Effects of Teacher Performance Measures on Student Achievement

Yoon Jae Ro*

For the most recent version, click this link

October 31, 2019

Abstract

In 2019, 34 states require objective measures of student growth in teacher evaluation. By exploiting nationwide data and the difference in timing of the policy adoption across states and districts, I find that adopting the Student Growth Measures (SGMs) decreases students' math scores, and the size of the effect gets bigger with time. This unexpected adverse effect of the policy is driven by the deterioration among previously high-performing districts and schools. Furthermore, by exploring the policy of providing Value-Added (VA) measures to teachers on student performance in Ohio, I find the distribution of students' performance shifted downward in schools with VA policy, suggesting that the VA is detrimental to high-performing students. These results show that SGMs have an unwanted effect on student achievement, undermining the performance of students at the top end of the distribution.

JEL Codes: I12, I18, I21

*Ph.D. Candidate, Department of Economics, UC Riverside, 900 University Ave, 3127 Sproul Hall, Riverside, CA 92521, USA. Email: yoon.ro@email.ucr.edu. All errors are my own.

1 Introduction

There is strong evidence that having a quality teacher is a critical factor in student achievement (Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Chetty et al., 2014). As a result, policymakers and educators have long been interested in finding accurate and efficient ways to identify effective teachers and to improve teaching quality. However, teacher evaluation systems have historically struggled to adequately identify effective teachers. A study from the New Teacher Project (TNTP) in 2009 highlighted the discrepancy between formal teacher evaluation ratings and the true distribution of teacher effectiveness, noting that 99 percent of teachers are rated satisfactory when the districts use the binary setting (Weisberg et al., 2009). Federal interventions gave impetus to focus on teacher evaluations. The first to lay this foundation was the No Child Left Behind (NCLB) Act, which required states to set up a standardized assessment and to rate schools based on the fraction of students demonstrating "proficiency." The federal Race to the Top (RTTT) competition and Elementary and Secondary Education Act waivers pushed federal involvement in public education policy further by creating strong incentives for states to require the evidence of student learning in teacher evaluations.

Pushed by those incentives, states rushed to adopt a new teacher evaluation systems comprised of multiple measures of teacher performance, including metrics based on students performance on standardized tests: Student Growth Measures (SGMs).[1] Indeed, the use of SGMs in teacher evaluation rapidly expanded over the past decade. In 2009, only 15 states required objective measures of student growth in teacher evaluation, while in 2015, this number increased to 43 states (NCTQ, 2019). And even among the remaining states without implementation, many large school districts have adopted the SGMs in their evaluation.

The purpose of this paper is to systematically examine the nature and consequences

[1]There are several ways to measure student growth. Statistical methods such as student growth percentile (SGP) method, and value-added (VA) methods estimate a teachers impact on student achievement using students prior achievement. This paper treats these two approaches as functionally equivalent. An alternative way to measure the performance is student learning objectives (SLO) that sets a classroom-specific achievement growth targets set by individual teachers.

of the inclusion of student performance data into teacher performance evaluation and how students may be differentially impacted by this policy. Although SGMs are now widely used, at times with considerable cost to states and school systems, little research has investigated how informing teachers of their performance would affect students' outcomes across the distribution of prior district, school, or student performance. The lack of causal evidence about the impact of SGMs adoption on students is mainly due to rigorous data requirements and the unavailability of relevant policy rollout. For example, all school districts in a state often implement the new system at the same time. And until recently, there was no clear way to make comparisons across state-specific standardized tests.

This paper overcomes the limitations by exploiting the different timings of the adoption of SGMs as one of the performance measures across states and districts. Each state implemented SGMs in different years. For example, Pennsylvania started using the value-added (VA) measures in 2006, while Florida began using the VA in 2011. I construct an original information data set detailing each state's implementation policy. I gathered the information on whether the state had run a pilot program before the statewide implementation, when and how the SGMs are adopted and further included in teacher evaluation, and which type of SGMs are used. In this paper, I focus on whether teachers received their performance measures based on SGMs rather than focusing on their inclusion in actual summative rating on evaluation. This groundwork allows me to use the precise timing of the adoption of SGMs across states, which occurred between 2006 and 2015, to provide the evidence on how the policy impacts student performance, in conjunction with nationwide data.

I use difference-in-differences (DID) and event studies to examine how monitoring teachers' contributions to student performance on standardized exams and releasing this information to teachers and administrators affects student achievement. However, the effects of incorporating SGMs into teacher evaluations may not be the same for all districts and all students. Examining only the average effect will mask the distributional effects of the policy. Thus, I examine the heterogeneity across the distribution of prior district, school, and

student performance. I start by exploring how the policy may differentially affect student achievement by the prior district's and school's performance. Additionally, I examine how a student's performance distribution within a school has changed after the policy. I also examine the heterogeneous effect of policy across individual student performance.

I examine these issues in three different settings and data sets to provide a complete analysis of these effects: the Stanford Education Data Archive (SEDA), Ohio School Report Cards, and the data from the North Carolina Education Research Data Center (NCERDC). The SEDA contains district-level average test scores in math and reading of 3rd to 8th graders of all states in the U.S. from the academic years 2005-2006 to 2014-2015. I link the SEDA to the information data set, which includes the implementation details and the list of pilot districts and exploit the different timing of districts adopting the SGMs between 2006 and 2015 to examine the policy impact on student performance. Using the SEDA instead of state-specific data enables me to rely on consistent measures of student achievement that are more nationally representative.

