ESSAYS IN THE ECONOMICS OF EDUCATION

By

Hwanoong Lee

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics – Doctor of Philosophy

2019

ABSTRACT

ESSAYS IN THE ECONOMICS OF EDUCATION

By

Hwanoong Lee

This dissertation comprises three essays on the Economics of Education. Its ultimate focus is to understand how different agents in the education market respond to releasing information about teacher and school performance and how public interventions influence human capital accumulation.

The first essay "The Effect of Releasing Teacher Performance Information to Schools: Teachers' Response and Student Achievement" examines the effects of releasing teacher value-added (VA) information on student performance in two settings; in the first, VA data was released to all potential employers within the district, while in the second, only the current employer received the data. I find that student achievement increased only in the district where the VA scores were provided to all potential employers. These effects were driven solely by improved performance among ex-ante less-effective teachers; the null effects in the other setting, however, were driven by moderate declines in performance among ex-ante highly-effective teachers and small improvements among less-effective teachers. These results highlight the importance of understanding how the design features of VA disclosure translate into the productivity of teachers.

The second essay "The Role of Credible Threats and School Competition within School Accountability Systems: Evidence from Focus Schools in Michigan" studies the impact of receiving accountability labels on the student achievement distribution under No Child Left Behind (NCLB) waivers. Using a sharp regression discontinuity (RD) design, I examine the achievement effects of Focus (schools with the largest achievement gaps) labels and find that schools receiving the Focus label improved the performance of low-achieving students relative to their barely non-Focus counterparts, and they did so without hurting high-achieving students. The positive achievement effects

for Focus schools were entirely driven by Title 1 Focus schools that faced financial sanctions associated with being labeled the following year. There is no evidence of an achievement effect associated with the Priority label. Next, I examine heterogeneous effects by looking at the number of alternative nearby schooling options. I find that when schools are exposed to a competitive choice environment, receiving the Focus label increased math test scores across the scoring distribution, while schools located in an uncompetitive choice environment improved the test scores of low achievers only. This evidence may suggest the importance of incorporating credible sanctions and school choice options into the school accountability system to maximize the effectiveness of the system on student achievement.

Finally, the third essay "The Effects of School Accountability Systems Under NCLB Waiver: Evidence from Priority Schools in Michigan" investigates the impact of receiving Priority labels on the student achievement distribution under No Child Left Behind (NCLB) waivers. Using a sharp regression discontinuity (RD) design, I examine the achievement effects of the Priority (schools with the lowest performance) label and find no evidence of an achievement effect associated with the Priority label. Next, I examine whether assigning the Priority label induced the changes in the composition of students. I define several key measures of student composition and find no evidence that the Priority designation influenced the student composition of schools.

To my dear son Luo

ACKNOWLEDGMENTS

At the culmination of my dissertation completion, I am much indebted to so many for their continuous help and support. First, I thank my advisor, Professor Todd E. Elder. I always regard myself as very fortunate to know him and to have been under his instruction. He carefully guided and inspired me through each step of my research, and invested countless hours reading and correcting my many drafts. Without his guidance and encouragement, this dissertation could never have been completed. Besides, he was incredibly generous with his valuable time and willingness to talk with me about my personal and academic issues. Throughout my career, I shall remain deeply grateful to him.

My dissertation committee member, Professors Scott A. Imberman have been superb. I thank him for his guidance, time and encouragement throughout the process of my dissertation and throughout my MSU years. Professor Imberman not only read many of my papers, but also always were willing to meet me and to attend my presentations at the Red Cedar Conference, at my Dissertation Proposal, and at several formal/informal seminars, then to give me extremely helpful comments and suggestions. His insights and fine minds always furthered my search for better answers to intriguing questions.

I am also indebted to Professor Soren Anderson, Mike Conlin, Joshua Cowen, Steven Haider, Stephen Woodbury, and Ben Zou for their insights and suggestions.

My fellow students were invaluable resources of personal and academic support. How gratifying it is to recall those days of studying together and discussing problem sets, research, and life with Hyeongjun Bang, Yeon Choi, Hyunsoo Kim, Soojin Kim, Soobin Kim, Kyu Yub Lee, and Jaemin Ryu.

Without financial support from the Department of Economics, my graduate study could not have been completed successfully. Thanks to gracious offers. And special thanks to Lori Nichols, Jay Feight of the Department of Economics for their administrative help and support during the last six years.

Finally, my sincerest gratitude goes to my loving wife, Sun Young Lee, for her love, support, and patience. And, as always, I owe incalculable debt to my father, Dea Yeol Lee, and mother, Hyunsook Lee, for their love and endless sacrifices enabling me to achieve my dreams.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

The Effect of Releasing Teacher Performance Information to Schools: Teachers' Responses and Student Achievement

1.1 Introduction

Recently, school districts and states across the country have begun to use value-added (VA) methodologies to evaluate teachers. As of 2015, 43 states required teacher VA measures to be included in their teacher evaluations (Dorety and Jacobs 2015), and VA scores have been actively used for retention, assignment, and compensation decisions across states. For example, 10 school districts in 7 states used VA measures to identify highly effective teachers and provided bonuses if those teachers transferred to schools serving the most disadvantaged students (Glazeman et al., 2013).[1] The Houston Independent School District implemented the ASPIRE incentive pay program, which is a rank-order tournament based on teachers' VA measures, and District of Columbia Public Schools used the VA score as one component of the teacher evaluation system and introduced performance-based incentives and sanctions based on the evaluation score. The application of VA measures is a particularly important step to leverage teacher quality, as recent evidence shows that these measures serve as an unbiased estimator of teachers' effects on student achievement (Chetty et al., 2014a) and that high VA teachers have large positive effects on students' future outcomes (Chetty et al., 2014b).

The value of teacher VA information, however, is not limited to its direct use in various personnel decisions. If VA information is provided to teachers, intrinsically motivated teachers may

[1] This intervention is known to participants as the Talent Transfer Initiative (TTI)

learn about their ability to improve student achievement and may change their instructional methods to improve their performance. If principals receive the VA information, they can recognize the effective teachers, modify their managerial decisions regarding teacher assignments, and provide adequate support to less-effective teachers (Rockoff et al., 2012). In addition, the VA information can be provided to both individual teachers and their principals or can be provided in a more public manner, enabling all other potential employers to access this information. Releasing performance information publicly to all potential employers may create extrinsic incentives for teachers who value their reputation or self-image. Hence, providing productivity information would be a useful policy tool, especially for the public sector, where the compensation structure is more rigid than in the private sector. Previous research, however, has concentrated on how VA measures based incentive pay programs influence student achievement (Brehm, Imberman, and Lovenheim, 2017; ; Adnot et al. 2016; Imberman and Lovenheim, 2015; Dee and Wyckoff, 2015; Glazeman et al., 2013), and these studies have been unable to separate the impact of VA disclosure from the impact of financial incentives embedded in the incentive pay program. Since the VA measures of teachers are readily available across school districts nationwide due to recent teacher evaluation reforms, understanding the impact of accessing this information on student achievement is essential.

In this paper, I conduct a two-pronged empirical analysis of the impact of releasing teacher VA information on student achievement by examining the unique policy changes in two urban districts in North Carolina. In 2000, Guilford County Schools (hereafter, Guilford) decided to receive annual teacher performance measures that basically showed the teachers' contribution to student achievement gains each year, while Winston-Salem/Forsyth County Schools (hereafter, Winston-Salem) decided to provide the same measures to schools in 2008. Since Guilford shared the VA information with all principals in K-12 public schools within the district while Winston-Salem provided the information to current principals only, the marginal benefit of increasing effort level for teachers in Guilford, on average, may be higher than teachers in Winston-Salem because Guilford teachers already knew that principals of their important outside options could access their VA data. Therefore, analyzing two natural experiments provides insight into whether providing the performance information publicly is crucial to create incentives for teachers to respond to the VA information and consequently improve their effectiveness in terms of student achievement.

The first empirical analysis evaluates the mean effects of releasing VA information on students' academic achievement. One common challenge for policy evaluations is that the policy change might be endogenous to unobserved aggregate-level shocks. For example, the treated districts may adopt VA information after experiencing negative shocks to student achievement. To address the endogeneity issues for each policy change, instead of using a simple difference-in-difference model, I use 3rd graders in a treated district as a placebo group because their teachers did not receive the VA information or implement the synthetic control method to isolate the unobserved district-level shocks that would confound the estimated effects. My findings across the numerous specifications and variations in modeling choices provide consistent evidence that distributing VA information improved student achievement in math test scores by approximately 0.096 standard deviation (SD) in Guilford. However, for Winston-Salem, I find little evidence that adopting teacher VA scores influenced student achievement, as all estimates from different specifications are close to zero and are not significantly different from zero at any conventional level. The estimates for Winston-Salem thus provide a cautionary note that VA policies do not automatically translate into student achievement gains.

The mean achievement effects could be driven by two potential mechanisms: (1) teachers increase their effort level and consequently increase their impacts on student achievement gains and (2) principals use this information strategically to improve the average test scores of their schools by assigning more students to highly effective teachers or laying off less-effective teachers. To determine which mechanisms influenced the different results in the two districts, the second component of my analysis concentrates on whether providing VA information to teachers changed the teachers' impact on student achievement gains. Using education production functions that carefully control for student-, classroom-, and school-level variables, I measure how the effects of VA policies vary across the distribution of teacher quality, as measured by pre-policy VA scores.

My findings across the various specifications show that the mean effects mask substantial heterogeneity with respect to teachers' pre-policy effectiveness. Specifically, the release of VA scores compressed the subsequent distribution of measured teacher quality in both districts. The decline in the performance gap between the best and worst teachers was largest among math teachers in Guilford, where the impact of a one standard-deviation increase in teacher quality declined by 0.070 SD (measured across the distribution of student test scores) after teachers received their VA scores. For math teachers in Winston-Salem, the benefit of having one standard deviation-higher VA teachers decreased by 0.038 SD. There was no measurable effect on the teacher performance distribution for reading teachers in either district. I also show that the positive mean achievement effects in Guilford were mostly driven by an increase in productivity among less effective teachers while the performance of highly effective teachers remained the same. The null effect on the average achievement in Winston-Salem, however, was driven by moderate declines in performance among highly effective teachers and small improvements among less effective teachers.

To test whether the positive achievement effect in Guilford was driven by other potential mechanisms, I estimate the impact of releasing information on classroom assignments and teacher turnover. I find no evidence that highly effective teachers were given larger classes or assigned students with higher underlying test score growth when teacher VA information was available to principals. The estimates for teacher turnover or switching into nontested grades are very small and statistically insignificant, which further suggests that the improved performance of teachers in Guilford drove the results.

My paper is closely related to that of Rockoff et al. (2012), who examine a pilot experiment that provided teacher VA information to principals in New York City, and to the papers of Bergman and Hill (2018) and Pope (2015), who study the release of teacher VA measures by the *Los Angeles Times* in Los Angeles. The prior research, however, provides a limited understanding of how teachers and principals respond to VA disclosure. The effect of VA disclosure in my context is highly policy relevant compared to these studies because both teachers and principals can access the VA information under recent teacher evaluation reform. The pilot experiment discussed by Rockoff et al. (2012) provided the VA data only to principals, making it very difficult to generalize the findings from this study to a setting where the VA data are provided to both principals and teachers.

Additionally, the findings from examining the public release of VA information through the *L.A. Times* have limited implications for settings in which teachers receive their performance measures privately. For example, Bergman and Hill (2018) report that higher-rated teachers had classroom scores approximately 0.2 SD higher than teachers rated one level lower after the release of VA information. However, most of these effects are driven by positive student and teacher sorting, which may complicate the analysis of the effects of VA disclosure on teacher performance; the student sorting within schools is expected to be small in my context since parents cannot access the VA scores. Furthermore, both Pope (2015) and Bergman and Hill (2018) only use test score variations taught by ex-ante worst and best teachers within the district. My analysis uses between district variations as well as within-district variations, which allow me to determine the average achievement effect of releasing VA data.

Finally, Bates (2017) also examines the provision of teacher VA information in both Guilford and Winston-Salem, but his paper focuses on the labor market consequences of providing this VA information. Exploiting the fact that the VA information was not provided to principals in other districts, he develops an asymmetric employer learning model and tests its ability to predict teacher mobility. He finds that less-effective teachers in treated districts were more likely to switch into the districts where principals were uninformed about the VA measure, while highly effective teachers in treated districts were more likely to switch into preferred schools within the treated districts.

The remainder of this paper is organized as follows. Section 2 details how the VA information was provided to schools, and I discuss the data in Section 3. Section 4 examines the mean achievement effect of providing VA information, and Section 5 examines the heterogeneous achievement effect based on the teachers' initial productivity level. I discuss other potential mechanisms for my results in Section 6, and Section 7 concludes the paper.

1.2 Background

In the 2000-2001 school year, Guilford contracted with Statistical Analysis Systems (SAS) to receive a report that provides measures of teacher effectiveness, and Winston-Salem decided to adopt the same measures as of the 2007-2008 school year. Teacher effectiveness in this report is estimated for each teacher using the Education Value-Added Assessment System (henceforth EVAAS) by SAS. The system is rooted in the Tennessee Value-Added Assessment System (TVAAS) model developed by Dr. William Sanders and colleagues. This model simultaneously estimates the teacher effects for separate subjects, grades, and years by estimating a set of linear mixed models that regress the full set of student test scores on indicator variables of the teachers a student had in the current and two previous years as well as indicators for subject, grade, and years.[2] SAS calculates the VA measure every year, which is interpreted as each teacher's impact on student test scores in a given year compared to the impact of the average teacher in the district.

The VA information is estimated for teachers who teach subjects for which the state of North Carolina requires multiple-choice end-of-grade (EOG) assessments or end-of-course (EOC) assessments. For 3rd to 5th grade students, the tested subjects include math and reading, and for 6th to 8th grade students, the tested subjects include math, reading, science, and social science. While it is possible to estimate the teacher VA measure for 3rd grade teachers since the-beginning-of-grade (BOG) tests were administered from 1997 to 2005, SAS did not provide VA measures for 3rd grade teachers. Once the VA score is estimated using the EOG and EOC test scores, the principals and teachers can access the VA information when the new academic year begins (late September or early October).

Figure 1.1 shows an example of how the VA information was presented in the EVAAS teacher report. The top panel contains the mean student score in the EOG math test as well as the mean predicted score, which is the mean expectation score based on the student's performance on previous tests, assuming these students were taught by average teachers within the district. The teacher

[2] See Ballou, Sanders, and Wright (2004) for a detailed description of the model

VA score, which is labeled the "Teacher Effect" in the report, is estimated by comparing the mean student score and the mean predicted score, and the standard error provides the basis for constructing a confidence interval around the Teacher Effect. Finally, in the last column, teacher effects are categorized as "Above," "NDD," and "Below." "Above" ("Below") indicates teachers who were (not) effective in improving student achievement compared to average teachers in the district, and "NDD" indicates teachers whose influence on student achievement was not significantly different from that of average teachers. Finally, the EVAAS teacher report also presents the teacher effects at different achievement levels, which is shown in the bottom panel. This figure is intended for diagnostic purposes because teachers may want to explore ways to improve instruction for the students making less progress. Green bars show the progress of students in the current school year, and the red interval is a 95 percent confidence interval.

The EVAAS teacher reports were accessed through the EVAAS software. Teachers and school administrators could access the teacher reports based on the level of access. One notable difference between the two districts is that Guilford allowed principals to access the VA reports of all teachers within the districts, while Winston-Salem permitted principals to access the reports of teachers in their schools only. This distinction may have influenced the teachers' responses to the VA data differently for two reasons. First, teachers in Guilford may have perceived the VA adoption as increased scrutiny of their performance because all principals within the district could access their performance information every year. Such access may provide extrinsic motivation to improve for teachers who were concerned about their reputation or self-image (Benabou and Tirole, 2006; Goldhaber and Hannaway, 2004).

Second, releasing performance information publicly to all potential employers may create extrinsic incentives for Guilford teachers who want to switch to preferred schools, as prior studies have found that teachers in low-performing schools are more likely move to high-achieving schools (Jackson, 2009; Hanushek et al., 2005; Hanushek, Kain, and Rivkin, 2004). Theoretically, however, teachers in Winston-Salem who want to switch into the preferred schools may also have the extrinsic motivation to improve their performance because they would anticipate potential employers to require the VA report in the hiring process. If a teacher does not share the VA report in the hiring process, principals might assume that the teacher is only as good as the average teachers who do not share the VA score, and consequently teachers whose scores are above the average would likely provide their VA report. The average scores of teachers who do not share the information would drop further until finally, all potential candidates would submit their reports. In practice, however, teachers in Winston-Salem may take time to rationally expect that they will be required to provide the VA information during the hiring process, or principals may not ask to see the VA score if they do not value this score.

One remaining concern is that both districts adopted the VA information in the early years, when the VA methodologies were rarely used, and it thus is likely that both teachers and principals felt unclear about the VA measures. Recent research surveying teachers who received VA scores demonstrated that teachers felt confused about what the new information meant and developed a negative attitude toward VA measures (Davis et al., 2015; Thomas 2014). However, there were district-level efforts in both districts to support teachers and principals. The districts provided a series of professional development seminars to help teachers and administrators understand what teacher VA reports provide. Furthermore, Guilford monitored the principals' evaluations of each teacher and sent a notification to principals if their subjective ratings of teachers were not consistent with the VA ratings.

1.3 Data and Sample

To assess the impact of providing VA scores on student performance, I use the matched student and teacher records covering the period from the 1995-1996 through the 2010-2011 school years from the North Carolina Education Research Data Center (NCERDC). The data contain the BOG and EOG test scores of 3rd to 5th graders in math and reading and various student and teacher characteristics. Student characteristics include grade, gender, race, parent education, and 6 categories of exceptional status, including special needs and gifted status. For teacher characteristics, the data include gender, race, highest degree earned, and years of teaching experience.

The primary objective of my research is to determine whether student academic achievement is influenced when a teacher receives VA information; thus, it is important to identify teachers who are eligible to receive the VA information. Although the North Carolina data include codes that link students and teachers, until 2006, the teacher codes indicated the proctors who administered the EOG test. For elementary classrooms, the proctor was likely to be the classroom teacher, but I restrict my sample to ensure that I match students to their classroom teachers correctly.[3] First, I eliminate student-year observations in cases in which proctors for the given subject did not have the given subject list in the school activity report (SAR) in the given year.[4] Additionally, I remove teachers who were co-teaching, had a teaching assistant or had fewer than ten students in a given year.

To carry out some econometric specifications, I require that a student had current test scores as well as lagged test scores. Hence, I drop a student-year observation if the student had data for only a single year. The only exception is 3rd graders because the BOG test, which is administered to assess second grade knowledge and skills, is available. As discussed in more detail below, including 3rd graders is critical to my identification strategy because it allows me to control for district-specific shocks that could influence adopting the VA policy and student achievement. The pretest data for 3rd grade are available only from 1997 to 2005, which covers the period from before to after Guilford began providing the VA measure; however, the pretest was not administered when Winston-Salem adopted the VA information. Hence, I generate two samples because different identification strategies are required to evaluate these two cases. I use the entire 1997-2011 period to evaluate the policy change in Winston-Salem but only the 1997-2005 period to evaluate Guilford.

Summary statistics of certain key variables for the Guilford and Winston-Salem samples are shown in Table 1.1. The table compares the means and SD of Guilford and the rest of the districts in the first four columns and compares them to those of Winston-Salem and the rest of the districts

[3] In describing these data, Jackson (2013) mentions that "discussions with education officials in North Carolina indicate that tests are always administered by the student's own teachers when teachers are present"

[4] I can match 88 percent (91 percent) of student-year observations for math (reading) to proctors who have a valid math (reading) class in the SAR. In most cases, a valid math teacher is also a valid reading teacher; however, for approximately 3 percent of student-year observations, the teacher was valid for only either math or reading. I exclude these student-year observations, though including these observations does not change any of the results.

in the last four columns. On average, Guilford and Winston-Salem are not representative of North Carolina. The table shows that Guilford and Winston-Salem had a higher proportion of black students (approximately 41 percent and 34 percent, respectively) than the rest of North Carolina, the lowest proportion of white students (approximately 49 percent and 52 percent, respectively), and the highest proportion of black teachers (approximately 24 percent and 20 percent, respectively). For the remaining districts, however, the proportion of black students was just over 27 percent, white students accounted for more than 60 percent of the population, and the proportion of black teachers was approximately 13 percent.

1.4 The Effect of Providing the Value-Added Information on Student Achievement

In this section, I evaluate the impact of adopting the VA information on student performance. As data availability does not allow a single empirical strategy to be used for both Winston-Salem and Guilford, I use two different empirical strategies. Section 4.1 describes a difference-in-difference-in-difference (DDD) approach to evaluate the policy change in Guilford and explains a synthetic control method for Winston-Salem. In Section 4.2, I provide estimates across the numerous econometric specifications for both districts.

1.4.1 Empirical Strategy

My first objective is to estimate the impact of adopting VA information on student achievement. Since almost all fourth and fifth grade teachers in treated districts received VA scores, I can employ a difference-in-difference (DID) model that compares the test score gains in a treatment district relative to the control districts when the policy was implemented. One concern with this approach is that the policy would be endogenous to unobserved district-level shocks. If a district systematically adopts the VA scores after experiencing negative persistent shocks on student achievement, the DID approach would underestimate the benefit of the policy. However, if a district receives

VA information after one or two bad years, the test scores would likely revert to the mean, and the DID approach would spuriously capture the positive impact even if the policy did not have a causal impact on student achievement.

In Figure 1.2, I present the average math and reading scores of 4th and 5th graders in Guilford, Winston-Salem, and the rest of districts. I also plot the 95 percent upper and lower bounds of the mean test scores relative to the one year prior to the year when the districts decided to adopt the VA measure. Figure 1.2 shows some evidence of a prepolicy trend before both districts adopted the VA scores. For Guilford, while the math and reading scores of all other districts were relatively stable during the prepolicy period (1997-1999), Guilford experienced test score drops in this period. The math score of Guilford in 1999 is significantly different from the math scores in 1997 and 1998, and the reading score in 1999 is significantly different from the reading scores in 1997. For Winston-Salem, in panels C and D, the pretrends are somewhat stable but provide systematic evidence in reading scores. The reading scores in Winston-Salem declined from 2003 to 2007, and the test scores for 2003 to 2005 are significantly different from the score in 2007.

To address the pretrends in Guilford, instead of using school level data, I use test scores and demographic characteristics of individual students to isolate the causal impact of providing VA information on student test score from the changes in composition of students who took the test. Next, I exclude the small districts in terms of student enrollment from the comparison districts. The exclusion of small districts is motivated by the fact that Guilford is the third largest school district in North Carolina and schools in Guilford are mostly located in urbanized areas. Also, when I implement permutation-based inference, small districts with volatile outcomes do not provide information to measure the relative rarity of estimating a large treatment effect for a large district where outcomes were stable in the prepolicy period. In Table 1.8, I report the means of key variables of Guilford and a set of comparison districts using the student enrollment ranking. I construct comparison districts using 60 large districts, 30 large districts, and 15 large districts.[5]

[5] I do not further limit the number of large districts to construct the comparison districts that were similar in size to Guilford as this limits the range of confidence levels I can perform for the placebo-based inference from permuting the treatment status over comparison districts.

Across the columns, the comparison districts are not identical to Guilford, but the comparison set that includes 15 large school districts is more comparable than the other comparison groups in most variables. For example, for Comparison Districts 1 and 2, four of the p-values lie below 0.05, but that number is three for Comparison District 3. I thus prefer using the 15 large districts as a comparison group rather than using all districts, but I also report the robustness of estimates across all sets of comparison districts given their relative similarity.

In Figure 1.3, I display the estimates from the event study model, where a treated dummy variable is interacted with a series of indicator variables for time relative to the one year prior to the year when Guilford (Winston-Salem) adopted VA reports. Each of the points in this figure indicates the test score differences between treatment and control districts conditional on various student-level controls, and all 95 percent confidence intervals use standard errors clustered at the school level. Panels A and B (Panels C and D) show the estimates using 15 large districts for Guilford (Winston-Salem). The panels A and B suggest that the concern of endogenous policy changes in a treated district is minimal once I control for student-level covariates and use the 15 large districts. In panel A, the estimate for math in 1997 is positive and becomes negative in 1998. Both estimates are small and not statistically distinguishable from zero, and then the estimates become positive and significant immediately after the policy change. In panel B, I find little evidence of pretrends in reading, because the estimates in 1997 and 1998 are small (-0.016 and -0.014 SD, respectively) and statistically insignificant.

Although I do not find any evidence of pre-trends in Guilford, it is still possible that unobserved district-level shocks or policies that influenced student achievement gains may coincide with the adoption of the VA data. To address this concern, I include 3rd graders as another comparison group and exploit another round of differencing. The 3rd graders in Guilford are an ideal comparison group as SAS did not provide VA measures for 3rd grade teachers, while I can use 3rd graders to control any district-level shocks that would coincide with adopting the VA data. For example, a disruptive event, such as a tornado, that affects Guilford (which did not have a differential effect across grades) would not lead to bias once I include 3rd graders. The DDD approach identifies the impact of providing the VA measure by comparing the changes in student achievement between 3rd graders and 4th/5th graders within Guilford relative to changes in achievement in the comparison districts conditional on observable student characteristics. The main empirical model is as follows:

$$\begin{aligned} y_{it} = \sum_{g=3}^{5}(Y_{it-1} \times Grade_g)\gamma_{1g} + X_{it}\gamma_2 + \delta_d \times \delta_t + TG_g \times \delta_t \\ + \beta_0 TD_d \times TG_g + \beta_1 TD_d \times TG_g \times Post_t + \varepsilon_{it} \quad (1.1) \end{aligned}$$

The dependent variable y_{it} is student i's EOG math and reading scores in year t, which are normalized by grade and year. A vector of lagged math and reading scores, Y_{it-1}, entered on the right-hand side allow the correlation of previous test scores with current scores. Given that the lagged scores for 3rd graders are from the BOG test instead of the-end-of-grade-2 test, the role of my lagged achievement measure may change by grade level. Thus, I interact the lagged math and reading score (Y_{it-1}) with grade indicators ($Grade_g$). X_{it}, is a vector including the observable student characteristics, including grade, gender, race, and the 6 categories of exceptional status, such as special needs, and gifted status. Additionally, the triple-difference model includes a full set of district (δ_d), year (δ_t), and treated grade (TG_g) fixed effects, and all of the two-way interactions to control the unobserved common shocks. $TG_g \times \delta_t$ controls for any unobserved shocks that affected all students in the treated grade across districts in a given year, and an unobserved local shock that influences all students in Guilford is captured by $\delta_d \times \delta_t$. $TD_d \times TG_g$ shows the average test score differences between treated and non-treated grades in treated districts (TD_d). The parameter of interest is β_1 which shows the impact of providing VA information on student achievement gains. Finally, because of the likelihood that errors are correlated across students within schools and within schools over time, for all specifications, I provide the standard errors in all the analyses that are clustered at the school level.

To check whether using student level data and limiting the comparison districts correct for the endogenous policy changes in Winston-Salem, panels C and D in Figure 1.3 presents the analogous results for Winston-Salem. These panels suggest some evidence of pretrends for both math and reading. Specifically, the estimates in year 2004 for both math and reading are negative (-0.07 and -0.05 SD, respectively) and statistically distinguishable from zero, and then the estimates gradually increase up to year 2007.[6] Unlike Guilford, however, I am not able to control for the pretrends in Winston-Salem using the triple model in Eq.(1), as the BOG tests were not administered during the sample period. Instead, I construct a control group using the synthetic method proposed by Abadie et al. (2010) to address the pretrends. The key insight of their method is that if one can create a control group using the weighted average of comparison districts in North Carolina that closely follows the outcome trajectories of Winston-Salem in the prepolicy period, this control group can be used as the counterfactual for Winston-Salem that would have occurred in the post-treatment period. Note that I cannot use the synthetic control method to evaluate the policy change in Guilford. The reason is that the applicability of the method requires a sizable number of prepolicy years, but I only have three prepolicy years for Guilford while I have eleven prepolicy years for Winston-Salem

To obtain the optimal weight for Synthetic Winston-Salem, I choose the optimal weight that solves the following equation:

$$W^*(V) = argmin_w \, (X_0 - \sum_{d=1}^{D} w_d . X_d)' V \, (X_0 - \sum_{d=1}^{D} w_d . X_d) \tag{1.2}$$

where $X_d \, (K \times 1)$ indicates a vector of predictor variables in donor district d, and $V \, (K \times K)$ is the predictor importance matrix.[7] Once I obtain a set of optimal weights for donor districts, my estimator for the treatment effect in year t is as follows:

[6] In Figure 1.8, I display the analogous results that use all districts for both Guilford and Winston-Salem. Regardless of the choice of comparison districts, the estimated results are qualitatively similar except for the estimates in panel B; I find some evidence of pretrends in reading when I use all districts, although the estimates in 1997 and 1998 are small.

[7] Follow the recommendation in Abadie et al. (2010), the importance matrix V is chosen to minimize the distance of a vector of the pretreatment outcome trajectories between the treatment and all donor districts. I choose the matrix V by solving the joint optimization procedure canned in the synthetic package in Stata.

$$\hat{\alpha}_t = y_t^0 - \sum_{d \in D} w_d^*(V) y_t^d \tag{1.3}$$

To compare to the estimator used in Guilford, my preferred estimate is a DID estimate that compares the average difference between the treatment and the synthetic control districts before and after the year when Winston-Salem decided to adopt the VA information.

$$DD_{ws} = (\frac{1}{T - T_0} \sum_{t \geq 2008} \hat{\alpha}_t) - (\frac{1}{T_0} \sum_{t < 2008} \hat{\alpha}_t) \tag{1.4}$$

The remaining challenges to implementing the synthetic control is selecting the pretreatment outcomes and predetermined variable as predictors. When the number of pretreatment years is finite, Ferman et al. (2017) show that the estimated results are sensitive to the choice of predictors, but there is little definitive guidance in the synthetic control literature to select the set of predictors. I follow Dube and Zipperer (2015) in selecting predictors as their method provides a transparent way of choosing predictors and thus minimizes specification searching. I use four different choice sets, which differ in whether I use all pretreatment outcomes or biannualized pretreatment outcomes, and in whether I include the pretreatment means of demographic characteristics. Once I defined the predictor sets, I use the cross-validation procedure that first fits the model using the pretreatment period and then evaluates the model performance based on the post-treatment period (i.e., using out-of-sample) to select the optimal predictor set. Clearly, using all pretreatment outcomes would maximize the pretreatment outcome fit in Eq.(3). However, this condition does not necessarily mean that this choice minimizes the prediction errors of the outcomes in the post-treatment period, which is the ultimate goal of interest to increase the reliability of the model performance. To compare the predictability of "synthetic controls" of donor districts in the post-treatment period, I use the average RMSPE (root mean square prediction error) in the postpolicy period for each choice of predictors and select the optimal predictors that minimize this quantity.

Table 1.2 shows the average RMSPE of all possible combinations for predictor sets in the pre- and post-treatment periods with 60 large donor districts. Details on how to select donor districts which consist of the donor pool to construct synthetic controls is given in Appendix C. Predictor set 1 in column (1) includes the set of average test scores in the pretreatment period in a given subject, and predictor set 2 in column (2) includes the biannual average test outcomes in a given subject. In column (3), I use both the biannual average math and reading scores, and in column (4), I include the biannual average math and reading scores with controls, including the pretreatment means of percent black, white, female, gifted, special education students, and student enrollment variables. The table shows that using annual mean test scores yields the smallest RMSPE in the post-treatment period regardless of which predictor sets are used. One explanation for this is that prior test scores can be regarded as a sufficient statistic to predict future test scores and thus adding additional covariates would not be beneficial.[8] Nevertheless, I report DID estimates with other predictor sets as the improved predictability from using all pretreatment outcomes are somewhat small compared to predictability when using other predictor sets.

1.4.2 Results

1.4.2.1 Guilford

Figure 1.9 shows the estimates using the event study model with covariates, where the treated grade (TG in Eq.[1]) interacts with a series of time indicators relative to the year prior to the year when Guilford adopted VA reports. The point estimates indicate the positive achievement effects on student math achievement as the achievement effect remains small and unchanged in the pretreatment period and then increases steadily starting the year in which teacher VA information was provided to the treated district. The figure shows little evidence of achievement effect for reading, however.

Results from the more parametric DDD model of Eq.(1) confirm the positive achievement effects on student achievement in math. In column (1) of Table 1.3, I estimate the simple DID

[8] Kaul et al. (2018) recommend using only the last observed outcome as a predictor instead of using all previous outcomes when other covariates are important to predict outcome variables. The basic idea behind this recommendation is that using all outcome variables as separate predictors renders all other covariates irrelevant. I calculate the average RMSPE in the postpolicy period with the last lag and the covariates. This average RMSPE is much larger than the average RMSPEs reported in Table 1.2, suggesting that prior test scores are more important than other demographic variables in predicting future test scores.

model that compares the test score changes of the TG in Guilford and all districts to illustrate how the endogeneity of the policy change would affect the estimates. Column (2) presents an estimate provided by another DID model that compares the test score changes in the TG and nontreated grade within Guilford. Column (6) contains my preferred estimates, which calculate the DDD model that includes 15 large districts as a comparison district. Finally, in columns (3) through (6), I report the estimate results using different comparison districts to reveal whether the estimates are sensitive to the choice of comparison districts. Across columns (3) to (6) in panel A of Table 1.3, there is evidence that providing VA information increases achievement gains in math conditional on various student observables. The point estimate in column (6) indicates that the achievement gap between the treated and nontreated grades in Guilford relative to the gap in the comparison districts widens by 0.096 SD after Guilford adopted the VA measures for 4th and 5th grade teachers. To relate this estimate to prior literature, the effect of providing the VA score to teachers and principals was approximately two times greater than the effect of providing VA information to the principal only (0.053 SD; Rockoff et al., 2012) and similar to the impact of the teacher evaluation reform in Cincinnati public schools (0.112 SD; Taylor and Tyler, 2012).

The point estimates in column (1) and column (6) are statistically indistinguishable, and the magnitudes of estimates are remarkably similar, which suggest that any confounding factors that influenced all grades in Guilford were not driving my results. Furthermore, the point estimates in column (2), which exploit the variations within Guilford, are comparable to the point estimates in columns (3) to (6). This indicates that any statewide policies within comparison districts that have differential impacts on the TGs and non-TGs would not influence the achievement effects.

Panel B shows the estimates using the reading test score as an outcome variable. Across all specifications, I do not find any evidence that the policy change increases the achievement gains in reading. The estimates, while positive, are all small and insignificant even at a 10 percent significance level. This finding may not be surprising, however, as prior studies examining similar policy interventions do not find achievement effects for reading, although these studies do report positive achievement effects for math (Taylor and Tyler, 2012; Rockoff et al., 2012). One explanation of the null effect in reading is that the teachers' effect on reading achievement is less varied than the teachers' effects on math achievement (Hanushek and Rivkin, 2010; Rivkin et al., 2005; Rockoff, 2004); thus, providing VA measures for reading would have smaller returns.

One may concern that the statistical inference that I used here is unlikely to be valid as I have only one treatment district. Conley and Taber (2011) show that the standard errors of estimates using the common methods in DID analysis, such as cluster-robust standard errors, will be severely biased with only one or two treated groups as the common inference methods rely on the large number of treated districts. To address this concern, I follow the non-parametric permutation test discussed in Chetty et al. (2009) for the triple interaction term, β_1 in Eq.(1). Since this method does not make any parametric assumptions, the inference would not be biased even if the number of the treated group is small. I define a "placebo triple" as consisting of one district with two treated grades out of three. I then estimate the triple interaction term using Eq.(1), assuming that the placebo triple is the treated group. Next, I repeat this procedure for all possible permutations of districts and TGs. For example, I repeat this procedure 330 times when I have 110 districts and 3 grades. Given sufficient "placebo triples," this approach produces a distribution of estimates of treatment effects under the null hypothesis of zero treatment effects. I obtain the p-values of the estimates by calculating the proportion of estimates that are larger than the estimates reported in Table 1.3.

The only assumption of the permutation test is that the distribution of the vector of observed outcomes is invariant with respect to reassignment of treatment status, which implies the distributions of DDD estimates from the treated districts and those of the control districts are identical.[9] In Figure 1.10, I displays the distribution of the average math score by the number of students. This figure clearly shows that the distribution of the average math score in small districts (fewer than 200 students) is more dispersed than the distribution of the score in 15 large districts; the test statistic from the two-sample Kolmogorov-Smirnov test is 0.241, and I can reject the null hypoth-

[9] This is formally called the symmetry assumption in a randomization test. See Canay et al. (2017) who discuss the property of the symmetry assumption in detail. In addition, see Hahn and Shi (2017) who discuss this assumption in the context of the synthetic method.

esis of the equal distribution at any statistical significance level. Using the distribution of the DDD estimates with all districts therefore biases the statistical inference as the distribution of the DDD estimates for Guilford is unlikely to be identical to the distribution that includes small districts.

To address this concern, my preferred specification uses the set of 15 large districts as a control group, but I conduct a set of permutation tests with the different comparison districts defined in Table 1.3 to evaluate how the failure of this assumption would affect the permutation inference. Figure 1.4 illustrates the results of the permutation tests by plotting the empirical cumulative distribution of the placebo effects for the math test scores (specifications from [3] to [6] of Table 1.3). The vertical lines indicate the estimates reported in Table 1.3. Overall, the obtained p-values from the permutation tests in panels B, C, and D confirm that the policy change led to unusually high test score gains (the corresponding p-values are 0.071, 0.022, and 0.021, respectively), even though these p-values are larger than the p-values using cluster-robust standard errors. One exception is panel A, which includes all small districts. The p-value of the treatment effect is 0.12, and the large treatment effect of Guilford is not rarely large when I use this distribution. However, the empirical distribution obtained from using all districts would not be the distribution of DDD estimates in Guilford. The placebo estimates from small districts are unusually large in absolute terms, which widen the empirical distribution of the placebo effects. This effect is confirmed in Figure 1.4, as the distributions of the placebo effects narrow as I drop the small districts in panel B and the midsized districts in panel C.

