something, at least, is better than knowing nothing (Lazear, 2006).

matriculating into a four-year college (practically and statistically significant) only for schools located in rural areas. Note, however, that strong conclusions cannot be drawn regarding the effect of the policy on other post-secondary outcomes, such as employment, because the estimates are estimated imprecisely. The fact that the accountability system in Korea is geared toward underperforming students may explain why we do not observe strong and significant effects on post-secondary outcomes for urban schools. If the system puts additional emphasis on promoting the achievement of students in general, we may see more improvement among average or above-average students because these students are more likely to graduate high school and matriculate into college than underperforming students. This study is limited in the sense that we do not conduct a formal test of mechanism variables because the necessary data are not available. Future studies should attempt to establish the formal causality of mechanism variables to derive additional policy implications regarding the use of school accountability systems.

Appendix

[Table A.1] RD estimates using student-level data

Variable	Bandwidth choice					
	$h = 1.5$	$h = 2.0$	$h = 2.5$	$h = 3.0$	$h = 3.5$	$h = 4.0$
			Panel A. Baseline achievement			
Reading	−0.883	2.674	2.127	2.375	2.696	4.292
	(0.419)	(0.392)	(0.380)	(0.435)	(0.292)	(0.531)
	[2,538]	[3,553]	[4,288]	[4,828]	[5,526]	[6,743]
Math	−0.692	−0.152	0.341	−0.696	−0.837	−1.520
	(0.832)	(0.844)	(0.749)	(0.129)	(0.602)	(0.708)
	[2,624]	[3,646]	[4,383]	[4,925]	[5,624]	[6,846]
English	2.470	−1.004**	−1.491	−1.598	−0.362	−0.005
	(0.879)	(0.036)	(0.187)	(0.323)	(0.338)	(0.336)
	[2,626]	[3,649]	[4,385]	[4,922]	[5,622]	[6,843]
Total	−0.971	0.122	−0.223	−0.262	0.482	1.692
	(0.783)	(0.803)	(0.684)	(0.737)	(0.694)	(0.142)
	[2,530]	[3,543]	[4,274]	[4,809]	[5,506]	[6,721]
			Panel B. Post-treatment effect			
Reading	6.803***	4.722***	6.729***	6.825***	6.106***	7.395***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	[2,475]	[3,452]	[4,120]	[4,612]	[5,276]	[6,412]
Math	8.862***	6.954***	6.747***	6.265***	6.406***	5.790***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	[2,490]	[3,472]	[4,145]	[4,644]	[5,312]	[6,450]
English	5.651***	4.238***	5.935***	6.317***	5.368***	6.583***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	[2,491]	[3,473]	[4,147]	[4,648]	[5,318]	[6,458]
Total	9.333***	6.883***	8.592***	8.592***	7.843***	8.647***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	[2,467]	[3,443]	[4,108]	[4,599]	[5,263]	[6,399]

Notes: Variables in percentile rank. *p*-values—based on the robust nonparametric standard errors estimator proposed by Calonico, Cattaneo, and Titiunik (2014)—are presented in parentheses. *h* denotes the size of bandwidth (in the percentage format). Effective sample size (i.e., the number of observations within a corresponding bandwidth) in brackets. RD estimates are obtained from running a local linear regression using the uniform kernel function (the estimates rarely change if the triangle kernel function is used instead). *, **, and *** indicate statistical significance at the 10, 5, and 1 percent level, respectively.

References

Akerlof, G. A., and R. E. Kranton (2005), "Identity and the Economics of Organizations," *Journal of Economic Perspectives*, 19(1), 9–32.

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

Black, S. E. (1999), "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*, 114(2), 577–599.

Borba, J. (2003), "California's Immediate Intervention/Underperforming Schools Program (II/USP): An Assessment of Principals' Perceptions after One Year," *Educational Research Quarterly*, 27(1), 45–62.

Burgess, S., C. Propper, H. Slater, and D. Wilson (2005), "Who Wins and Who Loses from School Accountability? The Distribution of Educational Gain in English Secondary Schools," The Centre for Market and Public Organisation Working Paper, Department of Economics, University of Bristol.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2019), "Regression Discontinuity Designs Using Covariates," *Review of Economics and Statistics*, 101(3), 442–451.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014), "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82(6), 2295–2326.

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

Cameron, A. C., and D. L. Miller (2015), "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 50(2), 317–372.

Cattaneo, M. D., B. R. Frandsen, and R. Titiunik (2015), "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate," *Journal of Causal Inference*, 3(1), 1–24.

Cattaneo, M. D., R. Titiunik, and G. Vazquez-Bare (2017), "Comparing Inference Approaches in RD Designs: A Reexamination of the Effect of Head Start on Child Mortality," *Journal of Policy Analysis and Management*, 36(3), 643–681.

Chakrabarti, R. (2013a), "Impact of Voucher Design on Public School Performance: Evidence from Florida and Milwaukee Voucher Programs," *B.E. Journal of Economic Analysis and Policy: Contributions*, 13(1), 349–394.