In addition, I use school-level data from Ohio to examine the impact of providing value-added (VA) measures to teachers on student performance. Ohio provides the quasi-experimental setting to explore these questions on account of the policy variation brought by the pilot program. The Ohio Department of Education (ODE) had run a pilot of this policy involving 139 school districts before the statewide implementation. The key feature of the pilot was that teachers in pilot schools received the VA information while teachers in control schools did not. I construct a panel of school-level proficiency data from the Ohio School Report Cards for the academic years 2005-2006 to 2016-2017. A nice feature of this school-level panel data is that the performance categories of each school can be broken down into five ordered categories. The within-state variation of VA adoption in Ohio and school-level data allows me to examine not only the overall effects of the policy but also the distributional effects across schools. I additionally use student-level data from North Carolina to connect my analysis to the existing literature exploring this question.

The provision of teacher-level effectiveness measures, such as SGMs, introduces different incentives for which may have a differential impact on students. First, SGMs may enhance student learning by providing the ground for more rigorous evaluation systems. Principal-agent theory suggests that if a supervisor monitor over employees, an agent's work effort will increase. One thinking of school as firms, the education authorities want to implement a system that is designed to induce more effort from teachers. This claim is well supported in the literature since even subjective evaluations appear to improve teacher performance and student achievement (Taylor and Tyler, 2012; Jacob and Lefgren, 2008). Thus, it is reasonable to expect that the policy could generate a positive impact on average students' performance, as the stated aim of providing SGMs is giving valuable feedback to teachers and improving their teaching quality. However, it is possible that such expectations will not be met. There could be unintended consequences on students as teachers strategically react to the policy. For instance, since the SGMs are necessarily tied to students' test results, teachers may redirect their effort level depending on the students' initial ability. Teachers may focus less on high-achieving students since there is less scope of achieving higher growth in their test scores. Neal and Schanzenbach (2010) showed that after the NCLB, teachers would focus on the students who are at the margin of passing, rather than students who were already proficient or those far from becoming proficient. Also, providing the VA measures to teachers increase the mobility of highly effective teachers, especially to the high-performing schools, could lead to a positive or negative impact on students. Thus, it is valuable to investigate how providing information about the effectiveness of individual teachers ultimately affects students' performance.

The results from all analyses convey a consistent story: providing teachers with the performance score based on SGMs negatively impacts average student performance in math, and this is driven by the deterioration in the performance of students at the top. Students in school districts with teachers accessing their performance information performed significantly lower on math tests than those in control districts by 0.016 SD. Looking at the time-varying

effects, by the fourth year of implementation of the policy, the students' test scores decreased by 0.083 SD. This negative effect is due to the deterioration in students' performance from the previously high performing school districts.

When looking at the school-level performance, I find no significant effects that can be attributed to the policy on overall proficiency level in reading or math in Ohio. Also, there are no differences between high-performing districts and low-performing districts. However, I find a significant shift in the share of students at the more advanced levels towards the proficiency margin for both reading and math. This result shows the evidence of student distribution is being shifted downward or compressed to the mean. The students at the top of the distribution negatively affected by the policy.

However, it is still unclear whether compression of the student performance distribution within the school necessarily points us to the evidence of teachers' redirecting their effort from the students with certain initial ability. In the Appendix B, I provide an additional analysis of the distributional effect of the adoption of VA by exploiting the variation coming from two school districts in North Carolina. I find that within the same classroom, during the same year, previously high-achieving students were adversely affected in their math scores.

This study contributes to the existing literature by providing policy-relevant information on the impact of teacher accountability system on student achievement. First, to my knowledge, this is the first paper to use nationally representative data to examine the causal effects of the adoption of the SGMs on student performance. Many studies in previous literature focus on examining the impact of the school accountability system, such as NCLB, on student performance (Dee and Jacob, 2011; Reback et al., 2014; Ladd, 2012). While there was an effort to isolate the causal effects of NCLB by using the comparison between own accountability systems to the national program, the overall test score effects of NCLB are inconclusive. For example, Dee and Jacob (2011) found that NCLB led an increase in math score for 4th-grade students while Reback et al. (2014) and Ladd (2012) found that the significance goes away as they manipulate the sample years. Neal and Schanzenbach

(2010) showed that there is an overall gain in student achievement while showing this effect originates from the students close to the proficiency margin. In my paper, I provide the first causal evidence of the impact of providing teacher-level performance measures on students by exploiting regional variation and nationwide data. In addition, I showed that there could be a differential impact on students, even when teachers have less incentive to focus on the proficiency threshold.

Second, my paper is closely related to several papers that focus on how information and evaluation influence teachers and, consequently, students performance. Bergman and Hill (2018) and Pope (2019) looked at the effect of LA Times ratings on teachers. Both papers find that teachers with low rating improved their performance when informed of their rating based on VA scores. However, one study found that the public rating did not affect overall test scores of students, while the latter found a beneficial impact on students test scores. It should be noted that positive student and teacher sorting drive these mixed results of the effects. Unlike these studies, I focus on the teacher-level VA information given to teachers rather than to parents. This feature limits the possibility of sorting, which gives less biased estimates in my study.

Lastly, two papers explore the direct effect of the adoption of the VA in North Carolina. Lee (2019) found that the release of VA information increased students math scores by 0.1 standard deviations and this effect is driven by low VA teachers improvement in their performance. Bates (2019) exploited the variation in principals access to the VA information and found that the release of VA information promoted within-district mobility for high VA teachers and out-of-district mobility for low VA teachers. These two papers found positive point estimates on the overall effects of the policies. However, both studies heavily rely on the variation brought by just two districts introducing the possibility on underestimates of the cluster-robust standard errors. Thus, the provided estimates require a caution in drawing the statistical inference. I provide evidence of this bias by conducting a randomization test in Appendix B.

The remainder of this paper is organized as follows. Section 2 provides a more detailed discussion of the policy background and describes the data and the sample for the analysis. Section 3 explains the empirical framework. Section 4 presents the main results, and robustness checks, and Section 5 concludes.