1.4.2.2 Winston-Salem

Figure 1.11 shows the location of donor districts underlying the Synthetic Winston-Salem with a preferred predictor set.[10] The figure shows that many of the chosen districts are not contiguous to Winston-Salem, which may suggest that the conventional procedure to select comparison groups based on geographical proximity would not be appropriate to capture pretrends of outcome variables.

[10] For interested readers, I report the combination of districts and weights underlying the Synthetic Winston-Salem for math and reading in Table 1.9.

Figure 1.5 plots the average math and reading test scores for Winston-Salem and Synthetic Winston-Salem from 1997 to 2011. The vertical line indicates 2008, the year when Winston-Salem decided to receive VA information. Panel A shows little evidence that providing VA information increases student math achievement in Winston-Salem. The two time series closely match each other in the preperiod, and this pattern generally persists during the postperiod. To obtain a sense of the significance of the treatment effect, I conduct the placebo exercise as if the treatment occurred for each of the 60 donor districts. Panel B shows the difference in the math test scores between the "treated" district and its synthetic control. I highlight the effects in the actual treated district, Winston-Salem, in black, while the rest of the placebos are plotted in gray. The placebo exercises indicate that the actual test score differences in the postperiod for Winston-Salem are very small relative to the differences for donor districts. Panel C shows the average reading test scores from 1997 to 2011 for both Winston-Salem and Synthetic Winston-Salem. Similar to the average math test scores, the average test scores of Synthetic Winston-Salem closely follows those of Winston-Salem in the preperiod, and there is little difference between the two in the post treatment period. When compared to the placebo effects from the donor districts depicted in panel D, again, the average effects for Winston-Salem are small.

Since the test score differences of the donor districts in the prepolicy period are relatively large compared to those of Winston-Salem, the large test score differences for the donor districts in the post period would not be informative to evaluate the significance of the treatment effect for Winston-Salem. That is, the large postperiod gap between the "treated" district and its synthetic control among donor districts would be spuriously created by lack of fit in the prepolicy period. For this reason, in Figure 1.12, I provide the different versions of this figure, and for each version, I include the placebo districts with a certain level of pretreatment RMSPE cutoffs. This figure shows that the treatment effects in Winston-Salem remain relatively small compared to the effects in the donor districts regardless of which pretreatment RMSPE cutoff is used.

In Table 1.4, I report the DID estimates for Winston-Salem using four different predictor sets and its p-value calculated from the placebo exercise. I obtain two-tailed p-values of the estimate by calculating the proportion of the absolute value of the estimates from the donor districts that are larger than the absolute value of the DID estimates in Table 1.4. I also report the DID estimates from the DID model with 15 large comparison districts in column (5) to compare them to DID estimates from the synthetic control methods. Table 1.4 provides further evidence that adopting teacher VA scores does not influence student achievement. The average achievement effects for math from my preferred specification in column (1) is almost zero, -0.005 SD, and insignificant, and the estimate for reading, while positive, is small and insignificant. It is interesting that the estimate of math from the DID model in column (5) is positive and statistically significant, but the small positive effect disappears once I correct the pretrends.

Column (2) through column (4) demonstrate the estimates with different choices of predictors. The small achievement effects for both math and reading are robust regardless of which predictor sets are used to calculate the optimal weights, which indicates that the zero effects are not driven by the choice of predictors. Overall, I show little evidence that providing VA information improves student achievement in Winston-Salem although I acknowledge that the empirical strategy I used here has a limited ability to detect small treatment effects.

1.5 Which Teachers are Responding More to Value-Added information?

There are several possible reasons for the lack of observed achievement effects in Winston-Salem, whereas the policy changes in Guilford increase student achievement. One explanation is that high-VA teachers in Winston-Salem may reduce their effort once they learn about their productivity, and consequently, their influence on student achievement gains decreases. This effect would generate small treatment gains even if the productivity of teachers with low-VA scores in Winston-Salem increases when the VA score is available. However, it is also possible that teachers in Winston-Salem do not respond to the policy change regardless of their initial VA level. I thus estimate the extent to which providing VA information changes the teachers' performance by exploring the heterogeneous impacts based on the teachers' initial productivity level. In Section 5.1, I discuss the analysis sample and explain the empirical methods. Section 5.2 provides the estimate results, and I check the robustness of the estimates in Section 5.3.

1.5.1 Analysis Sample and Empirical Strategy

To determine whether providing VA information has differential impacts on teacher performance, I track the achievement gains of students taught by a set of teachers in the treated districts before and after the policy change. I limit my analysis to two years prior to the year when teachers first accessed the performance information to two years after the policy change for three reasons.[11] First, to avoid conditioning teacher VA scores that could have been influenced by the policy change, the VA measure should be estimated using an out-of-sample period. Second, to minimize the mean reversion problem, the "out-of-sample" should not use the last years prior to the VA adoption. This is because VA score is measured with error, certain teachers who are identified as high VA teachers may have good students by chance and they are less likely to have good students in the subsequent year. Hence, if I were to identify highly effective or less-effective teachers using the last two years prior to the adoption of the VA information, I may find a spurious relationship between the adoption and the teachers' impact on student achievement, even if providing the VA information has no causal impact on student achievement. Finally, I examine only student achievement gains two years after the policy change to minimize teacher attrition from my sample.

While numerous specifications can be used to estimate teacher effects, I use the following lagged achievement model.[12]

[11] For Winston-Salem, I include three years prior to the year when teachers received the VA information to better evaluate the pretrends. This is motivated by evidence that highly effective and less effective teachers in Winston-Salem responded differently when they knew they would be evaluated using the end-of-test score. However, including only two years prior to the policy change does not change any of the results.

[12] Other than the lagged achievement model, there are numerous approaches that can be used to estimate teacher effects. For example, the average residual approach is widely used in several papers including Horvath (2015), Chetty et al. (2014a), and Kane et al. (2013). However, Guarino et al. (2015) argue that a proper strategy for teacher effects largely depends on the mechanism of student-teacher assignments. Their simulation evidence shows that the lagged achievement model is more robust than the other approaches if students are sorted into teachers based on their prior test scores. Since Horvath (2015) reports that more than 30 percent of elementary schools in North Carolina systematically sort students to teachers based on lagged test scores, I prefer to use the model that includes lagged test scores.

$$y_{it} = Y_{ij-1}\beta_0 + X_{it}\beta_1 + C_{ijt}\beta_2 + S_{ist}\beta_3 + \mu_t + \mu_g + \mu_j + \varepsilon_{it} \tag{1.5}$$

where, y_{ijt}, represents student i's math or reading test score in teacher j's class in the given year t. The vector of lagged math and reading scores, Y_{ijt-1}, enter the right-hand side, and, X_{it}, are the same student-level covariates that were used in the previous section. I also include a vector of class-level controls, C_{ijt}, such as the class means of prior-year test scores in math and reading, and class size. S_{ist} denotes a vector of the school-level controls, including school-grade means of prior-year test scores in math and reading, the percent white, percent black, and the percentage of gifted students. The term μ_t is a set of year effects to control for the year-specific shocks, μ_g is a grade fixed effect, and μ_j is a teacher fixed effect. I do not include a school fixed effect here because I want to calculate the teacher VA measures that are comparable across schools and grades; instead, I use a set of school-level controls to capture the school effects.[13]

I estimate teacher fixed effects using the student data in grades 3-5 from 1995 through 1998 for Guilford and from 2002 through 2005 for Winston-Salem. The teacher VA estimates, μ_j, are normalized, and then linked to teachers in the 1999-2003 data for Guilford and 2006-2011 data for Winston-Salem.[14] I further limit the sample by dropping teachers who are in the sample for only a single year, acknowledging that the estimated teacher fixed effects are noisy measures of true teacher effectiveness with a small number of student observations.

In Table 1.5, I contrast descriptive characteristics of teachers who had the estimated VA measures and teachers who did not. On average, teachers in my sample are more experienced than teachers who do not have the VA measure. This arrangement is expected because teachers with VA measures began their careers at least before the sample period, while new teachers who were hired in the sample period do not have the VA measure. These differences may limit general-

[13] Using within-school variation to identify teacher effects would attribute all the test score differences across schools to the school fixed effects even if the test score differences are due to teachers. Nonetheless, in Table 1.10, I reports the main results with alternative VA specifications.

[14] The VA estimator from the out-of-sample period is an unbiased estimator to predict teachers' impact on student achievement for the given year only if the given year is close to the out-of-sample period because the test scores from more recent classes are more precise predictors of current teacher quality (Chetty et al., 2014a). The VA measure from the out-of-sample period in my context (at least two years before the intervention) is likely a noisy measure of the teacher quality for the year when teachers first received their performance information.

izing to less-experienced teachers. However, as I can track the majority of teachers in Guilford and Winston-Salem (approximately 55 percent and 61 percent of teachers, respectively), the findings from this analytic sample are helpful to understand the mechanisms of the mean achievement effects that I documented above.

To examine whether the less-effective teachers were more responsive to the performance information, I first track the impact of one SD higher scoring teachers on student achievement gains over time using a value-added model (VAM) that controls for student characteristics, classroom characteristics, and school characteristics as follows:

$$y_{it} = \alpha + \sum_{k=-2}^{2} \beta_k year_{T+k} \times VA_j + \sum_{k=-1}^{2} year_{T+k} + Y_{it-1}\gamma_1 + X_{it}\gamma_2 + C_{it}\gamma_3 + S_{ist}\gamma_4 + \theta_s + \varepsilon_{it} \tag{1.6}$$

The variable $year_{T+k}$ is a set of year dummy variables equal to one if year t is equal to $year_{T+k}$, where $year_T$ indicates the first year when teachers were actually receiving the VA information. The key explanatory variable is VA_j which is the estimated teacher VA score using the out-of-sample period. The variable θ_s is a school fixed effect and ε_{it} is an idiosyncratic error term. The dependent variable and other student-, class-, school level controls remained the same as before. The parameter of interest is β_k, which shows the benefits of having one SD higher VA teachers on student achievement gains in a given year.

The inclusion of school fixed effects is motivated by the finding in Bates (2017) that providing VA information to both districts influences the teacher sorting within a district. It thus is likely that the teacher VA scores are endogenous to unobserved school quality. If high VA teachers move to a higher performing school after the policy change, comparing high- and low-VA teachers across schools would attribute the test score difference to teachers even if the impacts of high- and low-VA teachers are similar after the policy change. Hence, I prefer to include school fixed effects and a set of time-varying school-level controls, which allows me to identify the impact of one-standard-deviation-higher VA teachers on student achievement gains using the within-school variation.[15]

The DID model in Eq.(6), however, may fail to isolate the causal impact of providing teacher VA information if other statewide policies had differential effects on high- and low-VA teachers. For example, the North Carolina Bonus Program offered \$1,800 when middle and high school teachers in mathematics, science, or special education agreed to teach in high-poverty or low-achieving schools from 2002 to 2004, which might have deferentially affected high- and low-VA teachers in Guilford. In addition, when Winston-Salem adopted VA information, many districts, including Winston-Salem, implemented strategic staffing similar to the North Carolina Bonus Program.

To evaluate how much the reduced performance gaps in the treated districts was attributed to the adopting VA measures, I estimated the DDD model that compares the difference in the change in teacher performance between highly effective and less effective teachers in the treated district to differences in other districts as follows:

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 VA_j + \beta_2 TD_d \times VA_j + \beta_3 VA_j \times I(year_t = T-1) + \beta_4 VA_j \times I(year_t \geq T) \\ &+ \sum_{k=-2, k\neq -1}^{k=2} \delta_k year_{T+k} \times TD_d \times VA_j + Y_{it-1}\gamma_1 + X_{it}\gamma_2 + C_{it}\gamma_3 + S_{ist}\gamma_4 + \delta_d \times \delta_t + \varepsilon_{it} \quad (1.7) \end{aligned}$$

The model includes an indicator variable, $I(year_t = T-1)$, which is equal to 1 when the year t is the adopting year, and $I(year_t \geq T)$, which is equal to 1 if the year t is equal to the years when teachers were actually receiving the VA information. This parameterization is motivated to distinguish the adopting year from prior years, because highly effective and less effective teachers may respond differently when teachers know they will be evaluated using the-end-of-test score. The parameter β_4 picks up the impact of other policies that are common across 15 large districts, and the set of parameters (δ_k) from the three-way interaction terms demonstrate how the impact of one-standard-deviation-higher VA teachers on student achievement gains were evolved differently in the treated

[15] In Figure 1.13, I also show the estimates using between school variation as well as within-teacher variation. The results are similar regardless of the model specifications.

districts.[16] All other notations are the same as before.

Eq.(6) and (7) are identified under the assumption that the degree of student and teacher sorting within a school conditional on the observable student characteristics was not affected when teacher VA data was provided. As I show below, I find some evidence that positive student and teacher matching would be strengthened when VA information was available in Guilford, but the bias from positive sorting would operate in the opposite direction of the results. My preferred estimates are thus most likely a lower-bound on the true effects. One may also be concerned that less-effective teachers may be more likely to attrite from my sample after the policy change because they are more likely to stop teaching in the treated district or switch to nontested grade levels when the VA score is available.[17] However, I do not find any significant relationship between the timing of attrition and teacher VA scores. I will return to this issue in more detail in Section 6.2.

1.5.2 Results

Figure 1.6 show the estimated coefficients of the interaction terms between the year and the teacher VA measure in Eq.(6) from 1999 to 2003 for Guilford and 2006 to 2011 for Winston-Salem. The vertical lines in this figure show the year when the teachers first received the information. Each of the points in all panels represents the impact of teachers with one SD higher VA scores than the mean on student achievement gains, and all 95 percent confidence intervals use standard errors clustered at the school level. The year 2000 in panels A and B represent how teachers scoring one SD higher affected student achievement gains when they knew that the EOG test scores in 2000 would be used for the VA score, and the years from 2001 to 2003 show how the effects change when teachers actually receive the VA information based on prior performance. Panel A shows that the impact of having teachers with one SD higher VA measure on student achievement gains falls from 0.251 to 0.170 once teachers receive their VA information, and I can reject the null hypothesis

[16] Since these coefficients are normalized to one year prior to the year when teachers first received the VA information, the $T-1$ year triple interaction term ($year_{T-1} \times TD_d \times VA_j$) is the omitted variable. However, for Winston-Salem, the $T-2$ year triple interaction term ($year_{T-2} \times TD_d \times VA_j$) is the omitted variable, because the evidence indicates that high- and less- effective teachers in Winston-Salem responded differently in $T-1$ year.

[17] Chingos and West (2011) reported that high-VA teachers are less likely to be assigned to a low-stakes teaching position.

that the two estimates are equal under the 5 percent significance level. In addition to the change seen in the first year of providing VA information, the estimates in years 2002 and 2003 indicate that the causal effects may grow and persist for at least three years. However, for reading in panel B, I find little evidence that providing VA information reduces the productivity gaps among reading teachers. The estimates decline slightly from 0.114 to 0.087 in 2001 and to 0.075 in 2002, but the estimate returns to the original level in 2003.[18]

Panels C and D shows the analogous results for Winston-Salem. For both math and reading, I find that providing teacher VA scores decreases the impact of one-standard-deviation-higher VA teachers on student achievement gains. However, the figure indicates some evidence of anticipatory treatment effects, which is in contrast with the near-zero anticipatory estimated effects for Guilford. Specifically, the policy change influences the teacher quality distribution when teachers first receive the VA scores in Guilford, whereas the performance gap began to decline one year prior to the receiving year in Winston-Salem. One possible explanation is that other state-level policies that had differential effects on high- and low-VA teachers confounded the treatment effect. In Figure 1.14, I display the estimated coefficients of the triple interaction terms in Eq.(7). Consistently with the results in Figure 1.6, there is a clear decrease in the performance gap between high- and low-VA teachers after the year 2000 in Guilford. Interestingly, the performance gap in Winston-Salem still began to fall off the year when the district decided to adopt VA reports, making it unlikely that other state-level factors caused the different patterns between the two districts.

Although the above results indicate that providing VA scores to teachers and school administrators in both districts corresponded to a squeezing of the teacher quality distributions, the estimates do not provide useful information regarding which parts of the distribution this compression occurs. Thus, I use the teacher VA quartile rank instead of the normalized VA score in Eq.(6). Also, instead of using year indicators that are interacted with teacher quartile ranks, I interact a postpolicy period indicator with teacher quartile ranks to increase statistical precision. Figure 1.7 presents

[18] Figure also suggests that the out-of-sample approach used to construct the VA measure minimizes mean reversion. If regression to the mean were still to occur, then the impact of one-standard-deviation-higher than the mean VA teachers on student achievement gains would steadily decrease even before the policy change. However, for both math and reading, each of the points is quite stable until the year when teachers first receive the VA information.

point estimates and 95% upper and lower bounds from the regression equation with the quartile rank. The coefficients in the given quartile represent how much a teacher in the indicated quartile increased student achievement gains between pre- and post-policy periods. The figure suggests that the impacts of providing VA measures on the teacher quality distribution are qualitatively similar in both math and reading in both districts; the performance of less-effective teachers improves after the VA adoption while the performance of highly effective teachers remains unchanged or declined.

Panel A shows that the compression of math teacher quality in Guilford was mostly driven by an increase in productivity among less-effective teachers, while the performance of highly effective teachers remained unchanged. I find that the impact of the bottom-two-quartile teachers on student achievement gains increased significantly by approximately 0.145 and 0.139 SD after the policy intervention. The benefit of having top quartile teachers, however, barely changed (by 0.012 SD). On the other hand, in panel C, the compression of math teacher quality in Winston-Salem was driven by the moderate declines in performance among top quartile teachers and by small improvements among those in the bottom two quartiles. The impact of top quartile teachers on student achievement gains declined by 0.078 SD after the VA measure was introduced, while the impact of the bottom two quartile teachers increased slightly, approximately 0.023 and 0.033 SD respectively. The panel C provides a compelling explanation of why I find the zero achievement effects of VA information in Winston-Salem. The improved performance of less effective math teachers is canceled out by the decreased performance of highly effective teachers. Finally, panels B and D show the estimated results for reading teachers. Although most estimates in both panels are small and are not statistically significantly different from zero, the patterns of estimates are qualitatively similar to the results for math teachers. The point estimates for bottom two quartiles are small and positive, and the 4th quartiles, while small, are negative.

These results show that low-VA teachers in Guilford had high test score gains when the district adopted VA information, while low-VA teachers in Winston-Salem had small test score gains after the policy change. I argue that the results should not be interpreted as indicating that the low-VA teachers in the two districts responded to the VA information equally, but teachers in Winston-Salem may not know how to improve student achievement. For example, if low-VA teachers in Winston-Salem are disproportionately inexperienced teachers whereas low-VA teachers in Guilford are mostly experienced teachers, then the teachers in Winston-Salem may not know how to improve student achievement. However, in Figure 1.15, I show that conditioning on a rich set of teacher characteristics does not substantially change the estimates. Thus, it is not reasonable to think that low-VA teachers in Guilford knew how to improve student achievement but low-VA teachers in Winston-Salem, who are observationally identical to their counterparts, did not.

1.5.3 Robustness Checks

To evaluate whether the baseline results are robust to how I construct teacher VA measures, Table 1.10 compares my baseline estimates for both districts (reported in columns [1] and [4], respectively) to estimates with alternative VA measures. To facilitate the comparison between alternative measures, instead of reporting a set of the estimated coefficients of triple interaction terms (δ_k) in Eq.(7), I estimate the more parametric model that the postpolicy period indicator is interacted with the two-way interaction term ($TD_d \times VA_j$).[19] The triple interaction term shows how the impact of one-standard-deviation-higher VA teachers on student achievement gains changed in the treated districts when the districts adopted the VA report relative to the that of changes in the control districts.

Columns (2) and (5) of the table include school fixed effects in the baseline specification for estimating teacher VA measures. Since adding school fixed effects in the VA specification attributes all the test score differences across schools to school fixed effects other than teachers, this VA measure may underestimate the variation in teacher quality especially when highly effective teachers

[19] Specifically, I estimate the following model.

$$\begin{aligned} y_{it} = {} & \beta_0+\beta_1 VA_j+\beta_2 TD_d\times VA_j+\beta_3 VA_j\times I(year_t{=}T-1)+\beta_4 VA_j\times I(year_t\geq T)+ \\ & \delta_1 TD_d\times VA_j\times I(year_t{=}T-1)+\delta_2 TD_d\times VA_j\times I(year_t\geq T)+y_{it-1}\gamma_1+X_{it}\gamma_2+C_{it}\gamma_3+S_{ist}\gamma_4+\delta_d\times\delta_t+\varepsilon_{it} \end{aligned}$$

The estimated parameter reported in Tables 1.10 and 1.11 is δ_2. All notations are the same as in Eq.(7).

are sorted into higher performing schools (Hanushek et al., 2005). When I use this VA measure to estimate the DDD model, all estimates in columns (2) and (5) become attenuated because I am actually adding measurement error. The estimates for math teachers are still qualitatively similar to the baseline; however, the evidence indicates that the baseline estimates for reading are sensitive to the VA specification. In columns (3) and (6), I use the Empirical Bayes (EB) method to shrink the noisy VA measure with the small number of students toward the sample mean to yield efficient VA estimates.[20] Whether I use EB estimates or the estimated teacher fixed effects, the results are qualitatively the same. Finally, when I construct the VA measure, the out-of-samples for both districts include the same number of years of student data for comparison purposes. In column (7), I use all available years from 1995 to 2005 to construct the VA measure for Winston-Salem to better evaluate how the number of years of student data used affects my results. I use the EB method instead of the lagged achievement model since the number of students across teachers varies more as the number of years used increases. The estimates of both math and reading in column (7) are slightly larger than the estimates in column (6). This finding may indicate the existence of measurement errors in my VA measure, as I use a limited period of data. The measurement errors, however, would attenuate the estimates, which means that my baseline estimates are most likely a lower-bound on the true effect.

To evaluate how teacher attrition affects my results, in Table 1.11 of the Appendix A, I compares the baseline estimates for both districts (again, reported in columns [1] and [4]) to the estimates from the sample that include teachers present in all years of my analysis sample (reported in column [2] and column [5]). I find little evidence to suggest that my findings for math teachers are driven by attriters. The estimates for reading teachers, however, shows that the baseline estimates are sensitive to including attriters. Finally, in columns (3) and (6), I report the estimates from the DID model that compares high- and low-VA teachers within the treated districts to quantify how much the reduced performance gaps reported from the DID model were attributed to other factors that were common across districts. Comparing the point estimates from the DDD and DID model

[20] See Kane and Staiger (2008) who outline the procedure to compute the EB estimates.

suggests that approximately 36 percent (38 percent) of the reduction reported in the DID model for math (reading) can be explained by other factors that are common across the districts.[21]

1.6 Other Potential Mechanisms

In this section, I examine other potential mechanisms for the overall achievement effects that I document for Guilford other than teachers' responses to VA information because principals may use VA information strategically to improve the average test scores of schools. In Section 6.1, I examine whether highly productive teachers had more students in their classes and were reassigned to students with higher underlying test score growth than their counterparts in other districts. Section 6.2 explores whether teachers with low productivity are more likely to leave the treated district or be reassigned to low-stakes teaching positions when the district adopts the VA information.

1.6.1 Do principals strategically match teachers and students?

Prior studies report that effective teachers are more likely to have larger classes (Barrett and Toma, 2013), and that many schools systematically match students and teachers based on students' prior test scores (Horvath, 2015; Clotfelter et al., 2006). It is thus possible that less-effective teachers have fewer students in their classes than high-scoring teachers and that principals match students with higher achievement gains to highly effective teachers if principals use VA information to maximize school-level performance.

To confirm this mechanism in my context, I first collapse student-year observations into teacher-year observations and use the average previous math and reading test scores of classes, the number of students, and the average parental education for each teacher as outcome variables. I investigate whether these outcome variables are more associated with VA measures when districts adopted the VA measure by using the DID model and DDD models. These regression models are similar to

[21] Panel A of Table 1.12 shows the corresponding p-values of the baseline estimates in Table 1.11 using the non-parametric permutation test. The panel A clearly shows that the statistically significance estimates reported for both districts are robust to how I compute standard errors.

the models discussed above, but I add additional teacher characteristic variables including experience, number of years of schooling, gender, and race to better isolate the association between the effectiveness of teachers and classroom composition.[22] [23]

Table 1.6 displays the estimated results of the relationship between various classroom compositional outcomes and teacher VA measures. Columns (1) to (4) use the Guilford sample, and columns (5) to (8) use the Winston sample. I present estimates from the DID model in odd columns and estimates from the DDD model in even columns. Panel A shows how teachers with one SD higher VA measure than the mean in a given subject are differentially associated with class size when the VA information was provided to treated districts. To consist with positive achievement effects in Guilford, either high VA teachers in Guilford had more students after the VA adoption or low scoring teachers had fewer students. The panel A shows little evidence that the changes in class size with teacher VA scores are likely to be a primary channel. The estimates for both districts are rather negative, though no estimates are statistically significant at the 10 percent level.

Panels B to D display the estimated results that show whether systematic teacher and student matching emerges with teacher effectiveness when the VA information is provided to principals. For Winston-Salem, the estimates show little evidence that principals strategically used the VA information to match teachers and students; all estimates are small and not significantly different from zero. The estimated coefficients for Guilford, however, provide some evidence of positive student and teacher sorting when principals received the VA information. The estimates of the average parents' education are positive and statistically significantly different from zero regardless of the choice of models and subjects, although these estimates are economically small. For example, math teachers with one standard deviation higher VA had students whose parental education were 0.179 years higher (e.g., column [2] in panel D) than that of the parents of students of average teachers when the VA information was provided to principals. Furthermore, I find that the

[22] Since EVAAS teacher reports are available for principals and teachers when the new academic year begins (late September or early October), the first year that principals can use this information for teacher-student matching is one year after when teachers first receive the VA information. I thus redefine the postpolicy period and sample accordingly.

[23] I did not use the average parental education as an outcome variable for analyzing Winston-Salem, because the information regarding parental education is only available up to 2007.

statistical significance of the estimates are sensitive to how I compute p-values. The corresponding p-values using the permutation test are reported in panel B of Table 1.12 in Appendix A. These p-values are larger (0.125 and 0.250 respectively) than the p-values using cluster-robust standard errors, and the estimated results are not statistically significant at the 10 percent level.

1.6.2 Are less-effective teachers more likely to exit the treated districts?

Although the VA information was not used for high-stakes personnel decisions such as teacher evaluation or pay, adopting the VA information would influence teachers' decisions to leave the treated districts. Low-VA teachers may react negatively to this additional scrutiny and attempt to leave the treated districts entirely. Hence, if teachers with low VA scores are more likely separated from the treated districts and are then replaced by better quality teachers, receiving VA information indirectly influences student achievement. Indeed, Bates (2017) finds robust evidence that less-effective teachers who teach 4th to 8th grade students in Guilford are a full percentage point more likely to leave the district than less-effective teachers in the control districts when the VA information was adopted, whereas he finds little evidence for differential teachers' exiting out of the treated districts in Winston-Salem.

However, his findings may not be applicable to understanding the impact of teacher turnover on student achievement in my context because his analysis includes all math teachers from 4th to 8th grades in elementary and middle schools, while I am focusing on 4th and 5th grade teachers. Moreover, the estimates from Bates rather show the long-run mobility effects of providing VA information, while the proper teacher mobility effects needed to understand the potential mechanism in my context are short- or medium-run effects (from 1 years to 3 years after the VA adoption).[24] Finally, he does not consider exploring other types of teacher mobility, including quitting teaching or switching to non-tested grades. These types of mobility may have influenced student achievement in the treated districts as well if the relationships between teacher quality and these mobility changed when the treated districts adopted the VA data.

[24] For example, the post-treatment period of my analysis for Guilford covers 2001 to 2003, whereas the period for Bates spans from 2001 to 2011.

Table 1.7 examines any evidence of differential teacher movements across the distribution of teacher quality, as measured by pre-policy VA scores, when the VA information was provided. I examine various types of teacher mobility using the DID and DDD models that I discussed in previous section. In particular, using teacher-year-level data, I estimate a linear probability model with DID or DDD specifications to predict a binary outcome that equals 1 if teacher j initiates the given type of mobility in year t. I first examine whether less effective teachers were more likely to attrite from my sample when treated districts adopted the VA information. If the adopting VA information caused less-effective teachers to attrite the sample for some reasons in Guilford more than in Winston-Salem, this may explain the differential treatment effects between the two districts. However, panel A show little evidence that there was significant relationship between the timing of attrition and teacher VA scores in both districts. The estimates for Guilford in columns (1) to (4) are small, while negative, and are statistically insignificant.

Since switching to non-tested grades may be more relevant to the effect of providing VA data than quitting or retiring teaching, especially for tenured teachers, panel B limits to include switching to non-tested grades as an outcome variable.[25] The point estimates in panel B show how the higher VA score in the treated districts was differentially associated with the probability of switching to non-tested grade in next year when these districts adopted the VA data. I find little evidence that adopting the VA report had a systematic effect on the possibility of less effective math teachers switching to nontested grades in both districts.

Finally, panel C shows analogous results in whether teachers exit from current districts and then move to other districts. I found little evidence that adopting VA information influences teacher exit from the treated districts differently according to teacher quality; none of the estimates are statistically significantly different from zero.[26] I thus rule out the explanation that low-rated teachers leaving the district more frequently or switching more often into nontested grades influenced the

[25] Approximately 47 percent (43 percent) of the attrition in my sample is related with quitting or retiring teaching (switching to non-tested grades). The remaining reasons for the attrition are: (1) switching to districts in North Carolina that are not considered in my sample, and (2) temporary leaving.

[26] Another potential reason why my finding is in contrast to the finding in Bates is that he uses all available years of data, i.e., regardless of whether the year of data is pre- or post-policy, to calculate teacher VA measure. I take a contrasting approach and use the out-of-sample to construct teacher VA measure.

average test scores in treated districts.[27]

1.7 Conclusion

Publishing teacher VA information to the public would be a useful tool for school districts because publicly-available VA scores may exert social and peer pressure to improve teacher performance. However, it is difficult to publish VA scores because teacher unions such as the American Federation of Teachers are outspoken in their opposition to the release of VA data. On the other hand, providing VA information to teachers and school administrators can be readily implemented without incurring potential social costs because the recent teacher evaluation reforms in school districts across the country require VA information as one component of the teacher evaluation system. Hence, it is important to examine whether providing VA information privately to teachers or principals would increase student performance.

This paper examines how providing teacher VA information to teachers and principals affects student academic performance. Since one district provides VA information to all potential employers within the district and the other district releases this information to the current employer only, examining two natural experiments allows us to understand whether providing the performance information in a more public way matters. Using the matched students and teacher data, which enables me to track the achievement of students whose teachers are eligible to receive VA information, I estimate the average achievement effects in each case separately. Across myriad specifications and variations in modeling choices, I show that adopting VA information raised student math achievement in Guilford, but I do not find any achievement effects in Winston-Salem. The effect sizes of providing VA information to teachers and principals that I find for Guilford are approximately 0.1 SD, which is roughly half of the effect of reduced class size found in the Tennessee STAR experiment (Krueger, 1999). However, given that the cost of providing VA in-

[27] Sartain and Steinberg (2016) find that the new evaluation information of teachers leads to a higher likelihood of teachers' out-of-district movement among nontenured teachers with low performance. Unfortu- nately, I cannot test this finding with my data because the out-of-sample approach that I take here excludes most nontenured teachers, and thus, the estimated results should be interpreted with caution.

formation is considerably lower than reducing class sizes because it only needs the existence of preexisting student data, providing VA information to schools would be a cost-effective policy for improving student achievement.

The overall achievement effects that I document for Guilford, however, do not necessarily suggest that Guilford teachers responded to the new information and improved their performance, as principals may lay off less-effective teachers or assign more students to highly effective teachers. Hence, I investigate whether providing VA information affects the teachers' impact on student achievement gains. Using a VA model that includes student-, classroom-, and school-level controls, I find evidence that the achievement gains of students taught by one SD higher VA math teachers decreased by 0.070 SD when teachers received their VA information in Guilford. The benefit of having math teachers who were one SD VA higher than the mean declined by 0.038 SD in Winston-Salem, which suggests that teachers' performance may not change much if the VA information is provided privately.

One caveat of my research is that we may not expect the same benefit that I find for Guilford in future regimes, for two reasons. First, it is important that principals value VA scores. Based on a Bayesian learning framework, teachers would not have any incentive to respond to VA data if principals do not use VA data to infer teachers' ability. A recent simulation study by Steinberg and Kraft (2017) argues that weights for VA scores are important in determining teachers' summative evaluation ratings. Their simulation evidence shows that the teacher proficiency rate increases from 45 percent to 85 percent if the weights for VA measures decrease from 91 percent to 46 percent. If districts calculate teachers' summative ratings with less weight given to VA scores and principals use these summative ratings to infer teachers' ability, then teachers would not have extrinsic motivation to improve their performance. Second, my findings may not have external validity for settings where VA information is formally used for high-stakes personnel decisions such as teacher promotions or tenure decisions. Hence, future work is needed to understand whether the same benefits can be realized when such performance information is provided to teachers and principals in a high-stakes environment.

APPENDICES

APPENDIX A TABLES

Table 1.1: Summary Statistics for Guilford, Winston-Salem, and the Rest of Districts

	Guilford (1997-2005)				Winston-Salem (1995-2011)			
	Guilford		Rest of NC		Winston-Salem		Rest of NC	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Unit of observations: Student-year								
Female	0.507	0.500	0.501	0.500	0.500	0.500	0.501	0.500
White	0.493	0.500	0.643	0.479	0.520	0.500	0.599	0.490
Black	0.408	0.491	0.274	0.446	0.336	0.472	0.280	0.449
Hispanic	0.030	0.171	0.039	0.192	0.098	0.297	0.061	0.240
Gifted	0.188	0.390	0.123	0.329	0.173	0.378	0.130	0.336
Special Edu.	0.113	0.317	0.087	0.281	0.092	0.290	0.085	0.279
Parent Edu (less than HS)	0.468	0.499	0.541	0.498	0.316	0.465	0.379	0.485
Math score	0.070	1.017	0.072	0.964	0.074	1.031	0.056	0.973
Reading score	0.057	1.007	0.062	0.964	0.050	1.023	0.044	0.973
Observations	96,414		1,569,084		120,089		3,115,906	
Unit of observations: Teacher-year								
0-3 years' experience	0.226	0.418	0.223	0.416	0.186	0.389	0.236	0.425
4-10 years' experience	0.261	0.439	0.261	0.439	0.270	0.444	0.278	0.448
11+ years' experience	0.513	0.500	0.515	0.500	0.543	0.498	0.486	0.500
White	0.747	0.435	0.850	0.357	0.796	0.403	0.853	0.355
Black	0.236	0.425	0.137	0.344	0.198	0.399	0.133	0.339
Advanced Degree	0.277	0.448	0.277	0.448	0.355	0.479	0.285	0.451
Observations	4,871		89,122		5,809		143,968	

Notes: The table shows the Summary statistics of certain key variables for Guilford and Winston-Salem Sample. The table compares means and SD of Guilford and the rest of districts in first four columns and compares that of Winston-Salem and the rest of districts in the last four columns.

Table 1.2: Average Pre- and Post-treatment RMSPE by Predictor Sets and Donor Districts.

RMSPE	(1)	(2)	(3)	(4)
Panel A. Math Test Score				
Pre-treatment	0.033	0.037	0.035	0.036
Post-treatment	0.105	0.112	0.110	0.107
Panel B. Reading Test Score				
Pre-treatment	0.029	0.034	0.033	0.032
Post-treatment	0.090	0.095	0.096	0.093
Predictors				
Annul outcomes	Y			
Biannual outcomes		Y		
Both biannual outcomes			Y	Y
Other controls				Y

Notes: The average RMSPE is the square root of the mean of all donors' MSPEs for both in pre- and post-treatment period. Predictor set 1 includes the set of average test scores in the pretreatment period in a given subject, and predictor set 2 includes the biannual average test outcomes in a given subject. Predictor set 3 uses both the biannual average math and reading scores, and predictor set 4 includes the biannual average math and reading scores with controls, including the pre-treatment means of percent black, percent white, percent female, percent gift, percent special education students, and student enrollment variables.

Table 1.3: The Effect of Providing VA Information on Student Achievement in Guilford

	DID strategy		DDD strategy			
	Across Districts	Within Guilford	All Districts	Compari-son1	Compari-son2	Compari-son3
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Math Score						
$TD_d \times Post_t$	0.108***	0.091***				
	(0.013)	(0.025)				
$TG_g \times TD_d$			0.099***	0.102***	0.097***	0.096***
$\times Post_t$			(0.026)	(0.026)	(0.026)	(0.027)
Panel B. Reading Score						
$TD_d \times Post_t$	0.014	0.006				
	(0.008)	(0.016)				
$TG_g \times TD_d$			0.012	0.014	0.010	0.013
$\times Post_t$			(0.016)	(0.016)	(0.017)	(0.017)
Comparison District	Y		Y	Y	Y	Y
Comparison Grade		Y	Y	Y	Y	Y
Observations	1,075,784	96,414	1,665,498	1,463,410	1,119,190	780,842

Notes: The table presents the baseline estimates from Eq. (1). In column (1) and (2), I use a simple DID model. Column (1) uses all school districts as the control group and column (2) uses non-treated grade in Guilford as the control group. For column (3) to (6), I report estimates from the DDD model with different comparison districts. Clustered Standard errors are shown in parentheses. Statistically significant at *** 1%, ** 5%, and *10%.

