Assessing the Short-term Impact of the New York City Renewal Schools Program

Isaac M. Opper, William R. Johnston, John Engberg, Lea Xenakis

RAND Education and Labor

For more information on this publication, visit www.rand.org/pubs/working_papers/WR1303.html

Abstract

In this internal report prepared for the New York City Chancellor's Office, we estimate short-term impacts of the New York City Renewal Schools Program based on the first two full years of program implementation. We utilize a novel method of multiple rating regression discontinuity design (MRRDD) that leverages the selection criteria used to select schools as Renewal or non-Renewal Schools. Our analysis suggests that the Renewal Schools program is helping to improve student attendance while reducing chronic absenteeism, while also increasing the amount of credits earned among high schools students. We also found evidence suggesting that among high school students, the program impact was strongest at schools with greater levels of student economic needs.

Table of Contents

Figures

Tables

Abbreviations

ELA	English Language Arts
ENI	Economic Need Index
MRRDD	Multiple Rating Regression Discontinuity Design
NYCDOE	New York City Department of Education
RS	Renewal Schools
RSCEP	Renewal Schools Comprehensive Education Plan

1. Introduction

The RAND Corporation has been commissioned to perform a preliminary analysis of the impact of the New York City Renewal Schools (RS) Program. This analysis combines a variety of data sources to assess the short-term effects of the program on student outcomes, through the 2016-17 school year.

The primary goal of the add-on task is to determine whether students of the Renewal Schools are doing better academically, socially, and emotionally than they would be doing had their schools not been designated as Renewal Schools. Designation as a Renewal School led to several mandated changes in school programming and the availability of additional resources. The report specifically addresses three research questions:

1. What is the impact of the Renewal School initiative on student outcomes?
2. Does the impact differ among schools or students with different characteristics?
3. Does the impact differ among schools with varying amounts of measurable implementation activities?

In order to determine if schools are doing better than they otherwise would have been in the absence of the program, we compare their outcomes in the 2015-16 and 2016-17 school years to the outcomes that we predict would have occurred for those schools in the absence of the Renewal School designation.[1] To simplify notation, we will often refer to the 2015-16 school year simply as 2016. These predicted outcomes are based on the pre-designation characteristics of each Renewal School and on the outcomes of a strategically chosen set of comparison schools.

In Figure 1 below, we present an updated theory of action that will guide our analysis of the Renewal Schools Program. The left-hand portion, in grey, outlines the key supports and structures that we will study to determine program implementation.[2] The two columns on the right, in blue and orange, represent key outcome domains.

---

[1] Although the RS was rolled out in 2014-15, it was not fully implemented until 2015-16. We will therefore focus our estimates on the effect of the program in 2015-16 and 2016-17, although we do discuss results on the effect in 2014-15 briefly in Section 3.

[2] Due to limited variation in administrative data related to the implementation domain of standards aligned instruction, particularly related to schools' implementation of iReady assessments and use of Strategic Data Check-ins for Regents planning, we did not include a measure of this domain in our final analysis.

Figure 1. Renewal Schools Analysis Theory of Action

Program Implementation	▶	Leading Indicators	▶	Lagging Indicators
Assess the needs of Renewal Schools – **Needs Assessment** ▶ Provide leadership development to support inclusive leadership focused on instruction – **Leadership Development** ▶ Develop strong family-community ties – **Parent Community Ties**; Develop programs to improve the capacity of educators – **Professional Capacity**; Create safe, welcoming, simulating, nurturing, and student-centered learning communities – **Student Centered Learning Climate**; Provide curricula support and develop standards-aligned instruction – **Standards Aligned Instruction**	▶	Student Attendance; Student Behavior; Instructional Core; School Culture	▶	Academic Achievement; Educational Attainment; College Readiness

Source: Authors' adaption of information presented by NYCDOE:
http://schools.nyc.gov/AboutUs/schools/RenewalSchools/default

We examine how schools are functioning on several outcomes at the student- and school-level. We refer to these outcomes as either "leading" or "lagging" indicators. Leading indicators provide early signals of program progress toward measurable growth or improvement, which allow organizations, like the DOE, to improve based on data. Lagging indicators represent more distal, final outcomes that confirm trends, but do not necessarily inform program improvements (Foley et al., 2008).

Leading Indicator Domains

1. Student attendance
2. Student behavior
3. The instructional core (curriculum, pedagogy, and assessment)
4. School culture

Lagging Indicator Domains

1. Academic achievement
2. Educational attainment
3. College readiness

In the sections that follow, we describe the methodology used to answer each of the research questions, which is then followed by a detailed presentation of study results. We conclude with a consideration of study limitations and offer some directions for future work.

2. Methods and Data

Methods

To answer our first research question regarding the overall impact of the Renewal Schools program, we implemented a multiple rating regression discontinuity design (MRRDD). Our implementation used the *binding-score* MRRDD method, as outlined in Reardon & Robinson (2012). This method relies on the fact that Renewal Schools were chosen using a pre-specified formula, based on a number of school-level indicators, which are described in more detail below.

We can use the formula to identify which schools barely missed the eligibility cut-off and use them as comparison schools. By definition, these schools have very similar values of the criteria variables (referred to as running or rating variables in the MRRDD context) to the Renewal Schools that just made the eligibility cutoff. Just as importantly, the two sets of schools also are likely to be nearly identical in other ways, both on measures that we can observe in the data and measures that we cannot observe. Thus, any subsequent difference in student outcomes between the treated and comparison schools can be ascribed to their Renewal school status.

The criteria variables differ slightly depending on whether the school is an elementary/middle school or a high school. For elementary/middle schools, they include the percent of students who are proficient in math and English language arts (ELA). For high schools, they include graduation rates. A brief summary for elementary and middle schools is as follows, although the basic idea is the same for high schools.

In order to be designated a Renewal School, an elementary or middle school must:

3

1. Fall in the bottom quartile in percentage proficient or above in the ELA and Math exams in 2012, 2013 and 2014 (note: this rule comprises six criteria, one for each of two tests in each of three years, which we refer to as 1.a. – 1.f.),
2. Fall in the bottom three quartiles in adjusted growth percentile values for 2013-14 (note: this is called the "beat the odds" criterion),
3. Have a recent NYCDOE Quality Review[3] rating as proficient or below, and
4. Designated as a Focus or Priority school by the NY State Department of Education[4]

A school must meet all nine criteria to be designated as a Renewal School. In addition, the Chancellor was given discretion to add or remove schools from the criterion-based list. The Chancellor added four schools to the criterion-based list to form the final group of Renewal Schools.

Our implementation of the MRRDD estimator divided the criteria into two types. The first seven criteria, described in items 1 and 2 above, are continuous measures and expressed in terms of being below a percentile cutoff based on the distribution of a particular characteristic over all elementary and middle schools. For any one of these seven criteria, two schools are similar if they had the same or nearly the same percentile values for the criterion. The remaining two criteria, described in items 3 and 4 above, are categorical. These categories are very coarse, and it was not possible to determine the similarity of any two schools more precisely than whether or not they shared the same category.

Therefore, we began by limiting our analysis to schools that meet the conditions described by items 3 and 4. This left us with 204 elementary and middle schools, all of which had recent Quality Reviews of proficient or below and were designated either a Priority or Focus school. Of this group, 65 also met all of the criteria listed in items 1 and 2, making them Renewal Schools.

The next step was to define a set of schools that were "near" the cutoff for the criteria listed in items 1 and 2. The following pictures help describe our method. For simplicity, we focus on two criteria in our visualization although it generalizes to the seven criteria in a straightforward fashion.

In Figure 2 we present a simplified example, showing the percentile values for some **made-up data for some hypothetical schools** in the distribution of 2014 Math and 2014 ELA proficiency rates. We use made-up data for this example in order to protect the identity of actual NYC schools and to simplify the presentation. The horizontal and vertical lines at the 25th percentiles of each measure divide the graph into four quadrants. All schools in the lower left quadrant, indicated with a letter r or R (depending on whether or not they were used in the MRRDD, which is explained in more detail below), were designated as Renewal Schools. The remaining schools, indicated with the letter n or N (again depending on whether they were

---

used in the MRRDD), exceeded the 25th percentile in at least one of the two measures and, therefore, did not qualify to become Renewal Schools.

Figure 2. Hypothetical schools near the cutoff for 2014 Math and 2014 ELA proficiency
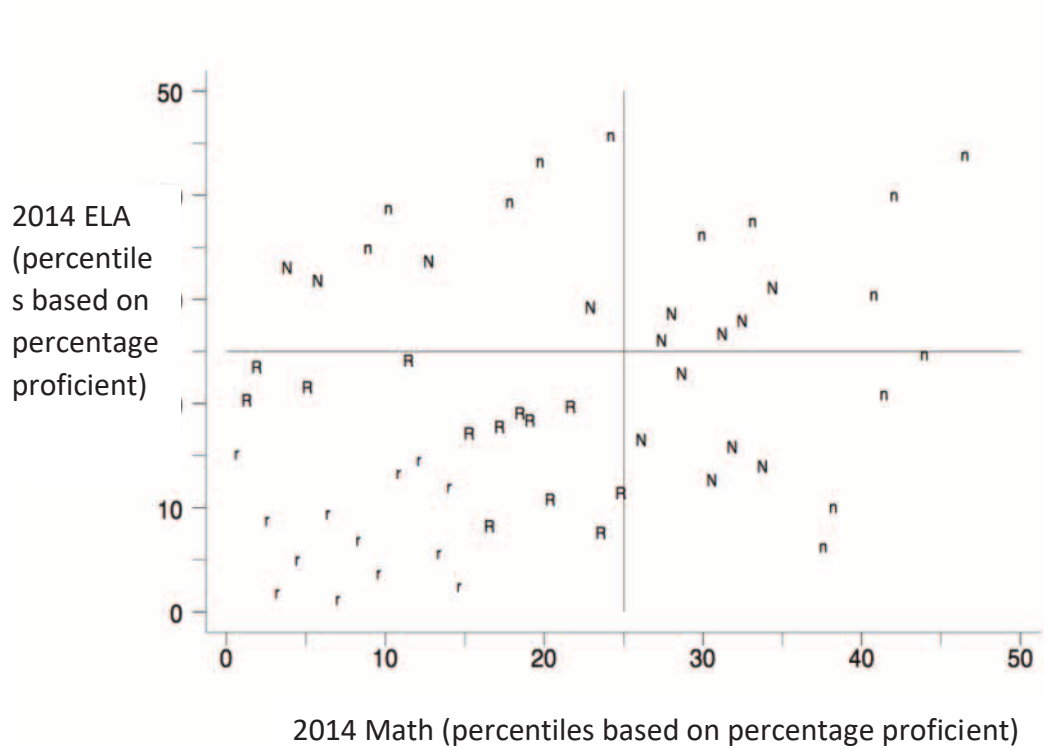
2014 ELA (percentiles based on percentage proficient)