2 Background and Data

The quality of a teacher is one of the most important factors found to promote students' immediate learning and even their long-term outcomes, such as job earnings (Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Chetty et al., 2014). Also, there is a substantial variation in teacher quality in raising students achievement on standardized tests (Rivkin et al., 2005). As a result, policymakers and educators have long been interested in finding accurate and efficient ways to identify effective teachers and to improve teaching quality. In theory, teacher evaluations are used to inform teachers of their performance and guide their professional development. However, teacher evaluation systems relied heavily on subjective measures and failed to differentiate the heterogeneity in teachers effectiveness. Recognizing the teachers role in education function, and failure of the teacher evaluation system, the education policy in the U.S. moved toward focusing on how to discern the quality teachers effectively.

Moreover, federal government intervention encouraged the states to focus on developing more rigorous evaluation systems. After the NCLB laid the foundation for states to reform teacher evaluation systems, the federal Race to the Top competition created strong incentives for states to make specific changes. Among the directed changes was the requirement to develop the high-stakes system comprised of multiple measures of teacher performance, including metrics based on students performance on standardized tests and increasing the frequency of the evaluation.

Driven by those incentives and growing recognition of the importance of considering teachers' contributions to students, states rushed to adopt the new system with student growth data. In 2009, only 15 states required objective measures of student growth in teacher evaluations; by 2015, this number increased to 43 (NCTQ, 2019). Among the many methods used for teacher effectiveness measurement, Value-Added (VA) models, and Student Growth Percentile (SGP) models are widely used. While conceptually similar, the two models differ in the estimation method. The VA models compare a students predicted performance to the average performance of a given teachers students. The SGP models compare students progress to that of other students with similar past performance.[2] In this analysis, I do not discern between the VA measures and the SGP since both use the student test scores in measuring teacher effectiveness.

Yet, there has been some push to delay linking teacher evaluations with test scores. Many states that passed such legislation faced delay due to many reasons, including pushback from teacher unions, and technical difficulty in developing the new rubrics. Thus, the actual implementation date can be different from the time of passage of the law. Also, some states even dropped the requirement of objective measures of student growth since 2015. Thus, currently, as of 2019, 34 states require teacher evaluations to include objective measures of student growth, down from a high of 43 in 2015(NCTQ, 2019).

To confirm the exact policy implementation year and its detail, I compiled data on the state's teacher evaluation system by a systematic search and outreach process. The details of this policy are shown in Table A1 which presents the year of actual implementation for each state as well as the set of states without such policy.[3] I began by reviewing the State Teacher Policy Database from the National Council on Teacher Quality that contains the detail information on state laws, rules, and regulations of the teaching profession. I then reviewed information on state education agency websites to verify policy implementation details. In addition, I searched for research papers, reports, and news articles to research

[2]In 2015, 15 states used VA measures, and 19 states use SGP models.

[3]Note that Washington, DC is excluded both from Table A1 and from this analysis.

whether the implementation details are different from the passed legislation. Lastly, for the states where I couldn't find reliable information, I directly contacted the agency to request such information. My rigorous search produced data on the detailed information of a complete set of teacher evaluation systems of 50 states with great attention to student data use. This information is crucial in conducting this research because omitting this could bias the results.

I want to point out several finding that this research produced. First, the research revealed that many states stalled the implementation of the new teacher evaluation systems. This means that even after the passage of legislation to link student performance on standardized test scores and teacher performance evaluation, the actual use of these measures was put on hold for some states. Second, many states ran a pilot of a new teacher evaluation system before the statewide implementation. Even though the pilot does not necessarily use the student growth measure in calculating the summative rating, many pilot states provide the measure to teachers. Third, there is vast heterogeneity in types, and the percentage of SGMs are used.

I link this information data set to the SEDA to conduct nationwide analysis. The SEDA includes a range of detailed data on educational outcomes in school districts and counties across the United States. Mainly, SEDA contains the district-level average test score in math and reading of 3rd to 8th graders of all states in the U.S. from 2008-2009 to the 2014-2015 school year. The most useful feature of the SEDA comes from providing the test scores in common metrics across states and districts, which allows the researcher to compare the test scores across the state, district, and year. In addition to the test score, the data includes a rich set of district characteristics such as gender and ethnicity/race composition, and percentage of students who have English Learner status, Special Education status, and Free Lunch status.

In this paper, I refer to state or district adopting policy when teachers start to receive their performance measures that use SGMs. Many states have started distributing the

performance measures that are tied to the students test score growth before incorporating those measures in their final ratings. I focus on the provision of the performance information to teachers rather than the summative ratings based on those measures. Also, I only consider the treatment status when a state or a district adopted the policy by 2015. There are two reasons for this. First, the SEDA contains the test scores only till the 2014-2015 school year and allows me only to examine the policy until these years. More importantly, many states repealed the use of SGMs in their teacher evaluation after 2015.

Although I focus on releasing the information on teacher effectiveness to teachers, some states use those performance measures in the final evaluation rating, and further tie it to the reward and sanction. The most common way to reward teachers is by performance pay for teachers. Teachers effort level can change with the monetary incentive. To control this, I link the information regarding the rewards for teachers from the Schools and Staffing Survey (SASS). The SASS is conducted by the Department of Education every few years and surveys a stratified random sample of teachers who provide information on their background, compensation, attitudes, school activities, and teaching methods. I am particularly interested in one variable the data contains: how much percentage of school districts indicated that they used financial incentives to reward excellence in teaching. Unfortunately, SASS is available in three waves: 2003,2007 and 2011. Thus, I use 2003 as a baseline and use the other two years to control how much growth in this percentage has happened in the states.