Table 1.4: The Effect of Providing VA Information on Student Achievement in Winston-Salem

	Synthetic Control Methods				DID
	(1)	(2)	(3)	(4)	(5)
Panel A. Math Score					
DID estimates	-0.005	0.015	0.005	-0.013	0.029*
P-value from two-tailed test	0.905	0.810	0.889	0.873	0.064
Panel B. Reading Score					
DID estimates	0.020	0.023	-0.009	-0.012	0.01
P-value from two-tailed test	0.823	0.790	0.920	0.920	0.312
Predictors					
Annul outcomes	Y				
Biannual outcomes		Y			
Both biannual outcomes			Y	Y	
Other controls				Y	

Notes: The table shows the DID estimate from Eq.(5) for Winston-Salem using four different predictor sets and its p-value calculated from placebo exercise. The two tailed p-values are obtained by calculating the proportion of the absolute value of estimates from donor districts that are larger than the absolute value of DID estimates. Predictor set 1 includes the set of average test scores in the pretreatment period in a given subject, and predictor set 2 includes the biannual average test outcomes in a given subject. Predictor set 3 uses both the biannual average math and reading scores, and predictor set 4 includes the biannual average math and reading scores with controls, including the pretreatment means of percent black, percent white, percent female, percent gift, percent special education students, and student enrollment variables.

Table 1.5: Teacher Characteristics

	Years experience	Graduate degree	Female	African-American	White
	(1)	(2)	(3)	(4)	(5)
Panel A. Teachers in Guilford					
Analysis sample (n=209)	16.202	0.321	0.938	0.756	0.225
Remainder of district (n=168)	6.762	0.250	0.869	0.756	0.232
Difference t-test p-value	0.000	0.134	0.022	0.999	0.868
Panel B. Teachers in Winston-Salem					
Analysis sample (n=174)	15.908	0.362	0.891	0.810	0.184
Remainder of district (n=111)	7.645	0.306	0.874	0.811	0.171
Difference t-test p-value	0.000	0.334	0.664	0.992	0.785

Notes: The table contrast descriptive characteristics of teachers who had the estimated VA measures and teachers who did not. P-values are obtained from the t-tests that compare the sample mean of the analysis sample and the remainder sample. The summary statistics shown in this table use data from the school year 1998-1999 for Guilford and 2005-2006 for Winston-Salem.

Table 1.6: Estimates of the Relationship between Classroom Composition and Teacher VA

	Guilford				Winston-Salem			
	Math VA		Reading VA		Math VA		Reading VA	
	DID	DDD	DID	DDD	DID	DDD	DID	DDD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Classroom Size								
$VA \times Post$	-0.395	0.086	-0.307	0.171	-0.087	0.231*	-0.196	0.303**
	(0.319)	(0.163)	(0.294)	(0.159)	(0.350)	(0.120)	(0.576)	(0.136)
$VA \times Post$		-0.534		-0.524		-0.303		-0.499
$\times Treat$		(0.360)		(0.339)		(0.368)		(0.587)
Panel B. Lagged average math score								
$VA \times Post$	0.030	-0.010	0.029	-0.008	-0.012	-0.020	-0.023	-0.020
	(0.040)	(0.014)	(0.032)	(0.017)	(0.038)	(0.013)	(0.064)	(0.014)
$VA \times Post$		0.040		0.035		-0.002		-0.012
$\times Treat$		(0.044)		(0.036)		(0.040)		(0.065)
Panel C. Lagged average reading score								
$VA \times Post$	0.036	-0.004	0.044	-0.012	-0.000	-0.019	-0.043	- 0.026**
	(0.032)	(0.013)	(0.030)	(0.015)	(0.035)	(0.012)	(0.051)	(0.013)
$VA \times Post$		0.039		0.056*		0.015		-0.026
$\times Treat$		(0.034)		(0.032)		(0.038)		(0.054)
Panel D. Parents' education								
$VA \times Post$	0.147**	-0.039	0.133*	-0.025				
	(0.059)	(0.035)	(0.073)	(0.037)				
$VA \times Post$		0.179***		0.145*				
$\times Treat$		(0.068)		(0.082)				
observations	847	6810	847	6810	811	7298	811	7298

Notes: The table report the estimates that show how teachers with one SD higher VA measure than the mean in a given subject are differentially associated with various measures of classroom composition when the VA information was provided in treated districts. Clustered Standard errors are shown in parentheses. The unit of observations is a teacher-year and statistically significant at *** 1%, ** 5%, and * 10%.

Table 1.7: Estimates of the Relationship between Teacher Turnover and Teacher VA

	Guilford				Winston-Salem			
	Math VA		Reading VA		Math VA		Reading VA	
	DID	DDD	DID	DDD	DID	DDD	DID	DDD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Exit Sample								
$VA \times Post$	-0.017	0.013	-0.030	0.006	0.006	-0.012	0.015	-0.014
	(0.033)	(0.013)	(0.028)	(0.014)	(0.028)	(0.013)	(0.037)	(0.013)
$VA \times Post$		-0.027		-0.030		0.015		0.028
$\times Treat$		(0.035)		(0.031)		(0.030)		(0.039)
Panel B. Switch into Non-tested Grade								
$VA \times Post$	-0.020	0.007	-0.020	-0.002	0.004	-0.002	-0.011	-0.007
	(0.019)	(0.009)	(0.019)	(0.010)	(0.022)	(0.010)	(0.030)	(0.010)
$VA \times Post$		-0.025		-0.016		0.006		-0.002
$\times Treat$		(0.021)		(0.021)		(0.023)		(0.030)
Panel C. Leave district								
$VA \times Post$	0.004	0.001	0.007	0.002	0.017*	0.003	0.016	-0.001
	(0.010)	(0.003)	(0.006)	(0.003)	(0.010)	(0.003)	(0.010)	(0.003)
$VA \times Post$		0.003		0.004		0.013		0.018
$\times Treat$		(0.009)		(0.007)		(0.010)		(0.011)
observations	1261	9968	1261	9968	862	7693	862	7693

Notes: The table report the estimates that show how one SD higher VA scores in treated districts are differentially associated with teacher turnover. Clustered Standard errors are shown in parentheses. The unit of observations is a teacher-year.

Table 1.8: Average of Key Demographic Variables for Guilford and Comparison Districts

	Guilford	All	Compar-	Compar-	Compar-	P-value		
		Districts	ison 1	ison 2	ison 3			
	(1)	(2)	(3)	(4)	(5)	(1)-(3)	(1)-(4)	(1)-(5)
% female	0.507	0.501	0.502	0.503	0.503	0.280	0.345	0.382
% white	0.493	0.643	0.643	0.645	0.615	0.000	0.000	0.000
% black	0.408	0.274	0.271	0.265	0.285	0.000	0.000	0.000
% hispanic	0.030	0.039	0.040	0.039	0.041	0.283	0.328	0.251
% gifted	0.188	0.123	0.125	0.130	0.148	0.000	0.000	0.000
% special education	0.113	0.087	0.086	0.085	0.083	0.109	0.102	0.0823
Parental education	0.468	0.541	0.534	0.516	0.484	0.001	0.001	0.266
(% less than high schools)								
Number of students	10915.61	3941.92	4440.18	5473.63	6870.90	0.000	0.000	0.000
# of Comparison Districts		109*	60	30	15			

Notes: The table reports the means of key variables of Guilford and a set of comparison districts. P-values are obtained from the t-tests that compare the mean difference of given characteristics between Guilford and the comparison districts. The number of students is defined by the number of student-year observations in each year-district cells. *Note that I exclude additional seven districts from any of comparison district sets as these districts do not have complete the BOG test scores.

Table 1.9: District Weights for Synthetic Winston-Salem

Districts	Weight
Panel A. Math Test Score	
Union	0.310
Sampson	0.267
Granville	0.120
Burke	0.087
Surry	0.043
Lincoln	0.038
McDowell	0.034
Rutherford	0.025
Randolph	0.010
Panel B. Reading Test Score	
Wilkes	0.217
Granville	0.215
Henderson	0.14
Johnston	0.135
Surry	0.087
Richmond	0.084
New Hanover	0.048
Chatham	0.047
Onslow	0.022

Notes: The table reports the combination of districts and weights chosen underlying the synthetic Winston-Salem with a preferred predictor set for both math and reading test scores.

Table 1.10: Robust Checks with Alternative VA Specifications

	Guilford			Winston-Salem			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Math Score							
VA × *Post* × *Treat*	-0.070***	-0.067***	-0.062***	-0.038**	-0.016	-0.032**	-0.042***
	(0.025)	(0.020)	(0.020)	(0.018)	(0.015)	(0.015)	(0.015)
Panel B. Reading Score							
VA × *Post*	-0.017	-0.002	-0.023*	-0.034*	0.008	-0.034**	-0.036**
	(0.025)	(0.013)	(0.013)	(0.020)	(0.014)	(0.016)	(0.017)
VA Specifications							
Baseline (BL)	Y			Y			
BL with school FE		Y			Y		
EB			Y			Y	
EB using all years							Y
student-year observations	143761	143761	143761	153113	153113	153113	153113

Notes: The table compares the baseline estimates to the estimates with alternative VA measures. The alternative VA specifications include the baseline VA specification with school fixed effects, Empirical Bayes (EB) method, and the EB method with using all available years. Clustered Standard errors are shown in parentheses. Statistically significant at *** 1%, ** 5%, and * 10%.

Table 1.11: P-values of Baseline Estimates Using Permutation Methods

	Guilford		Winston-Salem	
	Math	Reading	Math	Reading
Outcome variables	(1)	(2)	(3)	(4)
Panel A. Teacher Performance				
Test score	-0.070***	-0.017***	-0.038***	-0.034***
P-value from two-tailed test	0.000	0.000	0.000	0.000
Panel B. Classroom Composition				
Classroom size	-0.534	-0.524	-0.303	-0.499
P-value from two-tailed test	0.563	0.250	0.375	0.188
Lagged average math score	0.040	0.035	-0.002	-0.012
P-value from two-tailed test	0.250	0.563	0.875	0.875
Lagged average reading score	0.039	0.056	0.015	-0.026
P-value from two-tailed test	0.438	0.313	0.688	0.438
Parents' education	0.179	0.145		
P-value from two-tailed test	0.125	0.250		
Panel C. Teacher Turnover				
Leave district	0.003	0.004	0.013	0.018
P-value from two-tailed test	0.500	0.688	0.188	0.375
Switch into non-tested grade	-0.015	-0.041	0.005	0.030
P-value from two-tailed test	0.688	0.188	0.813	0.500

Notes: The table reports p-values for all triple difference terms in Table 1.6, Table 1.7, and Table 1.11 using the non-parametric permutation tests. The two-tailed p-value of the given estimate is obtained by calculating the proportion of the absolute value of estimates from placebo districts that are larger than the absolute value of the given estimate. Panel A through C correspond to Table 1.7 through 9. Statistically significant at *** 1%, ** 5%, and * 10%.

APPENDIX B FIGURES

Figure 1.1: Example of EVAAS Teacher Report for End-of-Grade Test

SAS ®EVAAS® Teacher Value-Added Report for 2006
Guilford County Schools

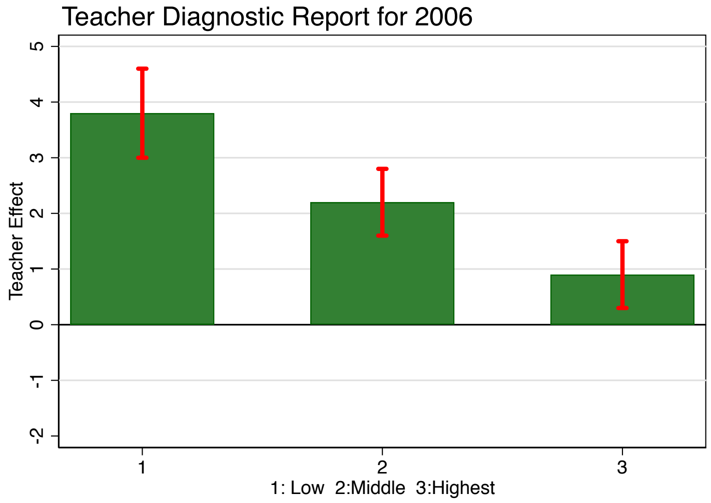
School:

Teacher:

Subject: End of Grade Math, Grade 5

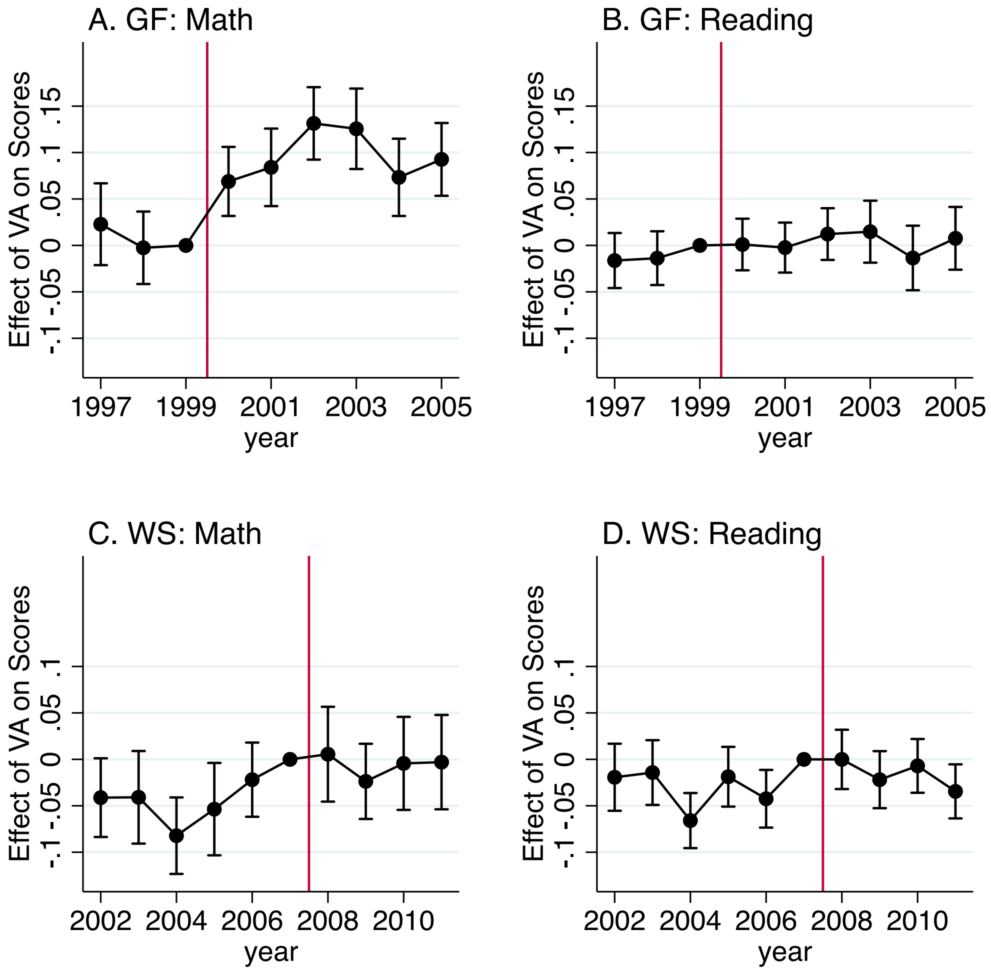
Year	N	Mean Student Score	Mean Score %tile	Mean Pred. Score	Pred. Score %tile	Teacher Effect	Effect Std Error	Teacher vs Comparison Avg
2006	21	349.9	25	351.9	32	-1.9	0.5	Below

Estimates are from multivariate, longitudinal analyses using all available test data for each student (up to 5 years). The analyses were completed via SAS®EVAAS® methodology and software, which is available through SAS Institute Inc. EVAAS, SAS, and all other SIS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. In the USA and other countries, ® indicates USA registration. Other brand and product names are trademarks of their respective companies.


Teacher Diagnostic Report for 2006

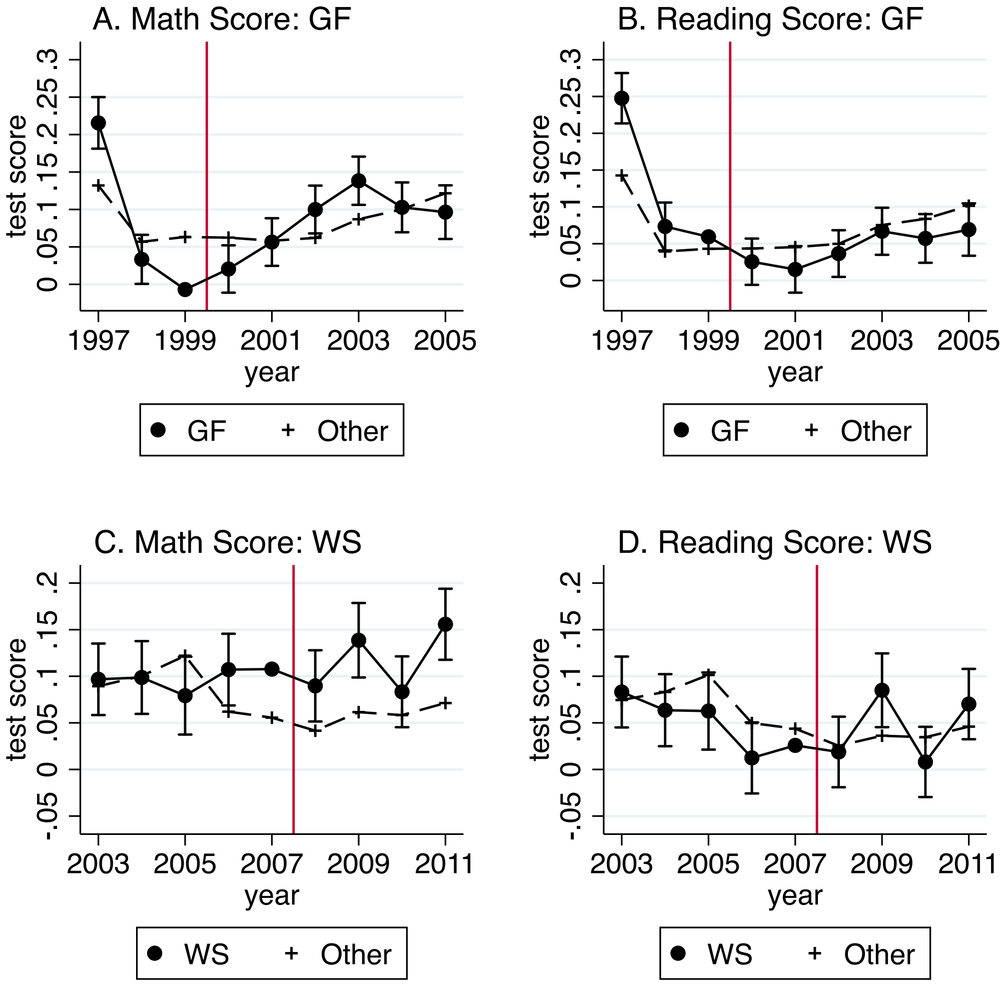
Source: Author's reproduction using a copy of a Guilford County Schools' Value-Added Report.

Figure 1.2: Effect of Providing VA Information on Test Scores: Using DID Model



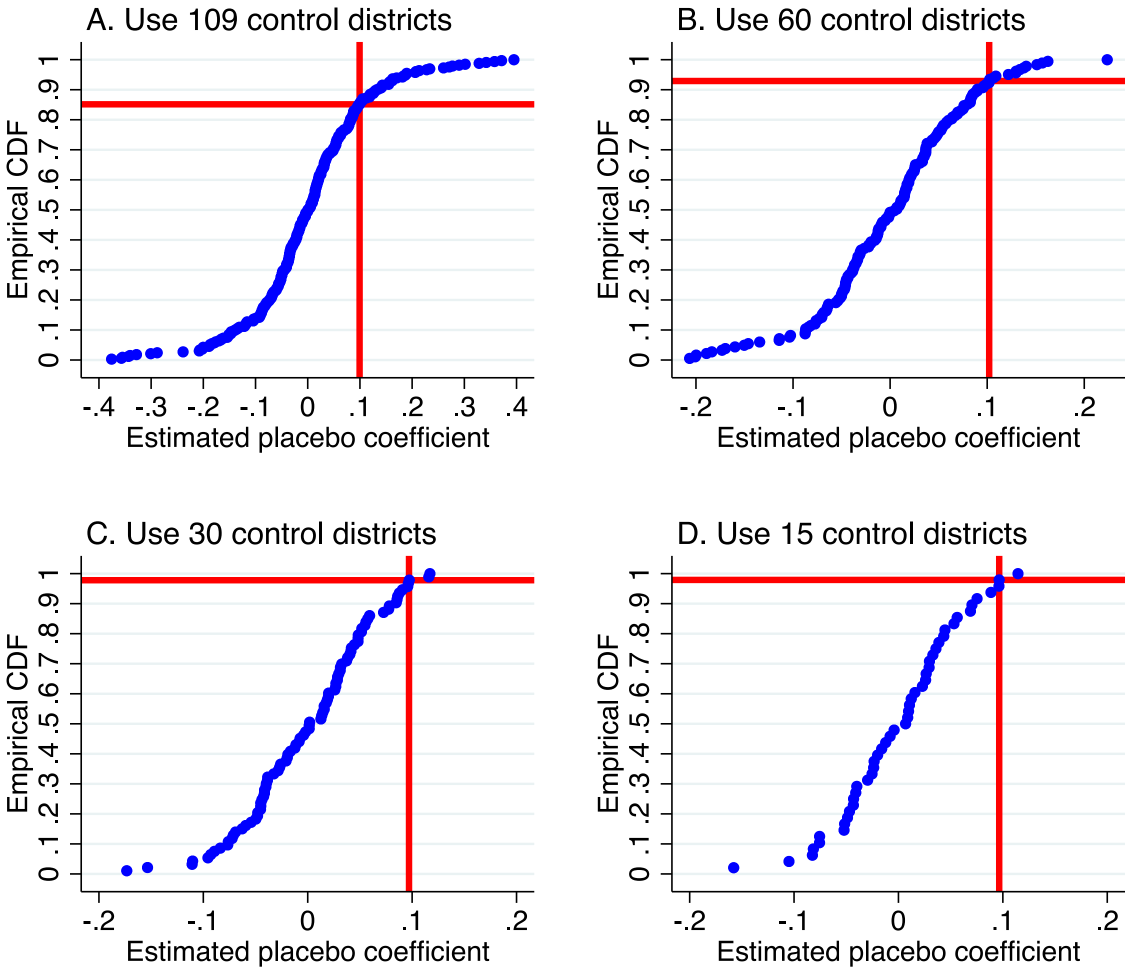
Notes: The figure shows estimates from the event study model that uses 15 large school districts as comparison districts. The vertical lines in both panels show one year prior to the year when treated districts decided to adopted VA measures. This model includes controls for lagged student test scores, race, 6 categories of exceptional status, and gift status. Each point in Figure 1.4 indicates the test score differences between treatment and control districts. All 95 percent confidence intervals use standard errors clustered at the school level.

Figure 1.3: Student Average Test Score by School Districts



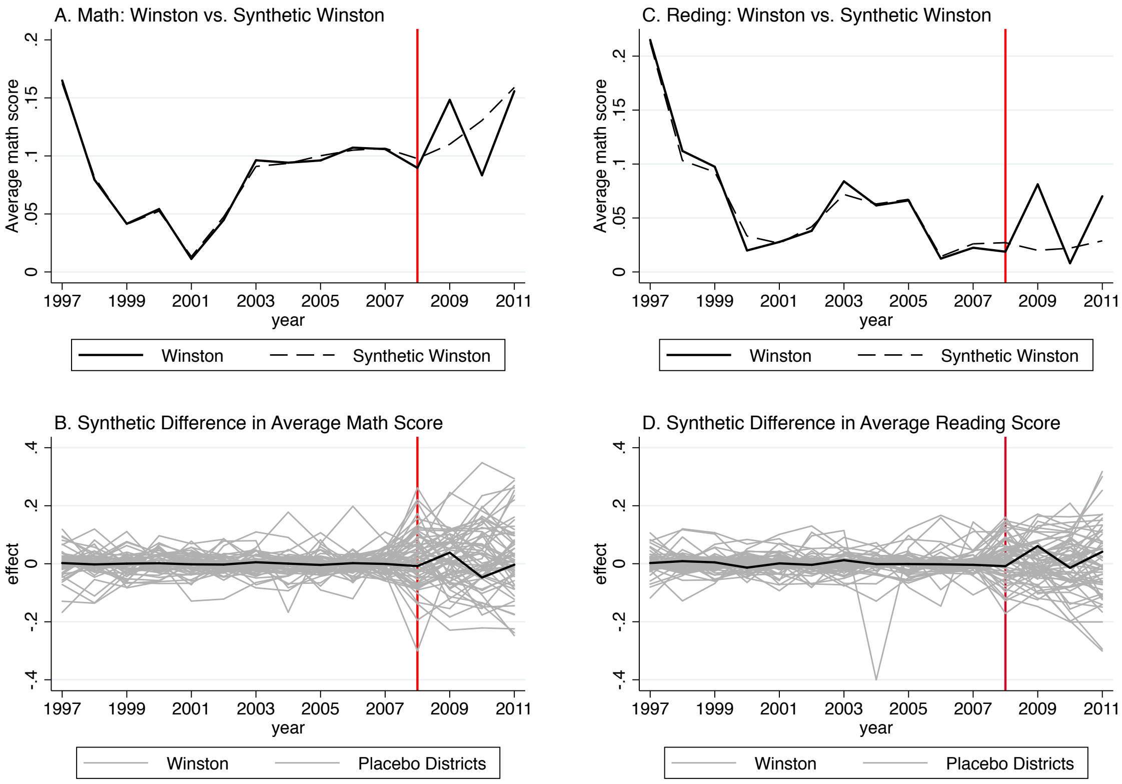
Notes: The figure shows the average math and reading score of 4th and 5th graders in Guilford, Winston-Salem, and the rest of districts. All 95 percent confidence intervals shows the upper and lower bound of mean test scores relative to the one year prior to the year when districts decided to adopt value-added measures.

Figure 1.4: Distribution of Placebo Estimates of Math Score by Number of School Districts



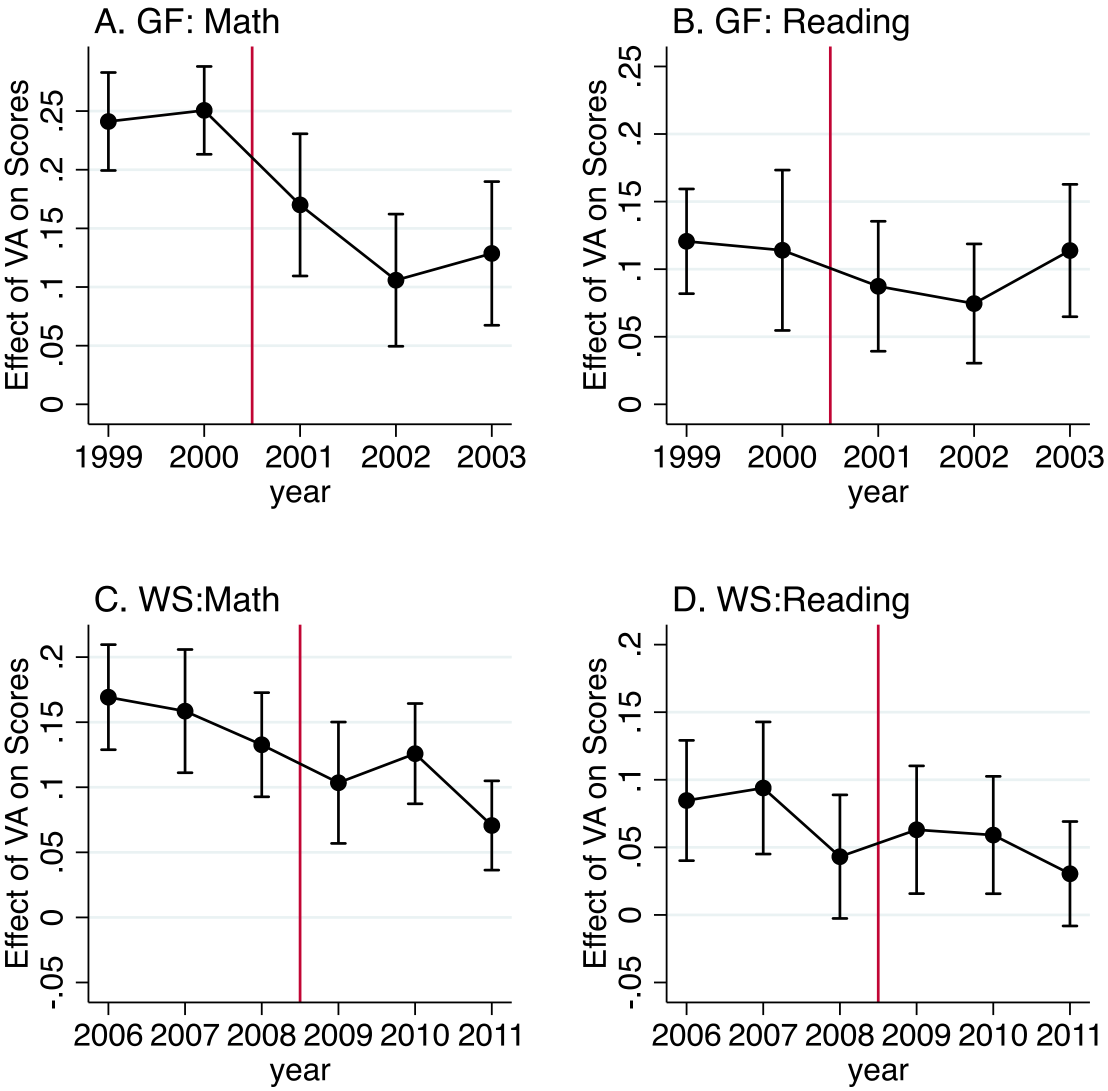
Notes: The figure shows the empirical cumulative distribution of placebo effects for math test scores. In each panel, I define a placebo triple consisting one district with two treated grades out of three. The vertical lines indicate the estimates reported in Table 1.3 and the horizontal lines show the corresponding p-values.

Figure 1.5: Average Math (Reading) Score in Winston-Salem, 1997-2011



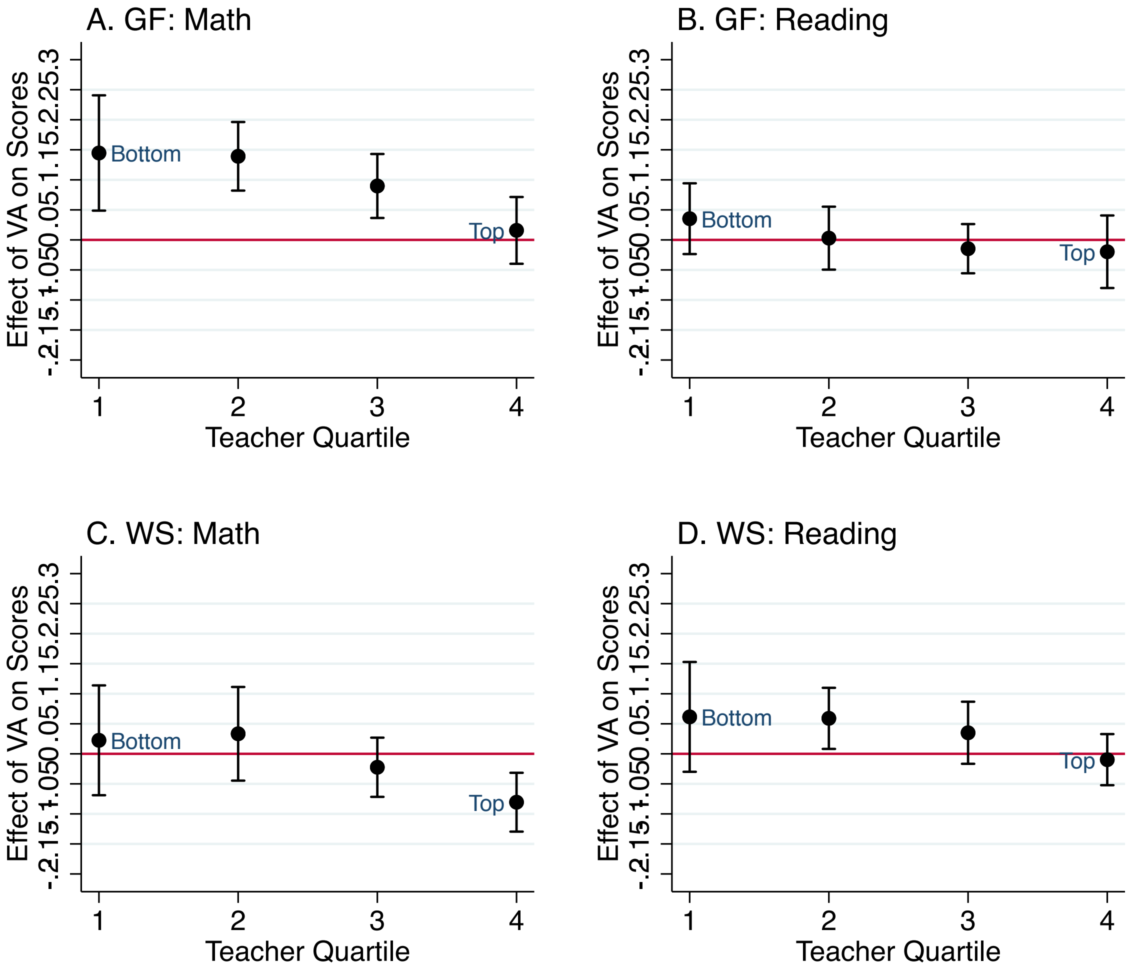
Notes: Panel A (panel C) plots the synthetic control estimates of average math (reading) scores for Winston-Salem from 1997 to 2011. The solid line plots the actual average math test score for 4th and 5th graders in Winston-Salem, while the dotted line plots the synthetic control estimate. The vertical dashed line indicates 2008, the year when Winston-Salem decided to adopt the value-added information. Panel B (panel D) plots the results of a permutation test of the significance of the math (reading) score difference between "treated" districts and its synthetic control. The solid dark line plots the difference for Winston-Salem, and the light gray lines plot the difference using other school districts.

Figure 1.6: The Impact of Teacher Quality on Student Achievement



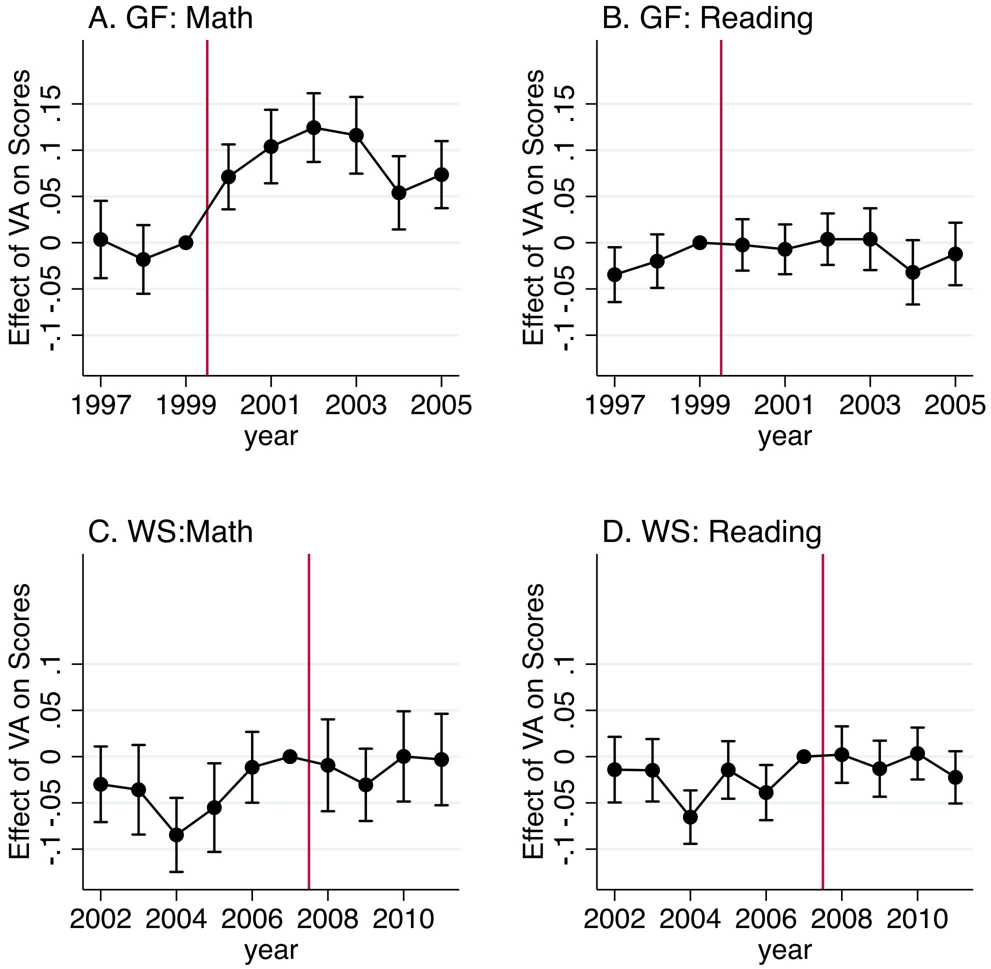
Notes: The figure plots the estimated coefficients of the interaction terms between the year and teacher VA measure from 1999 to 2003 for Guilford and from 2006 to 2011 for Winston-Salem. The vertical lines show the year when teachers first received the information. Each of the points represents the impact of teachers scoring one SD higher than the mean on student achievement gains. All 95 percent confidence intervals use standard errors clustered at the school level.