2014 Math (percentiles based on percentage proficient)

Notes: A school must be lower than 25th percentile in both reading and math to become a Renewal School. Each letter represents a school: r=Renewal School not used in MRRDD; R=Renewal School used in MRRDD; N=non-Renewal School used in MRRDD; n =non-Renewal School not used in MRRDD. *__These data are made up for demonstration purposes and do not reflect actual NYC schools.__*

For any two schools, the closer together they are in this graph, the more similar they were in terms of the 2014 math and ELA performance. Schools that are right next to each other on the graph are virtually identical in this respect. Although they likely differ from each other in other respects such as performance in other years, demographics, and so forth, it isn't possible to say how they will differ. Although this example uses made-up data, the process was the same for our analysis with the actual NYC schools data.

The schools identified with a capital R or N are within ten percentile points of the eligibility cutoff defined by the two line segments (also called the boundaries) that border the lower left quadrant. For Renewal Schools near the cutoff, indicated with an R, an increase of ten percentile points or less in one of the measures would make them ineligible for RS status. Likewise, for the non-RS schools indicated with an N, a decrease in one or both of the measures by 10 percentile points or less would make them eligible to be an RS. It is these two groups of schools near the boundaries that are used in the MRRDD estimation of RS impact, because they are the most similar to members of the other group.

In fact, the closer a school is to a boundary, the more likely it is to be very similar to a school on the opposite side of the boundary. Therefore, we calculated the distance of each school from the nearest boundary. This information is presented in Figure 3, in which each letter indicating a school within ten percentile points of a boundary has been replaced with its distance from that boundary.

Figure 3. Distance to eligibility boundary for hypothetical schools within 10 percentile points

2014 ELA (percentiles based on percentage proficient)

2014 Math (percentiles based on percentage proficient)

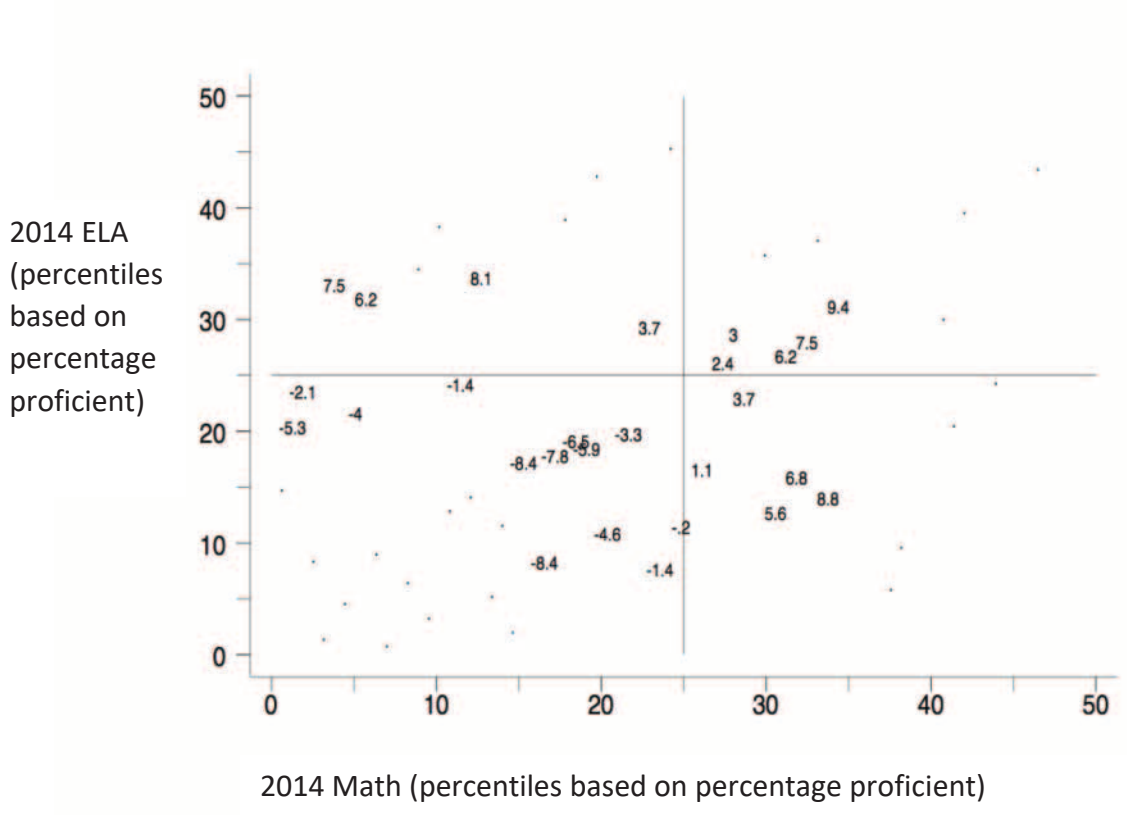Notes: A school must be lower than 25th percentile in both reading and math to become a Renewal School. Each dot represents a school that is more than 10 percentile points from the eligibility boundary. Schools within 10 percentile points of the eligibility boundary are labeled with their distance from the boundary. *__These data are made up for demonstration purposes and do not reflect actual NYC schools.__*

As a next step in our hypothetical example we use this distance information in conjunction with an outcome of interest to estimate the impact of the program on the outcome. We expect schools with less negative or more positive values of this distance to the boundary to have better outcomes, because they had better performance in 2014 math and ELA. However, we also hypothesize that Renewal Schools would have better outcomes due to the impact of the Renewal School program.

Figure 4 provides an example of this estimation step for 2017 ELA proficiency outcome, continuing to use the made-up data for imaginary schools. Each of the points within ten percentile points of the boundary in Figure 2 is plotted, using the distance indicated in Figure 2 as the horizontal coordinate. The vertical coordinate is the 2017 ELA proficiency, also in percentiles. On each side of the boundary, a non-sloping regression line is estimated, with the distance between the two regression lines is representing the estimated impact of the program.[5] In this constructed example using made-up data, the program increased the 2017 ELA performance 2.5 percentile points above what it would have been in the absence of the program.

---

[5] Another common approach to RDs is to allow the regression lines on each side of the boundary to slope, i.e. including the running variable as a control variable in the RD regression. Unlike most RD regressions, we controlled for a range of covariates (including those that comprise of the running variable) in the ridge regression, which minimizes the importance of the also including the running variable directly as an additional control. We opted against including a sloping line for this reason and because our simulations suggested that allowing the regression lines to slope causes an overfitting of the data and therefore worse estimates. This decision, however, only had a meaningful impact on the estimates of the RS program on high schools; the effect on elementary/middle schools is similar regardless of which methodology is used.

Figure 4. 2017 ELA Percentile and distance from RS eligibility cutoff (hypothetical example)

Distance from the nearest cutoff (in percentiles)
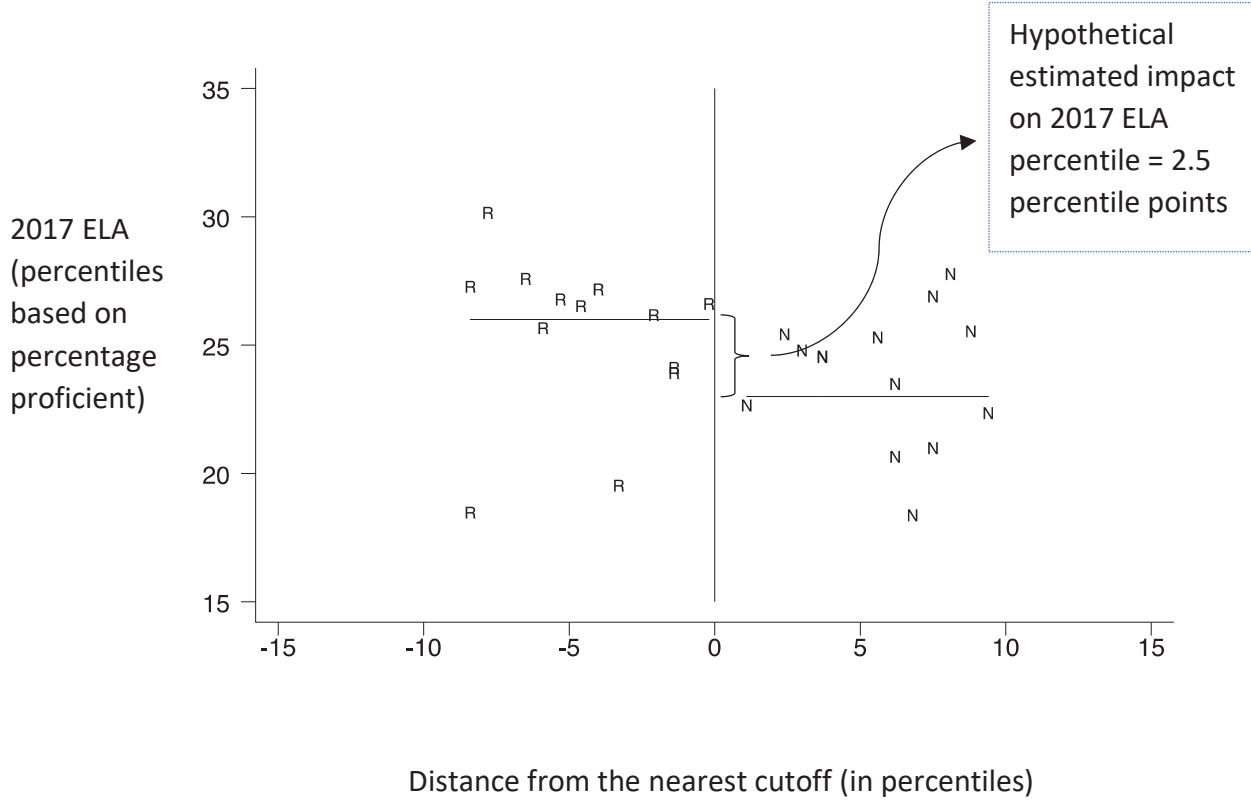
Notes: Each letter represents a school: R=Renewal School used in MRRDD; N=non-Renewal School used in MRRDD. ***These data are made up for demonstration purposes and do not reflect actual NYC schools.***