Descriptive statics for the sample by policy adoption status are reported in Table A2. Data includes 11994 school districts from 50 states, which results in 330443 of observation. Note that none of the non-adopted states have won the Race to the Top, and districts in adopted states are more likely to use teacher compensation than non-adopted states. One caveat of SEDA is that test scores for the grade-subject pairs are missing. For example, Arizonas 8th-grade math scores are not reported in 2009, 2010, and 2015 due to the technical issues. However, as I am using district-level average test scores across grades for each subject, the bias coming from the missingness is minimal. Still, it would be a concern if this is a

non-random missing for certain states. To check whether the states missing test scores for some grades is causing bias, I check the sensitivity of the coefficient estimates as I drop each grade from 4 to 8. The results are robust to this sensitivity test.

2.1 Ohio

While the SEDA and linked information uniquely offer the opportunity to examine the nationwide impact, I also explore the policy variation in Ohio. Ohio is one of the states that started a new teacher evaluation system, under the Race to the Top initiative, called the Ohio Teacher Evaluation System (OTES). The significant alteration in teacher evaluation in Ohio was shifting the teacher evaluation process from looking at how teachers do to what students learn in the classroom. Previously the teacher evaluation focused on the component where the supervisors conduct formal (and informal) observations on teacher performance in the classroom. With the introduction of OTES, teachers are now partially evaluated based on student performance measured by student growth.

The OTES is comprised of two categories of measurements: Teacher Performance on Standards and VA score. Each category makes up 50% of the teachers Final Summative Rating. The Teacher Performance on Standards is an assessment by administrators through classroom observation. The VA score is provided to the teachers by SAS through their Education Value-Added Assessment System (henceforth EVAAS). Until the 2013-2014 school year, EVAAS calculates teacher-level VA based on Ohio Achievement Assessments (OAA) results for grades 4 through 8 in reading and mathematics. For teachers whose VA is not available (for example, for teachers not teaching reading or mathematics for grade 4-8, or simply data not available), the school districts may use other assessments provided by national testing vendors and approved for use in Ohio.

This paper explores a variation introduced by a pilot program where 139 of 611 school districts participated during the 2011-2012 school year before the statewide implementation of OTES in 2013. The purpose of the pilot was mainly to inform the teachers and principals

of the new components of OTES. Each pilot district could choose one of four approved teacher evaluation models they planned to implement. The four models either implemented OTES or developed the evaluation system to align to the OTES, which both either include a student growth measure or not. However, regardless of the evaluation models that they choose to pilot, the major part of the pilot was providing teacher-level VA (before only providing school-level VA was provided). Teachers teaching grades 4 to 8 in either Reading or Mathematics had access to their VA information during the pilot year. The information on the pilot is gathered through personnel email correspondence with the Ohio Department of Education. Through this process, a list of pilot participants was provided, and I confirmed that the teachers in pilot schools actually received the VA score. It is reasonable to use the pilot program as the policy variation the same as previous analysis, and I am focusing on the provision of the VA scores to teachers.

To evaluate the VA policy impact on student performance in Ohio, I mainly use the school report cards that provide the public records of each schools and district's performance information. From the publicly available data in the Ohio Department of Education (ODE) websites, I compiled a school-level panel from the 2005-2006 school year through the 2016-2017 school year. The data contains various measures of student performance as well as school characteristics. The primary outcome I use in the analysis is the school-level performance of students in math and reading. For each grade and subject, the percentage of students who are proficient in each school is provided. This proficiency level can be further broken down into five ordered categories: advanced, accelerated, proficient, basic, and limited. This last feature of the data makes it possible to examine the distributional impact of the policy by exploring the within-school variation.

To avoid any bias in estimates confounded by the effect of schools closing or opening, I require schools to have the proficiency level reported throughout the entire sample year. If a school has been closed due to their poor performance, including it in the analysis would bias the coefficient of interest. Thus, I limit the sample to those who have such information

on each year of the sample period.[4] In the end, there are 597 pilot schools out of the total of 2108 schools in the analysis sample. The data also includes student characteristics of schools such as the composition of students' gender and race by each grade, as well as the percentage of limited English proficiency students and economically disadvantaged students.

Summary statistics of certain key variables for the pilot and the control school in Ohio samples are shown in Table A3. The table presents mean pre-treatment school-level characteristics for the pilot schools and the rest of the schools in Ohio. On average, pilot schools have a higher proportion of black students, a higher percentage of students with limited English proficiency, and economically disadvantaged students than the rest of Ohio. However, this will not be a problem in implementing the empirical method I use in the analysis as the difference between the two groups stays stable throughout the year.

3 Empirical Strategy

As stated, this paper aims to evaluate the impact of releasing student growth measures on student performance by exploiting the variation across states, districts, and schools. All analyses mainly use the differences-in-difference (DID) method utilizing the different timing of adoption with different data sources. In this section, I describe the empirical strategies for each data set.

Mainly, I exploit within-state (or within-district), and cross-cohort differences in exposure to the policy driven by cross-state (or cross-district) variation in the timing of when or whether states (or districts) adopted the policy in a difference-in-difference framework. This involves comparing the differences in average student outcomes before and after the adoption of the individual teacher student-outcome based performance measures within states (or districts) that adopt the policy against changes over the same time frame in states (or districts) that do not adopt the policy. I estimate:

[4] I exclude charter schools from the analysis since charter schools operate teacher evaluations differently than public schools.

$$Y_{ist} = \beta_o + \alpha_i + \lambda_t + \theta_g + \beta_1 D_{ist} + X'_{ist}\gamma + \epsilon_{ist} \quad (1)$$

I also consider the district-level variation within state. Since some districts adopt the objective teacher evaluations ahead of the rest of the state, considering this variation gives more power to identify the causal effects. Y_{ist} is the mean student achievement in district i in state s in year t. In the SEDA, the student achievement is given as district average by each grade and subject. D_{ist} is an indicator for whether a district currently release the student growth measures to teachers. The district fixed effects control for variation in outcomes that are common across students within a district, and the year fixed effects account for national shocks that impact all students in the same year. I also control for the proportion of students of ethnicity/race (Black, Asian, Hispanic), free lunch status, English limited status in each district. These controls are in the vector X in equation (1). All standard errors are clustered at the district level as the level of treatment assignment (Abadie et al., 2017).