Figure 1.7: The Impact of Teacher Quality on Student Achievement by Teacher Quartile Ranking



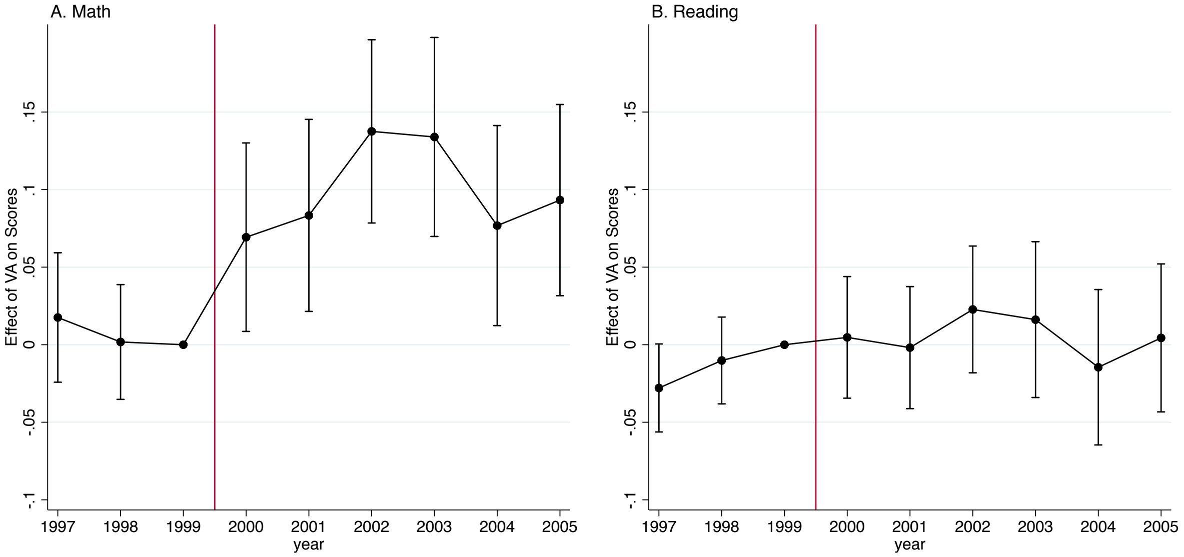
Notes: The figure plots the estimated coefficients of the interaction terms between the postpolicy indicator and teacher quartile ranking for Guilford and Winston-Salem. The coefficients in the given quartile represent how much a teacher in the indicated quartile increased student achievement gains when the VA reports were provided. All 95 percent confidence intervals use standard errors clustered at the school level.

Figure 1.8: Effect of Providing VA Information on Test Scores: Using DID Model with All Districts



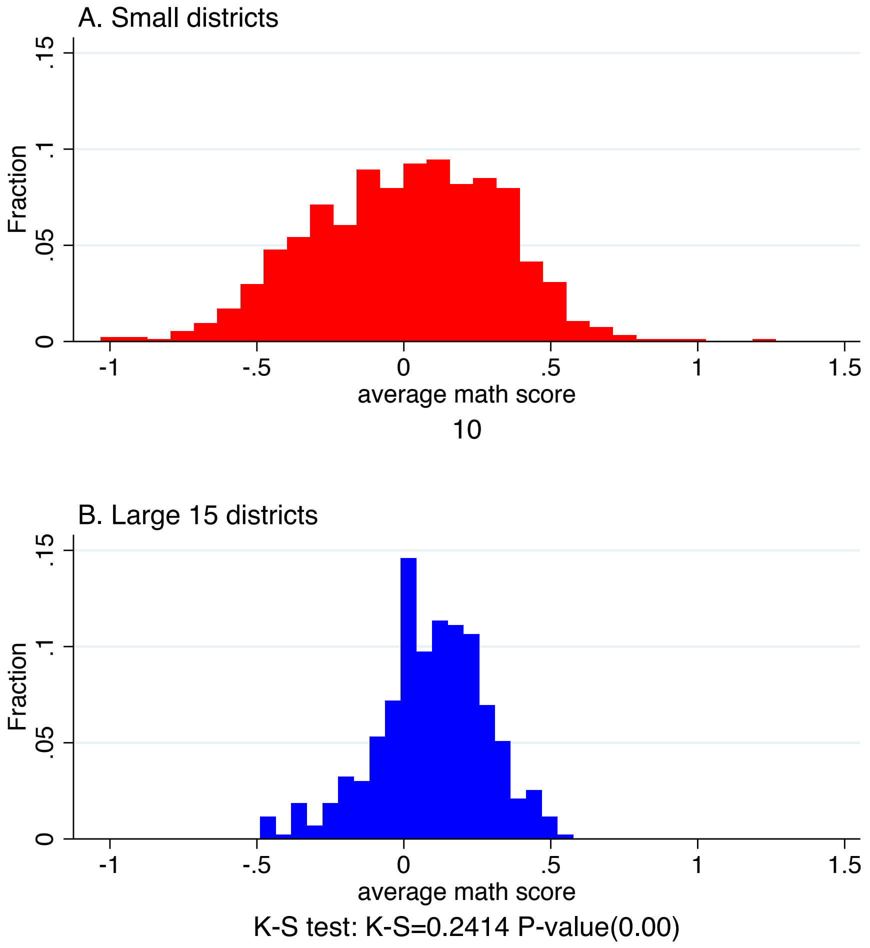
Notes: The figure shows estimates from the event study model that uses all school districts as comparison districts. The vertical lines in both panels show one year prior to the year when treated districts decided to adopted VA measures. This model includes controls for lagged student test scores, race, 6 categories of exceptional status, and gift status. Each point in this figure indicates the test score differences between treatment and control districts. All 95 percent confidence intervals use standard errors clustered at the school level.

Figure 1.9: Effect of Providing VA on Test scores by Years in Guilford: Using DDD Models



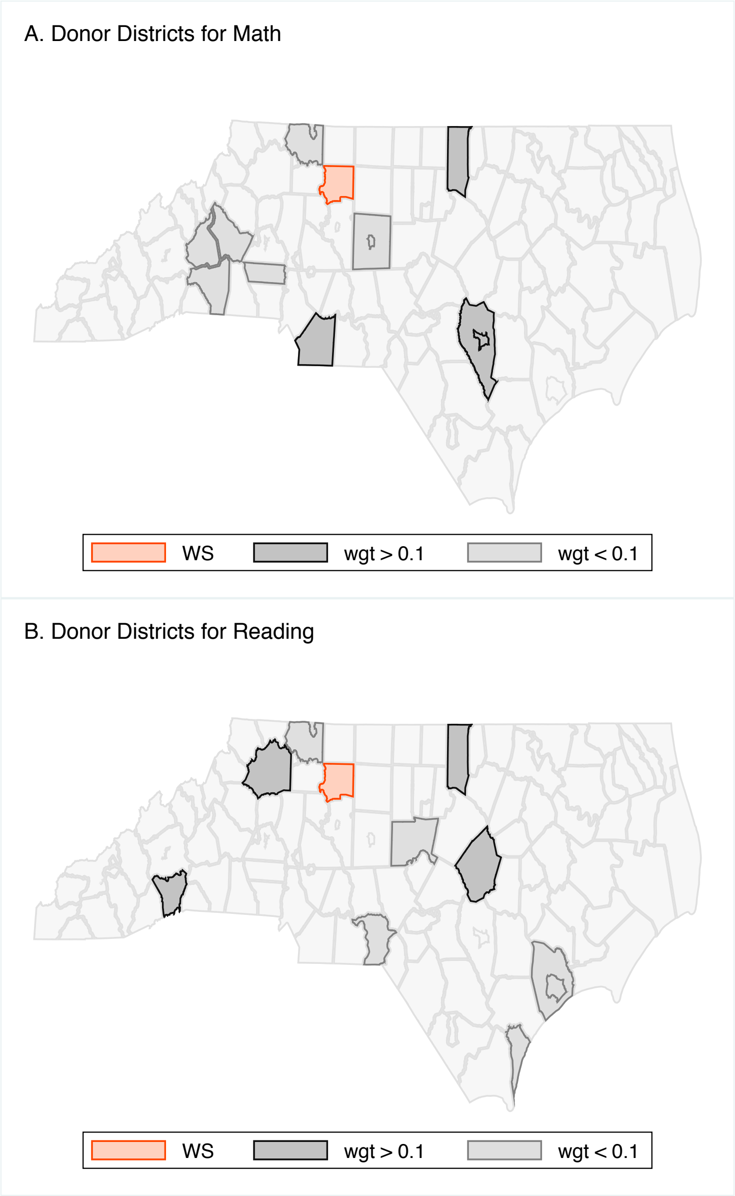
Notes: The figure plots the estimated coefficients using the event study model, where the treated grade in Guilford interacts with a series of time indicator from the following equation: $y_{it} = y_{it-1}\gamma_1 + X_{it}\gamma_2 + \delta_d \times \delta_t + TG_g \times \delta_t + \beta_0 TD_d \times TG_g + \sum_{t=1997}^{t=2005} \beta_t TD_d \times TG_g \times year_t + \varepsilon_{it}$. This model includes controls for lagged student test scores, race, 6 categories of exceptional status, and gift status. Each point (β_t)in this figure represents the effects of providing VA information on student achievement. All 95 percent confidence intervals use standard errors clustered at the school level.

Figure 1.10: The Distribution of the Average Math Score by the Number of Students



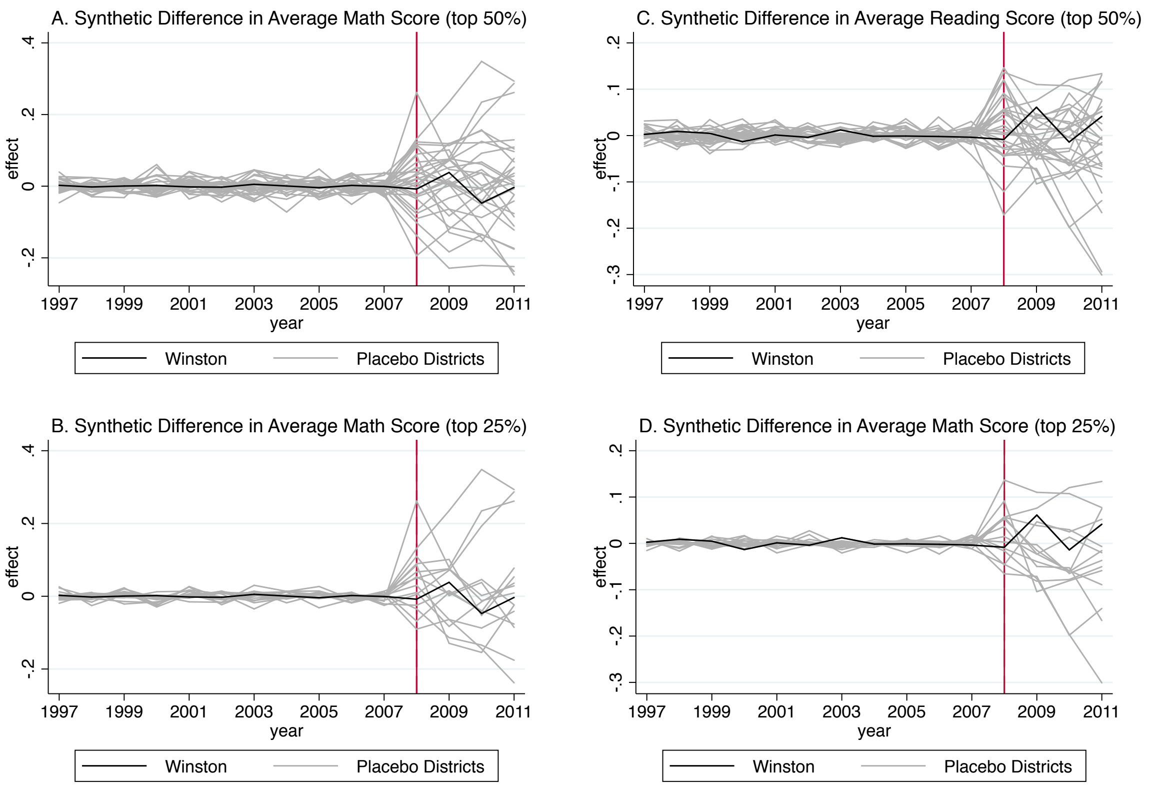
Notes: The figure shows the distribution of the average math score in small districts (the mean enrollment is smaller than 200) and the distribution of the score in large 15 districts. The test statistics (K-S) is calculated from the two-sample Kolmogorov-Smirnov test.

Figure 1.11: The Location of Donor Districts underlying the Synthetic Winston-Salem



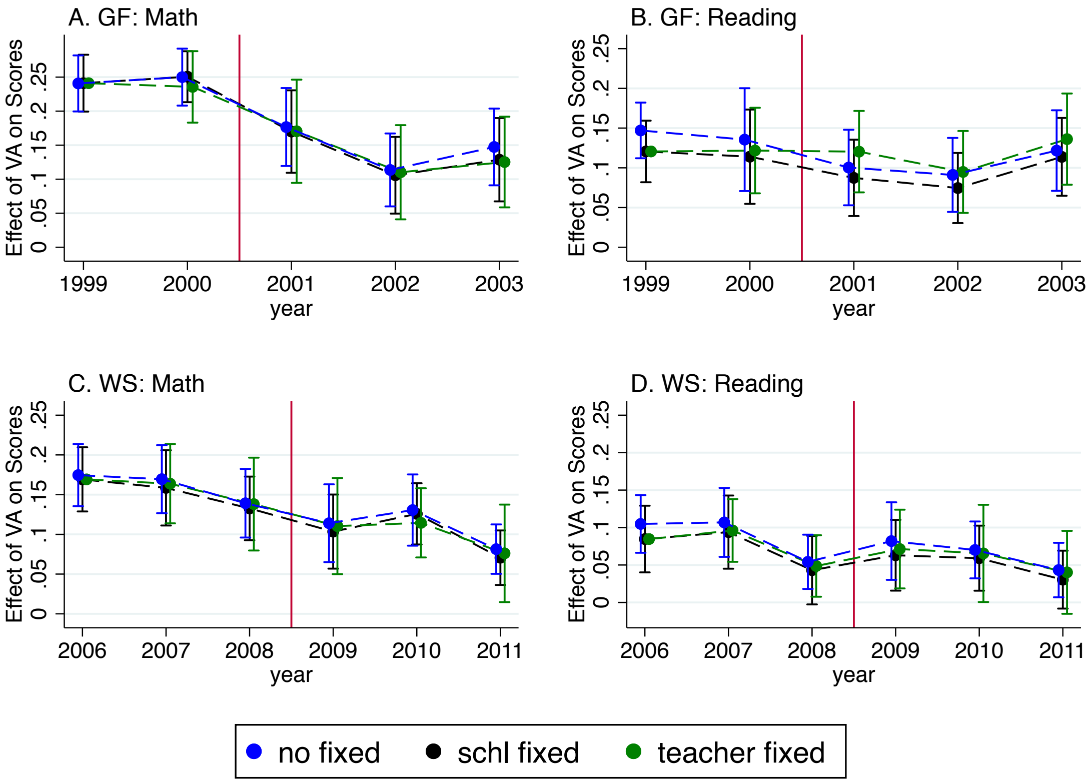
Notes: The figure shows the location of donor districts underlying the Synthetic Winston-Salem with a preferred predictor set. The districts colored dark gray are the donor districts with more than 10 percent weights, the gray districts have less than 10 percent, and the light gray have no weight. Winston-Salem is colored red.

Figure 1.12: Synthetic Difference in Average Math and Reading Score by RMSPE Cutoff



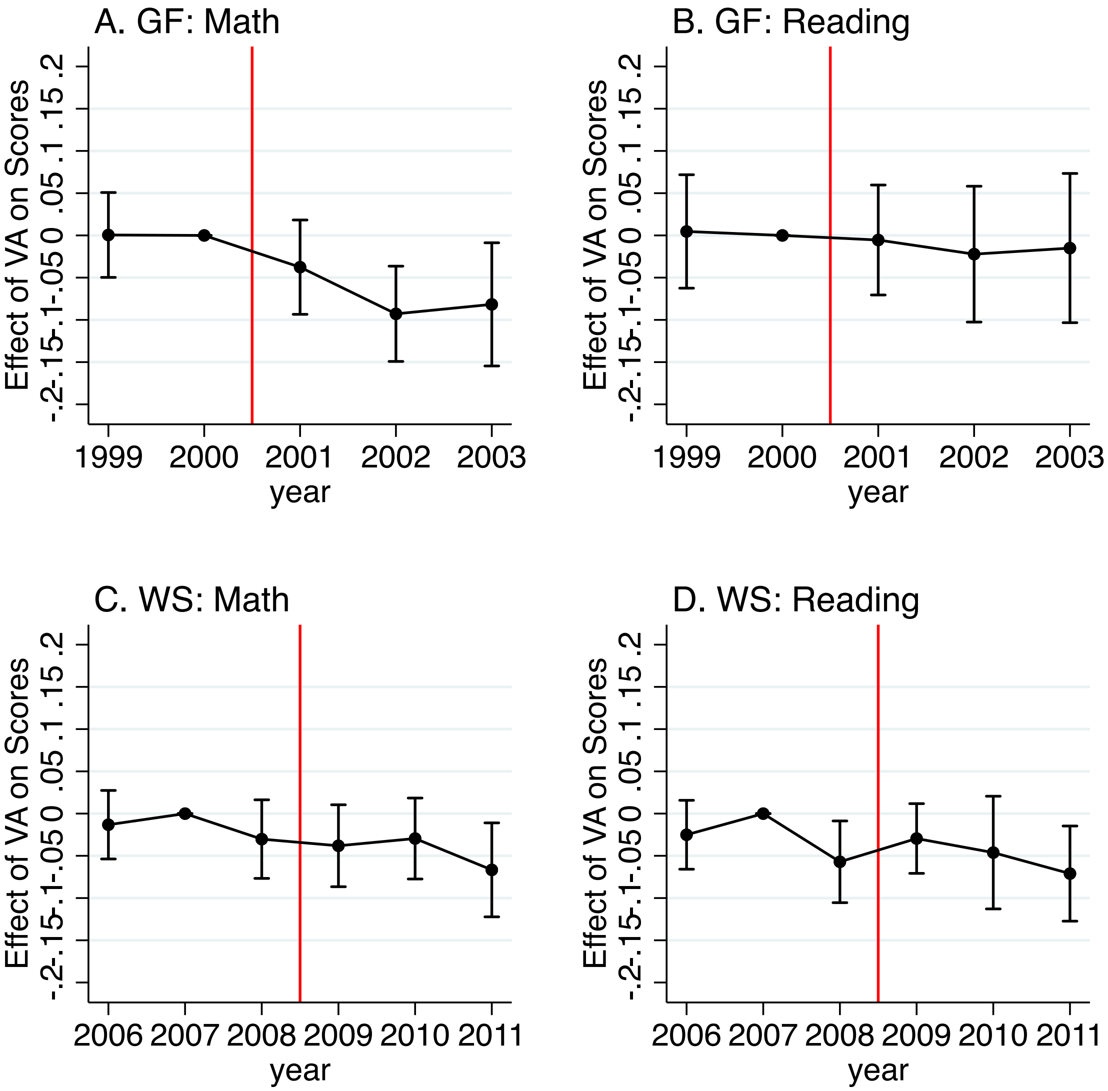
Notes: The figure plots the results of a permutation test of the significance of the difference between "treated" districts and its synthetic control for math score. The solid dark line plots the difference for Winston-Salem, and the light gray lines plot the difference using other school districts. The panel A (panel C) includes donor districts whose RMSPE are less than 5 times of the RMSPE of Winston-Salem and panel B (panel D) includes donor districts whose RMSPE are less than 2.5 times of the RMSPE of Winston-Salem.

Figure 1.13: The Impact of Teacher Quality on Student Achievement by Model Specifications I



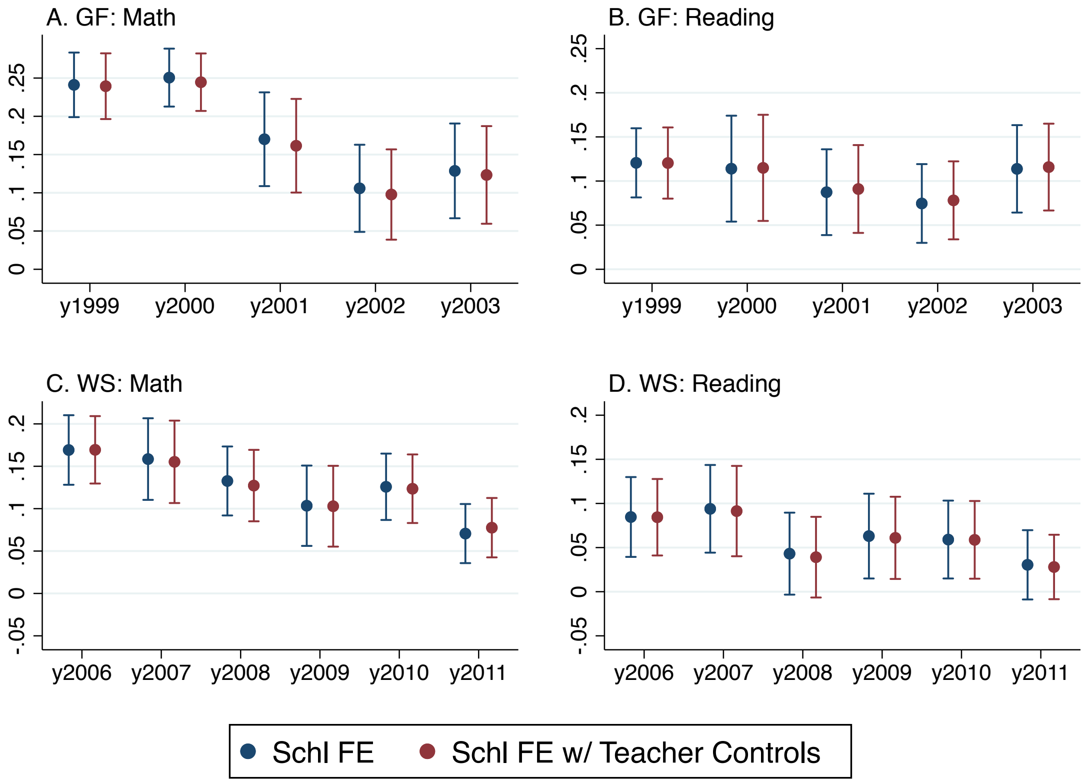
Notes: The figure plots the estimated coefficients of the interaction terms between the year and teacher VA measure from Eq.(6) with school fixed effects, without school fixed effects, and with teacher fixed effects. Each of the points represents the impact of teachers scoring one SD higher than the mean on student achievement gains. All 95 percent confidence intervals use standard errors clustered at the school level. The estimate with teacher fixed effects in 1999 is not shown because it is the omitted year when teacher fixed effects are included. The estimates with teacher fixed effects are normalized to the baseline estimate (with school fixed effects) in 1999.

Figure 1.14: The Impact of Teacher Quality on Student Achievement by Model Specifications II (Using DDD Model)



Notes: The figure plots the estimated coefficients of the interaction terms between the year and teacher VA measure from 1999 to 2003 and from 2006 to 2011 for Winston-Salem. The vertical lines show the year when teachers first received the information. Each of the points represents the impact of teachers scoring one SD higher than the mean on student achievement gains. All 95 percent confidence intervals use standard errors clustered at the school level.

Figure 1.15: The Impact of Teacher Quality on Student Achievement by Model Specifications II



Notes: The figure plots the estimated coefficients on the interaction terms between the year and teacher VA measure for both Guilford and Winston-Salem. Each blue dot shows the estimated coefficients from the baseline model, and each red dot represents the estimated coefficients from the model that includes various observable teacher characteristics such as experience, number of years in schooling, gender, and race. Each point represents the impact of one SD higher VA teachers on student achievement gains. All 95 percent confidence intervals use standard errors clustered at the school level.

APPENDIX C SYNTHETIC CONTROL METHOD

Here I describe why and how I exclude the number of small donor districts when I implement the synthetic control method to evaluate the policy change in Winston-Salem. The first reason is that student test scores in small districts are generally more volatile, making it difficult to find a convex combination of donor districts to closely match the outcome trajectories of small districts in the prepolicy period. The synthetic controls with poor fits are less likely to be informative to gauge the relative rarity of treated district as the lack of fits in the prepolicy period would artificially create the outcome gaps between donors and their synthetic controls in the postpolicy period. Second, more importantly, when small districts are used to construct the synthetic controls for a given district, it would worsen the predictability of the model in the postpolicy period relative to that when using the synthetic controls without small districts. The reason is that I may risk matching the outcomes of treated districts with noise measures of prepolicy outcomes when I include small districts. However, excluding small districts from the donor pool may produce a poor outcome fit in the prepolicy period, since it reduces the size of the donor pool when solving the optimization problem. I therefore remove small districts from the donor pool if excluding small districts decreases the average RMSPE in both pre- and post-treatment period

The below table shows the average RMSPE of all possible combinations for predictor sets and the donor districts in the pre- and post-treatment periods. Predictor set 1 in column (1) includes the set of average test scores in the pretreatment period in a given subject, and predictor set 2 in column (2) includes the biannual average test outcomes in a given subject. In column (3), I use both the biannual average math and reading scores, and in column (4), I include the biannual average math and reading scores with controls, including the pretreatment means of percent black, percent white, percent female, percent gifted, percent special education students, and student enrollment variables. This table confirms the intuition that using small districts may produce poor fits in both pre- and post-treatment periods; removing small districts in the donor pool improves the model fits in both math and reading regardless of the choice of predictor sets in pre- and post-treatment period. Since excluding additional small districts does not further decrease RMSPE in the pretreatment period, I remove 66 small districts from the donor pool and use 60 large districts in the main analysis.

RMSPE	(1)	(2)	(3)	(4)
Panel A. Math Test Score				
A.1 Using all donor districts				
Pre-treatment	0.034	0.052	0.050	0.046
Post-treatment	0.136	0.138	0.137	0.131
A.2 Using 60 large districts				
Pre-treatment	0.033	0.037	0.035	0.036
Post-treatment	0.105	0.112	0.110	0.107
Panel B. Reading Test Score				
B.1 Using all donor districts				
Pre-treatment	0.033	0.049	0.045	0.042
Post-treatment	0.098	0.101	0.101	0.100
B.2 Using 60 large districts				
Pre-treatment	0.029	0.034	0.033	0.032
Post-treatment	0.090	0.095	0.096	0.093
Predictors				
Annul outcomes	Y			
Biannual outcomes		Y		
Both biannual outcomes			Y	Y
Other controls				Y

REFERENCES

REFERENCES

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association 105 (490), 493–505.*

Ballou, D., W. Sanders, and P. Wright (2004). Controlling for student background in value-added assessment of teachers. *Journal of educational and behavioral statistics 29 (1), 37–65.*

Barrett, N. and E. F. Toma (2013). Reward or punishment? class size and teacher quality. *Economics of Education Review 35, 41–52.*

Bates, M. (2017). Public and private learning in the market for teachers: Evidence from the adoption of value-added measures. *Working paper.*

Benabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American economic review 96(5), 1652–1678.*

Bergman, P. and M. J. Hill (2018). The effects of making performance information public: Regression discontinuity evidence from los angeles teachers. *Economics of Education Review 66, 104–113.*

Brehm, M., S. A. Imberman, and M. F. Lovenheim (2017). Achievement effects of individual performance incentives in a teacher merit pay tournament. *Labour Economics 44, 133–150.*

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017). Randomization tests under an approximate symmetry assumption. *Econometrica 85(3), 1013–1030.*

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review 104(9), 2593–2632.*

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the impacts of teach- ers ii: Teacher value-added and student outcomes in adulthood.*American Economic Review 104(9), 2633–79.*

Chetty, R., A. Looney, and K. Kroft (2009). Salience and taxation: Theory and evidence. *American economic review 99(4), 1145–77.*

Chingos, M. M. and M. R. West (2011). Promotion and reassignment in public school districts: How do schools respond to differences in teacher effectiveness? *Economics of Education Review 30(3), 419–433.*

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of human Resources 41 (4), 778–820.*

Conley, T. G. and C. R. Taber (2011). Inference with difference in differences with a small number of policy changes. *The Review of Economics and Statistics 93(1), 113–125.*

Davis, C. R., L. Bangert, A. N. Comperatore, and M. Smalenberger (2015). Teacher and principal perceptions of the north carolina educator evaluation system.

Dee, T. S. and J. Wyckoff (2015). Incentives, selection, and teacher performance: Evidence from impact. *Journal of Policy Analysis and Management 34 (2), 267–297.*

Doherty, K. M. and S. Jacobs (2015). State of the states 2015: Evaluating teaching, leading and learning. *National Council on Teacher Quality.*

Dube, A. and B. Zipperer (2015). Pooling multiple case studies using synthetic controls: An application to minimum wage policies *Working paper.*

Ferman, B., C. Pinto, and V. Possebom (2017). Cherry picking with synthetic controls.*Woking paper.*

Glazerman, S., A. Protik, B.-r. Teh, J. Bruch, and J. Max (2013). Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment. ncee 2014-4004. *National Center for Education Evaluation and Regional Assistance.*

Goldhaber, D. and J. Hannaway (2004). Accountability with a kicker: Observations on the florida a+ accountability plan. *Phi Delta Kappan 85 (8), 598–605.*

Guarino, C. M., M. D. Reckase, and J. M. Wooldridge (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy 10(1), 117–156.*

Hahn, J. and R. Shi (2017). Synthetic control and inference. *Econometrics 5 (4), 52.*

Hanushek, E. A., J. F. Kain, D. M. O'Brien, and S. G. Rivkin (2005). The market for teacher quality. *Working paper*

Hanushek, E. A., J. F. Kain, and S. G. Rivkin (2004). Why public schools lose teachers. *Journal of human resources 39(2), 326–354.*

Hanushek, E. A. and S. G. Rivkin (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review 100 (2), 267–71.*

Horvath, H. (2015). Classroom assignment policies and implications for teacher value-added estimation. *Working paper.*

Imberman, S. A. and M. F. Lovenheim (2015). Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. *Review of Economics and Statistics 97(2), 364–386.*

Jackson, C. K. (2009). Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation. *Journal of Labor Economics 27(2), 213–256.*

Jackson, C. K. (2013). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics 95 (4), 1096–1116.*

Kane, T. J., D. F. McCaffrey, T. Miller, and D. O. Staiger (2013). Have we identified effective teachers? validating measures of effective teaching using random assignment. *In Research Paper. MET Project.*

Kane, T. J. and D. O. Staiger (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Kaul, A., S. Kloßner, G. Pfeifer, and M. Schieler (2018). Synthetic control methods: Never use all pre-intervention outcomes together with covariates. *Working paper.*

Krueger, A. B. (1999). Experimental estimates of education production functions. *The quarterly journal of economics 114(2), 497–532.*

Pope, N. (2015). The effect of teacher ratings on teacher performance. *Working paper.*

Rivkin, S. G., E. Hanushek, and J. Kain (2005). Teachers, schools, and academic achieve- ment. *Econometrica 73(2), 417–458.*

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review 94 (2), 247–252.*

Rockoff, J. E., D. O. Staiger, T. J. Kane, and E. S. Taylor (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review 102(7), 3184–3213.*

Sartain, L. and M. P. Steinberg (2016). Teachers' labor market responses to performance evaluation reform: Experimental evidence from chicago public schools. *Journal of Human Resources 51(3), 615–655.*

Steinberg, M. P. and M. A. Kraft (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researche 46 (7), 378–396.*

Taylor, E. S. and J. H. Tyler (2012). The effect of evaluation on teacher performance. *American Economic Review 102(7), 3628–51.*

Thomas, T. R. (2014). The Use of EVAAS Teacher Reports in Teacher Evaluation: Teacher Attitudes in Ohio's Public Schools. *Ph. D. thesis, Ohio University.*

CHAPTER 2

The Role of Credible Threats and School Competition within School Accountability Systems: Evidence from Focus Schools in Michigan

2.1 Introduction

Following the release of *A Nation at Risk* in 1983, the reforms of K-12 public education to improve the performance of chronically struggling schools and students became central in the policy debate in the United States. In the 1990s, some states and school districts enacted a test-based school accountability system to address this concern. The policy effort culminated in the federal No Child Left Behind (NCLB) Act, which expanded test-based accountability across the country by requiring all states to implement the accountability system. Under NCLB, all students in third to eighth grades were required to take annual tests in reading and math, and states were required to set "Adequate Yearly Progress" goals based on the percentage of students expected to show proficiency on state-created and state-mandated math and reading tests. Prior research examining NCLB and similar test-based state accountability systems generally finds positive impacts on student achievement (Ahn and Vigdor, 2014; Chakrabarti, 2014; Rouse et al., 2013; Dee and Jacob, 2011; Rockoff and Turner, 2010; Chiang, 2009; West and Peterson, 2006; Hanushek and Raymond, 2005). However, there is evidence of unintended negative consequences, such as manipulating test participation (Chakrabarti, 2013a; Cullen and Reback, 2006; Jacob, 2005) and changing student answers (Jacob and Levitt, 2003).

Because NCLB also set a 2014 deadline for 100% student proficiency on math and reading, and because Congress failed to make changes and reauthorized NCLB in advance of that deadline, Secretary Duncan and President Obama announced in the fall of 2011 that states could apply for waivers from certain NCLB requirements. The waivers expressed the Obama administration's vision of a reauthorization of NCLB focusing on "college-and-career-ready" criteria. However, the waiver-driven reforms also emphasized implementing a "differentiated accountability" system for states. To receive the waiver, each state was required to develop and implement an accountability system that would identify its lowest-performing schools ("priority" schools), as well as schools with the largest achievement gaps ("focus" schools). Once identified, priority schools were subject to interventions that were compatible with federal school turnaround principles, and focus schools were required to implement data-driven interventions that were appropriate for the particular environment of each school. Given that previous research has concentrated on a set of school accountability systems developed under NCLB, or on the accountability system in place before the NCLB era, and given that many features of school accountability systems under the NCLB waiver continue to be requirements under the recent reauthorization of NCLB, understanding the impact of specific features of the accountability system under the NCLB waiver is important for future discussions about the design features of school accountability systems.[1]

In this paper, I focus on the state of Michigan and explore the incentives and responses of schools that received Focus labels. I exploit a unique feature of school accountability in Michigan: the first year of Focus designation was regarded as a preparation year during which Focus schools were to isolate the effects of the stigma or financial incentives attached to these labels from the set of interventions associated with these labels in consecutive years. One particular concern of the Focus designation was that schools may have concentrated on the bottom 30 percent of students to improve their test scores, while diverting attention from the top 30 percent of students, as the Focus assignment was based on the achievement gap between the bottom 30 percent and top 30 percent of students within a school. I thus examine the impact of receiving one of these labels not only on the mean achievement of a school but also on the student achievement distribution.

[1] The reauthorization of NCLB is called the Every Student Succeed Act (ESEA). See Klein (2016) for an overview of school accountability systems under the ESSA.

To credibly identify the impact of both labels on various student achievement outcomes, I exploit discontinuous variations in the assignment of schools to the Focus labels. Using a sharp regression discontinuity (RD) design, I compare the mean or given quantile outcomes (e.g. 10 percentile, 25 percentile) within the school-grade distribution in Focus schools that were barely above the assignment cutoff to that of outcomes in schools just below the cutoff. My findings across the numerous specifications and variations in modeling choices provide consistent evidence that receiving the Focus label improved the math test scores of low-achieving students relative to their non-Focus counterparts by approximately 0.053 to 0.068 standard deviations. More importantly, I do not find any evidence that receiving the Focus label led to a decrease in the performance of high-achieving students, suggesting that the accountability system under the NCLB waiver may have increased the performance of low-achieving students without hurting high-achieving students.

The average achievement that I document for Focus schools, however, masks some heterogeneity. Because Title 1 schools with the Focus label received a set of financial sanctions from year 2 onward, and because they could be waived for these sanctions if they were not identified as Focus schools the following year, I first examine whether Focus schools with financial incentives responded more to this label.[2] I find that the positive achievement effects for Focus schools were entirely driven by Title 1 Focus schools that faced financial sanctions associated with being labeled as Focus schools the following year, whereas the achievement effects for non-Title 1 Focus schools are small and insignificant.

These results show that simply assigning the Focus labels in year T may not lead to increased student achievement in the following year unless the follow-up financial incentives are coupled with the labels. However, it is still possible that schools surrounded by many nearby alternative schools may have had a strong incentive to respond to accountability labels. The reason for this possibility is that if parents regard a Focus label as an indicator of low quality, Focus schools in such a competitive choice environment may be at higher risk of losing students if they are

[2] A Title 1 school is a school receiving federal funds for Title 1 students. At least 40% of students in a given school must enroll in the free and reduced lunch program to be eligible for receiving Title 1 funds from the U.S. Department of Education.

re-assigned as Focus schools the following year. Using the total number of alternative nearby schooling options for a given public school, including public schools in the same district, charter schools, and public schools in bordering districts that accept students from the given public school through the inter-district-school-choice (IDSC) program, I find that receiving the Focus label increased math test scores across the scoring distribution when schools were surrounded by many competitors. However, schools in an uncompetitive choice environment improved the test scores of low-achievers only. This finding may suggest the potential benefits of combining school choice programs and school accountability systems to prevent schools from performing educational triage on their students.