_____ (2013b), "Accountability with Voucher Threats, Responses, and the Test-Taking Population: Regression Discontinuity Evidence from Florida," *Education Finance and Policy*, 82(2), 121–167.

_____ (2014), "Incentives and Responses under No Child Left Behind: Credible Threats and the Role of Competition," *Journal of Public Economics*, 110, 124–146.

Chay, K. Y., P. J. McEwan, and M. Urquiola (2005), "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools," *American Economic Review*, 95(4), 1237–1258.

Chiang, H. (2009), "How Accountability Pressure on Failing Schools Affects Student

Achievement," *Journal of Public Economics*, 93(9–10), 1045–1057.

Conley, T. G., and C. R. Taber (2011), "Inference with "Difference in Differences" with a Small Number of Policy Changes," *Review of Economics and Statistics*, 93(1), 113–125.

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

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

Deming, D. J., S. Cohodes, J. Jennings, and C. Jencks (2016), "School Accountability, Postsecondary Attainment and Earnings," *Review of Economics and Statistics*, 98(5), 848–862.

Diamond, J., and J. Spillane (2004), "High-Stakes Accountability in Urban Elementary Schools: Challenging or Reproducing Inequality?" *The Teachers College Record*, 106(6), 1145–1176.

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

Figlio, D. N. (2006), "Testing, Crime and Punishment," *Journal of Public Economics*, 90(4–5), 837–851.

Figlio, D. N., and L. W. Kenny (2009), "Public Sector Performance Measurement and Stakeholder Support," *Journal of Public Economics*, 93(9–10), 1069–1077.

Figlio, D. N., and H. F. Ladd (2015), "School Accountability and Student Achievement," In: Ladd, H. F. & Goertz, M. F. (Ed.), Handbook of Research in Education Finance and Policy, 194–210.

Figlio, D. N., and M. E. Lucas (2004), "What's in a Grade? School Report Cards and the Housing Market," *American Economic Review*, 94(3), 591–604.

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

Gelman, A., and G. Imbens (2019), "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." *Journal of Business & Economic Statistics*, 37(3), 447–456.

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

Hart, C., and D. N. Figlio (2015), "School Accountability and School Choice: Effects on Student Selection across Schools," *National Tax Journal*, 68(3), 875–899.

Hastings, J. S., and J. M. Weinstein (2008), "Information, School Choice, and Academic Achievement: Evidence from Two Experiments," *Quarterly Journal of Economics*, 123(4), 1373–1414.

Jacob, B. 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, B., and S. D. Levitt (2003), "Rotten Apples: An Investigation of the Prevalence and

Predictors of Teacher Cheating," *Quarterly Journal of Economics*, 118(3), 843–877.

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

Krieg, J. M. (2008), "Are Students Left Behind? The Distributional Impacts of the No Child Left Behind Act," *Education Finance and Policy*, 3(2), 250–281.

Kuziemko, H., R. W. Buell, T. Reich, and M. I. Norton (2014), "Last-Place Aversion: Evidence and Redistributive Implications," *Quarterly Journal of Economics*, 129(1), 105–149.

Lazear, E. P. (2006), "Speeding, Terrorism, and Teaching to the Test," *Quarterly Journal of Economics*, 121(3), 1029–1061.

Lee, D. S. (2008), "Randomized Experiments from Non-Random Selection in U.S. House Elections," *Journal of Econometrics*, 142(2), 675–697.

Lee, D. S., and T. Lemieux (2010), "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48(June 2010), 281–355.

Lemke, R. J., C. M. Hoerandner, and R. E. McMahon (2006), "Student Assessments, Non-Test-Takers, and School Accountability," *Education Economics*, 14(2), 235–250.

MacKinnon, J. G., and M. D. Webb (2016), "Wild Bootstrap Inference for Wildly Different Cluster Sizes," *Journal of Applied Econometrics*, 32(2), 233–254.

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

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

Reback, R. (2008), "Teaching to the Rating: School Accountability and the Distribution of Student Achievement," *Journal of Public Economics*, 92(5–6), 1394–1415.

Rockoff, J., and L. J. Turner (2010), "Short-Run Impacts of Accountability on School Quality," *American Economic Journal: Economic Policy*, 2(4), 119–47.

Rouse, C. E., J. Hannaway, D. Goldhaber, and D. Figlio (2013), "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure," *American Economic Journal: Economic Policy*, 5(2), 251–281.

Roy, V., and F. Kochan (2012), "Factors that Facilitated an Alabama School Assistance Teams Success in a Low-Performing School," *International Journal of Educational Leadership Preparation*, 7(1), 1–14.

Sims, D. P. (2008), "Strategic Responses to School Accountability Measures: It's All in the Timing," *Economics of Education Review*, 27(1), 58–68.

West, M., and P. Peterson (2006), "The Efficacy of Choice Threats within School Accountability Systems: Results from Legislatively Induced Experiments," *Economic Journal*, 116(510), C46–C62.

Woo, S., S. Lee, and K. Kim (2015), "Carrot and Stick?: Impact of a Low-Stakes School Accountability Program on Student Achievement," *Economics Letters*, 137(510), 195–199.