Adopting the individual teacher effectiveness measure can change teachers effort as teachers become more aware of the policy over time. This can generate a time-varying treatment effect based on the length of exposure to the policy. The source of variation comes from the fact that each state adopted the student growth measure in different years. Thus, I employ an event study model that examines how outcomes changed among students who were differentially exposed to the policy that had been in place for different lengths of time based on which state, district and in which grade they were in.

Equally important, this event study model allows me to inspect the evidence of the key difference-in-differences assumption. Conditional on the controls in the model, the variation in policy exposure comes from two sources. The first is within-state (or district) differences in exposure over time driven by the year of the policy adoption. The second is cross-state (or district) variation in the timing of when or whether states adopted the policy. The assumption underlying the identification of parameters is that the policy should not be endogenous to unobserved state-level shocks. That is, the decision of whether and when to

adopt the policy must be uncorrelated with any prior trends in outcomes. For example, if a state adopted policy after having a negative trend in student test scores, the policy estimates would spuriously capture the positive impact even if the policy did not have causal impact on students. I estimate an event study model as follows:

$$Y_{ist} = \beta_o + \alpha_i + \lambda_t + \theta_g + \sum_{\tau=-k}^{K} \beta_\tau D_{ist}^{\tau} + X_{ist}'\gamma + \epsilon_{ist} \tag{2}$$

The variables used are same as previous equation except that the DID estimator is replaced with the event study indicators. The variable $D_{ist}^{\tau} = I\left(t - t_{0i} = \tau\right)$ is an indicator equal to one for being τ time periods relative to is initial treatment (t_{0i}) with τ ranges from -7 to 7 and $\tau = 0$ being the year of initial treatment. For example, if a district adopted the VA measure in year 2012, it will have a relative time of -1 for the year 2011 and 1 for the year 2013. This variable takes value of zero in states that have never been had a VA measure adoption. β_τ can be interpreted as estimates for pre-trends (for $\tau \leq 0$) as well as time-varying treatment effects (for $\tau > 0$). I omit the estimate for the $D_{ist}^{0} = I\left(\tau = 0\right)$ such that all β_τ estimates are relative to the year of adoption. Equation (2) also includes grade (θ_g), district (α_i), and year (λ_t) fixed effects. Standard errors are clustered at district level. The parameters of interest in equation (2) are β_1 to β_7, which show the time-varying effects of the policy among students who are first exposed to this policy in relative years 1 to 7. Also, the β_{-7} to β_{-1} estimates in equation (2) serves as a test of the assumption that there is no selection.

I also conduct several robustness checks. First, the existence of alternative policies that were implemented concurrently with the adoption of student growth measures can be a threat to the identification. One policy that can directly affect teachers behavior and thus have impact on student outcome is Teacher Incentive Pay. Financial incentives may have positive impact on students achievement by improving teachers effort level. However, it can have no impact if teachers were teaching at their highest effort level, or not knowing how to increase student achievement and a negative impact if teachers cheat. I control this by

using the Schools and Staffing Survey (SASS). I include an additional control variable that indicates the percentage of teachers receiving performance pay tied to the student test scores. I am also including the indicator of the Race to the Top winner states where 18 states and D.C. won awards that ranged from $17 million to $700 million. Race to the Top promotes states to develop and adopt standard assessment system with a statewide longitudinal data, evaluation system of teachers and principal based on performance. Since the Race to the Top encourages states to have performance based and standardized assessment system, winning the award can show the states' interest in improving their student achievment. Indeed, winners implemented more education related policies.

In addition, I examine the sensitivity of the results to outliers: whether a particular state is driving the effect. I estimate equation (1) 50 times, each time excluding a different state from the sample. Lastly, I use the state-level adoption years instead of using the district-level variation. The impact would be different to teachers when the state officially adopts the student growth measures instead of the local education agency adopting the policy. I provide the results of all robustness checks in the next section, along with the results from the main analyses.

The effects of providing student growth measures to teachers may not be the same for all students. In order to examine whether the differential impacts on student performance, I relax the linearity assumption in equation (1) and (2) and consider the heterogeneous achievement of districts. I explore whether the districts have differentially impacted by the policy depending on their initial performance. The policy may differentially affect the schools that were previously high or low performing. Teachers in different schools might exert different levels of effort to improve their students' performance. The teachers in high performing schools were already having a student with relatively higher test scores where the additional growth coming from students might be small. This can make teachers put less effort, which led to the null effect of the policy. In contrast, low-performing schools can be experiencing higher growth in students' test scores since most of the students have a large

room to increase. I estimate the following specification:

$$Y_{ist} = \beta_o + \alpha_i + \lambda_t + \theta_g + \beta_1 D_{ist} + \beta_2 D_{ist} \times I(H_{ist-1}) + X'_{ist}\gamma + \epsilon_{ist} \tag{3}$$

Where $I(H_{ist-1})$ is indicator of initial district achievement; and all other variables are defined as in equation (1). Indicator variable of previously high-achieving districts uses districts' performance of one year before the policy adoption year. The parameter of interest is β_2 which measures the differential impacts of the VA policy on students test scores of initially high-performing districts relative to that of low-performing districts. In addition to the heterogeneity analysis using difference-in-differences method, I also present the results using the event study models. These are estimated using equation (3) except $I(H_{ist-1})$ is interacted with the event study indicators.