To test whether the positive achievement effect documented for Focus schools was driven by other potential mechanisms, I examine whether the changes in the composition of students, which may be induced by the accountability labels, can explain my results. Using several key measures of student composition, I find no evidence that the Focus designation influenced the student composition of schools. Finally, I examine whether the Focus designation had a persistent impact on students who enrolled in a Title 1 Focus school in the year that school was facing financial threat as the induced gains among low-achieving students may have disappeared rapidly if schools taught to the test as a reaction to the threat; I find some evidence that attending Focus schools that faced financial threats in the first year persistently increased students' math test scores in the following year.

This paper contributes in a number of ways to the prior research examining the effect of school accountability on student achievement. First, to my knowledge, this paper is the first to examine how competitive pressure created by charter schools and open enrollment programs influences the efficacy of school accountability systems. While Chakrabarti (2014) examines the heterogeneous treatment effect of receiving an "F" grade on student score distributions by looking at the number of non-failing public schools within a district, she only considers the school choice options created by accountability regimes. Second, my paper extends the few prior studies that evaluate the efficacy of school accountability systems under NCLB waivers (Bonilla and Dee, 2017; Dee and Dizon-Ross, 2017; Hemelt and Jacob, 2017). While Dee and Dizon-Ross (2017) and Hemelt and Jacob (2017) find a null effect of Focus school interventions in Michigan and Louisiana, respectively, Bonilla and Dee (2017) find that the Focus reform in Kentucky improved math and reading achievement. These studies, however, have been unable to separate the effects of receiving "Focus" labels from the set of interventions associated with these labels. My analysis provides the first evidence of the labeling effects and finds key measures in potential treatment heterogeneity.

Finally, simultaneously to and independently from my paper, Hemelt and Jacob (2017) estimate the impact of the Focus reforms in Michigan. They find robust evidence that Focus school interventions in Michigan had no significant effect on mean test scores and no heterogeneous effect across students' score distribution. This result is in contrast to my findings that the Focus label improved both average math test scores and the performance of low-scoring students. The reason is that while my paper uses test scores in the immediate year in order to focus on the effect of the stigma or financial threats attached to accountability labels, the authors do not examine these effects, as they use student test scores from year 2 onward. Furthermore, they do not explore the heterogeneous treatment effects of financial incentives or competitive pressure induced by school choice programs in Michigan, which have important implications for policy makers seeking to better design future school accountability systems.

The remainder of the paper is organized as follows. Section 2 details the accountability systems, charter schools, and open enrollment programs in Michigan. Section 3 discusses the data. Section 4 presents the empirical strategy, and Section 5 presents the main results. I discuss potential mechanisms for my results in Section 6, and Section 7 concludes the paper.

2.2 Background

2.2.1 Differentiated Accountability under NCLB Waivers and Focus Schools in Michigan

The U.S. Department of Education announced in the fall of 2011 that states could apply for waivers from certain conditions of NCLB. Waivers were applied in particular to the requirement that schools and districts must achieve the unrealistic goal of 100% student proficiency on math and reading tests by 2014. To receive the waiver, each successful application for flexibility had to include two key components as part of a state accountability system: First, the accountability system had to adopt "college-and career-ready" criteria in at least reading and math, and second, in each state, the accountability system had to identify the schools with the lowest performance ("Priority" schools) and those with the largest achievement gaps ("Focus" schools).

Under this differentiated accountability system, Focus schools were identified as being among the schools with the largest achievement gaps, including at least ten percent of Title 1 schools. The intervention required for Focus schools was not specifically prescribed since states were only required to implement the data-driven interventions deemed appropriate for the particular environment of each school (U.S. Department of Education, ESEA Policy Document, 2012). Thus, it is not surprising to see broader heterogeneity in the types of interventions attached to the Focus label across states.

Like many other states, the Michigan Department of Education (MDE) submitted an NCLB waiver application, which was approved in July 2012. The new school accountability system (known as the Michigan School Accreditation and Accountability System) announced its first Focus school list in August 2012. Once identified, Focus schools received a set of four year-long interventions.

The interventions required that MDE and its districts to provide multiple types of support for Focus schools. Specifically, District Improvement Facilitators and district administrators set up a data-driven environment in every Focus school so that principals and teachers could identify 1-2 major factors in their teaching practices that could reduce the achievement gap; they also characterized the district-level infrastructure needed to implement the identified teaching practices. Finally, if Focus schools received Title 1 funding from the federal government, a set of financial sanctions was attached to the Focus label. In year 2 (2013-2014 school year) of their Focus status, Focus schools had to reserve 10 percent of their Title 1 building-specific allocation to improve their school system; they needed to set aside 10 percent of their Title 1 district-specific allocation in year 3 (2014-2015 school year) to provide "choice and transportation" to parents in Focus schools or to improve the school system.[3]

One unique aspect of the school accountability system under the NCLB waiver in Michigan is that the MDE considered the 2012-2013 school year as a preparation year, during which Focus schools were expected to develop plans of action that are aligned with Federal requirements. This unique feature has enabled me to use year-1 test scores to isolate the stigma effects from the set of interventions attached to the Focus labels. Note that the effects of the Focus interventions reported in prior research are combined effects of the labels and the interventions. (Bonilla and Dee, 2017; Dee and Dizon-Ross, 2017; Hemelt and Jacob, 2017). Since the year-1 test was administered in early-to-mid October and the accountability results were released to the public in early-to-mid August, one might think that using the year-1 test scores would not be sufficient to capture the label effects. However, as I discuss in more detail below, prior research finds that school accountability systems can improve student achievements in failing schools within a short time window (Rockoff and Turner, 2010).

To determine the list of Focus schools each year, the MDE mainly used "top-to-bottom" (TTB) ranking. The TTB index was a weighted average of three subject-specific achievement indexes: two-year average level scores, two-year average growth in scores, and the two-year average achievement gap (hereafter, AGI) between the bottom 30 percent and top 30 percent of students

[3] The 10 percent set-aside funds were mainly used to support professional development with regard to teaching through multi-tiered systems of support or through targeted instructions for low scoring students; or to provide time and space for daily (or weekly) teacher collaboration (MDE, Frequently Asked Questions about Michigan's Focus Schools, 2014).

within a school. Schools below the 5th percentile in the TTB ranking were identified as Priority schools, and Focus schools were identified as schools in the bottom 10th percentile of the AGI distribution.

Prior research has already effectively documented that schools game the accountability system. Schools are more likely to concentrate on high-stakes subjects or grades when the school accountability system is based on specific test scores in those subjects or grades (Chakrabarti, 2014). When the accountability system requires schools to pass their proficiency standard in at least one subject, they may also focus on a specific subject in which it is easy to raise test scores (Chakrabarti, 2013b; Goldhaber and Hannaway, 2004). Furthermore, Richardson (2015), Ladd and Lauen (2010), Neal and Schanzenbach (2010), and Reback (2008) find that schools tend to improve the test scores of students who are on the margin of passing the cutoff when the accountability systems are based on proficiency rates at the level of test scores. One possible gaming behavior is that Focus schools may have concentrated on the bottom 30 percent of students to improve their test scores and diverted their attention from the top 30 percent of students. I thus examine the impact of receiving Focus labels not only on the mean achievement but also on students' test scores across the achievement distribution.

One remaining concern is that schools may not have had an incentive to avoid being labeled Focus in year 2 because year 1 was regarded as a preparation period, and schools received the associated interventions regardless of whether they received one of the labels in year 2. However, at least two factors may have increased schools' incentives to respond to the labels. First, principals and teachers in Focus schools may have regarded the labels as social stigmas and tried not to be assigned the labels the following year given that the list of Focus schools was publicly available and received local media attention. Goldhaber and Hannaway(2004) surveyed principals and teachers and found that they viewed a school grade of "F" as a social stigma under the accountability system in Florida. Furthermore, when schools were surrounded by many alternative nearby schools, the stigma effect may have been even larger than for schools with only a few nearby options. The reason is that the labels may have signaled the low quality of schools to parents, and thus schools in competitive choice environments may have been more at risk of losing students. Second, when Title 1 Focus schools were removed from the list the following year, they received a waiver for a set of financial sanctions, as mentioned above; this waiver may have created financial incentives for Focus schools to respond to the Focus designation. Specifically, a Title 1 Focus school did not need to set aside 10 percent of its Title 1 building-specific funds (year 2 requirement) if they were not identified as a Focus school the following year.

2.2.2 School Choice Programs in Michigan

Since NCLB was enacted in 2002, school districts have been required to offer open enrollment options within districts; these options are intended to help students in repeatedly failing schools to make Adequate Yearly Progress (AYP), with the goal of having all students pass the state-set proficiency benchmarks. The accountability system in Michigan under the NCLB waiver similarly offered intra-district open enrollment programs for students who were enrolled in Focus schools. Specifically, in the 2012-2013 school year, Focus schools had to offer students the option to attend non-Focus schools within a district, and they also had to cover transportation costs. These schools, however, were not required to provide choice and transportation costs from the 2013-2014 school year onward.

Focus schools may have faced choice threats from alternative school choice options other than those options embedded in the accountability system. Specifically, during the sample period of my study (from the 2011-2012 to the 2012-2013 school year), the percentage of public school students in Michigan enrolled in charter schools was approximately 8.0 percent, and the percentage of public school students enrolled in neighboring districts by participating in the inter-district school choice program (IDSC) was approximately 7.8 percent (Cowen et al., 2015).[4] [5]

[4] Michigan was one of the first states in the U.S. to create competition from various school choice programs. The law creating charter schools was passed by the Michigan State Legislature in December 1993 and the law providing which created the inter-district school choice (IDSC) program was enacted in 1994.

[5] The IDSC program in this study refers to the choice program under Section 105. Section 105 allows schools to enroll students who reside in other local school districts within the same intermediate school district (ISD), and Section 105c allows schools to enroll students who reside within contiguous intermediate school districts. Note that the participation in IDSC program is voluntary, so some districts may not participate in the IDSC program. Also, each participating district determines caps on non-resident enrollment though most districts in the state nominally had

Figure 2.1 maps the distribution of public schools that received Focus labels, as well as charter schools during the 2011-2012 school year across Michigan. The figure clearly shows that Focus schools and charter schools are spread across Michigan, although charter schools are concentrated in large cities such as Detroit, Grand Rapids, Flint, and Lansing and their suburbs. The broad heterogeneity of the competitive pressure (induced by charter school penetration) for Focus schools can be seen on the figure. Figure 2.2 displays the spatial variation in the share of transfer students using the IDSC program in the 2011-2012 school year. Again, school districts were not required to participate in the IDSC program, and therefore, some school districts did not have any students who enrolled in neighboring districts by participating in the program. Figure 2.2 shows broad variations in the percentages of students enrolled in bordering districts using the IDSC program conditional on a district participating in this program. The figure indicates that public schools experienced a heterogeneous choice environment due to the IDSC program. For example, it is possible for one Focus school to have been located in a school district where more than 20 percent of residential students were enrolled in neighboring districts, while the other Focus school was within a district where only a few students used the IDSC programs.[6]

2.3 Data and Sample

To identify the impact of school labeling on students' score distribution, I combine three datasets from various sources. The first dataset that is publicly available from the MDE website contains accountability results from 2011 to 2013 at the school level for every school in Michigan serving grades 3 to 8. The data file includes the school performance index (hereafter SPI, but officially called the TTB ranking), which is used to determine eligibility for Priority labeling, and the achievement gap index (AGI), which is the running variable for Focus schools. I re-center these two measures on zero and change the sign of the measures, giving schools with positive values on

accepted students (Cowen and Creed, 2015).

[6] Some districts may reject transfer applications if the number of transferring students exceeds their caps. It is thus possible that the actual competitive pressure that the district faces may be greater if the number of IDSC applicants is larger than the number of available slots.

the index one of the labels. The second data file for enrollment, school type (e.g. charter, magnet), Title 1 status, and geographic location comes from the Common Core of Data (CCD) of the National Center for Education Statistics (NCES).

To measure student achievement and demographic characteristics at the school-year cell level, I start with the individual-level administrative data provided by the MDE and Michigan's Center for Educational Performance and Information (CEPI). The data represent the universe of Michigan students in grades 3-8 from 2011-2014, and contain students' test scores in mathematics, reading, science, social studies, and writing. Additionally, the data include various demographic controls, such as grade, gender, race, special education, and free lunch status. Given that the unit of variation is the school-year, using school-year observations produces virtually identical results as using student-year observations. I thus aggregate student-year observations into school-grade-year observations by calculating the unconditional mean, SD, and various quantiles of the within-school-grade achievement distribution. Note that I aggregate to school-grade-year instead of school year to ensure that I correctly compare schools with different grade configurations.[7] Starting with student-year observations allows me to construct various quantile measures, which are not usually presented in the publicly available school-level or school-grade-level data.[8]

During the sample period, Michigan administered Michigan Educational Assessment Program (MEAP) exams in grades 3-8 in early -to -mid- October. The accountability results were released to the general public in early -to -mid- August, leaving three months for principals and teachers to prepare for the impending test. When Hemelt and Jacob (2017) evaluate the effect of accountability labeling on student achievement, they use the following year's fall exams to study the year 1 effect. For example, they use Fall 2013 exams to examine the year 1 effect of cohort 2012 instead of using the Fall 2012 exam. I take a contrasting approach and use the same year's exam to examine the year 1 effect for two reasons.[9] First, extensive MEAP exam preparation prior to taking

[7] The estimated results are similar when I use school-year observations.

[8] The MDE website provides the MEAP test results publicly since 2010. The data provides school-grade-year level information including the percentage of students in minimal, basic, proficient, and advanced levels for math, reading, science, social study, and writing.

[9] Brummet (2014), who examines the effect of school closing on student achievement using students in Michigan, took the same approach.

the test may have influenced student test scores. Rockoff and Turner (2010) examine the school accountability system in New York City, where schools only have four to six months to respond to their accountability grade. They find significantly improved student achievement in schools with a low grade, indicating that school administrators were able to improve student test scores within a short time window. Second, more importantly, when schools received one of the accountability labels, their incentives to improve test scores were more closely aligned with the current year's exams than with the following year's exams. The reason is that the current year's test results were used to estimate the accountability results for the following year. Therefore, Title 1 schools that received the Focus label in August 2012 had to reduce the achievement gap between the bottom and top 30 percent of students on the Fall 2012 exam to avoid the year 2 financial sanctions.

My sample includes schools serving grades 3 to 8 in 2013 because annual state administered tests are not available for other grades. Among these schools, most schools serving the tested grades are elementary or middle schools. I drop schools that serve some 3rd to 8th graders as well as students in higher grades, e.g. schools with grade configurations such as 6-12, because the SPI for these schools was required to incorporate the high school graduation rate, causing these schools to focus on different criteria. I further limit the sample by dropping schools serving special education students from my main analysis because these schools were subject to different mandates than other public schools. I also eliminate 2 schools with fewer than 50 students from my sample because these schools serve special populations.

I do not drop charter schools in my main sample because students in charter schools had to take the MEAP exams and because charter schools are subject to the same school accountability system as traditional public schools. My approach contrasts with the approach taken by Hemelt and Jacob (2017), who cast doubt on whether charter management organizations enforced the reforms required by the Focus labels. Since my analysis focuses on the short-run impacts of school accountability labeling on student test scores and not on the impact of a set of reforms in consecutive years (from 2 to 4 years after receiving the labels), it is not unreasonable to include charter schools in my main sample. Furthermore, prior research documents that parents at charter schools are more responsive to school accountability ratings (Hanushek et al., 2007) and that the effect of the accountability rating is larger at charter schools (Baude, 2015); the short-run impacts of the accountability rating may thus have been greater for charter schools. Nevertheless, I will show later that excluding charter schools has little impact on my estimates.

The summary statistics of certain key variables for the sample of schools in 2013 are shown in Table 2.1. The table compares the means of schools with the Focus labels and all schools; in the full sample, approximately 14.34 percent of schools received the Focus labels in 2013. Table 2.1 clearly shows that Focus schools were high performing schools and served relatively high-SES students compared to all K-8 schools. In terms of demographics, schools with a Focus label served a similar percentage of white students compared to all K-8 schools, but the percentage of Asian students in Focus schools was approximately three times higher than in all K-8 schools (approximately 7.45 and 2.71 percent, respectively). Overall, schools with Focus labels were academically better schools, but this performance concealed large achievement gaps between students in the bottom and top 30 percent.

2.4 Empirical Strategy

To examine the effect of the "Focus" labeling on various academic outcomes, I compare schools that were barely above the cutoff and thus were assigned one of the two labels to schools that were just below the cutoff and thus were not assigned Focus labels. Because MDE used two different running variables to designate Priority and Focus schools, it is possible that some schools were above both cutoff points. In this case, the MDE prioritized the "Priority" labeling, and therefore schools that qualified for both labels were designated as Priority schools. I dropped Priority schools to examine Focus labeling, and thus my research evaluates the impact of Focus labels with a "Sharp RD design" (Hahn et al., 2001).[10] I model the impact of receiving Focus labels on student achievement by the following regression equation:

[10] For interested readers, I plot the relationship between the Focus assignment and the AGI in Figure 2.6.

$$Y^m_{gst} = \alpha_0 + \alpha_1 I(r_s \geqq r*) + \alpha_2 r_{st-1} + \alpha_3 I(r_s \geqq r*) \times r_{st-1} + \sum_{g=4}^{8} \gamma_g G_g + X_{st-1}\beta + \varepsilon_{st} \tag{2.1}$$

$$where\, t = 2013$$

I use g index to show a specific grade (from 3 to 8) and s denotes a school. Y^m_{gst} represents various grade-school level measures (m) based on math and reading test scores in year t, including the mean, SD, and various quantiles in the within-school-grade achievement distribution. $I(r_s > r*)$ is a binary variable that takes one if schools' AGI score is greater than or equal to the cutoff, and r_{st-1} shows the running variable for AGI scores in year $t-1$. I control for a trend in running variables with a linear spline ($r_{st-1} + I(r_s \geqq r*) \times r_{st-1}$) because I estimate the model using observations near the cutoff. G_4 to G_8 are grade dummy variables as I pool the observations across grades.[11] Additionally, Eq (1) contains a vector of school-level control variables (X_{s2012}) reported in Table 2.1 to increase precision. As Table 2.8 shows, the point estimates without control variables are similar to the baseline. Finally, due to the likelihood that errors are correlated across grades within schools, for all specifications, I provide the standard errors in all the analyses that are clustered at the school level.

For bandwidth selection, I use an optimal bandwidth selection method proposed by Calonico et al.(2014) for each outcome measure and each different specification.[12] The optimal bandwidths based on their method range from 0.45 to 0.55 for Focus schools. In all tables that show the local RD estimates, I report the optimal bandwidth and the number of observations that are used to estimate the RD estimate α_1. I then estimate Eq.(1) with the given bandwidth using a triangular kernel to weight observations within the bandwidth since the triangular kernel is optimal for the boundary estimation given the optimal bandwidth (Imbens and Lemieux, 2008). Nevertheless, the choice of other kernels (uniform, Epanechnikov. etc) has little impact on the estimated results.

[11] The relationship between a running variable and outcome measures could be different across grades. I re-estimate the model that includes an interaction term between the grade dummies and the linear spline term. Including the interaction term does not change any of the results.

[12] Their method basically chooses the bandwidth that minimizes the asymptotic MSE (mean square error) of the RD point estimator.

The main identification assumption to estimate Eq. (1) is that each school does not have precise control over the running variables (Lee and Lemieux, 2010). Prior studies have documented evidence that schools may game the accountability system. (Figlio, 2006; Cullen and Reback, 2006; Jacob, 2005; and Jacob and Levitt, 2003). However, I argue that manipulating running variables in my context is unlikely since cutoff in AGI is not deterministic. The cutoffs were determined by the relative positions among all eligible schools, and principals did not have any reference points to predict the cutoffs for the Focus labels since the labels were exogenous shocks to them. Nevertheless, in Figure 2.5, I plot the distributions of the SPI score and AGI score; [13] the vertical black line in both panels A and B shows the cutoff. If schools had manipulated running variables, I would observe a large cluster on the left side of the cutoff. However, the figure clearly shows no discontinuity at the cutoff, and the test statistics proposed by Cattaneo et al. (2016) indicate that the discontinuity estimates are small and insignificant.

To further test the validity of the RD design, I explore whether pre-treatment covariates are continuous at the cutoff. If the unobserved school qualities were discontinuous at the cutoff, the pre-treatment variables would be different around the threshold. Table 2.2 shows the estimated coefficients for the baseline control variables reported in Table 2.1. I do not find evidence that Focus schools are not comparable to their barely non-Focus counterparts; all discontinuity estimates for various pre-assignment covariates are not statistically significant even at the 10 percent level.

Next, I analyze whether the responses of the Focus schools depended on receiving Title 1 funds. As discussed above, Focus schools that received Title 1 funds were threatened to set aside 10% of their school-level Title 1 fund allocations in the following year if they received Focus labeling two consecutive years. To examine whether different incentives conditional on Title 1 status influence the estimate, I first estimate Eq. (1) separately for Title 1 schools and non-Title 1 schools. To ensure that pre-assigned control variables are continuous within the Title 1 and non-Title 1 samples, I report the discontinuity estimates of the control variables in Table 2.9 Except for the percentage of Hispanic students for Title 1 Focus schools in panel B, the discontinuity

[13] I generate the density of the running variables across the cutoff by using the user written program in Stata.

estimates in both panels are never statistically significantly different from zero, which supports the validity of the RD design within the subsample. I should note that with a large number of multiple comparisons, I expect that a few outcomes can be statistically distinguishable from zero due to pure random variation.

To test the equality of the estimated treatment effects between non-Title 1 and Title 1 Focus schools, I attempt to disentangle the treatment effects for Title 1 schools with Focus labels by estimating the fully saturated model that interacts the Title 1 dummy variable with all control variables on the right-hand side of Eq. (1):

$$\begin{aligned} Y_{gst}^{m} &= \alpha_0 + \alpha_1 F_{st} + \alpha_2 F_{st} \times T_{st} + \alpha_3 r_{st-1} + \alpha_4 r_{st-1} \times T_{st} + \alpha_5 F_{st} \times r_{st-1} \\ &+ \alpha_6 F_{st} \times r_{st-1} \times T_{st} + \sum_{g=4}^{8} \gamma_g G_g + \sum_{g=4}^{8} \delta_g G_g \times T_{st} + X_{st-1}\beta_1 + T_{st} X_{st-1}\beta_2 + \varepsilon_{st} \end{aligned} \tag{2.2}$$

where, F_{st} indicates a binary variable that has value 1 if a school receives the Focus label in year t and T_{st} equals 1 if a school receives Title 1 funds in year t. All other notations are the same as in Eq. (1). The parameter of interest α_1 captures the effect of the pure labeling effect for Focus schools, and α_2 provides the impact of the financial threat. Thus, $\hat{\alpha}_2$ and the corresponding standard error is used to construct t-statistics that compare the coefficients between non-Title 1 and Title 1 schools in the subgroup analysis.

2.5 Results

In this section, I first evaluate the impact of receiving the "Focus" labeling on student performance distribution. Next, I attempt to separate out the effect of the financial threat and the stigma associated with receiving those labels. Finally, I then turn to exploring the heterogeneity in the effects of Focus labeling when the availability of alternative nearby schooling options is different.

2.5.1 The Impact of Labeling on Student Achievement

Figures 3 illustrate the effects of receiving Focus labels on various student math achievement outcomes, including the mean and the different points in the within-school-grade achievement distribution. The x-axis shows schools' AGI relative to the cutoffs that determined the Focus labeling for the 2012-2013 school year, and the y-axis describes the mean and various quantiles of the math test scores. To better visualize outcomes around the cutoff, I only include schools with an AGI within ± 0.5 from the cutoff. Each of the points in these figures indicates the average outcome measures of mean and various quantile outcomes collapsed into bins. Instead of equal-length bins, I use bins that contain the same number of observations since most schools with the Focus labels are concentrated around the cutoff. Furthermore, I follow Cattaneo et al. (2017) to choose the number of optimal bins in the sense that the overall variability of the binned means resembles the variability in the raw data.[14] This procedure ensures a large number of local means near the cutoff, which is useful for obtaining a graphical illustration of the variability of the data around the cutoff.

Figure 2.2.3 shows some evidence that schools with Focus labeling improved student math achievement across the scoring distribution. The vertical distances between the local means near the cutoff are positive for most outcome variables (see panels A to D). Table 2.3 presents the regression analog to these results. In panels A and B of Table 2.3, I first present the results that examine the impact of receiving Focus labeling on 2011 Fall test scores. Because students were taking this exam 9 months prior to the Focus designation, I expect the estimates of Focus school indicators across various outcome variables to be small and insignificant. This expectation is confirmed in panels A and B; all estimates in columns (1) through (7) are small and insignificant. Panel C (panel D) shows the estimated results that use math (reading) test scores from the school year 2012-2013. Unlike the estimates from the placebo exercises, I find that the Focus designation improved the average math test scores by 0.052 SD. To relate this estimate to prior literature, this magnitude was approximately half of the estimated effects of receiving an F grade on the following year's test

[14] Specifically, I generate the set of scatter plots in Stata using the rdplot package with the mimicking variance quantile-spaced option.

scores under the Florida accountability system (0.118 SD; Chiang, 2009), and the effect is more similar to the impact of receiving an F or D grade on the following year's exams in New York (0.10 and 0.05 SD; Rockoff and Turner, 2010). The smaller mean achievement effect of Focus labeling, however, is consistent with a shorter time window between the announcement of Focus labeling and the following year's exam.

Panel C of Table 2.3 also shows that there is minimal concern about educational triage in my context; in fact, the evidence shows that the accountability system under the NCLB waiver may have increased the performance of some students without hurting others. Specifically, the test scores of low-performing students (hereafter, students in the 10th and 25th percentiles in the school-grade achievement distribution) improved when their schools were designated as Focus schools, while I find no evidence that the estimated effects for high-performing students (hereafter, students in the 75th and 90th percentiles of the school-grade achievement distribution) were negative.

Panel D of Table 2.3 shows the estimates using the various outcome measures from the reading test scores. Across all columns, I do not find any significant evidence that receiving the Focus label influenced student achievement in reading. The estimates, while positive, are all small and insignificant, even at a 10 percent significance level. However, this finding may not be surprising if the achievement gap between the bottom 30 percent and top 30 percent of students is larger in math than in reading, making Focus schools concentrate more on math than reading. Table 2.10 shows the summary statistics of the AGI for both math and reading. The table indeed indicates that the mean of the AGI for math was larger than that of the AGI for reading among Focus schools. Another possible explanation is that math scores would have been much easier to improve than reading scores when schools were given limited preparation time for the test. This is consistent with the previous finding that when F schools had limited time prior to the state test, the impact of receiving an F grade on student achievement was approximately twice as large in math than in reading (Rockoff and Turner, 2010).

To check whether the significant positive effects in Table 2.3 are sensitive to bandwidth selection, I depict baseline estimates reported in panel C of Table 2.3 by bandwidths ranging from 0.1 to 0.6 point in Figure 2.7. I add 95 and 90 percent confidence intervals in all panels to check the robustness of the estimates visually. Figure 2.6 clearly indicates that the significant estimates are insensitive to bandwidth selection, as all point estimates stabilize at 0.3 points so that the solid lines in all panels are basically flat between 0.4 and 0.6 points. This exercise also suggests that the positive achievement effects of Focus labeling are not driven by unmeasured associations between the running variable and the underlying factors of student test scores.

2.5.2 Heterogeneous Effects by Receiving Financial Sanctions

As discussed in Section 2.1, Title 1 Focus schools received a waiver for financial sanctions if they were not labeled as Focus schools the following year, indicating that the treatment effects may have depended on the Title 1 status. The difference in incentives attached to the Title 1 status thus provides some insights into the role of "threat of financial sanction" and the stigma associated with being labeled a Focus school the following year.

Before accessing the heterogeneous treatment effects by Title 1 status, in Figure 2.4, I display the mean and the different percentiles of math scores observed in non-Title 1 schools and Title 1 schools to provide a visual summary. Panels A and B of Figure 2.4 show that the positive labeling effects documented in the previous section are mostly concentrated among Title 1 schools; the fitted line in panel A indicates that among schools that did not receive Title 1 funds, the mean scores of Focus schools are not very different from the mean scores of non-Focus schools. On the other hand, the fitted line in panel B shows a clear jump at the cutoff, and the binned means of Focus schools near the cutoff are less noisy than the binned means in panel A. Panels C-F present evidence on low-achieving students. While panels C and E show little evidence that assigning Focus labels to non-Title 1 schools improved the test scores of low achievers located in the 10th and 25th percentiles within the school-grade distribution, panels D and F indicate that Title 1 schools with Focus labels did focus on low-achieving students.

In panels I to L, I display the binned means of high-achieving students (at the 75th and 90th percentiles) in non-Title 1 and Title 1 schools. Panels J and L of Figure 2.4 provide some evidence that the improved performance of low achieving students in Title 1 Focus schools did not come at the expense of hurting high-performance students; the two panels do not show any discontinuities at the cutoff. In panels C and D of Table 2.4, I report the analogous results using reading test scores. The estimates in panels C and D are small and insignificant, providing little evidence that Title 1 Focus schools improved student reading test scores.

The nonparametric RD estimates that include baseline covariates are reported in panels A and B of Table 2.4 and tell an identical story. I find that assigning Focus labeling was only effective for schools facing financial sanctions. Furthermore, the estimates across the performance distribution from columns (3) to (7) indicate that the accountability system with proper consequences in fact increases the performance of low-performing students without harming high-performing students.[15] In panels C and D of Table 2.4, I report the analogous results using reading test scores. The estimates in both panels are small and insignificant, providing little evidence that Title 1 Focus schools improved student reading test scores.

Table 2.4 clearly shows significant evidence that receiving the Focus label influenced student math scores for Title 1 schools only. My preferred interpretation to this result is that the stigma that might have been linked with the Focus labels did not push schools to respond to the accountability system unless the follow-up financial incentives were coupled with the labels. This interpretation is in line with previous research (e.g. Hanushek and Raymond, 2005; Saw et al., 2017) which finds that giving an accountability grade without any consequences is not effective in improving student performance.

The preferred interpretation, however, needs to be viewed with caution for two reasons. First, while all coefficients on receiving the label for Title 1 schools are larger than the coefficients for non-Title 1 schools, it is only significantly different from the non-Title 1 coefficient for the 25th percentile outcome (see p-values at the bottom of panel B). The t-tests that compare the coefficients between non-Title 1 and Title 1 schools are based on the estimated coefficient and the standard error

[15] Figure 2.8 displays how the point estimates reported in panel A are sensitive to the bandwidth. The figure clearly indicates that the estimated results are robust to bandwidth selection.

of α_2 in Eq.(2). However, I should note that the fully saturated regression model is demanding of the data, making difficult to detect the moderate coefficient differences between Title 1 and non-Title 1 schools. Second, there are other potential reasons why Title 1 schools may have different responses than non-Title 1 schools. For example, Title 1 schools with the Focus label were low-performing schools (0.226 SD for math and 0.06 SD for reading) than non-Title 1 schools with the label (0.659 SD for math and 0.162 SD for reading). Thus, Title 1 schools may have been able to improve student test scores more easily than non-Title1 schools.

Given that prior research finds that charter schools are more responsive to the accountability rating (Baude, 2015) and that 162 out of 164 charter schools in my sample received Title 1 funds, it is possible that the heterogeneous treatment effects reported in Table 2.4 are simply driven by charter schools and not driven by increased financial incentives among Title 1 schools. To address this concern, Table 2.10 shows the point estimates without charter schools. The estimated results using both math and reading scores are consistent with the baseline estimates in Table 2.4.

Finally, I use an additional year of post-NCLB waiver data (2013-2014 school year) and examine whether the positive achievement effects are robust to this addition. I should note that the MDE administered a new assessment test (M-STEP) in the 2014-2015 school year. Since the new test was calibrated to a different scale and was taken in Spring 2015 rather than Fall 2014, I restrict the post-NCLB waiver data to the 2013-2014 school year. Next, I pool two post-program years and examine the impact of receiving the Focus label in year T (where $T = 2012, 2013$) on following-year outcomes in year T+1. In Table 2.11, I report point estimates with an additional year of data. Once again, schools facing financial threats showed improvements in average math scores with a rightward shift of the math scoring distribution, although point estimates are slightly attenuated compared to baseline estimates.

2.5.3 Heterogeneous Effects by Competitive Pressure

I find little evidence that simply assigning the Focus labels in year T did not spur increased student achievement in the following year. However, it is still possible that schools with many alternative schooling options nearby would have faced increased competitive pressure if they received one of the accountability labels. For example, if parents regarded a Focus label as an indicator of low quality, then a Focus school surrounded by many non-Focus schools in the same district, charters, or open enrollment options in other districts may have been at a higher risk of losing students than a Focus school with few alternatives, indicating that the labeling effects may have depended on the choice environment.

To access the heterogeneous labeling effects by competitive pressure, I first need to measure the degree of school competition from nearby schools.[16] Chakrabarti(2014) uses the number of schools that pass the benchmark for Adequate Yearly Progress (AYP) within a certain radius of AYP-failed schools under NCLB to examine whether the response of AYP-failed schools is contingent on the extent of school competition they faced. The reason is that if students switch to other schools because of the stigma attached to an "AYP-failed" label, they are less likely to choose another AYP-failed school. Her measure, however, may not depict the pressures that public schools face from alternative schooling options if nearby schools do not include the grades that a public school offers. For example, a K-5 public school with many nearby charter schools that offer grades 6 to 8 does not see increased alternative choice options as threats. To better reflect the pressure from alternative school options, I consider nearby schools as competitors of each public school given the accountability label only if (1) nearby schools do not receive the given label and (2) neighboring schools provide any of the grades taught in that public school.

Next, I geocode all public schools using their physical addresses and measure the crow's-flight distance between all public school pairs. I then count nearby competitors within a 2-mile and 5-mile radius of a given public school including public schools in the same district, charter schools, and public schools in neighboring districts that accept students from the given public school by opting into the IDSC.[17] With these measures in hand, I define a public school as facing less school

[16] Prior research that examines the impact of competitive pressure on student achievement generally uses the number of nearby schools, minimum distance to nearby schools, or the share of total enrollment in nearby schools (Chakrabarti and Roy, 2016; Figlio and Hart, 2014; Chakrabarti, 2013b; Imberman, 2011; Bettinger, 2005).

[17] I use the administrative enrollment history dataset for the universe of Michigan students to track whether a student used the IDSC program. With this information, I define that a student in a given public school, *s*, in a given district, *d*, can access public schools in neighboring districts if the neighboring districts allow more than 10 IDSC students from

competition if the number of nearby competitors is below the median. Similarly, a public school faces more school competition if the number of nearby competitors is above the median. I should note that I do not include the interaction term between the indicator of receiving Focus labeling and the competitive measures to examine the heterogeneous effects because estimating the parametric RD regression is likely severely biased under model misspecification.[18] I thus report nonparametric RD estimates within the two subsamples defined above. Finally, one may concern that my competition measure based on counts of alternative schooling options merely captures the degree of urbanity of a given region. However, I prefer the count-based measure of school competition because teachers are more likely to know how many alternative options nearby than how small or large those schools are relative to their school. Nevertheless, Table 2.13 reports the estimates in which I measure the number of students in alternative schools within a 2-mile and 5-mile radius of a given public schools, divided by the total number of students in all schools; the results are similar regardless of the choice of competition measures.

Before presenting the estimated results by number of alternative schooling options, I explore whether pre-treatment covariates are continuous at the cutoff within the two subsamples of schools that faced either few or many schooling options. In Table 2.12, I report the mean differences of various control variables reported in Table 2.1 near the cutoff using the nonparametric RD regression. The table indicates that Focus schools with few (many) alternatives are generally similar to non-Focus schools with few (many) alternatives across the pre-treatment control variables. This evidence supports the validity of the RD design within the subsample.

Table 2.5 shows the estimated results that consider all alternative options within a 2-mile and 5-mile radius of a given public school. Panel A shows the results using the full sample without charter schools to illustrate how the average treatment effect is different from the effect in the subsample. In panels B and C of Table 2.5, I contrast the estimated results in the two subsamples, which are defined using a 2-mile radius. For schools in the uncompetitive choice environment,

the given district (*d*).

[18] Hsu et al. (2016) use Monte Carlo simulation and show that the interaction term method is severely biased under model misspecification even when the parametric model is estimated using data close to the cutoff of the running variable.

Focus schools responded to a squeezing of the math scoring distribution. Columns (3) to (7) of panel B indicate that the compression of the math distribution was driven by the increased performance of low-achieving students, while the performance of high-achieving students was the same. On the other hand, for schools in the competitive environment, receiving the Focus label increased the average math score, but I do not find any evidence of the reduced variability of the within-math-score distribution; across the math-scoring distribution, receiving the Focus label shifts the math distribution to the right. This finding is interesting because prior research that exploits plausibly exogenous variation in charter school penetration finds that increased charter school penetration negatively influences the math and reading test scores of nearby public schools (Imberman, 2011)[19]

One explanation for this heterogeneous treatment effect is that when Focus schools were surrounded by many competitors, they may have been less likely to concentrate on low-achieving students only. The reason is that if they focused on low achieving students only, they were at greater risk of losing high-achieving students than Focus schools with few alternatives. The other explanation is that student mobility induced by the Focus label may have been heterogeneous due to the choice environment. Some parents may have seen the label as an indication that the school would soon improve; hence, more motivated students were more likely to be selected into Focus schools for which there were many alternative options. In the next section, I discuss whether the accountability rating is associated with student mobility in more detail. Panels D and E show the analogous results using the number of all choice options within a 5-mile radius. Whether I use the measure with a 2-mile or 5-mile radius, the results are qualitatively the same.