The example in Figure. 4 has been a simplified demonstration of the method using hypothetical data. The use of many selection criteria, rather than just the two used above, does not change the basic process. All of the criteria used in the present study were expressed in NYC-wide percentiles, which simplifies the process of combining the boundary distances into a single dimension.[6] Once we have reduced the boundary distances into a single dimension, we can run a traditional unidimensional (or single rating) regression discontinuity (RD as illustrated in Figure 4). For the single rating RD, we used a triangular weighting matrix and limited the schools under consideration to a specified "bandwidth" – i.e., the 10 percentile point distance from the boundary in Figure. 4. This ensures that the results are indeed driven by schools that either barely qualified for or barely missed out on becoming a RS. The bandwidths we used differed between the elementary school (ES)/middle school (MS) analyses and the high school (HS) analyses, reflecting the different number of schools of each type in NYC and participating in the RS program.[7]

In order to reduce the variation of the outcome around the regression lines due to other factors, we added a step to the estimation procedure. In this step, we first estimated a ridge regression of the outcome on all of the selection criteria; school demographics, outcome measures from 2012 to 2014, and a dummy for Renewal School status.[8] Then, we used the residual from the first stage regression as the outcome in the final regression discontinuity estimation step. Importantly, we calculated this residual ignoring the coefficient on the Renewal School dummy variable to ensure that this process does not bias the estimates. This last step improved the precision of the estimated impact, allowing us to identify statistically significant effects that otherwise would not be detected.

The main strength of the MRRDD methodology is that it allows for rigorous causal inference regarding the impact of the Renewal School program on the outcomes of interest. Regression discontinuity is regarded as the second best type of evaluation design with very good internal validity, only surpassed by a randomized control trial design. This strength is obtained by focusing on schools which are nearly identical, but happen to be on different sides of the boundary and, therefore, happen to have different Renewal School statuses.

The flip side of this strength is that, in theory, the estimated impact only pertains to these schools that are near the cutoff. In our case, an elementary Renewal School that is in the

---

[6] Mathematically, the seven measures are mapped into the single dimension by taking their maximum. This "binding score" captures how close RS were to becoming ineligible and how close non-RS were to becoming eligible.

[7] Because our observations are not completely independent, due to the fact that we have two observations per school (one for each year we include in the analysis), the traditional equations for the optimal RD bandwidth do not apply. In part because of this, we ensured that our choice of bandwidth and kernel weights do not affect the results through a number of robustness checks.

[8] For the ridge regression, we included all schools within 25 percentiles of the nearest cut-off, instead of the bandwidth of 10 used for the RD. We also did not include the kernel weights in the ridge regression like we do in the RD.

bottom ten percent of math and ELA proficiency in each of the three years used for eligibility determination is given little or no weight in the impact estimation. It is possible that the impact of the program for schools such as these – i.e., the most challenged schools – is different than for Renewal Schools that barely qualified for the program.

In practice, however, the cut-offs were set low enough that most of the Renewal Schools were included in the analysis. About sixty percent of elementary and middle Renewal Schools are within a ten percentile points of the boundary on at least one of the eligibility criteria, and therefore are included in our estimation procedure. Since there are fewer high schools, we use a larger bandwidth of twenty percentile points which means that all high school Renewal Schools are included in our estimates.

Probing Effect Heterogeneity by School Demographics and Program Year

Of course, it is possible that program impact differs among schools when grouped by factors other than their distance to the boundary. As suggested by RQ2, we divided our sample of Renewal and non-Renewal Schools based on the economic need measure developed by NYC DOE and examine whether program impact differs along this dimension. In addition to exploring whether the effect differed depending on the demographic make-up of the school, we also estimated the effect separately for 2016 and 2017 in order to explore how the effect differs based on the year. In this analysis, we estimated the effect of the program in 2015 in order to determine whether the program affected outcomes in the year it was rolled-out.

Probing Effect Heterogeneity by Level of Implementation

Another set of factors that is likely to be related to the effectiveness of the Renewal School program are the various implementation activities that we measured, in part, in our earlier report on the implementation of New York City Community Schools Initiative (Johnston, et al., 2017). However, these implementation domains—Needs Assessment, Leadership Development, Parent Ties, Professional Capacity, and Student-Centered Learning Climate— were measured after Renewal School program had begun and were likely to be affected by other factors, measured and unmeasured, that also affect the outcomes. Therefore, it is not possible to determine the extent to which these implementation activities actually cause improvements in outcomes. Because of this, our regressions of outcomes on these implementation activities should be interpreted as associations rather than causal estimates, whose correlation could reflect the shared causality driven by other factors. It is also important to note that since we only have measures of these implementation activities in the Renewal

Schools and have no information about whether or not the other NYC schools are undertaking similar activities, we excluded the non-Renewal Schools from this analysis and these analyses were not conducted within the MRRDD framework.

Nevertheless, we did attempt to control for measured pre-existing differences among the Renewal Schools before examining the relationship between outcomes and implementation activities. Therefore, we ran a ridge regression of the outcomes on covariates that include pre-program outcomes and school demographics, similar to the ridge regression that we ran prior to the MRRDD analysis. We used the residual from this regression in a comparison between schools with low implementation measures and high implementation measures. We did this one by one for the five implementation measures, and examined whether there are statistically significant differences in outcomes, after controlling for these covariates, between high and low implementers.

Data

We examined the impact of the of the Renewal School program using a variety of data sources. The data that we used primarily consists of administrative data obtained from the New York City Department of Education (NYCDOE). The data can be thought as belonging to one of four main groupings: school-level information, outcome data for students, New York City Quality Review rubric, and other student and school level outcome data.

School Level Information

First, we have detailed school-level information on the variables that determined whether schools qualified to become a RS. In addition to categorical data that indicated which of the qualifying criteria a school met, we had continuous information on the measures that determined whether a school met some of the qualifying criteria. For example, for elementary and middle schools, we had six variables that measure the percent of each school's students who were proficient in mathematics and English for the years 2012, 2013, and 2014. Similarly, for high schools, we had information on their graduation rates for the years 2012, 2013, and 2014. We also had a continuous "Beat the Odds" measure for both elementary/middle and high schools, which also determined whether schools were eligible to be a RS. Together, the categorical and continuous measures are what we use in our MRRDD.

In addition to these measures that determined whether schools were considered academically eligible to become a RS, we had information on whether schools met the other criteria, i.e. whether or not they were identified as priority or focus schools by the New York State Department of Education and whether they scored proficient or below on their most recent Quality Review rating.

Outcome Data for Students

The second broad category of data that we had was outcome data for students. While we have outcome data for students from 2012-2017, we only used the post-treatment outcomes in 2016 and 2017 as outcomes in our analysis. We use the 2012-2014 outcomes as controls in our ridge regression. For the most part, we ignore outcomes in 2015 since we consider this year to neither be a "treated" year or a "baseline" year.

We focused on six outcomes for elementary and middle schools: math test scores, ELA test scores, attendance rates, percent of students who are chronically absent (i.e. who are absent more than ten percent of the time), the average number of suspensions per student per year, and the average number of disciplinary incidents per student per year. For each variable, the student-level variables were aggregated to the school-year level by taking the average over everyone who attended the school in a particular year. For the math and ELA test score measures, this average only includes individuals in grades 3-8, since they are the only students for whom we have test scores.

For high school analyses, we had a total of ten measures that were used as outcomes. We aggregated six of these measures to the school-year level in the same way as for elementary and middle school measures, i.e. by taking the average over everyone who attended the school in a particular year. These six measures were: the attendance rate, percent of students who are chronically absent, the average number of suspensions per student per year, the average number of disciplinary incidents per student per year, the number of credits earned, and the dropout rate. The other four measures used were: the graduation rate, the percent of students who take the PSAT, the percent of students who take the SAT, and the average number of AP tests passed. Aggregating these measures to the school-year level is complicated by the fact that these measures can be thought of as cumulative; for example, taking the SAT in 11th grade likely decreased the chance that an individual would take the SAT in 12th grade. Thus, for all of the measures except for the percent of students who take the PSAT, we only focused on 12th grade students and calculated whether they had ever taken the SAT (regardless of the grade they took the SAT) or the total number of AP tests they had passed (regardless of the grades they passed these tests). Consequently, we calculated the school-year level variables by averaging over all of the 12th grade students, rather than over all students. Similarly, we calculated the school graduation rate as the fraction of 12th grade students who graduated from the high school in a given year. For the PSAT variable, we use the same approach as for the other variables, but focus on 11th grade students instead of 12th grade students. The cumulative nature of these outcomes suggests that any effect of the RS program may be slower to show up in these outcomes than for the other cross-sectional outcomes.