While cross-state analysis to measure the impact of the policy on student achievement is meaningful, I examine the effects of policy in more detail by exploiting the cross-school district variation in the adoption of VA measures in Ohio. As described earlier, Ohio ran a pilot of adopting VA measures before the statewide implementation. The 4th- to 8th-grade teachers in pilot schools received the VA information while teachers in control schools did not. I again use difference-in-differences approach to compare the student performance in the treatment schools relative to the control schools after the policy was implemented. In order for this approach to be valid, the policy should not be endogenous to unobserved district-level shocks. In addition to the graphical evidence of the parallel trend in pre-pilot periods Figure A.1, I estimate an event study specification that allows for a complete set of interactions between the indicator of treatment status and years. The result is represented in Figure A.2, which confirms the assumptions in using the specification, I proceed to the main empirical model as follows:

$$Y_{ist} = \beta_o + \alpha_i + \lambda_t + \theta_g + \beta_1 D_{is,1t} + \beta_2 D_{is,2t} + X'_{ist}\gamma + \epsilon_{ist} \tag{4}$$

Since the teachers teaching in grades 4 to 8 received the VA score, analysis is limited to 4th to 8th grade students. In order to accommodate the two-step roll-out of the policy including pilot and statewide implementation, there are two dummy indicators in the equation. $D_{is,1t}$ is a dummy variable that takes the value one for the treated schools after the OTES pilot in year 2011-2012, while $D_{is,2t}$ equals one for the treated schools after the OTES statewide implementation in year 2013-2014. Since the treated schools here are defined for the pilot participant schools, the initial pilot participants remain the same even after the statewide implementation. Thus, β_1 represents the immediate effect of the pilot and β_2 represents the long-term effect of the pilot. X_{ist} is the vector of student characteristics of schools that includes the share of limited English learners, economically disadvantaged students, gender, and race/ethnicity. In addition, the model includes a set of grade (θ_g), school (α_i), and year(λ_t) fixed effects. I provide the standard errors clustered at the district level as the pilot was assigned at district level.

Y_{ist} is the student performance variable of school i in district s and in year t. There are two sets of outcomes that I use in the analysis. First, to see the overall effect of the policy on student performance, I use the percentage of students that are proficient for each school, which I retrived from the Ohio Report Card. Using this outcome allows me to examine the overall policy impact on student outcome.

Second, to gain more insight into the overall effect, I examine how the policy affected the proportion of students in each performance level category. The schools report the number of students in ordered category levels: advanced, accelerated, proficient, basic, and limited. Examining only the change in percentage of students who pass the proficiency level of performance can mask the distributional effects of the policy. Thus, I estimate the same DID model (equation (4)) for five different outcomes, respectively: the percentage of students in each ordered category. Since I am estimating five separate regressions with the outcomes that are correlated with each other, the equation errors would be correlated. That is, the set of five equations has contemporaneous cross-equation error correlation since each proficiency

category is summed to one and change in one part is accompanied by a change in other parts by definition. For example, an increase in one group must be followed by a decrease in another group. To address this issue, I adjust the standard error of each regression by implementing a seemingly unrelated regression (SUR).

In addition to exploring the heterogeneous impact of policy by examining how distribution of students change, I investigate the heterogeneous impact of the policy across schools. To explore this heterogeneity, I ask to what extent the evaluation pilot differentially impacted achievement in pilot schools with different levels of prior achievement. I divide the schools into two groups (high- and low-performing) based on the average performance during the pre-policy period. Similar to the previous specification, I estimate the following:

$$Y_{ist} = \beta_o + \alpha_i + \lambda_t + \theta_g + \beta_1 D_{is,1t} + \beta_2 D_{is,2t} + \beta_3 D_{is,1t} \times I(High_{ist-k}) + X'_{ist}\gamma + \epsilon_{ist} \quad (5)$$

4 Result

This section describes the results from all the analyses, including the test for the identification strategy, main results, and robustness checks.

4.1 Nationwide

Figure 1 and Figure 2 show the full set of estimates using the event study model of equation (1) with including all control variables. I also overlay a linear fit for the pre- and post-treatment periods to see if there are differential pre-trends and if there are time-varying treatment effects. The visual evidence in Figure 1 supports the identification strategy: there is no evidence of differential trends in test scores in pre-treatment periods in math. The point estimates of pre-treatment periods are small and insignificant. As the school districts adopt the SGMs, the effect on math scores is small and remains unchanged for the first three

years. From the fourth year of adoption, the math scores decline as a function of exposure time. However, the presence of pre-trends is detected in Figure 2. Reading scores linearly declines as the exposure time to the policy while exhibiting the significant positive pre-trends in reading scores. It can be interpreted as the policy change appears to have an effect on the outcome before it is implemented. Thus, I only report the results for math scores in nationwide analysis.

The estimates shown in Table 1 confirms the adverse effects on student achievement in math shown in Figure 1. The estimates in column (3) that includes control variables as well as alternative policy controls indicate that there is negative impact on students with the adoption SGMs. Attending the schools in a district with SGMs decrease students' math score by 0.016 SD. The different timing of policy adoption can generate a time-varying treatment effects based on the length of the exposure to the policy. Table A4 reports the estimates from event study model. The results indicate that there is no immediate effect on the students as the estimates are insignificant in the early years of policy adoption. The effect grows and becomes significant, beginning four years of adoption. These results indicate that attending school in a state with this policy reduces students math achievement by approximately 0.083 SD after four years of adoption. The effect grows to 0.175 SD after seven years of adoption.

These results are somewhat striking since it shows the opposite results from the previous literature which estimates the effects of teacher evaluation on student performance. For example, two pieces of evidence from Cincinnati Public Schools show that subjective measures of teacher effectiveness promote student achievement (Kane et al., 2011; Taylor and Tyler, 2012). More recent evidence from Chicago public schools shows that schools that participated in a teacher evaluation program designed to improve classroom instruction through structured principal-teacher dialogue performed better in reading (Steinberg and Sartain, 2015). One reason for these contrasting results is that teachers with high performance scores may reduce their effort in teaching, and thus their influence on student achievement gains decreases. Also, teachers' strategic reaction to the policy may lag since it requires time for

teachers to understand the policy better. This can explain maybe the fact that there are no effects on test scores by the third year of implementation but starts declining from the 4th year of the implementation. After teachers learn about their productivity, teachers may reduce effort, and this translates into a negative impact on students' performance.