To assess whether Focus schools facing different types of schooling options responded differently, in Table 2.14, I shows the results using the number of nearby charter competitors or the number of nearby public schools in neighboring districts. Since only 20.7 percent (14.7 percent) of public schools had a charter competitor (a public competitor in the bordering district) within a

[19] Prior studies that used school and student fixed-effects strategies, however, do not find the negative achievement effects of charter school penetration. See for example, Zimmer and Buddin (2009), Booker et al. (2008), Bifulco and Ladd (2006), Sass (2006), and Buddin and Zimmer (2005).

2-mile radius, using the competitive measure with the 2-mile radius may not capture the choice environment that most public schools faced (see Figure 2.9). I thus use the competitive measure with a 5-mile radius from the public school for this analysis.

In panels B and C, I use the number of nearby charter competitors to see whether Focus schools behaved differently according to the number of nearby charter schools. Although point estimates in panels B and C are mostly not significant, the estimates tell the same story. Focus schools with few alternative charter schools improved the performance of low-scoring students, while the scores of high-performing students are comparable to those of scores in control schools near the cutoff. The estimates for schools with many nearby charter competitors are positive across the math scoring distribution, and these findings are qualitatively similar to the results that use all alternative schooling options. Panels D and E show the analogous results using the number of public schools in bordering districts within a 5-mile radius. The results clearly indicate that whether I use the competitive measure from nearby charter schools or public schools, I find the analogous evidence that Focus schools in the competitive environment improved their test scores across the scoring distribution, while Focus schools in the uncompetitive environment reduced the achievement gaps between high- and low-achieving students by increasing the performance of low-achievers.

2.6 Discussion

2.6.1 Does Student Mobility Drive the Results?

My preferred interpretation of the improved performance of Title 1 Focus schools is that Focus labeling with financial sanctions shifted schools' incentives toward concentrating on low-achieving students. In this section, I discuss whether the changes in student-body composition that may have been induced by the accountability labels can explain my findings, and I argue that the preferred interpretation best matches the results, although I acknowledge that I cannot rule out other potential explanations.

To examine whether receiving the Focus labels changed the demographic composition of students in Focus schools, I use several key measures of student composition as outcome variables and estimate the nonparametric RD regression. When Hemelt and Jacob (2017) explore the effects of the accountability labels on student composition, they examine a set of socioeconomic variables such as the percentage of black, Hispanic, economically disadvantaged students, and special education students. On top of these measures, I create math and reading quartile ranks based on students' lagged test scores (2011 Fall test), and use the percentage of students in each math and reading quartile rank for those enrolled in the 2012-2013 school year. I believe examining these variables best reflects the possible compositional changes induced by the accountability label. The reason is that whether students switched to other schools in response to the Focus labels may have depended on the level of student achievement. For example, the parents of low-achieving students may have viewed the Focus label as an indication that their schools were ineffective in teaching their children. On the other hand, some parents of low-performing students may have wanted to send their children to Focus schools if they expected that the label would cause Focus schools to devote more effort to low-scoring students.

In Table 2.6, I display the estimated results using several key measures of student composition. To calculate the percentage of students belonging to the given quartile rank, I use the lagged math and reading scores of students who enrolled in the year immediately following the announcement of Focus schools. In order to be consistent with positive achievement effects for Focus schools, either low achieving students in Focus schools moved away from the school, or high performers were selected into the school. The table shows little evidence of the compositional impact of the accountability labels, since I do not find any sizable or consistent changes in the composition of high- and low-scoring students, nor do I find changes in the racial or socioeconomic composition in the schools.

The overall null effects in Table 2.6 may mask some heterogeneity stemming from the choice environment around schools. When parents have many choice options near the current school, it may be less costly for them to change schools. In Table 2.15, I examines the heterogeneous effects of the Focus label on the composition of students in relation to the number of alternative choice options defined in the previous section. Regardless of whether public schools faced competitive pressure, I do not find any evidence that the Focus label influenced the student composition of the schools. One explanation of the null effect is that parents may not have had enough time to respond to the accountability labels. The public announcement of the list of Focus schools came in August 2012, and parents thus had a maximum of one month before the new school year began to search for and choose alternative schooling options.[20]

2.6.2 Medium-Run Effects of Receiving a Financial Threat

Given that the Focus label with its financial threat moderately improved student test scores despite limited preparation time, one may be concerned that the financial threat may have increased teaching to the test. Then, the achievement gains among low-achieving students may have disappeared rapidly if schools taught to the test as a reaction to the Focus label. It is thus interesting to see whether the educational reform had a persistent impact on students who enrolled in a Title 1 school in the year when the school faced a financial threat. Recall that the MDE administered the new assessment test in year 3 (2014-2015 school year), and the baseline sample includes students from third to eighth grades in year 1 (2012-2013 school year). I consider whether the threat effect translates into students' year 2 test scores (2013-2014 school year).

To examine the medium-run achievement effects of the threat, I use individual-level data to keep track of a set of students from year 0 (2011-2012 school year) to year 2. Next, I divide students into quartile ranks based on pre-determined math test scores (2011 Fall test). The reason is that the persistence of treatment effects may depend on students' initial test score levels. I include students in my sample if they are observed in the data three years in a row (from year 0 to year 2). Among fourth through seventh graders who had lagged math test scores and attended schools in my baseline sample in year 1, this restriction eliminates approximately 12.1 percent of

[20] My data allows me to observe whether students changed schools in the middle of the school year: a very small share of students switched schools or exited the public school system in Michigan.

students from my sample.[21] I then estimate the following local linear model by OLS

$$y_{is2013} = \alpha_0 + \alpha_1 F_{k2012} + \alpha_2 r_{k2012} + \alpha_3 F_{k2012} \times r_{k2012} + \gamma X_{ik2011} + \varepsilon_{is2013} \tag{2.3}$$

The dependent variable y_{ist} is student i's MEAP scores in year 2013. Note that I use the s index to show students' current school in year t, and the k index represents a school that students attended in 2012. The model includes an indicator variable, F_{k2012}, which is equal to 1 when the school that students attended in year 1 was a Focus school and r_{k2012} is the AGI of school k. Note that Eq. (3) includes a set of predetermined variables (X_{ik2011}) reported in Table 2.1 to avoid conditioning school-level controls that could have been influenced by the policy change, as students may have been induced to switch schools. Again, I provide the standard errors that are clustered at the school level. The parameter of interest is α_2, which shows the medium-run effects of the Focus label. I estimate Eq. (3) using the local RD method.[22]

Table 2.7 displays the estimated results of attending Focus schools in year 1 on students' math test scores in year 2. I report the initial effect of the Focus label using year 1 data in panels A, C, and E to show the extent to which the estimated treatment effect is transmitted into the next year. In column (1) of Table 2.7, I estimate Eq. (1) using all students regardless of their initial performance level. From columns (2) to (5), I use students who are in the given test-score-quartile rank. Panels A and B show point estimates that use students who attended all public schools regardless of Title 1 status. The estimates indicate that the immediate short-run improvement in math test scores from attending Focus schools in year 1 decreased by more than 60 percent in 1 year after exposure to the Focus label. This finding is roughly consistent to those in Hemelt and Jacob (2017) who find schools with Focus labels did not improved their average math test scores in year 2 relative to their barely non-Focus counterparts.

When I limit my attention to students who attended Title 1 Focus schools in year 1, however,

[21] The number of students who had lagged test scores in my base sample in the 2012-2013 school year is 409,610. Among those students, 49,360 students are not observed in the 2013-2014 school year. I exclude these students for this analysis.

[22] To better evaluate how the average achievement effects are decomposed by the test score quantile ranking, I use a bandwidth of 0.5 points for the local RD method. The optimal bandwidth for all specifications is close to 0.5 points.

I find some evidence that attending Focus schools that faced financial threats in the first year persistently increased math test scores by 0.12 SD in year 2. This medium-run impact is roughly equivalent in magnitude to the effects in year 1 (0.104 SD). It is interesting to see that the medium-run impact of the financial threat on student achievement is similar across the quartile rank. The analysis therefore indicates some evidence of the persistence of the threat effects for those students who remained in the public school system three years in a row.

Finally, in panels E and F, I prove point estimates that use students who attended non-Title 1 schools. Interestingly, the Focus designation improved average math scores immediately by 0.066 SD in year 1 (See, column 1 in panel E), however, the achievement effects in year 1 dropped to negative 0.1 SD in year 2. Therefore, the null achievement effects in year 2 reported in Hemelt and Jacob (2017) mask substantial heterogeneity with respect to Title 1 status; Title 1 Focus schools were still able to improve the performance of low-achieving students in year 2.

2.7 Conclusion

This paper has examined whether receiving school accountability labels under NCLB waivers influenced student achievement. Specifically, if principals responded to the Focus label by concentrating on low-achieving students, I would expect positive achievement gains in the performance of low-scoring students. I thus examine the achievement effects not only on mean test scores but also on various quantile outcomes within school-grade distributions. Using discontinuous variations in a Focus assignment, I non-parametrically estimate the RD model that compares schools that were just above the cutoff and schools that were barely below the cutoff. Across a myriad of specifications and variations in modeling choices, I show that receiving the Focus label raised the average math achievement as well as the performance of low-scoring students. The effect size of the Focus designation is roughly similar to half of the effects of receiving an F grade on the following year's test scores under Florida's A + accountability system. However, the smaller mean achievement effect is consistent with the fact that schools only had at most three months to prepare for the upcoming tests.

The average achievement that I document for Focus schools, however, masks some heterogeneity. I first examine whether Focus schools with financial incentives responded more to this label. I find evidence that whether Focus schools had financial incentives entirely drives the baseline estimate. This finding, however, should be interpreted with caution because inherent differences between Title 1 and non-Title 1 Focus schools may have caused different responses between the two groups. Next, I study the potential heterogeneous treatment effects depending on whether schools were exposed to school competition. Considering all types of school choice options available to parents, I find evidence that Focus schools only improved the performance of low-achieving students when schools were located in an uncompetitive choice environment, while Focus schools exposed to a competitive choice environment improved their math test scores across the scoring distribution. This may indicate that expanding school choice programs enhances the efficacy of school accountability systems.

This work presents an interesting complement to Hemelt and Jacob (2017) who find a null treatment effect of the Focus designation in the medium- or long-run even when Focus schools received a set of supportive interventions. My study shows that schools responded to the Focus label in the short run without supportive interventions, especially when financial incentives were attached to this label. Combining these two studies may suggest that providing continuing credible incentives to school administrators is essential for the efficacy of school accountability systems.

APPENDICES

APPENDIX A TABLES

Table 2.1: Summary Statistics by Accountability Labeling

	Focus schools (1)	All schools (2)
Accountability results		
Percent Focus	100.00	14.34
Focus running variable	0.48	-0.67
Average math score	0.33	-0.01
Average reading score	0.09	0.01
Student characteristics		
Percent free lunch	32.67	44.87
Percent special education	12.69	13.76
Percent black	11.24	16.85
Percent Hispanic	5.44	6.22
Percent white	70.61	70.26
Percent Asian	7.45	2.71
School characteristics		
Percent elementary	59.43	65.54
Percent K-8	8.54	9.65
Percent middle	32.03	24.81
Percent charter schools	6.76	7.45
Percent title1 schools	65.48	77.08
Percent magnet schools	11.39	14.19
Percent located in urban	28.83	21.49
Total enrollment	491.01	439.44
Number of schools	281	1959

Notes: The table shows the summary statistics of certain key variables for the sample of schools in the 2012-2013 school year. The table compares the means of schools with a focus label as well as all schools.

Table 2.2: Testing Validity of RD Analysis by Using Preprogram Characteristics

	Mean math	Mean reading	%Q1 Math	%Q2 Math	%Q3 Math	%Q4 Math
Focus Schools	-0.050	0.022	1.147	0.237	0.222	-3.187
	(0.056)	(0.025)	(1.471)	(1.003)	(0.791)	(2.053)
Observations	765	474	770	730	565	873
Bandwidth	0.583	0.382	0.588	0.555	0.437	0.649

	%Q1 Reading	%Q2 Reading	%Q3 Reading	%Q4 Reading	% Hispanic	% Black
Focus Schools	0.453	-0.255	-0.198	-0.188	-1.434	-1.551
	(1.386)	(0.729)	(0.602)	(1.729)	(1.065)	(2.615)
Observations	680	685	684	584	669	637
Bandwidth	0.515	0.518	0.517	0.454	0.508	0.490

	% Free lunch	% Special edu	Enrollment	Magnet	K5	Middle
Focus Schools	-0.708	0.398	0.727	-0.001	-0.077	-0.005
	(3.110)	(0.734)	(29.917)	(0.048)	(0.077)	(0.077)
Observations	690	623	742	572	749	674
Bandwidth	0.522	0.482	0.564	0.448	0.569	0.511

Notes: The table shows the estimated discontinuities of baseline control variables for Focus schools. The percentage of students in each math and reading quartile ranks based on 2011 Fall test. In addition to the indicator of passing running variables, each regression includes a linear spline of running variables. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.3: The Effect of Receiving Focus Labeling on Student Achievement

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean	SD	10th	25th	50th	75th	90th
			percentile	percentile	percentile	percentile	percentile
Panel A. 2011 Fall Test (Placebo): Math							
Focus Schools	-0.024	-0.010	-0.006	-0.012	-0.027	-0.038	-0.038
	(0.024)	(0.013)	(0.024)	(0.025)	(0.027)	(0.032)	(0.042)
Observations	2309	2057	1811	2182	1919	2235	2052
Bandwidth	0.573	0.509	0.457	0.536	0.482	0.552	0.508
Panel B. 2011 Fall Test (Placebo): Reading							
Focus Schools	-0.032	-0.011	-0.021	-0.028	-0.018	-0.032	-0.056
	(0.022)	(0.008)	(0.028)	(0.026)	(0.024)	(0.025)	(0.032)
Observations	1790	2364	2240	2303	2426	2066	1464
Bandwidth	0.451	0.587	0.554	0.569	0.601	0.510	0.379
Panel C. 2012 Fall Test: Math							
Focus Schools	0.052**	-0.005	0.053**	0.068**	0.057**	0.037	0.048
	(0.025)	(0.012)	(0.026)	(0.029)	(0.029)	(0.031)	(0.038)
Observations	2211	1985	1974	1747	2189	2087	1904
Bandwidth	0.550	0.497	0.493	0.438	0.546	0.516	0.478
Panel D. 2012 Fall Test: Reading							
Focus Schools	0.020	-0.005	0.024	0.024	0.023	0.021	0.006
	(0.022)	(0.011)	(0.034)	(0.027)	(0.027)	(0.022)	(0.031)
Observations	2081	1879	1949	2162	1759	2207	2162
Bandwidth	0.514	0.472	0.488	0.537	0.445	0.549	0.535

Notes: The table presents the estimated results of receiving Focus labels on various student achievement measures. In addition to the indicator of passing running variables, each regression includes a linear spline of running variables, grade dummies, and a set of pre-determined variables reported in Table 2.1. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.4: Heterogeneity of Receiving Focus Labeling by Title 1 Status

	(1) Mean	(2) SD	(3) 10th percentile	(4) 25th percentile	(5) 50th percentile	(6) 75th percentile	(7) 90th percentile
Panel A. Non-Title 1 Schools: Math							
Focus Schools	0.014	-0.003	0.010	-0.005	0.025	0.007	-0.002
	(0.043)	(0.020)	(0.051)	(0.051)	(0.051)	(0.054)	(0.069)
Observations	584	532	527	476	584	557	511
Bandwidth	0.550	0.497	0.493	0.438	0.546	0.516	0.478
Panel B. Title 1 Schools: Math							
Focus Schools	0.068**	-0.010	0.078**	0.104**	0.067**	0.044	0.063
	(0.029)	(0.014)	(0.033)	(0.035)	(0.034)	(0.037)	(0.045)
Observations	1627	1453	1447	1271	1605	1530	1393
Bandwidth	0.550	0.497	0.493	0.438	0.546	0.516	0.478
Test that Non-Title 1 = Title 1 (p-value)	0.289	0.771	0.256	0.073	0.484	0.569	0.442
Panel C. Non-Title 1 Schools: Reading							
Focus Schools	0.017	0.015	-0.012	0.022	-0.002	-0.010	0.050
	(0.046)	(0.020)	(0.059)	(0.051)	(0.049)	(0.046)	(0.059)
Observations	554	502	522	575	479	584	575
Bandwidth	0.514	0.472	0.488	0.537	0.445	0.549	0.535
Panel D. Title 1 School: Reading							
Focus Schools	0.019	-0.014	0.037	0.033	0.026	0.021	-0.004
	(0.026)	(0.014)	(0.041)	(0.033)	(0.033)	(0.026)	(0.036)
Observations	1527	1377	1427	1587	1280	1623	1587
Bandwidth	0.514	0.472	0.488	0.537	0.445	0.549	0.535
Test that Non-Title 1 = Title 1 (p-value)	0.969	0.227	0.464	0.855	0.623	0.571	0.427

Notes: The table presents the estimated results of receiving Focus labels by Title 1 status. In addition to the indicator of passing running variables, each regression includes a linear spline of running variables, grade dummies, and a set of pre-determined variables reported in Table 2.1. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.5: Heterogeneity of Receiving Focus Labeling by # of Alternative Schooling Options

	(1) Mean	(2) SD	(3) 10th percentile	(4) 25th percentile	(5) 50th percentile	(6) 75th percentile	(7) 90th percentile
Panel A. Full sample							
Focus	0.042*	-0.018	0.057**	0.078**	0.049*	0.003	0.019
	(0.022)	(0.011)	(0.025)	(0.025)	(0.026)	(0.029)	(0.035)
Observations	2009	1605	1558	1792	1909	1650	1862
Bandwidth	0.556	0.453	0.446	0.501	0.528	0.466	0.513
Panel B. Below-median number of all choice options (2 miles)							
Focus	0.031	-0.037**	0.084**	0.066**	0.037	-0.017	-0.019
	(0.030)	(0.015)	(0.034)	(0.033)	(0.034)	(0.039)	(0.047)
Observations	1372	1090	1049	1217	1306	1115	1272
Panel C. Above-median number of all choice options (2 miles)							
Focus	0.073	0.003	0.055	0.126*	0.081	0.043	0.080
	(0.047)	(0.021)	(0.059)	(0.055)	(0.058)	(0.064)	(0.063)
Observations	637	515	509	575	603	535	590
Panel D. Below-median number of all choice options (5 miles)							
Focus	0.035	-0.048**	0.101**	0.073**	0.044	-0.015	-0.038
	(0.034)	(0.015)	(0.038)	(0.037)	(0.038)	(0.044)	(0.050)
Observations	1222	980	944	1081	1155	1001	1123
Panel E. Above-median number of all choice options (5 miles)							
Focus	0.061	0.009	0.035	0.099**	0.066	0.032	0.090*
	(0.039)	(0.018)	(0.047)	(0.047)	(0.048)	(0.051)	(0.052)
Observations	787	625	614	711	754	649	739

Notes: The table presents the heterogeneous estimated results of receiving Focus labels that consider all alternative options within 2 mile (panel B and C) and 5 mile (panel D and E) radius in a given public school. Each regression includes the indicator of passing running variables, each regression includes a linear spline of running variables, grade dummies, and a set of pre-determined variables reported in Table 2.1. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.6: The Effect of Assigning the Focus Label on Student Composition in 2012-2013 school year

	%Q1 Math	%Q2 Math	%Q3 Math	%Q4 Math	%Q1 Reading	%Q2 Reading
Focus Schools	-0.139	0.190	-0.291	0.162	0.070	-0.196
	(0.922)	(0.621)	(0.943)	(1.038)	(0.842)	(0.824)
Observations	466	664	577	516	670	559
Bandwidth	0.379	0.513	0.454	0.409	0.516	0.441
	%Q3 Reading	%Q4 Reading	%Black	%Hispanic	%Free Lunch	%Special edu
Focus Schools	0.429	-0.397	0.387	0.085	-0.119	0.397
	(0.699)	(1.062)	(0.305)	(0.222)	(0.781)	(0.407)
Observations	769	584	581	694	679	626
Bandwidth	0.597	0.461	0.460	0.538	0.524	0.498

Notes: The table presents the estimated discontinuities using several key measures of student composition. The percentage of students in each math and reading quartile ranks based on 2011 Fall test. Each regression includes the indicator of passing running variables, each regression includes a linear spline of running variables, and a set of pre-determined variables reported in Table 2.1. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.7: Effect of Attending Focus Schools in Year 1 on Student Math Score in Year 1 & 2

	(1)	(2)	(3)	(4)	(5)
	Mean	1st quartile	2nd quartile	3rd quartile	4th quartile
Panel A. Year 1 All school (2012-2013 school year)					
Focus	0.074**	0.057*	0.066**	0.066**	0.075**
	(0.023)	(0.030)	(0.027)	(0.030)	(0.032)
Observations	125310	19925	30897	35314	39174
Panel B. Year 2 All schools (2013-2014 school year)					
Focus	0.025	0.044	0.024	0.012	-0.040
	(0.029)	(0.032)	(0.035)	(0.044)	(0.048)
Observations	125310	19925	30897	35314	39174
Panel C. Year 1 Title 1 schools only (2012-2013 school year)					
Focus	0.104**	0.103**	0.115**	0.098**	0.120**
	(0.029)	(0.036)	(0.034)	(0.041)	(0.039)
Observations	76720	13948	19893	21456	21423
Panel D. Year 2 Title 1 schools only (2013-2014 school year)					
Focus	0.120**	0.107**	0.092**	0.106**	0.100**
	(0.033)	(0.037)	(0.042)	(0.047)	(0.046)
Observations	77315	13948	19893	21456	21423
Panel E. Year 1 Non-Title 1 schools only (2012-2013 school year)					
Focus	0.066	0.011	0.021	0.028	0.066
	(0.045)	(0.059)	(0.045)	(0.051)	(0.056)
Observations	48590	5977	11004	13858	17751
Panel F. Year 2 Non-Title 1 schools only (2013-2014 school year)					
Focus	-0.100**	-0.029	-0.040	-0.105	-0.113*
	(0.043)	(0.056)	(0.061)	(0.077)	(0.061)
Observations	48590	5977	11004	13858	17751

Notes: The table shows the estimated results of attending Focus schools in year 1 on student math test scores in year 2. Panels A and B shows point estimates that use students attended all public schools and panels C and D limit students who attended Title 1 schools. The percentage of students in each math and reading quartile ranks based on 2011 Fall test. Each regression includes the indicator of passing running variables, each regression includes a linear spline of running variables, and a set of pre-determined variables reported in Table 2.1. Note that all estimated results are calculated from individual level data. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.8: Sensitivity Check: Without Controls

	(1) Mean	(2) SD	(3) 10th percentile	(4) 25th percentile	(5) 50th percentile	(6) 75th percentile	(7) 90th percentile
Panel A. Focus: Math							
Focus Schools	0.070	-0.014	0.074	0.119	0.082	0.056	0.042
	(0.070)	(0.013)	(0.055)	(0.077)	(0.079)	(0.082)	(0.074)
Observations	1858	2225	2111	1457	1832	1780	2176
Bandwidth	0.467	0.550	0.521	0.374	0.461	0.449	0.535
Panel B. Focus: Reading							
Focus Schools	0.064	-0.008	0.057	0.061	0.080	0.058	0.037
	(0.055)	(0.009)	(0.062)	(0.062)	(0.060)	(0.053)	(0.054)
Observations	1756	3040	1946	1911	1498	1824	1924
Bandwidth	0.439	0.746	0.487	0.479	0.385	0.460	0.483

Notes: The table compares the baseline estimates reported in Table 2.3 to the estimates without predetermined control variables. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.9: Mean Difference of Pre-determined Covariates by Title 1 Status

Panel A. Non-Title 1 Schools

	(1)	(2)	(3)	(4)	(5)	(6)
Lagged Outcomes	Mean math	Mean reading	%Q1 Math	%Q2 Math	%Q3 Math	%Q4 Math
Focus Schools	-0.001	-0.040	1.650	-0.963	-2.632	1.321
	(0.142)	(0.048)	(2.658)	(2.450)	(1.778)	(5.595)
Observations	169	166	176	195	127	176
Bandwidth	0.448	0.441	0.469	0.512	0.344	0.469
	%Q1 Reading	%Q2 Reading	%Q3 Reading	%Q4 Reading	% Hispanic	% Black
Focus Schools	0.062	-0.049	-1.456	1.548	2.371	2.084
	(2.298)	(1.568)	(1.209)	(3.881)	(1.519)	(3.952)
Observations	181	259	166	186	176	211
Bandwidth	0.484	0.671	0.436	0.496	0.470	0.558
	% Free lunch	% Special edu	Enrollment	Magnet	K5	Middle
Focus Schools	2.290	1.143	-121.004	0.008	0.035	-0.001
	(5.446)	(1.388)	(103.578)	(0.147)	(0.195)	(0.208)
Observations	214	188	193	152	255	229
Bandwidth	0.573	0.503	0.510	0.406	0.663	0.597

Panel B. Title 1 Schools

	Mean math	Mean reading	%Q1 Math	%Q2 Math	%Q3 Math	%Q4 Math
Focus Schools	-0.085	0.014	1.595	1.240	1.007	-4.434
	(0.089)	(0.032)	(2.557)	(1.557)	(0.973)	(3.401)
Observations	311	445	346	311	628	301
Bandwidth	0.353	0.484	0.394	0.353	0.666	0.337
	%Q1 Reading	%Q2 Reading	%Q3 Reading	%Q4 Reading	% Hispanic	% Black
Focus Schools	0.440	0.260	0.150	-1.039	-3.393**	-3.403
	(2.122)	(1.039)	(0.873)	(2.246)	(1.757)	(3.816)
Observations	402	521	535	373	462	483
Bandwidth	0.445	0.557	0.568	0.413	0.497	0.513
	% Free lunch	% Special edu	Enrollment	Magnet	K5	Middle
Focus Schools	-1.645	-0.421	66.747	0.023	-0.156	0.043
	(4.324)	(1.319)	(48.516)	(0.080)	(0.103)	(0.091)
Observations	402	346	307	443	511	411
Bandwidth	0.447	0.392	0.347	0.482	0.543	0.453

Notes: The table shows the estimated discontinuities of baseline control variables within Title 1 and non-Title 1 schools respectively. In addition to the indicator of passing running variables, each regression includes a linear spline of running variables. Cluster standard errors are reported. *Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.10: Sensitivity Checks: Excluding Charter Schools

	(1) Mean	(2) SD	(3) 10th percentile	(4) 25th percentile	(5) 50th percentile	(6) 75th percentile	(7) 90th percentile
Panel A. Non-Title 1 Schools: Math							
Focus	0.008	-0.001	-0.004	0.007	0.018	0.000	-0.003
	(0.045)	(0.021)	(0.054)	(0.046)	(0.054)	(0.057)	(0.065)
Observations	526	491	428	596	502	470	566
Bandwidth	0.498	0.467	0.406	0.564	0.477	0.440	0.533
Panel B. Title 1 Schools: Math							
Focus	0.047	-0.024	0.062	0.102**	0.043	-0.013	0.024
	(0.034)	(0.015)	(0.039)	(0.034)	(0.040)	(0.042)	(0.044)
Observations	1272	1173	1002	1464	1210	1093	1378
Bandwidth	0.498	0.467	0.406	0.564	0.477	0.440	0.533

Notes: The table compares the baseline estimates reported in panels A and B of Table 2.4 to the estimates without charter schools. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.11: Sensitivity Checks: Including Additional Years

	(1) Mean	(2) SD	(3) 10th percentile	(4) 25th percentile	(5) 50th percentile	(6) 75th percentile	(7) 90th percentile
Panel A. Math							
Focus	0.030	0.014	0.017	0.005	0.048	0.037	0.041
	0.026	0.012	0.029	0.031	0.030	0.032	0.042
Observations	803	567	813	552	753	720	572
Bandwidth	0.524	0.379	0.531	0.371	0.495	0.479	0.382
Panel B. Reading							
Focus	0.053**	0.000	0.045**	0.071**	0.053**	0.055**	0.058**
	0.018	0.008	0.017	0.021	0.021	0.024	0.027
Observations	2608	1783	2629	1742	2443	2373	1798
Bandwidth	0.524	0.379	0.531	0.371	0.495	0.479	0.382

Notes: The table compares the baseline estimates reported in panels A and B of Table 2.4 to the estimates with an additional year of data. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.12: Mean Difference of Pre-determined Covariates by Alternative Schooling Options

Panel A. Below-median number of all choice options (2 miles)

	(1)	(2)	(3)	(4)	(5)	(6)
Lagged Outcomes	Mean math	Mean reading	%Q1 Math	%Q2 Math	%Q3 Math	%Q4 Math
Focus Schools	-0.036	0.005	1.641	0.737	0.001	-1.760
	(0.099)	(0.035)	(2.015)	(1.616)	(1.358)	(4.025)
Observations	380	332	541	482	269	379
Bandwidth	0.459	0.407	0.627	0.563	0.331	0.456
	%Q1 Reading	%Q2 Reading	%Q3 Reading	%Q4 Reading	% Hispanic	% Black
Focus Schools	-0.824	-0.142	-0.098	0.771	-0.210	1.170
	(2.097)	(1.423)	(0.842)	(3.047)	(1.442)	(2.853)
Observations	327	390	455	328	338	284
Bandwidth	0.402	0.469	0.531	0.403	0.415	0.350
	% Free lunch	% Special edu	Enrollment	Magnet	K5	Middle
Focus Schools	-4.196	-0.654	12.446	-0.040	-0.268	0.330**
	(5.416)	(1.274)	(56.373)	(0.084)	(0.171)	(0.158)
Observations	287	335	399	405	305	261
Bandwidth	0.355	0.410	0.476	0.484	0.381	0.322

Panel B. Above-median number of all choice options (2 miles)

	Mean math	Mean reading	%Q1 Math	%Q2 Math	%Q3 Math	%Q4 Math
Focus Schools	-0.025	0.039	0.798	-1.365	1.004	-0.311
	(0.143)	(0.050)	(3.736)	(2.433)	(1.637)	(5.691)
Observations	180	203	180	204	205	178
Bandwidth	0.468	0.521	0.470	0.524	0.528	0.464
	%Q1 Reading	%Q2 Reading	%Q3 Reading	%Q4 Reading	% Hispanic	% Black
Focus Schools	0.195	0.454	-0.133	-0.481	-3.676	-5.129
	(3.591)	(1.664)	(1.430)	(3.488)	(2.801)	(6.873)
Observations	180	197	207	169	229	191
Bandwidth	0.472	0.507	0.530	0.445	0.579	0.487
	% Free lunch	% Special edu	Enrollment	Magnet	K5	Middle
Focus Schools	-2.246	1.162	-9.744	0.243	0.172*	-0.069
	(7.807)	(1.973)	(51.674)	(0.149)	(0.099)	(0.066)
Observations	176	216	161	110	170	165
Bandwidth	0.463	0.555	0.411	0.301	0.449	0.425

Notes: The table shows the estimated discontinuities of baseline control variables within the subsamples. Panel A (Panel B) contains schools that the number of nearby competitors is below (above) the median. Each regression includes the indicator of passing running variables and a linear spline of running variables. Cluster standard errors are reported. *Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.13: Heterogeneity of Receiving Focus Labeling by Market Share Based Competition Measure

	(1) Mean	(2) SD	(3) 10th percentile	(4) 25th percentile	(5) 50th percentile	(6) 75th percentile	(7) 90th percentile
Panel A. Full sample							
Focus	0.042*	-0.018	0.057**	0.078**	0.049*	0.003	0.019
	(0.022)	(0.011)	(0.025)	(0.025)	(0.026)	(0.029)	(0.035)
Observations	2009	1605	1558	1792	1909	1650	1862
Bandwidth	0.556	0.453	0.446	0.501	0.528	0.466	0.513
Panel B. Below-median number of all choice options (2 miles)							
Focus	0.046	-0.038**	0.114**	0.082**	0.040	-0.008	0.007
	(0.033)	(0.016)	(0.041)	(0.037)	(0.038)	(0.044)	(0.050)
Observations	1144	912	880	1014	1089	931	1060
Panel C. Above-median number of all choice options (2 miles)							
Focus	0.042	-0.005	0.022	0.080	0.061	0.017	0.029
	(0.042)	(0.018)	(0.048)	(0.049)	(0.050)	(0.054)	(0.056)
Observations	865	693	678	778	820	719	802
Panel D. Below-median number of all choice options (5 miles)							
Focus	0.009	-0.042**	0.079**	0.045	0.006	-0.039	-0.037
	(0.034)	(0.016)	(0.036)	(0.037)	(0.037)	(0.045)	(0.051)
Observations	1099	887	853	975	1042	903	1013
Panel E. Above-median number of all choice options (5 miles)							
Focus	0.075**	0.005	0.042	0.110**	0.091**	0.046	0.073
	(0.037)	(0.018)	(0.046)	(0.042)	(0.044)	(0.048)	(0.050)
Observations	910	718	705	817	867	747	849

Notes: The table presents the heterogeneous estimated results of receiving Focus labels with the different competition measure (alternative school penetration) within 2 mile (panel B and C) and 5 mile (panel D and E) radius in a given public school. Each regression includes the indicator of passing running variables, each regression includes a linear spline of running variables, grade dummies, and a set of pre-determined variables reported in Table 2.1. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.14: Heterogeneity of Receiving Focus Labeling by # of Charter and IDSC Options

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean	SD	10th	25th	50th	75th	90th
			percentile	percentile	percentile	percentile	percentile
Panel A. Full sample							
Focus	0.042*	-0.018	0.057**	0.078**	0.049*	0.003	0.019
	(0.022)	(0.011)	(0.025)	(0.025)	(0.026)	(0.029)	(0.035)
Observations	2009	1605	1558	1792	1909	1650	1862
Bandwidth	0.556	0.453	0.446	0.501	0.528	0.466	0.513
Panel B. Below-median number of charter schools (5 miles)							
Focus	0.029	-0.021	0.058	0.064*	0.029	-0.009	-0.007
	(0.031)	(0.016)	(0.036)	(0.033)	(0.035)	(0.041)	(0.049)
Observations	1213	978	940	1088	1153	1002	1127
Panel C. Above-median number of charter schools (5 miles)							
Focus	0.060	-0.011	0.058	0.095*	0.072	0.028	0.058
	(0.043)	(0.019)	(0.052)	(0.051)	(0.052)	(0.057)	(0.058)
Observations	796	627	618	704	756	648	735
Panel D. Below-median number of IDSC options (5 miles)							
Focus	0.035	-0.037**	0.076*	0.073*	0.034	-0.025	-0.008
	(0.037)	(0.017)	(0.040)	(0.039)	(0.045)	(0.050)	(0.055)
Observations	1112	864	830	987	1060	889	1030
Panel E. Above-median number of IDSC options (5 miles)							
Focus	0.056*	0.005	0.041	0.088*	0.070*	0.036	0.070
	(0.032)	(0.016)	(0.035)	(0.034)	(0.036)	(0.041)	(0.047)
Observations	897	741	728	805	849	761	832

Notes: The table presents the heterogeneous estimated results of receiving Focus labels that use the number of nearby charter schools (panel B and C) or the number of nearby public schools in neighboring districts (panel D and E). Each regression includes the indicator of passing running variables, each regression includes a linear spline of running variables, grade dummies, and a set of pre-determined variables reported in Table 2.1. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 2.15: The Effect of Assigning Focus Labels on Student Composition in 2012-2013 school year: # of Alternative Schooling Options

Panel A. Below-median number of all choice options (5 miles)

	(1)	(2)	(3)	(4)	(5)	(6)
	%Q1 Math	%Q2 Math	%Q3 Math	%Q4 Math	%Q1 Reading	%Q2 Reading
Focus Schools	0.287	1.747*	-1.203	-1.283	0.974	-1.504
	(1.279)	(0.836)	(1.196)	(1.164)	(1.025)	(0.935)
Observations	240	367	308	289	416	314
Bandwidth	0.347	0.506	0.432	0.414	0.567	0.445

	%Q3 Reading	%Q4 Reading	%Black	%Hispanic	%Free Lunch	%Special edu
Focus Schools	0.350	0.155	0.024	0.293	1.011	0.012
	(0.904)	(1.229)	(0.254)	(0.367)	(0.826)	(0.552)
Observations	375	379	331	354	388	378
Bandwidth	0.518	0.520	0.462	0.491	0.534	0.533

Panel B. Above-median number of all choice options (5 miles)