Tables 1 and 2 present the average outcomes in 2014, the last year before the program was rolled out, for Renewal Schools and the non-Renewal Schools used in the analysis for elementary/middle schools and high schools, respectively. The results illustrate that both groups of schools that we used in the analysis were lower performing than the average school

in New York City. It also shows that, even among schools close to the boundary, non-Renewal Schools had slightly better outcomes in 2014 than Renewal Schools; this motivates our use of the MRRDD approach rather than just comparing non-renewal to renewal schools. Finally, the summary statistics also give a good reference for the size of the estimated effects, discussed in the next section.

New York City Quality Review Rubric

Third, we used information from the New York City Quality Review rubric. In our analysis, we focus on the indicators that measure the Instructional Core and the School Culture. The Instructional Core is measured by three indicators: Curriculum, Pedagogy, and Assessment; School Culture is measured by two indicators: Positive Learning Environment and High Expectations. Each of the indicators themselves are given one of four ratings: underdeveloped, developing, proficient, or well-developed. We turned these ratings into quantitative measures by turning the ratings into a four point scale, where schools got a rating of one if they were underdeveloped and a rating of four if they were well-developed. We then used this four point scale as an outcome in our analysis.

Other Student and School Level Outcome Data

Fourth, we had a range of other student and school-level data that we employed as covariates in our regressions. The demographic information about the students that we employed included students' race, whether they were classified as English Language Learners, whether they were diagnosed with a learning disability, and whether they were in poverty. We also used baseline data on the schools, including their outcome data from 2012 to 2014, their proficiency-levels in 2009 and 2010, and their student growth percentile ratings in 2010 and 2011. Finally, to explore the heterogeneity of the RS effects, we used the 2014-2015 values for the school's Economic Need Index (ENI).[9] This variable is publicly reported by NYCDOE and is meant to judge how many students at the school are facing economic hardship.[10]

We also used information that we collected to capture how well the schools implemented certain aspects of the RS program. See Appendix A for a summary of implementation measures and data sources, and descriptive statistics. This implementation data was derived primarily from a RAND-developed school leader survey that was distributed to principals and Community School Directors at Renewal Schools in the fall of 2016. The survey included questions

---

[9] While we focus on the schools' ENI in 2014-15, less than 1% of the overall variation in the ENI occurs within-schools over time, rather than across-schools. Thus, the choice of year to focus on would not make a large difference in the findings. We have also conducted analysis which showed that there are no statistically significant changes in the demographic make-up of renewal schools in the past three years, regardless of whether measuring the demographic make-up of renewal schools using students' race, English language learner status, poverty status, disability status, or the schools' ENI.

[10] The Economic Need Index is determined by the following calculation: (% of students in temporary housing) + (% of students eligible for eligible for public assistance * 0.5) + (% of students eligible for free lunch * 0.5)

regarding the extent to which Renewal School programs were being implemented as intended and whether the program schools experienced particular barriers or demonstrated successes in the implementation process. Ninety-three percent of schools surveyed had at least one school leader complete the online survey, although individual survey items had lower responses rates, resulting in missing rates well above 3% for many implementation indicators.

In addition to the school leader survey, we leveraged administrative data sources to measure aspects of program capture implementation domains. These included information from the NYCDOE Quality Review Rubric as a means of capturing schools' use of teacher teams and leadership development. We also used information from the NYCDOE's Ladder of Engagement tracking system, or VAN data, which is derived from the Voter Action Network data tracking system for monitoring political campaign volunteerism. Specifically, we tracked the percentage of families who were on the step 1 of the ladder (the "on ramp"), which means they were recruited to attend an event at the school, such as an adult education class, a parent-teacher conference, a Community School Forum, or a school celebration where they sign a sign-in sheet. They could also get on a step by being recruited by another parent to fill out a commitment card in which they promise to engage with future school activities.

There were other data sources that we considered but ultimately did not use for the implementation analysis, including administrative records showing schools' use of iReady assessment, schools' implementation of Strategic Data Check-ins, schools' use of the Data Wise inquiry process, and the contents of the Renewal Schools Comprehensive Education Plans. These data sources were reviewed and considered for potential inclusion in this analysis, but due to the fact that much of this data had limited variation (as was the case for iReady assessments and Strategic Data Check-ins) or was qualitative in nature and difficult to quantify (as was the case for the RSCEPs), we ultimately omitted it from the present analysis.

3. Results

In this section we present the results from the analyses described in the previous sections. We first present the average treatment effects (RQ1) and conclude with results pertaining to the heterogeneity in effects based on student characteristics (RQ2) and measurable aspects of program implementation (RQ3).

Estimated Average Effects

As discussed in detail in the Methods section, we estimated the effect of the RS program using a multiple rating regression discontinuity approach, augmented with a ridge regression to improve the precision of the estimates. Figure 5 illustrates the approach, when using elementary/middle school attendance rates as the outcome of interest. As can be seen, going from the right to the left on the x-axis, there was a large jump in the attendance rate when schools moved from being barely ineligible to become a RS to being barely eligible to become a RS. This translates into a statistically significant effect, as shown in column (4) of Table 3; it suggests that RS improved their average student attendance rate by around 1.5 percentage points. Similarly, column (3) shows that RS status can be causally linked to a 5 percentage point reduction in the number of students who are chronically absent.

Table 3 also reports the estimated effects for each of the elementary/middle school outcomes. Columns (1) and (2) show that although the estimated effect of RS status on a school's average math and ELA test scores is positive, this effect is small compared to the overall uncertainty in the estimate and thus is not close to being statistically significant. Similarly, columns (5) and (6) show that the estimated effect of RS status on suspensions and disciplinary incidents was negative (i.e. becoming a RS reduced the number of disciplinary incidents and suspensions); however, this estimate was quite noisy and not close to being statistically significant.

In Tables 4a and 4b we present the estimated effects for each of the high school outcomes. Moving from left to right in Table 4a, we show that RS designation led to lower proportions of chronically absent students (a difference of approximately 5 percentage points between RS and non-RS), higher average attendance rates (a difference of approximately 2 percentage points), and 0.744 more credits earned. We also found that Renewal Schools had lower rates of student suspension and disciplinary incidents, although these estimates were not statistically significant. There were not statistically significant differences related to the number of suspensions and disciplinary incidents.

Additional results are presented in Table 4b, where we show that Renewal Schools had a slightly larger dropout rate (less than 1 percentage point) compared to non-Renewal Schools, and there were no statistically significant effects related to graduation rate, rates of students taking the SAT, PSAT and AP exams. It is important to note that the lack of statistically significant effects does not necessarily mean that the RS program had no impact; rather, the lower number of high schools that implemented the program means that the effect estimates at the high school level are considerably less precise than the estimates for elementary/middle schools. As a concrete example, the standard error (i.e. the uncertainty) in the estimated effect on the proportion of students who are chronically absent are 3.2 larger when focusing on high schools than when focusing on elementary/middle schools.

We did not find any statistically significant effect of the RS program on any of the indicators that make up the Instructional Core or School Culture, a finding that is shown in Table 5. Again,

it is worth emphasizing that our findings do not definitively show that the RS program did not have an impact on a school's Instructional Core or School Culture, rather our results indicate that we do not have enough statistical power to say whether or not there is an effect. Further complicating the interpretation of Table 5 is the fact that only a subset of the schools are rated according the quality review each year. If the schools that are rated are not randomly selected, the missing data would bias the estimated effect of the RS on the Quality Review.

Figure 5. Effect of Renewal Schools on Elementary and Middle School Attendance Rates
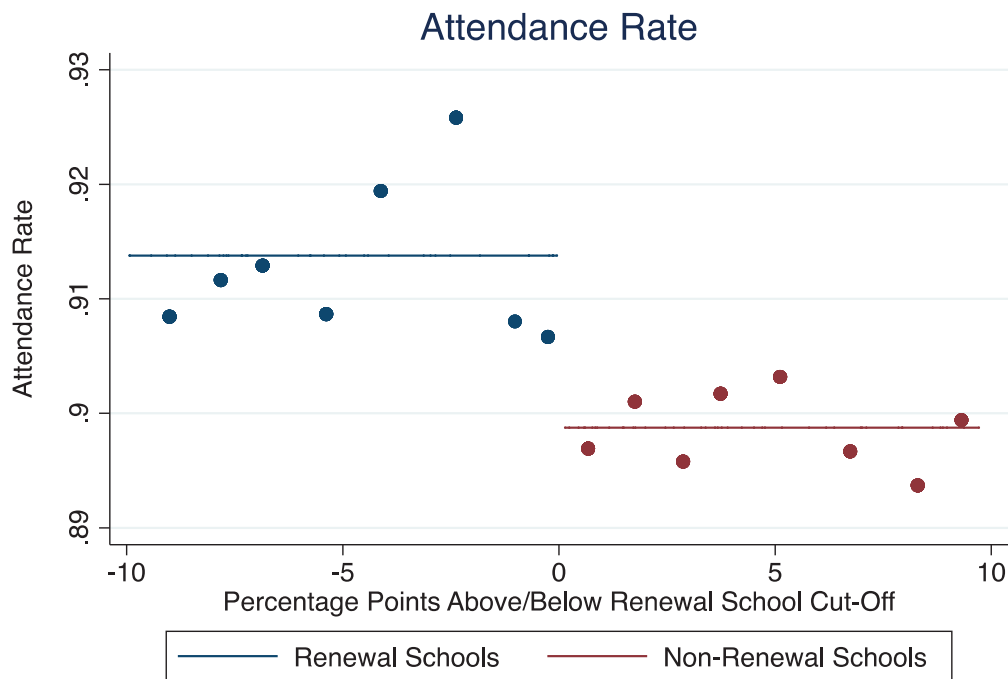
NOTE: Estimates based on data from SY2016 & SY2017.