While I cannot directly examine the change in the effort level of the teachers, I investigate the possible mechanism through examining the heterogeneous effects of the policy: how the policy differentially affects the school districts depending on their previous performance. Column (2) in Table 2 presents the estimates of heterogeneous treatment effects from equation (3). The students in relatively high-performing school districts experience a stronger negative impact on their math scores. The policy reduces math scores by 0.0461 SD in high-performing school districts relative to the low-performing school districts. This heterogeneous effect of policy on student performance further confirmed by the event study model, where the same specification is used except the indicator of high-performing districts is interacted with a series of relative time indicators in equation (2). Table A5 presents the full set of estimates from a single estimation equation. The estimates in column (4) represent the time-varying effects of the policy on high-performing districts relative to low-performing districts. Notably, there is an immediate negative effect on students math scores for the high-performing school districts. Attending school districts in states with the SGMs reduces the students math score by 0.038 SD in the first year. The effect grows to 0.115 SD three years after the policy adoption. However, the negative impact on student becomes small and insignificant from year 5, and it becomes positive in year 7.

The results are robust to a range of alternative specifications. First, the negative effect on the student test score could be spurious if one or many early-adopted states are driving the results. Thus, I examine the sensitivity of my findings to outliers by reestimating equation (1) 50 times for the outcome, each time dropping a different state from the analysis sample. The results of this exercise are reported in Figure A.3. The estimates are insensitive to excluding each one state. Second, to see whether early-adopted districts within the states

are driving the result, I estimate equation (1) by using the state adoption year instead of the district's adoption year. Both estimations of overall and heterogeneous effects are similar to the main analysis. The results of using state-level adoption year of SGMs are presented in Table A6 and Table A7.

4.2 Ohio

In order to further investigate the possible mechanism driving the results found in the nationwide analysis, I exploit the variation in the timing of adoption of VA in Ohio. The event study estimates shown in Figure A.2 not only confirm the parallel trend assumption in DID but also provide the preview of the results. There are no significant effects of VA on school performance in both subjects. Table 3 summarizes the average impact of the VA policy on student achievement from equation (4). Again, the results confirm that there is no impact on student performance measured in the percentage of students who are proficient or above.

Although I did not find any VA impact on the average performance level of schools, exploring the heterogeneous effects of the policy could give us more insight into the overall effect. Examining only the change in the percentage of students who meet the proficiency level of performance can mask the distributional effects of the policy. There are two ways to examine the heterogeneous effects of the policy. First, I examine how the policy affected the proportion of students in each performance level category: advanced, accelerated, proficient, basic, and limited. Treating these ordered categories as a separate measure of school performance, I estimate the DID model (equation (4)) for five different outcomes, respectively.

Table 4 reports the full set of results for both subjects. For both subjects, there is a decline in the percentage of students in the advanced category, while there is an increase in the proficient category. The percentage of the most advanced performance level decreased by 1.6 percentage points, and 0.9 percentage points decline in the accelerated performance level in Math. At the same time, there is a 0.7 percentage point increase in proficient level

and basic level, respectively. The results from reading scores exhibit a similar pattern to those from math scores, while the magnitude is much smaller. The percentage of students in advanced performance levels decreased by 0.53 percentage points, while the percentage of students in proficient performance level increased by 1.2 percentage points. Although it is difficult to identify whether a decrease in top category increases the middle category, the results confirm that the student distribution is compressed toward the middle. The fact that there is a significant decrease in the percentage of students who are ranked at the highest performance level serves as a piece of suggestive evidence that the VA is detrimental to the previously high-performing students.

Second, I examine the heterogeneous effect across schools by comparing the policy impact between high-performing schools and low-performing schools. As shown in Table 5, there are no differential impacts detected for high-performing schools relative to low-performing schools in both overall proficiency level and five categories of performance.

5 Conclusion

Many states in the U.S. initiated policies to encourage their public schools to effectively use any form that incorporates student growth data calculated from test scores. With the federal intervention NCLB and Race to the Top program, states rushed to use student growth measures as one component that measures teacher performance. The use of SGMs in teacher evaluation rapidly expanded over the past decade: from 15 states in 2009 to 43 states in 2015. Despite the importance of this policy, there have been very few opportunities to evaluate the impact of this policy. The rigorous data requirements and the unavailability of relevant policy rollout are the main reasons. This paper overcomes the limitation by exploiting different timing in the adoption of SGMs with three different settings and data and examines how teachers receiving their performance measures calculated from student test score growth affects student achievement and how students may be differentially impacted by this policy.

I first examine the impact of providing objective performance measures to teachers on overall student achievement. Using variation in the timing of policy adoption across states and districts, I find that providing SGMs to teachers leads to a decrease in average math scores. That is, attending schools in a district with SGMs decrease students math score by 0.016 SD. Exploring the time-varying treatment effects of the policy, I find that attending school in a district with VA policy reduces students math achievement by approximately 0.083 SD after four years of adoption, and it grows to 0.175 SD after seven years of adoption.

These results are surprising since the aim of providing objective measures of performance is encouraging teachers to put more effort into improving their performance, which could lead to an increase in student achievement. One possible explanation of opposite empirical findings from the expectation is that teachers may strategically react to the policy by focusing on students with certain initial academic achievement.