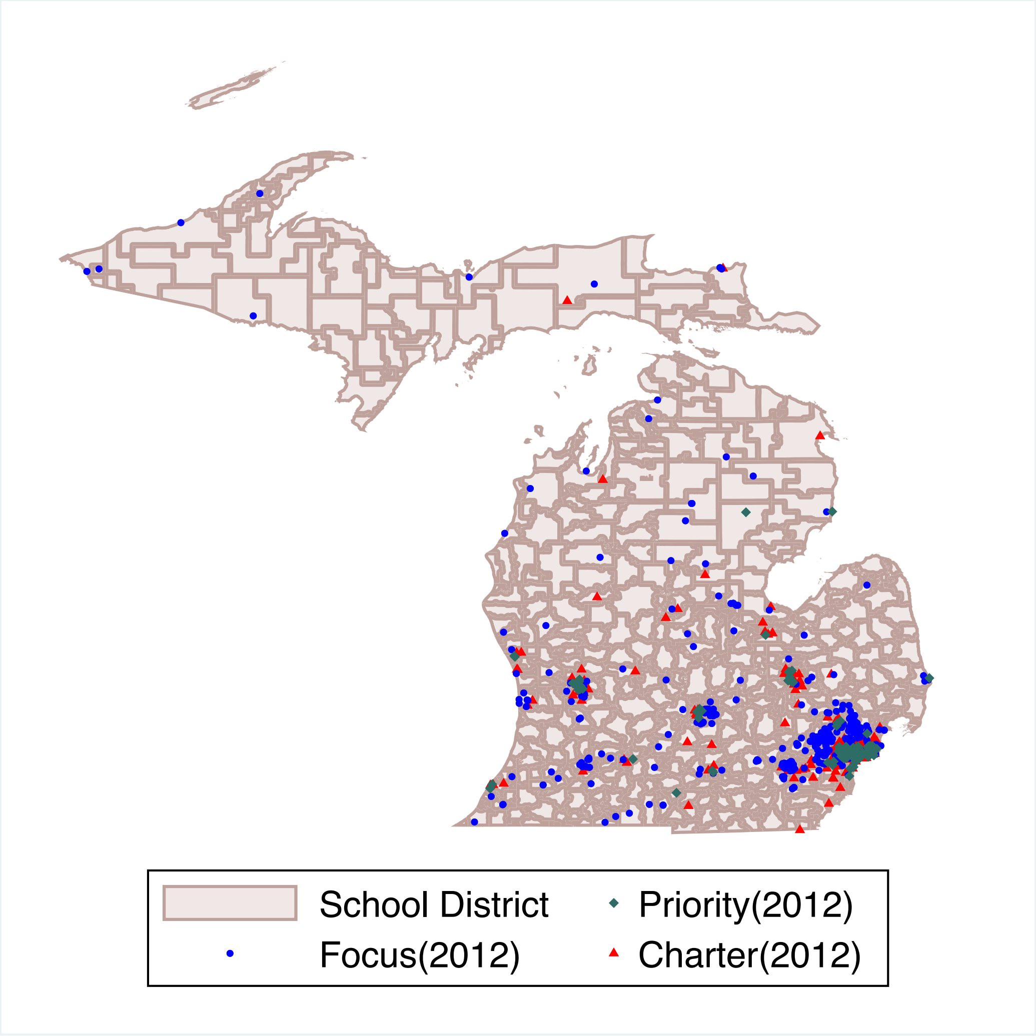
	%Q1 Math	%Q2 Math	%Q3 Math	%Q4 Math	%Q1 Reading	%Q2 Reading
Focus Schools	0.022	-1.749*	0.202	1.589	-1.567	0.788
	(1.523)	(1.017)	(1.834)	(1.852)	(1.305)	(1.461)
Observations	161	247	207	200	271	210
Bandwidth	0.347	0.506	0.432	0.414	0.567	0.445

	%Q3 Reading	%Q4 Reading	%Black	%Hispanic	%Free Lunch	%Special edu
Focus Schools	0.953	-0.043	0.504	0.113	-1.968	0.933
	(1.352)	(1.683)	(0.623)	(0.317)	(1.305)	(0.710)
Observations	255	255	218	237	260	255
Bandwidth	0.518	0.520	0.462	0.491	0.534	0.533

Notes: The table shows the heterogeneous effects of Focus label on the composition of students by the number of alternative schooling options. Each panel presents the estimated discontinuities using several key measures of student composition. The percentage of students in each math and reading quartile ranks based on 2011 Fall test. Each regression includes the indicator of passing running variables, each regression includes a linear spline of running variables, and a set of pre-determined variables reported in Table 2.1. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

APPENDIX B FIGURES

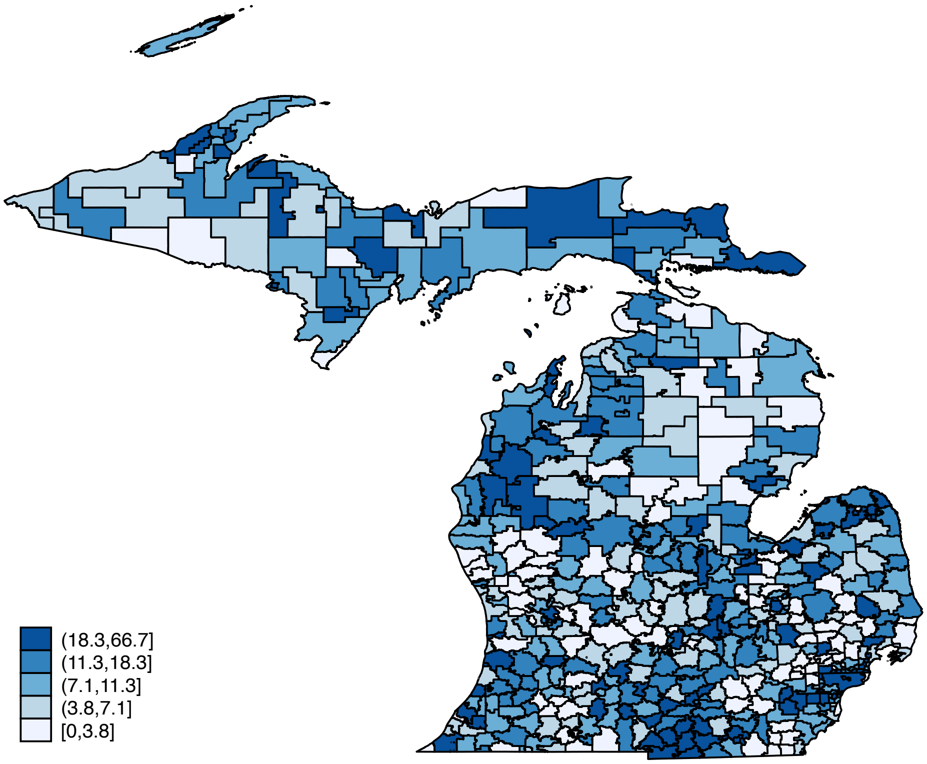
Figure 2.1: Geographic Distribution of Priority, Focus, and Charter Schools in 2012



‘

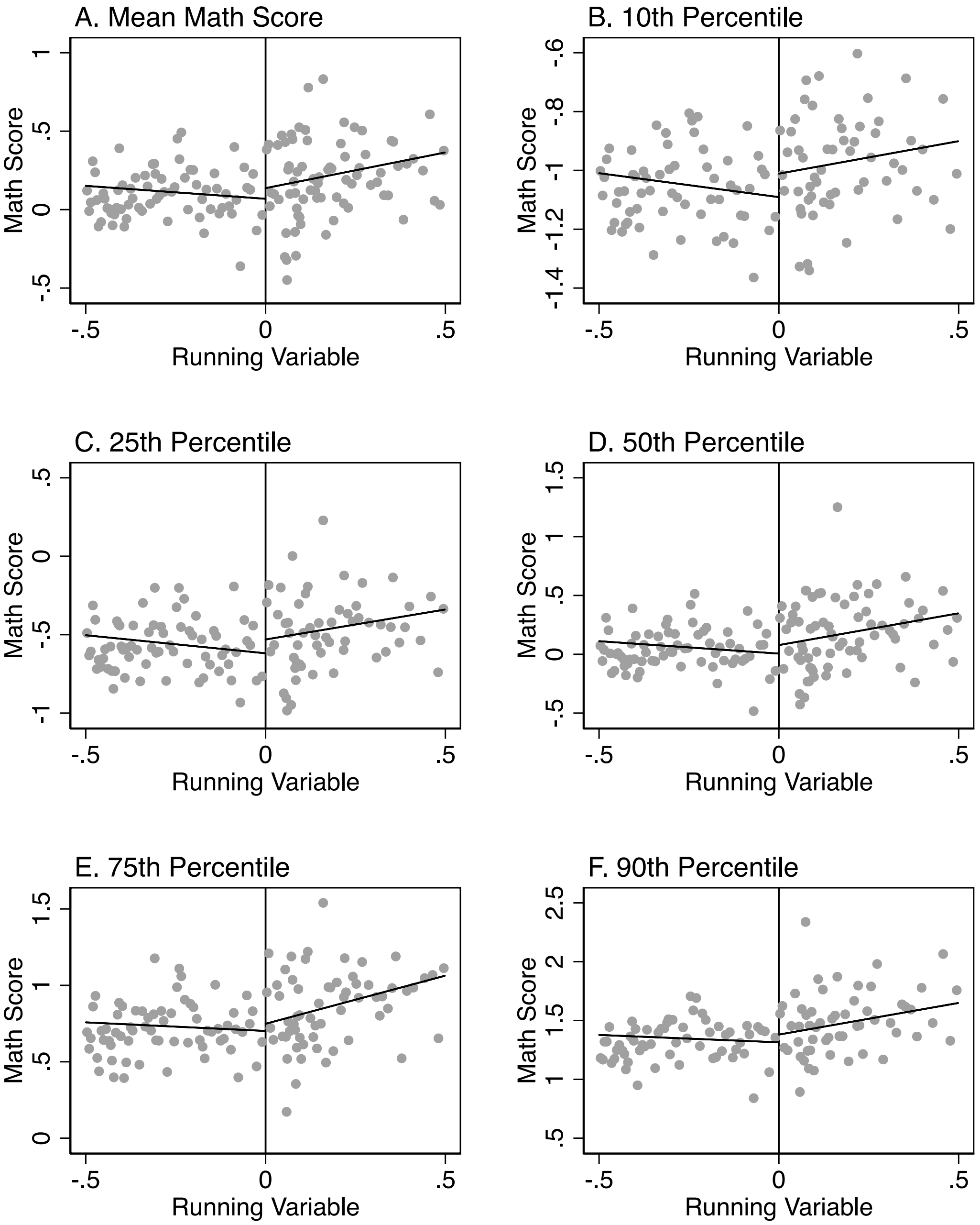
Notes: The figure maps the distribution of public schools that received Priority and Focus labels as well as charter schools during the 2011-2012 school year across Michigan.

Figure 2.2: Spatial Variation in Share of Transfer Students Using IDSC in 2012



Notes: The figure displays the spatial variation in the share of transfer students using the IDSC program in the 2011-2012 school year. Source: Author's estimations from data provided by the Michigan Department of Education

Figure 2.3: Effect of Receiving Focus Label on Math Achievement



Notes: The figure illustrates the effects of receiving Focus labels on various student math achievement outcomes, including the mean and the different points in the within-school- grade achievement distribution. The x-axis shows schools' AGI relative to the cutoffs that determine Focus labeling for the 2012-2013 school year, and the y-axis describes the mean and various quantiles of the math test scores. To better visualize outcomes around the cutoff, I only include schools with AGI within ± 0.5 from the cutoff. Each of the points in these figures indicates the average outcome measures of mean and various quantile outcomes collapsed into bins.

Figure 2.4: Heterogeneity of Focus Labeling Effect: Receiving Title1 Fund

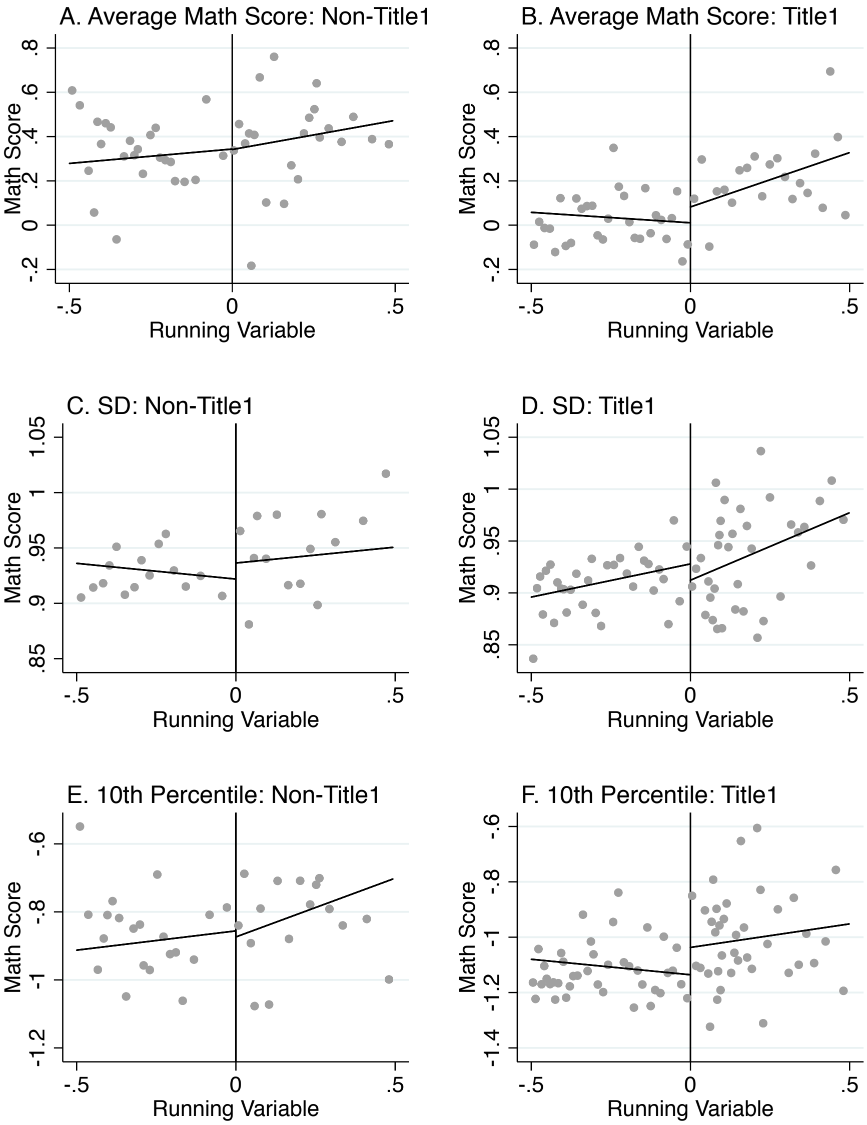
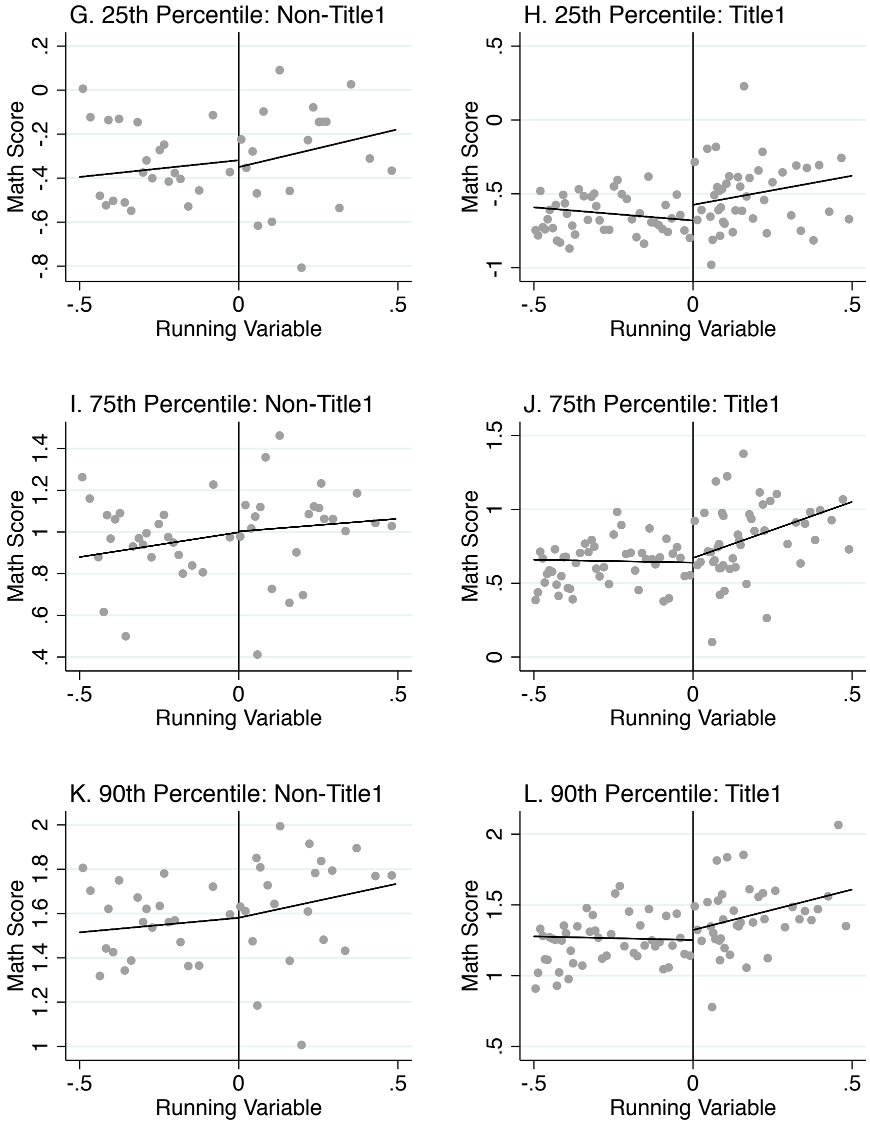
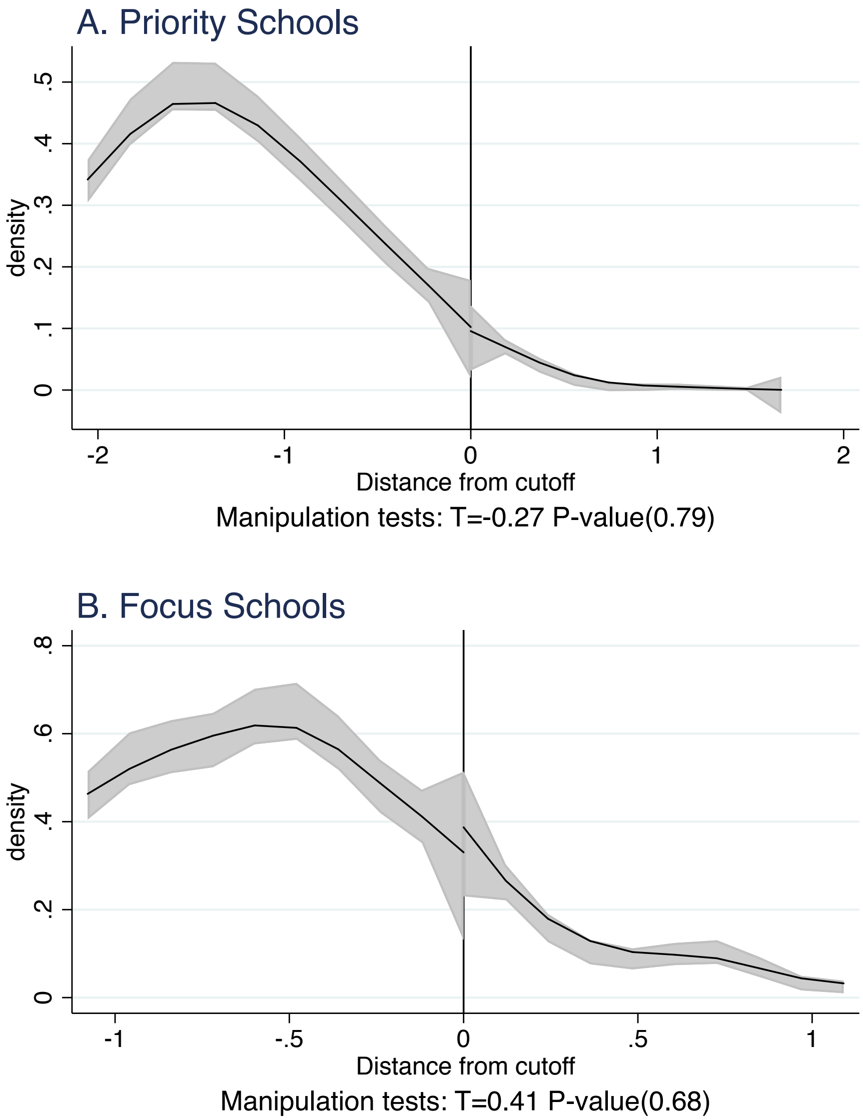


Figure 2.4: (cont'd)



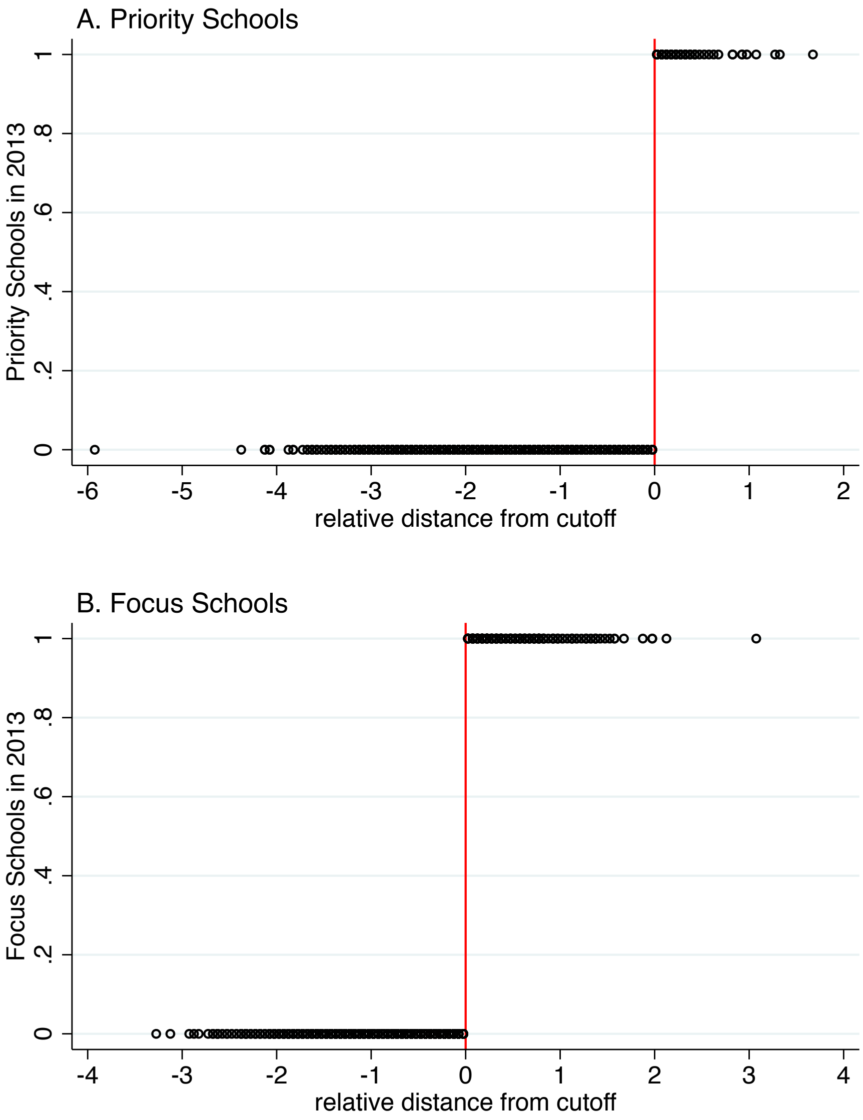
Notes: The figure display the mean and the different percentiles of math scores in the within-school-grade achievement distribution observed in non-Title 1 schools and Title 1 schools. The x-axis shows schools' AGI relative to the cutoffs that determine Focus labeling for the 2012-2013 school year, and the y-axis describes the mean and various quantiles of the math test scores. Each of the points in these figures indicates the average outcome measures of mean and various quantile outcomes collapsed into bins. Left (right) panels in Figure 2.4 shows the achievement outcomes for non-Title 1 (Title 1) schools.

Figure 2.5: Density of Running Variables across the Distance from Cutoff



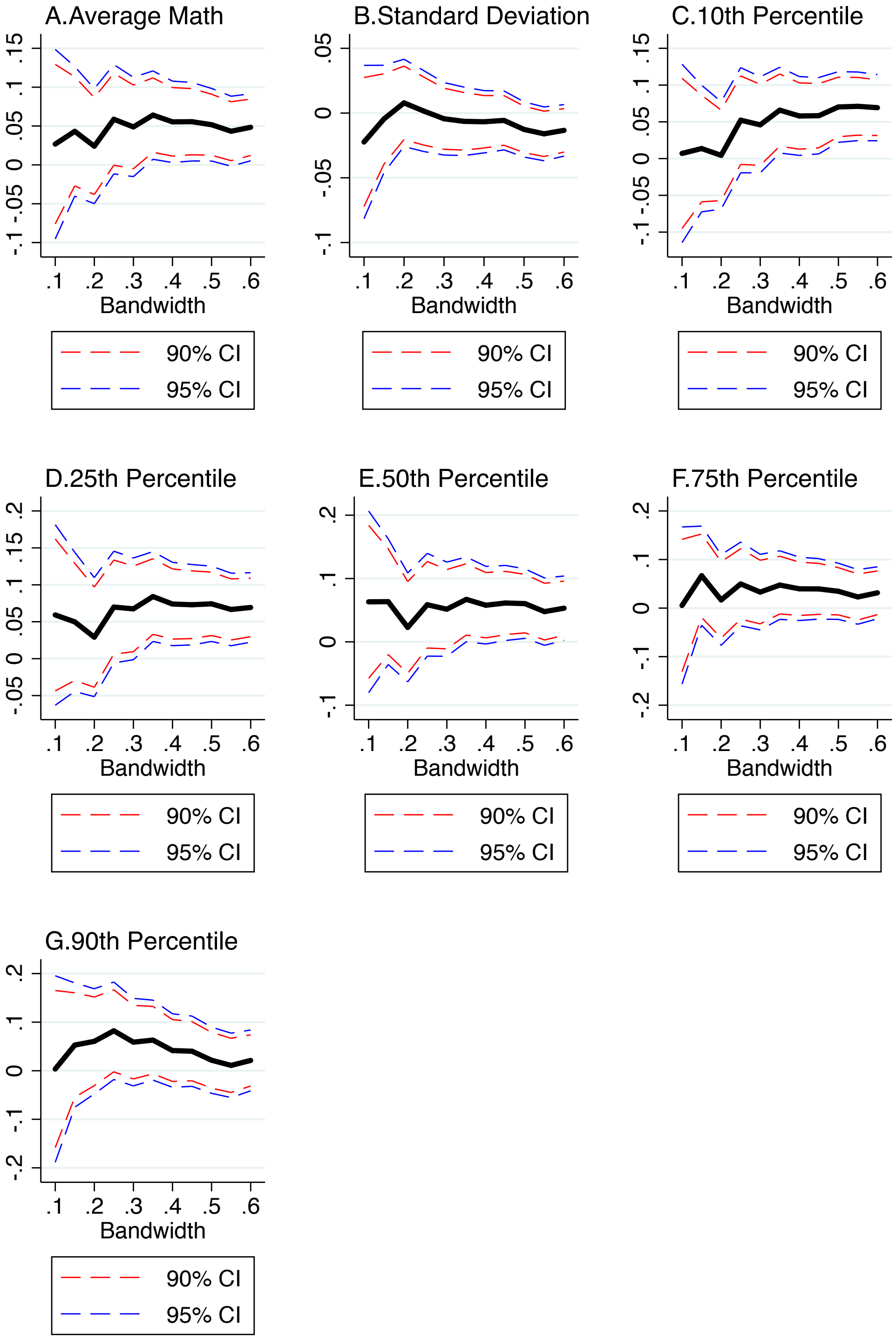
Notes: The figure shows the distributions of SPI score and AGI score. The vertical black line in both panels A and B shows the cutoff. The test statistics is calculated from the manipulation test proposed by Cattaneo et al. (2016).

Figure 2.6: Relationship between Treatment Status and Distance from Cutoff



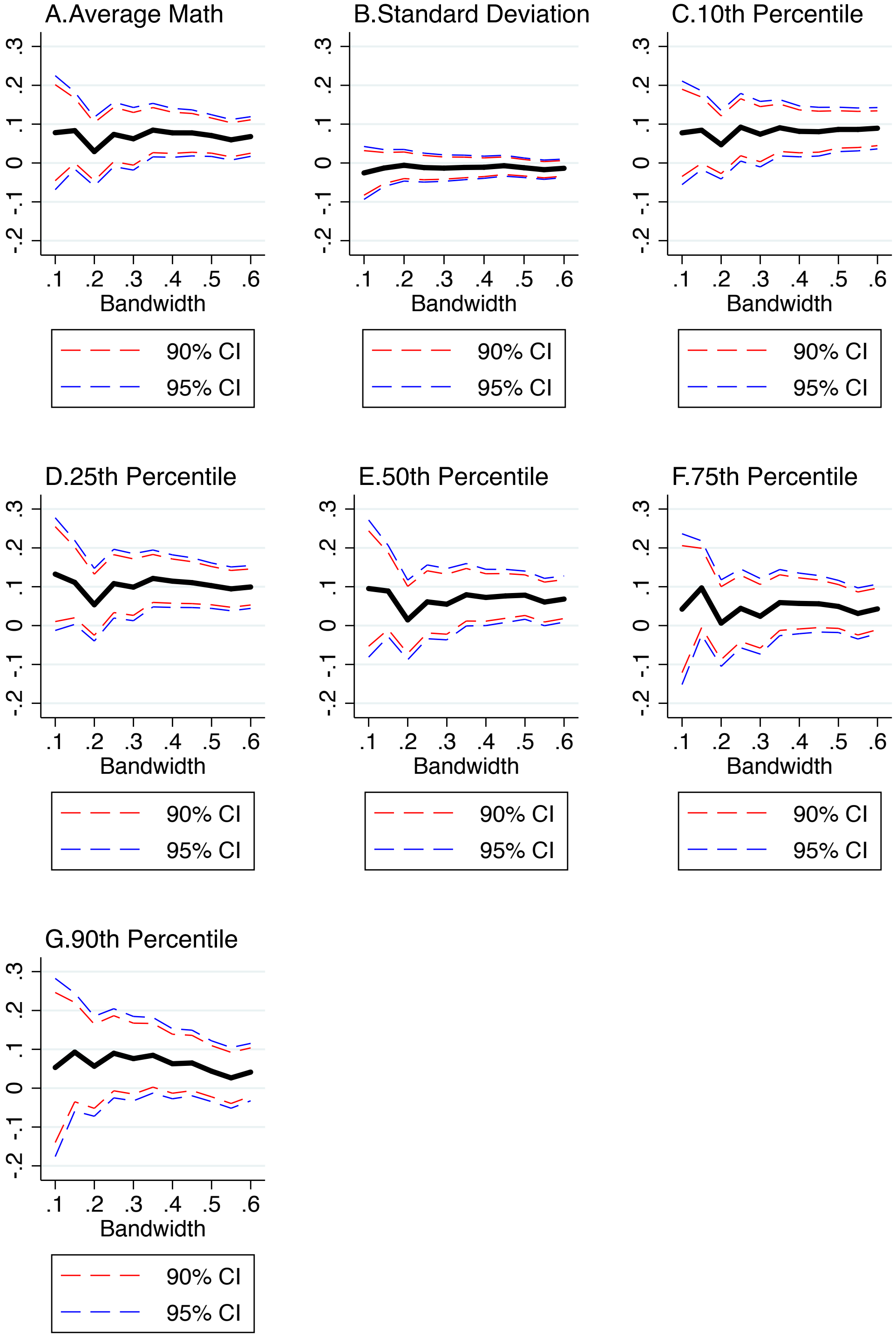
Notes: Panel A displays the relationship between Priority assignment and SPI and panel B shows the relationship between Focus designation and AGI. The vertical red line in both panels A and B shows the cutoff.

Figure 2.7: Sensitivity to Bandwidth, Focus Schools



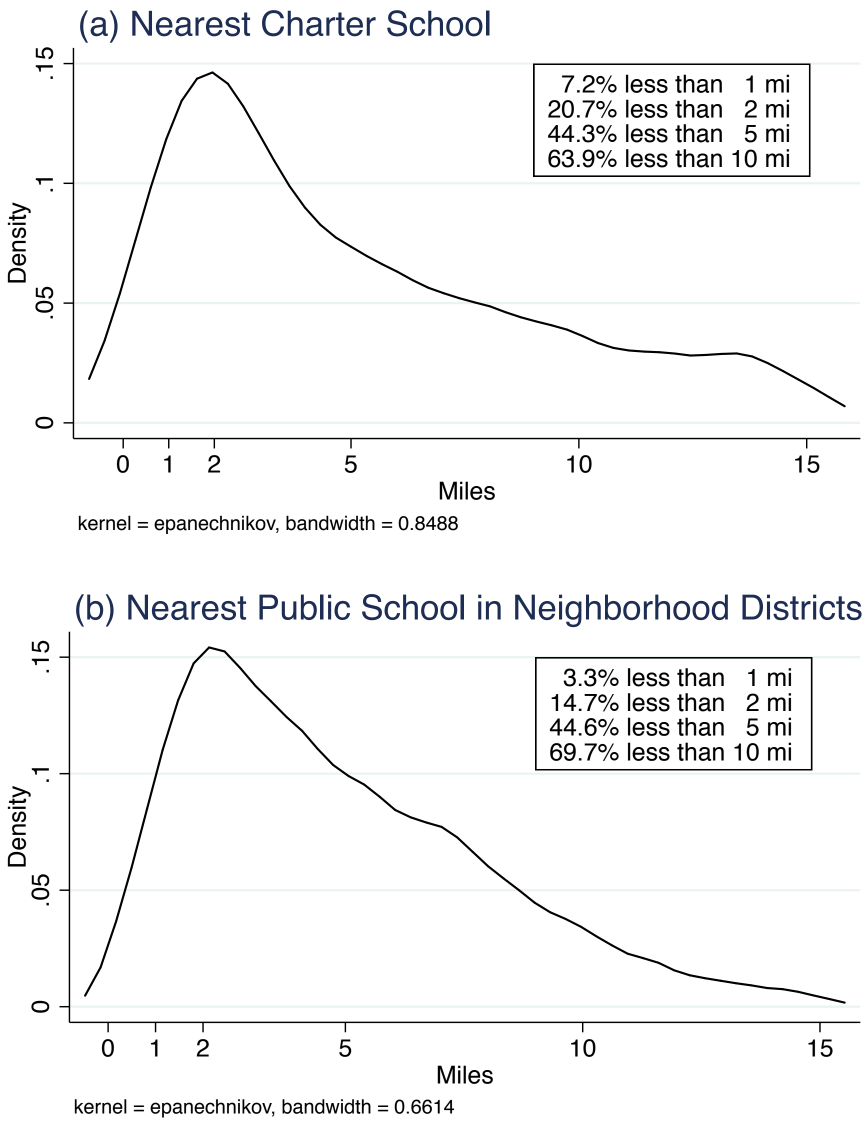
Notes: The figure depicts baseline estimates reported in panel C of Table 2.3 by bandwidths ranging from 0.1 to 0.6 point. The red (blue) dash line in all panels shows 90 (95) percent confidence intervals.

Figure 2.8: Sensitivity to Bandwidth, Title1 Focus Schools



Notes: The figure depicts baseline estimates reported in panel B of Table 2.4 by bandwidths ranging from 0.1 to 0.6 point. The red (blue) dash line in all panels shows 90 (95) percent confidence intervals.

Figure 2.9: Distribution of Distance between Public Schools and the Public Schools' Nearest Charter Schools



(a) Nearest Charter School

kernel = epanechnikov, bandwidth = 0.8488

(b) Nearest Public School in Neighborhood Districts

kernel = epanechnikov, bandwidth = 0.6614

Notes: Panel A shows the distribution of distance between public schools and the public schools' nearest charter schools in Michigan. Panel B displays the distribution of distance between public schools and public schools' nearest neighboring public schools in Michigan conditional on public schools have a public competitor in neighboring districts. Note that 23.9 percent of public schools do not have a public competitor in neighboring districts. To better visualize the distribution of the distance schools within 15 miles in panel A, I drop about 24.7 percent of schools with nearest charter competitor further than 15 miles away.

REFERENCES

REFERENCES

Ahn, T. and J. Vigdor (2014). The impact of no child left behind's accountability sanctions on school performance: Regression discontinuity evidence from north carolina. Technical report, National Bureau of Economic Research.

Baude, P. L. (2015). Should I stay or should I go: school accountability ratings, achievement, and school choice. Ph. D. thesis, University of Illinois.

Bettinger, E. P. (2005). The effect of charter schools on charter students and public schools. *Economics of Education Review 24(2), 133–147.*

Bifulco, R. and H. F. Ladd (2006). The impacts of charter schools on student achievement: Evidence from north carolina. *Education Finance and Policy 1(1), 50–90.*

Bonilla, S. and T. Dee (2017). The effects of school reform under nclb waivers: Evidence from focus schools in kentucky. Technical report, National Bureau of Economic Research.

Booker, K., S. M. Gilpatric, T. Gronberg, and D. Jansen (2008). The effect of charter schools on traditional public school students in texas: Are children who stay behind left behind? *Journal of Urban Economics 64(1), 123–145.*

Brummet, Q. (2014). The effect of school closings on student achievement. *Journal of Public Economics 119, 108–124.*

Buddin, R. and R. Zimmer (2005). Student achievement in charter schools: A complex picture. *The Journal of Policy Analysis and Management:24(2), 351–371.*

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82(6), 2295–2326.*

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2017). A practical introduction to regression discontinuity designs. itWorking Manuscript.