Testing the Validity of the Estimation Methodology

In this subsection, we discuss three approaches we used to ensure that the estimates reported in the figure above accurately reflect the true causal effect of the Renewal School program.

First, we conducted various robustness checks to ensure that the results did not depend on any of the assumptions we made. These robustness checks include: varying the bandwidth in the regression discontinuity estimation; running the RD without any additional controls, i.e. without residualizing the outcome via the ridge regression; and removing the outliers before conducting the analysis. While we do not include these results in this report, they are available upon request; the reason we do not include them here is that they are quite similar to the results reported in the previous section.

Second, we conducted the analysis using a different research methodology. In this case, we matched treatment schools to potential comparison schools based on the variables that determined whether a school qualified as a RS or not, i.e. for elementary/middle school we matched on the continuous proficiency measures from 2014 to 2016, the continuous "Beat the Odds" measure, and whether they would have otherwise qualified to be a RS. Because we then end up with a scenario where RS were matched to slightly better achieving control schools, we conducted a difference-in-difference analysis. Again, the results are available upon request and are broadly similar to the results presented in the previous section. The one major difference is that the difference-in-difference estimates suggest that RS positively affected math scores, while the RD estimates included here suggest that there is no effect of RS.

Finally, we exploit the fact that there are two additional criteria, in addition to the ones we use for the MRRDD, that schools needed to meet in order to be classified as a RS as placebo tests. To do so, we run the same specification as before using only schools that did not meet the other criteria; i.e. they either were not classified as a priority or focus school by the State of New York or they scored better than proficient in their quality review. For these schools, moving from being barely above the RS cutoff to being barely below the cutoff had no impact on their treatment status; this is because they did not meet the other criteria, so the cutoff we exploit is not binding. Thus, if our method is a valid approach, we would expect schools on either side of the cutoff to be quite similar. If, on the other hand, we find that there is a discontinuous jump at the cutoff even for schools that are not eligible to become a RS, we would be concerned in the approach. Table 6 shows that the placebo test worked well for elementary and middle schools, in the sense that when focusing only on schools that were not eligible to become a RS, there is not a discontinuous jump at the cutoff for any of the outcomes we use. This suggests that the estimates reported above were indeed valid estimates of the true causal effect of the RS program.[11]

Heterogeneous Effects by School Demographics and Program Year

The previous section focused on estimating the average effect of the RS program. In this section, we explore the extent of heterogeneity in these estimates. Specifically, we considered two sources of heterogeneity: heterogenous effects by year and heterogenous effects by demographic makeup of the student body.

We first estimated the effect separately for each year of the program: 2015, 2016, and 2017. This allows us to explore how the effect evolves over the time of the RS program. Note that the effect could evolve over time for multiple reasons. One reason is that the program matures and

---

[11] Tables A3a and A3b shows that the placebo tests worked less well for high schools, with a few of the estimates being statistically significant. It's worth noting that all of the statistically significant estimates suggest that being below the cut-off is negatively linked with improved outcomes. This implies that, if anything, the estimated effects of RS for high schools using the MRDDD methodology are downward biased, i.e., they would show the program to be less helpful than it actually is.

schools' fidelity of implementation may improve over time, so the effect of the RS program may grow accordingly. An entirely different reason is that in 2017 many students would have been exposed to the program for three years. If the effect on the students is cumulative, this could lead to larger observed effects in 2017 than in 2016 or 2015. In this report, we did not endeavor to understand why the effects might differ over time and instead just explore whether they do. As shown in Table 7, we find that among elementary and middle schools, the effects were larger in 2017 than in either 2015 or 2016, and the effect grew for every one of the outcomes we used. In fact, when we restricted our analysis to 2017, we found that there was a statistically significant decrease in the number of disciplinary incidents at elementary and middle school RS. To put this decrease into context, it suggests that the RS program led to a decrease of 40 incidents per year at a school with 250 students. This pattern was less clear for high school outcomes, with almost none of the outcomes becoming statistically significant at the 5% level in either of the years; this is shown in Tables A2a and A2b.

We also tested whether the estimated program effects differed based on the demographic makeup of the student body. To do so, we focus on one demographic measure, the schools' Economic Need Index (ENI), and divided schools into two based on this measure. Because of the goals of the RS program, all of the schools we include in the analysis scored highly on the ENI. This is seen in the Figure 6, which includes a histogram of the ENI scores of RS juxtaposed against the overall ENI distribution of all schools in New York City. We thus split the sample into high and low economic needs using the median ENI score of the schools included in our analysis; this ensures there are the same number of schools with higher economic need and schools with lower economic need. This process meant that even though we call them "higher" and "lower" economic need schools, it is likely more appropriate to consider them "very high" and "high" economic need schools. More precisely, for elementary and middle schools that had an ENI score of about 0.846 were considered to be above median and those below 0.846 were considered to be below median. This is true even though the median ENI for all elementary and middle schools in NYC is 0.718.

Table 8 reports the results of the heterogeneity analysis for elementary and middle schools. The first row reports the effect of the RS program on schools with an ENI score of above 0.846 and the second row reports the effect of the RS program on schools with an ENI below 0.846. In Table 8 we show that for elementary and middle schools, the RS program impact was present for schools with ENI rates above and below the median cutpoint. For high schools, however, we found that the effects of RS are concentrated among high ENI schools. As we should in Tables 9a and 9b there was an effect of the RS program for higher ENI schools for proportion chronically absent, average attendance rate, and credits earned. There were no statistically significant effects for schools with below-median ENI levels.

Figure 6. Economic Needs Index Distribution for Renewal and Non-Renewal Schools
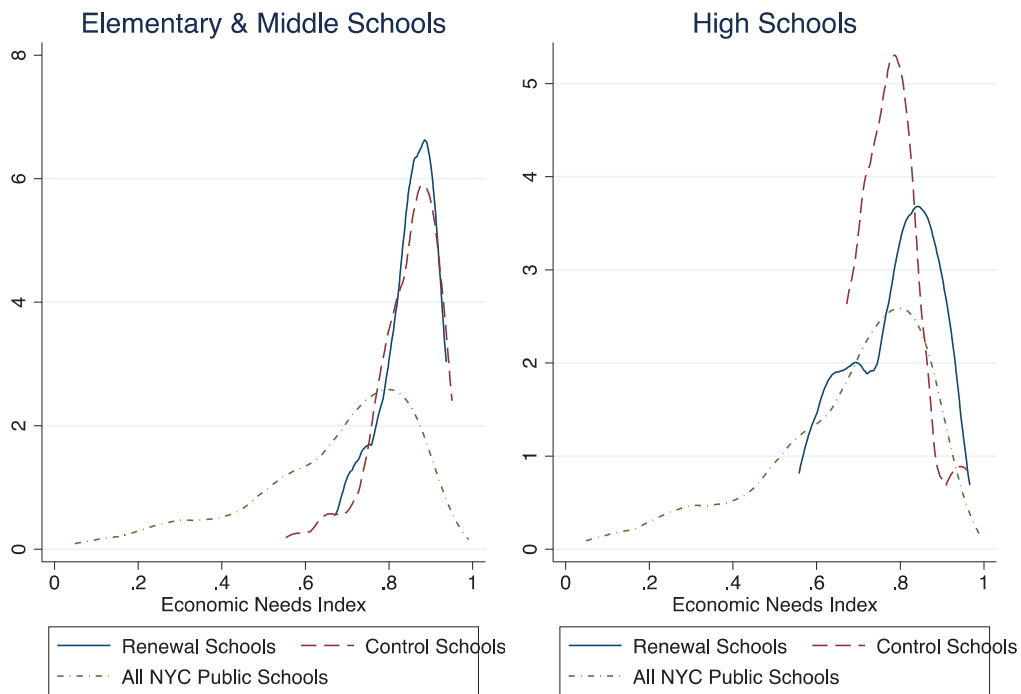
Note: The scale of the vertical axis allows the reader to calculate the fraction of schools in a specific interval. For example, the curve for All NYC Public Schools on the High Schools figure (i.e., right panel of Figure 6) is approximately at the 0.5 mark on the vertical axis for the points corresponding to Economic Needs Index of 0.3 to 0.4. These numbers can be used to calculate that 0.05 (or 5%) of NYC high schools have an ENI between 0.3 and 0.4: $(0.4 - 0.3) * 0.5 = 0.05$. For regions of the graph in which the curve is not horizontal, an average curve height can be used. For example, between 0.8 and 0.9 on the High Schools figure, the All NYC Public Schools curve varies between 1.5 and 2.5 on the vertical axis. Using 2 as the midpoint implies that 20% of the NYC high schools have ENI values in this range: $(0.9-0.8) * 2 = .20$.

Heterogeneous Effects by Level of Implementation

Finally, we turn our attention to RQ3 and explore the extent to which the measures we have collected on the level of RS implementation were related to the outcomes of interest. As we describe in the data section as well as Appendix A, the implementation measures were derived from a variety of data sources that were only available for the Renewal Schools, not the comparison schools. As shown in Table 10, for elementary and middle schools, we found that suspension rates were lower in schools with higher levels of implementation of needs assessment and professional capacity activities.

For high schools, as shown in Tables 11a and 11b, we found that the effect of the RS program on proportion chronically absent and average attendance rates was stronger for schools with higher levels of professional capacity implementation (columns 1 and 2). Furthermore, we found that higher levels of implementation of parent ties activities was associated with lower rates of suspensions and disciplinary incidents and, counterintuitively, lower levels of credits earned.

Additional counterintuitive findings are presented in Table 11b, where we show that greater levels of implementation if needs assessment and professional capacity activities led to a diminished program impact on proportion of students taking the SAT. Conversely, professional capacity implementation was associated with greater program impact on graduation rates. proportion chronically absent and average attendance rates. Overall, we do not find a clear pattern in the relationship between the various implementation scales and outcomes, which suggests that these findings should not be acted on without further investigation.

4. Conclusion

In conclusion, we attest that through our rigorous quasi-experimental methodology, the New York City Renewal Schools appear have some positive impact on important leading indicators of attendance and chronic absenteeism among students at all grade levels. These effects grew over the years of implementation, with some additional outcomes becoming significant in the second year. For high schools, the effects are stronger in schools with the highest levels of economic needs.

We also found mixed evidence that the impact of the RS program varied systematically with the schools' reported measures of implementation, but the pattern of results was inconsistent and warrants further investigation.

Although we have used methods that we think provide the strongest possible evidence regarding the impact of the RS program, this report nonetheless has some notable limitations. The use of a regression discontinuity approach means that we estimate the effects of the RS program on those schools who barely qualified. This means that we cannot rule out the possibility that schools who more easily qualified for the RS program had much larger or smaller effects than the ones reported here.

We are limited by the number of schools that are included in the RS program and the number of schools with similar characteristics in NYC. Many of our impact estimates are sufficiently large to be of interest, but the small size of our sample prevents our estimates from being precise enough to be confident in the actual size of the impact.

We also recognize that our implementation measures are mostly self-reported by the schools themselves and are likely to reflect existing differences among the schools. Without random assignment of implementation level, it is difficult to know whether our estimates of the differential impact of the RS program in relation to the level of implementation provide evidence of the mediation effect of implementation or are biased by differential levels of desirable reporting and by the omission of unmeasured school differences.

In spite of these limitations, we believe that the evidence regarding the positive impact of the RS program on attendance outcomes is very strong. Other possible impacts, such as the positive impact on credits earned among high school students and the differential impact based on schools' level of economic needs, suggest that the program is having a positive impact on lagging indicators of academic success, particularly among the most vulnerable students.

Tables

Table 1. Summary Statistics, Elementary and Middle Schools

	All NYC Elementary/Middle Schools	Renewal Schools Included	Non-Renewal Schools Included	Renewal Schools Not Included
Percent English Language Learners	14.0%	15.9%	16.7%	17.2%
	(0.128)	(0.101)	(0.087)	(0.101)
Percent in Poverty	76.0%	90.4%	89.7%	90.6%
	(0.20)	(0.06)	(0.07)	(0.088)
Percent with Disability	18.6%	24.4%	23.4%	24.6%
	(0.06)	(0.04)	(0.05)	(0.040)
Percent White	15.1%	2.1%	1.9%	1.1%
	(0.21)	(0.02)	(0.02)	(0.007)
Percent Asian	14.6%	3.1%	2.1%	1.7%
	(0.19)	(0.04)	(0.03)	(0.023)
Percent Black	30.3%	50.3%	39.3%	49.9%
	(0.30)	(0.23)	(0.22)	(0.269)
Percent Hispanic	38.4%	43.3%	55.6%	46.6%
	(0.26)	(0.24)	(0.22)	(0.277)
Proportion Chronically Absent	0.23	0.406	0.345	0.394
	(0.12)	(0.07)	(0.08)	(0.09)
Average Attendance Rate	0.917	0.87	0.892	0.882
	(0.03)	(0.04)	(0.03)	(0.03)
Average Math Test Score	-0.0551	-0.717	-0.621	-0.812
	(0.51)	(0.12)	(0.17)	(0.14)
Average English Test Score	-0.0291	-0.645	-0.554	-0.724
	(0.46)	(0.11)	(0.13)	(0.13)
Times Suspended	0.0456	0.141	0.0737	0.0909
	(0.08)	(0.10)	(0.07)	(0.06)
Number of Disciplinary Incidents	0.129	0.617	0.259	0.287
	(0.37)	(0.87)	(0.46)	(0.34)
Number of Schools	1,140	38	44	27

This table reports average outcomes and school demographics in 2014 for all the NYC elementary and middle schools, as well as the subset of renewal and non-renewal schools that are included in the analysis, i.e. the schools that are classified as a priority or focus school, score proficient or below on their quality review, and are within 10 percentiles of the boundary. The last column reports the average outcomes for renewal schools that were not included in the analysis.

Table 2. Summary Statistics, High Schools

	All NYC Elementary/Middle Schools	Renewal Schools Included	Non-Renewal Schools Included	Renewal Schools Not Included
Percent English Language Learners	14.0%	27.7%	20.7%	15.2%
	(0.206)	(0.244)	(0.281)	(0.059)
Percent in Poverty	77.7%	85.2%	82.0%	81.1%
	(0.134)	(0.095)	(0.106)	(0.029)
Percent with Disability	15.9%	18.7%	18.2%	18.8%
	(0.074)	(0.083)	(0.079)	(0.066)
Percent White	8.7%	2.0%	5.7%	7.0%
	(0.134)	(0.020)	(0.081)	(0.041)
Percent Asian	11.1%	9.0%	5.5%	18.3%
	(0.150)	(0.096)	(0.077)	(0.122)
Percent Black	37.8%	31.4%	33.9%	21.2%
	(0.264)	(0.212)	(0.227)	(0.220)
Percent Hispanic	41.0%	58.6%	53.9%	52.1%
	(0.238)	(0.254)	(0.192)	(0.121)
Proportion Chronically Absent	0.42	0.54	0.47	0.51
	(0.225)	(0.098)	(0.104)	(0.046)
Average Attendance Rate	0.83	0.78	0.80	0.80
	(0.095)	(0.050)	(0.032)	(0.022)
AP Tests Taken	0.23	0.10	0.10	0.25
	(0.535)	(0.122)	(0.093)	(0.227)
Proportion Who Take the SAT	0.62	0.52	0.54	0.61
	(0.246)	(0.117)	(0.203)	(0.070)
Proportion Who Take the PSAT	0.86	0.80	0.88	0.79
	(0.159)	(0.081)	(0.078)	(0.066)
Dropout Rate	0.05	0.08	0.05	0.05
	(0.046)	(0.031)	(0.017)	(0.015)
Graduate Rate	0.66	0.48	0.60	0.58
	(0.241)	(0.071)	(0.047)	(0.072)
Credits Earned	11.53	10.25	10.77	10.39
	(1.918)	(1.248)	(1.009)	(0.453)
Times Suspended	0.12	0.14	0.17	0.15
	(0.119)	(0.102)	(0.147)	(0.105)
Number of Disciplinary Incidents	0.33	0.27	0.29	0.28
	(0.435)	(0.228)	(0.235)	(0.113)
Number of Schools	382	26	16	3

This table reports average outcomes and school demographics in 2014 for all the NYC high schools, as well as the subset of renewal and non-renewal schools that are included in the analysis, i.e. the schools that are classified as a priority or focus school, score proficient or below on their quality review, and are within 20 percentiles of the boundary. The three schools listed as not being included in the analysis include the schools that didn't qualify to be a renewal school, but were added due to the chancellor's discretion.

Table 3. Main Effects, Elementary and Middle Schools

	(1) Average Math Score	(2) Average ELA Score	(3) Proportion Chronically Absent	(4) Average Attendance Rate	(5) Times Suspended	(6) Number of Disciplinary Incidents
Estimated Effect of Renewal School	0.0322 (0.0328)	0.0293 (0.0317)	-0.0508*** (0.00958)	0.0154*** (0.00388)	-0.00416 (0.0156)	-0.0673 (0.0414)
Schools Included	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle
Bandwidth of RD	10	10	10	10	10	10
Number of School-Year Observations	156	156	161	161	161	161

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only schools that were classified as "priority of focus" by the state of New York and were proficient or below in their review. Standard errors (in parenthesis) are clustered at the school level. The outcomes were first residualized using a ridge regression to improve precision of the estimates. Math and ELA test scores are measured in student standard deviation units and Times Suspended and Number of Disciplinary Incidents are measured per student per year.
* p<0.10 ** p<0.05 *** p<0.01.

Table 4a. HS Main Effect, High Schools (Panel 1)

	(1) Proportion Chronically Absent	(2) Average Attendance Rate	(3) Credits Earned	(4) Times Suspended	(5) Number of Disciplinary Incidents
Estimated Effect of Renewal School	-0.0508* (0.0308)	0.0213* (0.0116)	0.744** (0.352)	-0.0102 (0.0174)	-0.138 (0.118)
Schools Included	High Schools	High Schools	High Schools	High Schools	High Schools
Bandwidth of RD	20	20	20	20	20
Number of School-Year Observations	85	85	85	85	85

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only schools that were classified as "priority of focus" by the state of New York and were proficient or below in their review. Standard errors (in parenthesis) are clustered at the school level. The outcomes were first residualized using a ridge regression to improve precision of the estimates. Information on the number of disciplinary incidents at the school is only available for SY 2016-17.
* p<0.10 ** p<0.05 *** p<0.01.

Table 4b. HS Main Effect, High Schools (Panel 2)

	(1) Dropout Rate	(2) Graduate Rate	(3) Proportion Who Take the SAT	(4) Proportion Who Take the PSAT	(5) AP Tests Taken
Estimated Effect of Renewal School	0.00791* (0.00414)	0.0227 (0.0286)	0.0438 (0.0535)	0.0285 (0.0464)	0.00611 (0.0284)
Schools Included	High Schools	High Schools	High Schools	High Schools	High Schools
Bandwidth of RD	20	20	20	20	20
Number of School-Year Observations	85	85	85	85	85

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only schools that were classified as "priority of focus" by the state of New York and were proficient or below in their review. Standard errors (in parenthesis) are clustered at the school level. The outcomes were first residualized using a ridge regression to improve precision of the estimates. Information on the number of disciplinary incidents at the school is only available for SY 2016-17.
* p<0.10 ** p<0.05 *** p<0.01.