Cattaneo, M. D., M. Jansson, and X. Ma (2016). rddensity: Manipulation testing based on density discontinuity. *The Stata Journal (ii), 1–18.*

Chakrabarti, R. (2013a). Accountability with voucher threats, responses, and the test-taking population: Regression discontinuity evidence from florida. *Education Finance and Policy 8(2), 121–167.*

Chakrabarti, R. (2013b). Vouchers, public school response, and the role of incentives: Evidence from florida. *Economic inquiry 51(1), 500–526.*

Chakrabarti, R. (2014). Incentives and responses under no child left behind: Credible threats and the role of competition. *Journal of Public Economics 110, 124–146.*

Chakrabarti, R. and J. Roy (2016). Do charter schools crowd out private school enrollment? evidence from michigan. *Journal of Urban Economics 91, 88–103.*

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics 93(9-10), 1045–1057.*

Cowen, J., B. Creed, and V. Keesler (2015). Dynamic participation in inter-district open enrollment: Evidence from michigan 2005-2013. Working paper. *Education Policy Center, Michigan State University.*

Cullen, J. B. and R. Reback (2006). Tinkering toward accolades: School gaming under a performance accountability system. *In Improving School Accountability, pp. 1–34. Emerald Group Publishing Limited.*

Dee, T. and E. Dizon-Ross (2017). School performance, accountability and waiver reforms: Evidence from louisiana. Technical report, National Bureau of Economic Research.

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.*

Figlio, D. and C. M. D. Hart (2014). Competitive effects of means-tested school vouchers. *American Economic Journal: Applied Economics 6(1), 133–56.*

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics 90 (4-5), 837–851.*

Goldhaber, D. and J. Hannaway (2004). Accountability with a kicker: Observations on the florida a+ accountability plan. *Phi Delta Kappan 85(8), 598–605.*

Hanushek, E. A., J. F. Kain, S. G. Rivkin, and G. F. Branch (2007). Charter school quality and parental decision making with school choice. *Journal of public economics 91(5-6), 823–848.*

Hanushek, E. A. and M. E. Raymond (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management 24(2), 297–327.*

Hemelt, S. W. and B. Jacob (2017). Differentiated accountability and education production: Evidence from nclb waivers. Technical report, National Bureau of Economic Research.

Hsu, Y.-C., S. Shen, et al. (2016). Testing for treatment effect heterogeneity in regression discontinuity design. Technical report, Institute of Economics, Academia Sinica, Taipei, Taiwan.

Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics 142(2), 615–635.*

Imberman, S. A. (2011). The effect of charter schools on achievement and behavior of public school students. *Journal of Public Economics 95(7-8), 850–863.*

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. A. and S. D. Levitt (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics 118(3), 843–877.*

Klein, A. (2016). The every student succeeds act: An essa overview. *Education Week, 114–95.*

Ladd, H. F. and D. L. Lauen (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and management 29(3), 426–450.*

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of economic literature 48(2), 281–355.*

Neal, D. and D. W. Schanzenbach (2010). Left behind by design: Proficiency counts and test-based accountability. *The 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.*

Richardson, J. (2015). Accountability incentives and academic achievement: Distributional impacts of accountability when standards are set low. *Economics of Education Review 44, 1–16.*

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), 251–81.*

Sass, T. R. (2006). Charter schools and student achievement in florida. *Education Finance and Policy 1(1), 91–122.*

Saw, G., B. Schneider, K. Frank, I.-C. Chen, V. Keesler, and J. Martineau (2017). The impact of being labeled as a persistently lowest achieving school: Regression discontinuity evidence on consequential school labeling. *American Journal of Education 123(4), 585–613.*

West, M. R. and P. E. Peterson (2006). The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments. *The Economic Journal 116(510), C46–C62.*

Zimmer, R. and R. Buddin (2009). Is charter school competition in california improving the performance of traditional public schools? *Public Administration Review 69(5), 831–845.*

CHAPTER 3

The Effects of School Accountability Systems Under NCLB Waiver: Evidence from Priority Schools in Michigan

3.1 Introduction

As of January 1, 2016, 43 states and the District of Columbia had been granted waivers from certain unrealistic requirements embedded in No Child Left Behind (NCLB) Act. The waivers expressed the Obama administration's vision of a reauthorization of NCLB focusing on "college-and-career-ready" criteria. However, the waiver-driven reforms also emphasized implementing a "differentiated accountability" system for states. To receive the waiver, each application must include state accountability system that could identify their lowest performing schools ("priority" schools) as well as schools with the largest achievement gaps ("focus" schools). Every state has different assignment rules for these groups, and the consequences imposed on these schools that fail to fulfill the required goals also vary by states. In the case of Michigan, Priority schools were identified as schools in the bottom 5 percent of their "top-to-bottom" (TTB) ranking and Focus Schools were determined based on the two-year average achievement gap (hereafter, AGI) between the bottom 30 percent and top 30 percent of students within a school, and the Focus labels were assigned to schools that fall below 10 percentile of the AGI distribution.

Given that a few prior studies that evaluate the efficacy of school accountability systems under NCLB waivers particularly focus on the Focus reform (Bonilla and Dee, 2017; Dee and Dizon-Ross, 2017; Hemelt and Jacob, 2017), and given that these studies have been unable to separate the effects of receiving the accountability labels (both Priority and Focus labels) from the set of interventions associated with these labels, understanding the impact of receiving Priority label under the NCLB waiver is important to understand the efficacy of school accountability system under the recent reauthorization of NCLB. The reason is that the requirements under the recent reauthorization of NCLB for low-performing schools closely follow the requirements for Priority schools.

In this paper, I focus on the state of Michigan and explore the responses of schools that received a Priority label. Using a unique feature of school accountability in Michigan, that is, the first year of Priority designation was regarded as a preparation year during which Priority schools were, I am able to isolate the effects of the stigma attached to this label from the set of interventions associated with the label in consecutive years. Because Priority schools may have concentrated on the low achieving students while diverting attention from the high performing students, I examine the impact of receiving the label not only on the mean achievement of a school but also on the student achievement distribution.

To credibly identify the impact of Priority labels on various student achievement outcomes such as unconditional mean, SD, and various quantiles of the within-school-grade achievement distribution, I exploit discontinuous variations in the assignment of schools to the Priority labels. I use a sharp regression discontinuity (RD) design and compare the mean or given quantile outcomes (e.g. 10 percentile, 25 percentile) in Priority schools that were barely above the assignment cutoff to that of outcomes in schools just below the cutoff. I find little evidence of positive achievement effects on average math scores or specific performance measures at different quantiles of the within-school-grade achievement distribution; my null findings are robust across the numerous specifications and variations in modeling choices. Finally, to examine whether the changes in the composition of students, which may be caused by the Priority labels, could explain my null findings, I define several key measures of student composition such as the percentage of black, hispanic students, and the percentage of students belong to the given math and reading quartile ranks. Using these measures as outcomes, I find no evidence that the Priority designation influenced the student composition of schools.

The remainder of the paper is organized as follows. Section 2 gives an overview of the accountability systems and Priority schools in Michigan. Section 3 discusses the data. Section 4 presents the empirical strategy, and Section 5 shows the main results. Section 6 concludes the paper.

3.2 Background

The U.S. Department of Education announced in the fall of 2011 that states could apply for waivers from certain conditions of NCLB. Waivers were applied in particular to the requirement that schools and districts must achieve the unrealistic goal of 100% student proficiency on math and reading tests by 2014. To receive the waiver, each successful application for flexibility had to include two key components as part of a state accountability system: First, the accountability system had to adopt "college-and career-ready" criteria in at least reading and math, and second, in each state, the accountability system had to identify the schools with the lowest performance ("Priority" schools) and those with the largest achievement gaps ("Focus" schools).

Under this differentiated accountability system, Priority schools were identified as being among the lowest-performing schools, including at least five percent of the state's Title 1 schools. Once identified, Priority schools were subject to interventions that were compatible with federal school turnaround principles. Specifically, each Priority school had to select one of the following models: transformation, turnaround, restart, or close.

To receive the waivers, like many other states, the Michigan Department of Education (MDE) submitted an NCLB waiver application, which was approved in July 2012. The first Priority school list was announced in August 2012 and the MDE mainly used TTB ranking to determine the list of Priority schools each year. The TTB index was a weighted average of three subject-specific achievement indexes: two-year average level scores, two-year average growth in scores, and the two-year average achievement gap (AGI) between the bottom 30 percent and top 30 percent of students within a school. Schools below the 5th percentile in the TTB ranking were identified as Priority schools.

Prior research has already effectively documented that schools game the accountability system. Chakrabarti (2013) and Goldhaber and Hannaway (2004) find that schools tend to focus on a specific subject in which it is easy to raise test scores when the accountability system requires schools to pass their proficiency standard at least one subject. Furthermore, schools are more likely to improve the test scores of students who are on the margin of passing the cutoff when the accountability systems are based on proficiency rates at the level of test scores (Richardson, 2015; Ladd and Lauen, 2010; Neal and Schanzenbach, 2010; Reback, 2008). However, the complexity embedded in calculating the TTB index limited the ability of Priority schools to game the accountability system. The reason is that the MDE used all five subjects (math, reading, science, social science, and writing), and test scores across grades (from 3rd to 8th grade) were equally weighted in the calculations. Moreover, MDE used a weighted average of three achievement indexes that reflecting the level, growth, and gaps of test scores. Nevertheless, I examine the effect of receiving the Priority label not only on the average test scores but also on students' performance across the scoring distribution.

Once identified, Priority schools received a set of four year-long interventions. The interventions required that Priority schools develop an improvement plan based on one of four federal school turnaround models, and the MDE monitored the implementation of the selected model for each school. Among these four options, many Priority schools had implemented the "transformation" option which obligates schools to develop school leader effectiveness including replacing the principal (MDE, Frequently Asked Questions about Michigan's Priority Schools, 2014).[1] I should note that the MDE considered the 2012-2013 school year as a preparation year, during which Priority schools were expected to develop plans of action that are aligned with Federal requirements. This unique feature has enabled me to use year-1 test scores to isolate the stigma effects from the set of interventions attached to the Priority labels.

One remaining concern is that Priority schools may not have had an incentive to avoid being labeled in year 2 because year 1 was regarded as a preparation period, and schools had to implement the one of four federal school turnaround models regardless of whether they received one of the

[1] The turnaround model is similar to the transformation model but more harsh because this model requires schools to replace their principal and at least 50 percent of the school's staff.

labels in year 2. However, because principals and teachers in Priority schools (similar to receive "F" grade under the NCLB regime) may have regarded the labels as social stigmas and because the list of Priority schools was publicly available with local media attention, receiving the Priority label could have increased schools' incentive to respond the Priority label.[2]

3.3 Data and Sample

I combine two datasets from various sources to estimate the impact of Priority labeling on students' score distribution. The first dataset that is publicly available from the MDE website contains accountability results from 2011 to 2013 at the school level for every school in Michigan serving grades 3 to 8. The data file includes the school performance index (hereafter SPI, but officially called the TTB ranking), which is used to determine eligibility for Priority labeling, and the achievement gap index (AGI). I re-center SPI on zero and change the sign of the measures to give positive values for the Priority schools.

The second data file is the individual-level administrative data provided by the MDE and Michigan's Center for Educational Performance and Information (CEPI). The data represent the universe of Michigan students in grades 3-8 from 2011-2014, and contain students' test scores in mathematics, reading, science, social studies, and writing. Additionally, the data include various demographic controls, such as grade, gender, race, special education, and free lunch status. Given that the unit of variation is the school-year, using school-year observations produces virtually identical results as using student-year observations. I thus aggregate student-year observations into school-grade-year observations by calculating the mean, SD, and various quantiles within a school-grade achievement distribution. Note that the publicly available school-level performance data does not provide various quantile measures within the school-grade distribution.

My sample includes schools serving grades 3 to 8 in 2013 because annual state administered tests are not available for other grades. Among these schools, most schools serving the tested

[2] Goldhaber and Hannaway(2004) who surveyed principals and teachers found that they viewed a school grade of "F" as a social stigma under the accountability system in Florida.

grades are elementary or middle schools. I drop schools that serve some 3rd to 8th graders as well as students in higher grades, e.g. schools with grade configurations such as 6-12, because the SPI for these schools was required to incorporate the high school graduation rate, causing these schools to focus on different criteria. I further limit the sample by dropping schools serving special education students from my main analysis because these schools were subject to different mandates than other public schools. I also eliminate 2 schools with fewer than 50 students from my sample because these schools serve special populations.

When Hemelt and Jacob (2017) examined the impact of the labels, they dropped charter schools because they cast doubt on whether charter management organizations enforced the reforms required by the Focus or Priority labels. However, I do not exclude charter schools in my main sample for two reasons. First, students in charter schools had to take the MEAP exams, and charter schools are subject to the same school accountability system as traditional public schools. Furthermore, my analysis focuses on the short-run impacts of school accountability labeling on student test scores and not on the impact of a set of reforms in consecutive years (from 2 to 4 years after receiving the labels), so it is reasonable to include charter schools in my main sample.

Table 3.1 shows the summary statistics of certain key variables for the sample of schools in 2013. The table compares the means of schools with the Priority and all public schools in Michigan; in the full sample, approximately 3.11 percent received the Priority labels in 2013. Table 3.1 clearly shows that Priority schools were low-performing schools and served minority and low socioeconomic status (SES) students in urban settings. Specifically, in terms of demographics, schools with a Priority label served a similar percentage of Hispanic students compared to all K-8 schools, but the percentage of black students in Priority schools was approximately five times higher than in all K-8 schools (approximately 78.43 and 16.85 percent, respectively).

3.4 Empirical Strategy

To examine the effect of the Priority labeling on various performance measures, I use "Sharp RD design" (Hahn et al., 2001) that compares schools that were barely above the cutoff and thus were assigned the Priority label to schools that were just below the cutoff and thus were not assigned as Priority schools. I use the the following regression equation to examine the impact of receiving Priority labels on student achievement by:

$$Y_{gst}^{m} = \alpha_0 + \alpha_1 I(r_s \geqq r*) + \alpha_2 r_{st-1} + \alpha_3 I(r_s \geqq r*) \times r_{st-1} + \sum_{g=4}^{8} \gamma_g G_g + X_{st-1}\beta + \varepsilon_{st} \qquad (2.1)$$

$$where\, t = 2013$$

I use g index to show a specific grade (from 3 to 8) and s denotes a school. Y_{gst}^{m} represents various grade-school level measures (m) based on math and reading test scores in year t, including the mean, SD, and various quantiles in the within-school-grade achievement distribution. $I(r_s > r*)$ is a binary variable that takes one if schools' SPI score is greater than or equal to the cutoff, and r_{st-1} shows the running variable for SPI scores in year $t-1$. I control for a trend in running variables with a linear spline ($r_{st-1} + I(r_s \geqq r*) \times r_{st-1}$) as I estimate the model using observations near the cutoff. G_4 to G_8 are grade dummy variables because I pool the observations across grades. To increase precision, I include X_{s2012} which is a vector of school-level control variable shown in Table 3.1. The point estimates without control variables are similar to the baseline, however (see, Table 3.2). Finally, the standard errors in all the analyses and specifications are clustered at the school level to account the likelihood that errors are correlated across grades within schools.

For bandwidth selection, I use an optimal bandwidth selection method proposed by Calonico et al.(2014) for each outcome measure and each different specification. Their method basically chooses the bandwidth that minimizes the asymptotic MSE (mean square error) of the RD point estimator. The optimal bandwidths based on this method range from 0.17 to 0.3 points for Priority schools. Note that in all tables that show the local RD estimates, I show the optimal bandwidth and the number of observations that are used to estimate the RD estimate α_1. Then, I estimate Eq.(1) with the given bandwidth using a triangular kernel to weight observations within the bandwidth since the triangular kernel is optimal for the boundary estimation with the optimal bandwidth

(Imbens and Lemieux, 2008).

The main identification assumption to use the RD design is that each school does not have precise control over the running variables (Lee and Lemieux, 2010). Figlio (2006), Cullen and Reback (2006), Jacob (2005) and Jacob and Levitt (2003) have shown some evidence that schools may game the accountability system. However, manipulating running variables in my context is unlikely since cutoff points in SPI are not deterministic. Specifically, the cutoffs were determined by the relative positions among all eligible schools, and principals did not have any reference points to predict the cutoffs for the Priority labels because the labels were exogenous shocks to them. Nevertheless, I present the distributions of the SPI score and AGI score in Figure 3.1; the vertical black line in both panels shows the cutoff. Figure 3.1 clearly shows no discontinuity at the cutoff, indicating that schools did not manipulate both the SPI and AGI score. Note that the test statistics proposed by Cattaneo et al. (2016) are small and insignificant.

Finally, I explore whether pre-treatment covariates are continuous at the cutoff to further test the validity of the RD design. The pre-determined variables would be different at the cutoff if the unobserved school qualities were discontinuous around the threshold. Table 3.3 shows the estimated coefficients for the baseline control variables reported in Table 3.1. I do not find evidence that Priority schools are not comparable to their barely non-Priority counterparts; all discontinuity estimates for various pre-assignment covariates are not statistically significant even at the 10 percent level. However, I should note that the point estimates in the Priority sample are noisy, making difficult to detect the small mean differences between Priority and their barely non-Priority counterparts.

3.5 Results

3.5.1 The Impact of Priority Labeling on Student Achievement

Figures 3.2 illustrate the effects of receiving Priority labels on various student math achievement outcomes, including the average and the different points in the within-school-grade achievement distribution. The x-axis shows schools' SPI relative to the cutoffs that determined the Priority labeling for the 2012-2013 school year, and the y-axis describes the mean and various quantiles of the math test scores. I only include schools with an SPI within ± 0.5 from the cutoff to better visualize the behavior of outcomes near the cutoff. Each of the points in these figures indicates the average outcome measures of mean and various quantile outcomes collapsed into bins. Note that I use bins that contain the same number of observations instead of equal-length bins since most Priority schools are concentrated around the cutoff. Furthermore, I follow Cattaneo et al. (2017) to choose the number of optimal bins in the sense that the overall variability of the binned means resembles the variability in the raw data.[3]

Figure 3.2 presents little evidence of positive achievement effects on average math scores or specific performance measures at different quantiles of the within-school-grade achievement distribution. The vertical distances between the local means near the cutoff are close to zero for most outcome variables (see panels A to E). Furthermore, panel F of Figure 3.2 indicates that receiving Priority labeling may have harmed students in the 90th percentile within their school-grade score distribution, although the binned means to the right of the cutoff are far noisier than those of the binned means among non-Priority schools.

Table 3.4 presents the nonparametric estimates analogous to these results. For each panel, I report the optimal bandwidth for each outcome variable and the number of observations within the optimal bandwidth. In panels A and B of Table 3.4, I first present the results that examine the impact of receiving Priority labeling on 2011 Fall test scores. I expect the estimates of Focus school indicators across various outcome variables to be small and insignificant because students were taking this exam prior to the first Priority designation. Panels A and B confirm this expectation as all estimates across all columns are small and statistically insignificant.

Panels C and D confirms the null achievement effects of Priority labeling on the scoring distribution. All estimates in panel C are small (ranging from -0.043 SD to 0.033 SD) and insignificant. The estimates using the 2012 Fall reading test are reported in panel D of Table 3.4. Similarly to the

[3] This procedure ensures a large number of local means near the cutoff, which is useful for obtaining a graphical illustration of the variability of the data around the cutoff.

math results, there is no evidence that receiving Priority labeling led to an increase in student test scores across the scoring distribution. In addition, the estimate for students in the 75th percentile is moderately negative, although the effect is not statistically significant.

3.5.2 Does Student Mobility Drive the Results?

To examine whether receiving Priority labels changed the student-body composition in Priority schools, I first construct several key measures of student composition. Next, I use these measures as outcome variables and estimate the nonparametric RD regression. When Hemelt and Jacob (2017) explore the effects of the accountability labels (Priority and Focus) on student composition, they only examine a set of socioeconomic variables such as the percentage of black, Hispanic, economically disadvantaged students, and special education students. On top of these measures, I create math and reading quartile ranks based on students' lagged test scores (2011 Fall test), and use the percentage of students in each math and reading quartile rank for those enrolled in the 2012-2013 school year. I believe examining these variables best reflects the possible compositional changes induced by the accountability label. The reason is that whether students switched to other schools in response to the Priority labels may have depended on the level of student achievement. For example, the parents of low-achieving students may have viewed the Priority label as an indication that their schools devote more effort to low-scoring students.

In Table 3.5, I display the estimated results using the key measures of student composition. To calculate the percentage of students belonging to the given quartile rank, I use the lagged math and reading scores of students who enrolled in the year immediately following the announcement of Priority schools; I use the Priority sample for the 2012-2013 school year. Table 3.5 shows little evidence of the compositional impact of the accountability labels, since I do not find any sizable changes in the composition of high- and low-scoring students, nor do I find changes in the racial or socioeconomic composition in the schools; the discontinuity estimates are never statistically significant except in two cases (% students in the second quartile of reading test scores and % special education students). Note that with a large number of multiple comparisons, a few cases can be statistically different from zero merely by random variation.

One explanation of the null effect is that parents may not have had enough time to respond to the accountability labels. The public announcement of the list of Priority schools came in August 2012, and parents thus had a maximum of one month before the new school year began to search for and choose alternative schooling options.

3.6 Conclusion

This paper has examined whether receiving Priority labels under NCLB waivers influenced student achievement. Specifically, if teachers and principals responded to the label by focusing on low-achieving students, the positive achievement gains in the performance of low-scoring students would be expected. Hence, I examine the achievement effects not only on mean test scores but also on various quantile outcomes within school-grade distributions. Exploiting discontinuous variations in Priority assignments, I non-parametrically estimate the RD model that compares schools that were just above the cutoff and schools that were barely below the cutoff. I show that receiving the Priority label did not raise the average math achievement as well as the performance of low-scoring students. Also, I find little to no effect of receiving the label on student composition

This finding is generally consistent with Hemelt and Jacob (2017) who find a null treatment effect (in terms of school staffing, composition, and student achievement) of the Priority designation in the medium- or long-run. Combining these two studies may suggest some evidence of implementation infidelity. A more open line of questioning such as qualitative interviews for principals and teachers in Priority schools could also reveal the causal mechanisms of the null treatment effects.

APPENDICES

APPENDIX A TABLES

Table 3.1: Summary Statistics by Accountability Labeling

	Priority schools (1)	All schools (3)
Accountability results		
Percent Priority	100.00	3.11
Priority running variable	0.36	-1.49
Average math score	-0.72	-0.01
Average reading score	-0.29	0.01
Student characteristics		
Percent free lunch	84.63	44.87
Percent special education	16.57	13.76
Percent black	78.43	16.85
Percent Hispanic	5.72	6.22
Percent white	12.50	70.26
Percent Asian	0.55	2.71
School characteristics		
Percent elementary	50.82	65.54
Percent K-8	40.98	9.65
Percent middle	8.20	24.81
Percent charter schools	11.48	7.45
Percent title1 schools	98.36	77.08
Percent magnet schools	8.20	14.19
Percent located in urban	78.68	21.49
Total enrollment	417.43	439.44
Number of schools	61	1959

Notes: The table shows the summary statistics of certain key variables for the sample of schools in the 2012-2013 school year. The table compares the means of schools with priority and focus labels as well as all schools.

Table 3.2: Sensitivity Check: Without Controls

	(1) Mean	(2) SD	(3) 10th percentile	(4) 25th percentile	(5) 50th percentile	(6) 75th percentile	(7) 90th percentile
Panel A. Priority: Math							
Priority Schools	-0.003	-0.055	0.030	0.023	-0.013	0.002	-0.075
	(0.097)	(0.045)	(0.046)	(0.063)	(0.110)	(0.142)	(0.149)
Observations	345	440	400	363	338	348	401
Bandwidth	0.247	0.330	0.311	0.260	0.240	0.249	0.314
Panel B. Priority: Reading							
Priority Schools	-0.056	-0.032	-0.058	-0.047	-0.043	-0.114	-0.034
	(0.092)	(0.034)	(0.053)	(0.071)	(0.103)	(0.121)	(0.132)
Observations	400	414	449	440	500	432	400
Bandwidth	0.312	0.320	0.339	0.332	0.362	0.324	0.312

Notes: The table compares the baseline estimates reported in Table 3.4 to the estimates without predetermined control variables. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 3.3: Testing Validity of RD Analysis by Using Preprogram Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
Lagged Outcomes	Mean math	Mean reading	%Q1 Math	%Q2 Math	%Q3 Math	%Q4 Math
Priority Schools	-0.084	0.018	5.301	-1.071	-1.797	-2.009
	(0.074)	(0.051)	(4.132)	(1.459)	(2.288)	(2.205)
Observations	196	130	202	93	130	150
Bandwidth	0.493	0.354	0.512	0.253	0.355	0.400
	%Q1 Reading	%Q2 Reading	%Q3 Reading	%Q4 Reading	% Hispanic	% Black
Priority Schools	4.877	-0.783	-2.356	-1.370	-3.439	9.443
	(3.735)	(1.418)	(1.537)	(2.381)	(3.802)	(10.809)
Observations	133	126	102	120	138	131
Bandwidth	0.359	0.350	0.297	0.342	0.370	0.356
	% Free lunch	% Special edu	Enrollment	Magnet	K5	Middle
Priority Schools	2.659	0.777	8.819	0.193	-0.263	0.086
	(4.111)	(1.783)	(68.178)	(0.171)	(0.210)	(0.158)
Observations	126	88	112	98	101	119
Bandwidth	0.351	0.239	0.321	0.278	0.286	0.340

Notes: The table shows the estimated discontinuities of baseline control variables for Priority and Focus schools. The percentage of students in each math and reading quartile ranks based on 2011 Fall test. In addition to the indicator of passing running variables, each regression includes a linear spline of running variables. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

Table 3.4: The Effect of Receiving Priority Labeling on Student Achievement

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean	SD	10th	25th	50th	75th	90th
			percentile	percentile	percentile	percentile	percentile
Panel A. 2011 Fall Test (Placebo): Math							
Priority Schools	0.037	0.033	-0.034	0.025	0.039	0.082	0.049
	(0.033)	(0.029)	(0.024)	(0.025)	(0.033)	(0.054)	(0.077)
Observations	348	352	437	472	387	299	295
Bandwidth	0.229	0.229	0.306	0.322	0.267	0.209	0.208
Panel B. 2011 Fall Test (Placebo): Reading							
Priority Schools	0.059	-0.007	0.028	0.050	0.057	0.014	0.068
	(0.038)	(0.018)	(0.031)	(0.045)	(0.050)	(0.050)	(0.070)
Observations	355	420	570	383	383	359	390
Bandwidth	0.236	0.295	0.385	0.256	0.258	0.238	0.268
Panel C. 2012 Fall Test: Math							
Priority Schools	-0.008	-0.043	0.028	0.033	-0.033	-0.003	-0.025
	(0.059)	(0.023)	(0.042)	(0.045)	(0.078)	(0.081)	(0.086)
Observations	291	355	330	315	312	319	326
Bandwidth	0.212	0.259	0.240	0.221	0.221	0.227	0.233
Panel D. 2012 Fall Test: Reading							
Priority Schools	-0.039	-0.037	-0.030	-0.008	0.021	-0.091	0.004
	(0.050)	(0.026)	(0.041)	(0.047)	(0.066)	(0.056)	(0.068)
Observations	291	291	388	374	307	307	242
Bandwidth	0.213	0.212	0.308	0.289	0.219	0.220	0.177

Notes: The table presents the estimated results of receiving Priority labels on various student achievement measures. In addition to the indicator of passing running variables, each regression includes a linear spline of running variables, grade dummies, and a set of pre-determined variables reported in Table 3.1. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

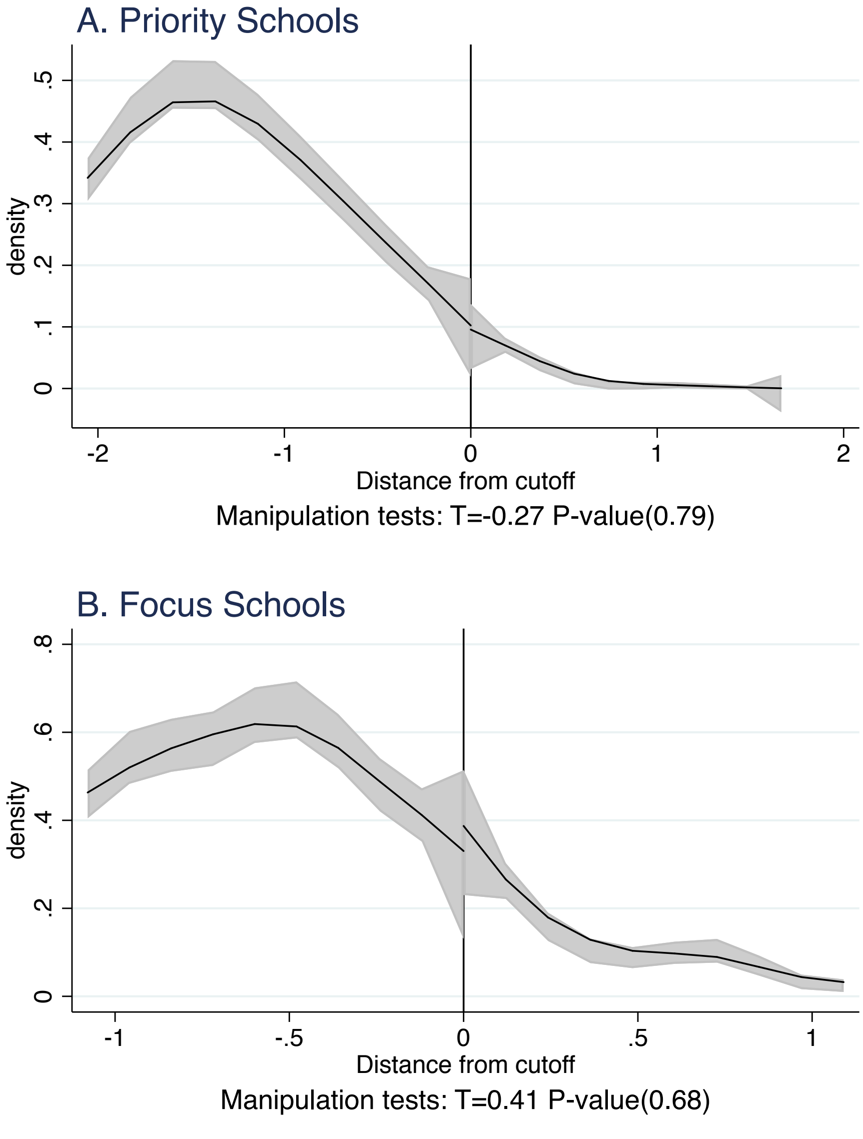
Table 3.5: The Effect of Assigning Priority Labels on Student Composition in 2012-2013 school year

	(1) %Q1 Math	(2) %Q2 Math	(3) %Q3 Math	(4) %Q4 Math	(5) %Q1 Reading	(6) %Q2 Reading
Priority Schools	1.754	-1.245	0.536	-0.962	1.946	-3.647*
	(2.460)	(1.631)	(2.225)	(1.204)	(1.983)	(1.898)
Observations	117	122	103	112	161	92
Bandwidth	0.342	0.352	0.314	0.329	0.423	0.255
	%Q3 Reading	%Q4 Reading	%Black	%Hispanic	%Free Lunch	%Special edu
Priority Schools	-0.641	1.066	-0.086	-0.333	-1.184	3.526**
	(1.256)	(2.419)	(1.466)	(0.653)	(2.157)	(1.079)
Observations	168	78	112	102	152	75
Bandwidth	0.444	0.216	0.329	0.307	0.412	0.231

Notes: The table presents the estimated discontinuities using several key measures of student composition. The percentage of students in each math and reading quartile ranks based on 2011 Fall test. Each regression includes the indicator of passing running variables, each regression includes a linear spline of running variables, and a set of pre-determined variables reported in Table 3.1. Cluster standard errors are reported. Statistically significant at *** 0.1%, ** 1%, * 5%, and + 10%.

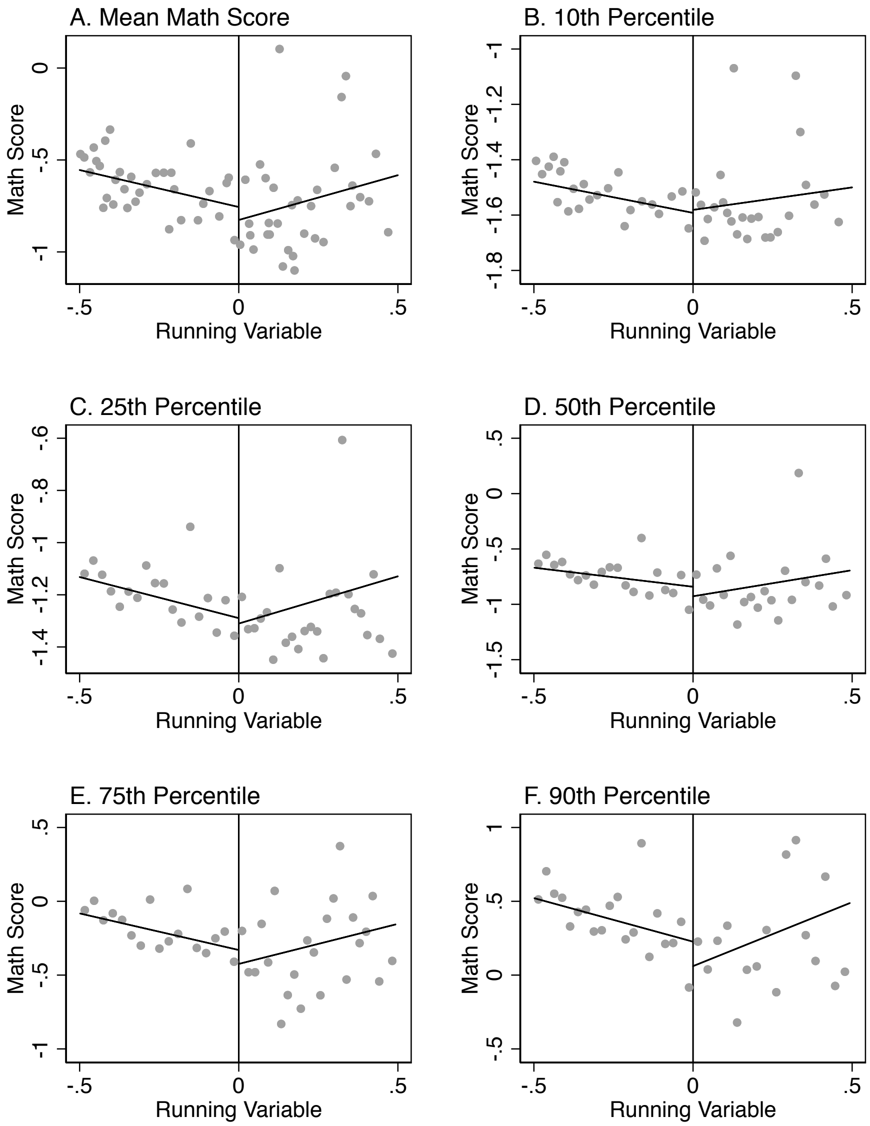
APPENDIX B FIGURES

Figure 2.1: Density of Running Variables across the Distance from Cutoff



Notes: The figure shows the distributions of SPI score and AGI score. The vertical black line in both panels A and B shows the cutoff. The test statistics is calculated from the manipulation test proposed by Cattaneo et al. (2016)

Figure 2.2: Effect of Receiving Priority Label on Math Achievement



Notes: The figure illustrates the effects of receiving Priority labels on various student math achievement outcomes, including the mean and the different points in the within-school- grade achievement distribution. The x-axis shows schools' SPI relative to the cutoffs that determine Priority labeling for the 2012-2013 school year, and the y-axis describes the mean and various quantiles of the math test scores. To better visualize outcomes around the cutoff, I only include schools with SPI within ± 0.5 from the cutoff. Each of the points in these figures indicates the average outcome measures of mean and various quantile outcomes collapsed into bins.

REFERENCES

REFERENCES

Bonilla, S. and T. Dee (2017). The effects of school reform under nclb waivers: Evidence from focus schools in kentucky. Technical report, National Bureau of Economic Research.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82(6), 2295–2326.*

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2017). A practical introduction to regression discontinuity designs. itWorking Manuscript.

Cattaneo, M. D., M. Jansson, and X. Ma (2016). rddensity: Manipulation testing based on density discontinuity. *The Stata Journal (ii), 1–18.*

Chakrabarti, R. (2013a). Accountability with voucher threats, responses, and the test-taking population: Regression discontinuity evidence from florida. *Education Finance and Policy 8(2), 121–167.*

Chakrabarti, R. (2013b). Vouchers, public school response, and the role of incentives: Evidence from florida. *Economic inquiry 51(1), 500–526.*

Cullen, J. B. and R. Reback (2006). Tinkering toward accolades: School gaming under a performance accountability system. *In Improving School Accountability, pp. 1–34. Emerald Group Publishing Limited.*

Dee, T. and E. Dizon-Ross (2017). School performance, accountability and waiver reforms: Evidence from louisiana. Technical report, National Bureau of Economic Research.

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics 90 (4-5), 837–851.*

Goldhaber, D. and J. Hannaway (2004). Accountability with a kicker: Observations on the florida a+ accountability plan. *Phi Delta Kappan 85(8), 598–605.*

Hemelt, S. W. and B. Jacob (2017). Differentiated accountability and education production: Evidence from nclb waivers. Technical report, National Bureau of Economic Research.

Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics 142(2), 615–635.*

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. A. and S. D. Levitt (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics 118(3), 843–877.*

Ladd, H. F. and D. L. Lauen (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and management 29(3), 426–450.*

Lee, H. (2019). The Role of Credible Threats and School Competition within School Accountability Systems: Evidence from NCLB Waiver. *Working manuscript*

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of economic literature 48(2), 281–355.*

Neal, D. and D. W. Schanzenbach (2010). Left behind by design: Proficiency counts and test-based accountability. *The 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.*

Richardson, J. (2015). Accountability incentives and academic achievement: Distributional impacts of accountability when standards are set low. *Economics of Education Review 44, 1–16.*

Rockoff, J. and L. J. Turner (2010). Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy 2(4), 119–47.*