Table 5. Quality Review Outcomes

	1.1 Curriculum	1.2 Instruction	1.4 Positive Learning Environment	2.2 Assessment	3.4 High Expectations	4.2 Teacher Teams and Leadership Dev.
Estimated Effect of Renewal Schools	0.0572 (0.0960)	-0.0101 (0.0963)	0.155 (0.126)	-0.0866 (0.0943)	-0.109 (0.111)	0.114 (0.110)
Schools Included	All Schools	All Schools	All Schools	All Schools	All Schools	All Schools
Bandwidth of RD	10	10	10	10	10	10
Number of Observations	270	270	182	270	270	270

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only schools that were classified as "priority of focus" by the state of New York and were proficient or below in their review. The outcomes were first residualized using data on the schools' quality reviews before 2014. Standard errors (in parenthesis) are clustered at the school level.
* p<0.10 ** p<0.05 *** p<0.01.

Table 6. Placebo Test Results, Elementary and Middle Schools

	(1) Average Math Score	(2) Average ELA Score	(3) Proportion Chronically Absent	(4) Average Attendance Rate	(5) Times Suspended	(6) Number of Disciplinary Incidents
Placebo RD Estimate	-0.0304 (0.0519)	-0.0540 (0.0461)	0.0124 (0.0289)	-0.00398 (0.00741)	0.00398 (0.0131)	0.0644 (0.0565)
Schools Included	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle
Bandwidth of RD	10	10	10	10	10	10
Number of School-Year Observations	154	154	156	156	156	156

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only schools that were classified as either not "priority of focus" by the state of New York or were not proficient or below in their review. Standard errors (in parenthesis) are clustered at the school level. The outcomes were first residualized using a ridge regression to improve precision of the estimates. The outcomes were first residualized using a ridge regression to improve precision of the estimates. Math and ELA test scores are measured in student standard deviation units and Times Suspended and Number of Disciplinary Incidents are measured per student per year.
* p<0.10 ** p<0.05 *** p<0.01.

Table 7. Yearly Results, Elementary and Middle Schools

	(1) Average Math Score	(2) Average ELA Score	(3) Proportion Chronically Absent	(4) Average Attendance Rate	(5) Times Suspended	(6) Number of Disciplinary Incidents
Estimated Effect in 2014-2015	-0.0474* (0.0286)	-0.0142 (0.0342)	-0.00900 (0.0105)	0.000801 (0.00434)	-0.00572 (0.0169)	-0.00765 (0.0452)
Estimated Effect in 2015-2016	-0.0132 (0.0410)	0.0339 (0.0478)	-0.0373** (0.0183)	0.0124** (0.00610)	0.00821 (0.0226)	0.0395 (0.0643)
Estimated Effect in 2016-2017	0.0858 (0.0663)	0.0753 (0.0592)	-0.0600*** (0.0151)	0.0205*** (0.00755)	-0.0129 (0.0280)	-0.162** (0.0697)
Schools Included	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle
Bandwidth of RD	10	10	10	10	10	10
Number of School-Year Observations	161	161	161	161	161	161

Each number comes from a separate regression, all of which include only schools that were classified as "priority of focus" by the state of New York and were proficient or below in their review. Standard errors (in parenthesis) are clustered at the school level. The outcomes were first residualized using a ridge regression to improve precision of the estimates. Math and ELA test scores are measured in student standard deviation units and Times Suspended and Number of Disciplinary Incidents are measured per student per year.
* p<0.10 ** p<0.05 *** p<0.01.

Table 8. Heterogeneity Results, Elementary and Middle Schools

	(1) Average Math Score	(2) Average ELA Score	(3) Proportion Chronically Absent	(4) Average Attendance Rate	(5) Times Suspended	(6) Number of Disciplinary Incidents
Estimated Effect on High Economic Need Schools	0.0722 (0.0479)	0.0550 (0.0471)	-0.0633*** (0.0103)	0.0144*** (0.00238)	0.00610 (0.0165)	-0.0622 (0.0479)
Estimated Effect on Low Economic Need Schools	-0.00859 (0.0472)	-0.00785 (0.0460)	-0.0345** (0.0159)	0.0146** (0.00671)	-0.0162 (0.0275)	-0.0179 (0.0567)
Schools Included	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle
Bandwidth of RD	10	10	10	10	10	10
Number of School-Year Observations	153	153	158	158	158	158

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only schools that were classified as "priority of focus" by the state of New York and were proficient or below in their review. Standard errors (in parenthesis) are clustered at the school level. The outcomes were first residualized using a ridge regression to improve precision of the estimates. Schools were classified as being in high economic need versus low economic need using their SY2014-15 measure of the Economic Needs Index as reported on the NYC DOE website. Both renewal and non-renewal schools were classified as being either high or low economic need, defined as being above or below the median of schools within 25 percentiles of the discontinuity. The "Estimated Effect on High Economic Need Schools" measures the effect of Renewal Schools on the high needs schools and "Differential Effect on Low Economic Need Schools" measures the differential effect between above median schools and below median schools. Math and ELA test scores are measured in student standard deviation units and the Number of Disciplinary Incidents is only available for SY 2016-17.

* p<0.10 ** p<0.05 *** p<0.01

Table 9a. Heterogeneity Results, High Schools (Panel 1)

	(1)	(2)	(3)	(4)	(5)
	Proportion Chronically Absent	Average Attendance Rate	Credits Earned	Times Suspended	Number of Disciplinary Incidents
Estimated Effect on High Economic Need Schools	-0.0614** (0.0284)	0.0310** (0.0132)	0.954** (0.399)	-0.0239 (0.0188)	-0.0431 (0.0354)
Estimated Effect on Low Economic Need Schools	-0.0742 (0.0523)	0.0163 (0.0164)	0.693 (0.443)	0.00659 (0.0256)	-0.154 (0.231)
Schools Included	High Schools	High Schools	High Schools	High Schools	High Schools
Bandwidth of RD	20	20	20	20	20

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only schools that were classified as "priority of focus" by the state of New York and were proficient or below in their review. Standard errors (in parenthesis) are clustered at the school level. The outcomes were first residualized using a ridge regression to improve precision of the estimates. Schools were classified as being in high economic need versus low economic need using their SY2014-15 measure of the Economic Needs Index as reported on the NYC DOE website. Both renewal and non-renewal schools were classified as being either high or low economic need, defined as being above or below the median of schools within 25 percentiles of the discontinuity. The "Estimated Effect on High Economic Need Schools" measures the effect of Renewal Schools on the high needs schools and "Differential Effect on Low Economic Need Schools" measures the differential effect between above median schools and below median schools. The Number of Disciplinary Incidents is only available for SY 2016-17.
* p<0.10 ** p<0.05 *** p<0.01

Table 9b. Heterogeneity Results, High Schools (Panel 2)

	(1)	(2)	(3)	(4)	(5)
	Dropout Rate	Graduate Rate	Proportion Who Take the SAT	Proportion Who Take the PSAT	AP Tests Taken
Estimated Effect on High Economic Need Schools	0.00610 (0.00565)	0.0542 (0.0413)	0.0528 (0.0422)	0.0296 (0.0612)	0.00650 (0.0287)
Estimated Effect on Low Economic Need Schools	0.00720 (0.00511)	-0.00480 (0.0240)	0.0678 (0.125)	-0.00831 (0.0615)	0.0360 (0.0596)
Schools Included	High Schools	High Schools	High Schools	High Schools	High Schools
Bandwidth of RD	20	20	20	20	20

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only schools that were classified as "priority of focus" by the state of New York and were proficient or below in their review. Standard errors (in parenthesis) are clustered at the school level. The outcomes were first residualized using a ridge regression to improve precision of the estimates. Schools were classified as being in high economic need versus low economic need using their SY2014-15 measure of the Economic Needs Index as reported on the NYC DOE website. Both renewal and non-renewal schools were classified as being either high or low economic need, defined as being above or below the median of schools within 25 percentiles of the discontinuity. The "Estimated Effect on High Economic Need Schools" measures the effect of Renewal Schools on the high needs schools and "Differential Effect on Low Economic Need Schools" measures the differential effect between above median schools and below median schools. The Number of Disciplinary Incidents is only available for SY 2016-17.
* p<0.10 ** p<0.05 *** p<0.01

Table 10. Implementation Results, Elementary and Middle Schools

	(1)	(2)	(3)	(4)	(5)	(6)
	Average Math Score	Average ELA Score	Proportion Chronically Absent	Average Attendance Rate	Times Suspended	Number of Disciplinary Incidents
Leadership Development	0.0292	-0.0181	-0.00474	0.00106	0.00652	-0.0598
	(0.0379)	(0.0396)	(0.0128)	(0.00466)	(0.0113)	(0.0568)
Needs Assessment	0.0447	0.0572	-0.0124	-0.00211	-0.0324*	0.00120
	(0.0377)	(0.0432)	(0.0188)	(0.00507)	(0.0195)	(0.0594)
Professional Capacity	0.0447	0.0572	-0.0124	-0.00211	-0.0324*	0.00120
	(0.0377)	(0.0432)	(0.0188)	(0.00507)	(0.0195)	(0.0594)
Student Centered Learning Climate	0.0374	0.0528	-0.0123	0.000925	-0.00423	-0.0898
	(0.0360)	(0.0371)	(0.0155)	(0.00436)	(0.0142)	(0.0639)
Parent Ties	0.0162	0.0243	-0.00330	-0.00414	-0.00275	-0.0704
	(0.0281)	(0.0363)	(0.0137)	(0.00431)	(0.0128)	(0.0589)
Schools Included	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle	Elem. and Middle

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only renewal schools. The coefficients show the difference between schools that implemented the program better than average and those that implemented it worse than average. The outcomes were first residualized using a ridge regression to improve precision of the estimates. Math and ELA test scores are measured in student standard deviation units. Standard errors (in parenthesis) are clustered at the school level.

* p<0.10 ** p<0.05 *** p<0.01

Table 11a. Implementation Results, High Schools (Panel 1)

	(1)	(2)	(3)	(4)	(5)
	Proportion Chronically Absent	Average Attendance Rate	Credits Earned	Times Suspended	Number of Disciplinary Incidents
Leadership Development	0.00739	-0.00613	0.496	-0.0231	-0.0747
	(0.0369)	(0.0129)	(0.491)	(0.0277)	(0.0574)
Needs Assessment	0.00561	0.00667	0.262	-0.0157	-0.0852*
	(0.0375)	(0.0123)	(0.533)	(0.0273)	(0.0436)
Professional Capacity	-0.0750*	0.0279**	0.666	-0.0402	-0.0530
	(0.0400)	(0.0136)	(0.465)	(0.0276)	(0.0543)
Student Centered Learning Climate	-0.0342	0.00412	0.552	0.0175	0.00697
	(0.0407)	(0.0137)	(0.522)	(0.0328)	(0.0569)
Parent Ties	0.0400	-0.00692	-0.872*	-0.0618***	-0.0876**
	(0.0444)	(0.0158)	(0.505)	(0.0228)	(0.0411)
Schools Included	High Schools	High Schools	High Schools	High Schools	High Schools

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only renewal schools. The coefficients show the difference between schools that implemented the program better than average and those that implemented it worse than average. The outcomes were first residualized using a ridge regression to improve precision of the estimates. Math and ELA test scores are measured in student standard deviation units. Standard errors (in parenthesis) are clustered at the school level.
* p<0.10 ** p<0.05 *** p<0.01

Table 11b. Implementation Results, High Schools (Panel 2)

	(1)	(2)	(3)	(4)	(5)
	Dropout Rate	Graduate Rate	Proportion Who Take the SAT	Proportion Who Take the PSAT	AP Tests Taken
Leadership Development	0.00193	-0.0103	0.00376	-0.0261	0.0517
	(0.00673)	(0.0300)	(0.0461)	(0.0356)	(0.0568)
Needs Assessment	0.00568	-0.0226	-0.0903**	-0.0238	0.0669
	(0.00527)	(0.0352)	(0.0377)	(0.0287)	(0.0476)
Professional Capacity	0.00360	0.0515*	-0.0655*	-0.0120	0.0784
	(0.00633)	(0.0299)	(0.0395)	(0.0286)	(0.0483)
Student Centered Learning Climate	-0.00124	0.0115	0.0333	-0.0374	0.0412
	(0.00866)	(0.0333)	(0.0584)	(0.0276)	(0.0568)
Parent Ties	0.0124*	-0.00151	-0.0326	0.0329	0.0487
	(0.00635)	(0.0304)	(0.0492)	(0.0289)	(0.0495)
Schools Included	High Schools	High Schools	High Schools	High Schools	High Schools

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only renewal schools. The coefficients show the difference between schools that implemented the program better than average and those that implemented it worse than average. The outcomes were first residualized using a ridge regression to improve precision of the estimates. Math and ELA test scores are measured in student standard deviation units. Standard errors (in parenthesis) are clustered at the school level.
* p<0.10 ** p<0.05 *** p<0.01

Appendix A. Supplemental Tables

Table A1. Implementation Data Summary

Variable	Min	Mean	Max	SD	% Miss
Needs Assessment	**0**	**2.329**	**3**	**0.717**	**18%**
School leader (SL) reports about RS targets and benchmarks leading to realistic yet rigorous goals[a]	0	2.361	3	0.793	15%
SL reports about RS targets and benchmarks leading to internal accountability with school[a]	0	2.271	3	0.741	18%
Leadership Development	**0.8**	**2.104**	**3**	**0.499**	**36%**
SL reports about RS supports helping with progress tracking using emerging data[a]	0	2.129	3	0.900	18%
SL reports about RS supports helping to adjust action plans toward meeting goals, targets and benchmarks[a]	0	2.085	3	0.982	16%
SL reports about RS supports amplifying distributed leadership practices[a]	0	2.085	3	0.915	31%
Quality Review Indicator 4.2 Teacher Teams and Leadership Development (2015-16)	1	1.845	3	0.570	1%
Quality Review Indicator 4.2 Teacher Teams and Leadership Development (2016-17)	1	2.158	3	0.463	11%
Parent Ties	**-5.237**	**0.071**	**1.662**	**1.107**	**31%**
Proportion of families that are at the "On Ramp" level in the OCS family engagement VAN system	0.015	0.187	0.594	0.119	2%
Collaboration capacity score from RAND's report on Community School implementation	-10.670	-0.039	2.990	2.183	29%
Professional Capacity	**1**	**2.066**	**3**	**0.652**	**28%**
SL reports about PD's effect on teacher improvement and student achievement[a]	1	2.016	3	0.660	26%
SL reports about PD's effect on curricular coherence[a]	1	2.127	3	0.707	26%
SL reports about PD's effect on teacher capacity for instructional leadership[a]	1	2.109	3	0.715	25%
Student Centered Learning Climate	**1**	**2.185**	**3**	**0.582**	**34%**
SL reports about RS supports helping retention of effective teachers[a]	0	1.789	3	0.995	33%
SL reports about Renewal Hour being sufficiently staffed[a]	0	2.423	3	0.822	16%
SL reports about Renewal Hour meeting students' academic and social-emotional needs[a]	1	2.217	3	0.764	19%

[a]Data derived from RAND's school leader survey, administered fall 2016. Items were on a likert scale items and coded so that 0 means "strongly disagree", 1 means "somewhat disagree", 2 means "somewhat agree" and 3 means "strongly agree.

Table A2a. Yearly Results, High Schools (Panel 1)

	(1)	(2)	(3)	(4)	(5)
	Dropout Rate	Graduate Rate	Proportion Who Take the SAT	Proportion Who Take the PSAT	AP Tests Taken
Estimated Effect in 2014-2015	-0.00429 (0.0133)	-0.0526 (0.0742)	-0.0227 (0.0700)	0.000834 (0.0281)	-0.0227 (0.0342)
Estimated Effect in 2015-2016	0.0256 (0.0168)	0.0582 (0.0820)	-0.0613 (0.117)	0.0269 (0.0444)	-0.00921 (0.0507)
Estimated Effect in 2016-2017	-0.0129 (0.0134)	-0.143* (0.0758)	-0.156 (0.114)	0.0467 (0.145)	-0.0957 (0.0933)
Schools Included	High Schools	High Schools	High Schools	High Schools	High Schools
Bandwidth of RD	20	20	20	20	20
Number of School-Year Observations	49	49	49	46	49

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only schools that were classified as "priority of focus" by the state of New York and were proficient or below in their review. Standard errors (in parenthesis) are clustered at the school level. The outcomes were first residualized using a ridge regression to improve precision of the estimates.

 * p<0.10 ** p<0.05 *** p<0.01

Table A2b. Yearly Results, High Schools (Panel 2)

	(1)	(2)	(3)	(4)	(5)
	Proportion Chronically Absent	Average Attendance Rate	Credits Earned	Times Suspended	Number of Disciplinary Incidents
Estimated Effect in 2014-2015	0.0126 (0.0491)	0.00494 (0.0181)	-0.0109 (0.624)	-0.0752 (0.0465)	-0.196 (0.124)
Estimated Effect in 2015-2016	0.0221 (0.0753)	-0.0166 (0.0212)	0.561 (0.955)	-0.0325 (0.0464)	-0.0650 (0.242)
Estimated Effect in 2016-2017	0.0758 (0.103)	-0.0351 (0.0301)	-0.0495 (0.906)	0.00225 (0.0492)	-0.292* (0.168)
Schools Included	High Schools	High Schools	High Schools	High Schools	High Schools
Bandwidth of RD	20	20	20	20	20
Number of School-Year Observations	49	49	49	46	49

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only schools that were classified as "priority of focus" by the state of New York and were proficient or below in their review. Standard errors (in parenthesis) are clustered at the school level. The outcomes were first residualized using a ridge regression to improve precision of the estimates.

* p<0.10 ** p<0.05 *** p<0.01

TA3a. Placebo Results, High Schools (Panel 1)

	(1)	(2)	(3)	(4)	(5)
	Dropout Rate	Graduate Rate	Proportion Who Take the SAT	Proportion Who Take the PSAT	AP Tests Taken
Placebo RD Estimate	-0.00260 (0.00803)	0.0209 (0.0370)	-0.0330 (0.0500)	-0.0126 (0.0384)	-0.0960* (0.0555)
Schools Included	High Schools	High Schools	High Schools	High Schools	High Schools
Bandwidth of RD	20	20	20	20	20
Number of School-Year Observations	85	85	85	85	85

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only schools that were classified as either not "priority of focus" by the state of New York or were not proficient or below in their review. Standard errors (in parenthesis) are clustered at the school level. The outcomes were first residualized using a ridge regression to improve precision of the estimates. The outcomes were first residualized using a ridge regression to improve precision of the estimates.

* p<0.10 ** p<0.05 *** p<0.01

TA3b. Placebo Results, High Schools (Panel 2)

	(1)	(2)	(3)	(4)	(5)
	Proportion Chronically Absent	Average Attendance Rate	Credits Earned	Times Suspended	Number of Disciplinary Incidents
Placebo RD Estimate	0.0321 (0.0483)	-0.00758 (0.0198)	-0.314 (0.453)	0.0456 (0.0320)	-0.0184 (0.0626)
Schools Included	High Schools	High Schools	High Schools	High Schools	High Schools
Bandwidth of RD	20	20	20	20	20
Number of School-Year Observations	85	85	85	85	85

Data from SY2015-16 and SY 2016-17 are included in the regressions, which included only schools that were classified as either not "priority of focus" by the state of New York or were not proficient or below in their review. Standard errors (in parenthesis) are clustered at the school level. The outcomes were first residualized using a ridge regression to improve precision of the estimates. The outcomes were first residualized using a ridge regression to improve precision of the estimates.

* p<0.10 ** p<0.05 *** p<0.01