It also implies that the impact of this policy may not be the same for all districts and all students. Thus, I examine the heterogeneity across the distribution of prior district, school, and student achievement. Investigating the heterogeneous effects of policy across the distribution of districts, I find that previously high-achieving districts suffer from a bigger negative impact on math scores. These adverse effects are larger for the upper end of the distribution of district performance. By further investigating the differential effects, I find that the evidence of the detrimental impact of the VA for the students at the top of the performance distribution in schools. In Ohio, I find the distribution of student performance in schools with VA shifting toward the mean. In both subjects, the percentage of the students in the top performance category decreases as the schools adopt VA.

These results show that SGMs have an unwanted effect on student achievement, undermining the performance of students in a top portion of the distribution. The contraction of the distribution of student achievement suggests that policy may alter the equity of education. Providing VA to teachers make it easier for teachers to move to other schools since the VA serves as a signal of productivity of teachers. Indeed, Bates (2019) showed that highly ef-

fective teachers move to higher-performing schools. When VA is provided, increased teacher mobility and generated transition costs from these movements can harm student achievement and even exacerbate the gaps between districts (Boyd et al., 2008). However, if the VA measures are targeted for improving the teaching quality and effort level of the low effective teachers rather than rewarding the highly effective teachers, the discrepancy in added effort level with VA measures can cause the contraction of distribution of student achievement.

References

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in chicago public schools. *Journal of Labor Economics*, 25(1):95–135.

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research.

Bates, M. D. (2019). Public and private employer learning: Evidence from the adoption of teacher value-added.

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

Boyd, D., Grossman, P., Lankford, H., Loeb, S., and Wyckoff, J. (2008). Who leaves? teacher attrition and student achievement. Technical report, National Bureau of Economic Research.

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

Dee, T. S. and Jacob, B. (2011). The impact of no child left behind on student achievement. *Journal of Policy Analysis and management*, 30(3):418–446.

Jacob, B. A. and Lefgren, L. (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of labor Economics*, 26(1):101–136.

Kane, T. J., Taylor, E. S., Tyler, J. H., and Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of human Resources*, 46(3):587–613.

Ladd, H. (2012). Commentary on dee and jacob. *Brooking Papers on Economic*.

Lee, H. (2019). The effect of releasing teacher performance information to schools: Teachers? response and student achievement. Job Market Paper.

NCTQ (2019). Measures of professional practice national results. *State Teacher Policy Database*.

Neal, D. and Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, 92(2):263–283.

Pope, N. G. (2019). The effect of teacher ratings on teacher performance. *Journal of Public Economics*, 172:84–110.

Reback, R., Rockoff, J., and Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under no child left behind. *American Economic Journal: Economic Policy*, 6(3):207–41.

Rivkin, S., Hanushek, E., and Kain, J. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2):417–458.

Rockoff, J. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review: Papers and Proceedings of the One Hundred Sixteenth Annual Meeting of the American Economic Association*, 94(2):247–252.

Steinberg, M. P. and Sartain, L. (2015). Does teacher evaluation improve school performance? experimental evidence from chicago's excellence in teaching project. *Education Finance and Policy*, 10(4):535–572.

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

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., and Morgan, K. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *New Teacher Project*.

Figures and Tables

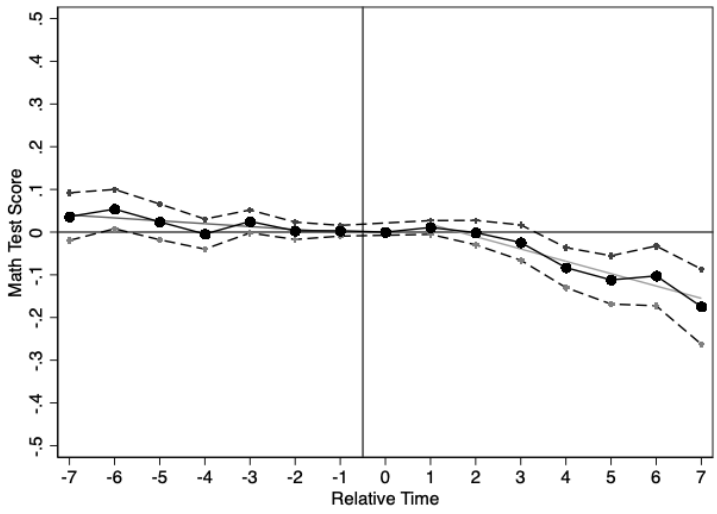


Figure 1: Event Study - Math

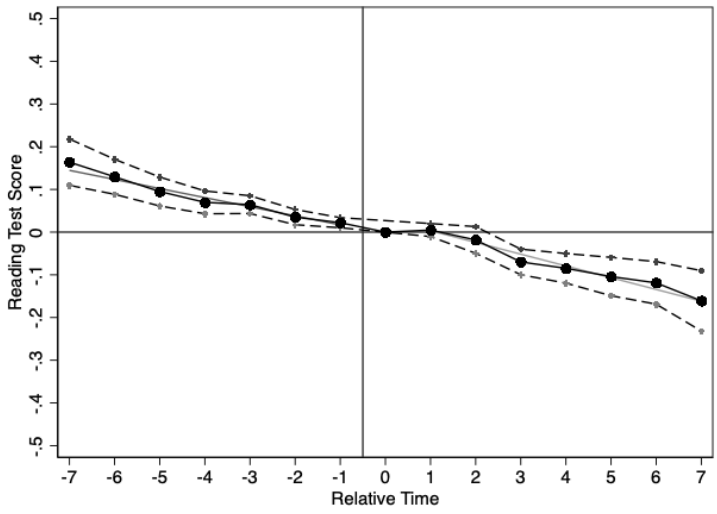


Figure 2: Event Study - Reading

Table 1: The overall effects of releasing SGMs to teachers on student math scores

	(1)	(2)	(3)
Relative Years to Policy Adoption	-0.0092**	-0.0059	-0.0063
	(0.0039)	(0.0041)	(0.0042)
Treated	0.0264***	0.0261***	0.0257***
	(0.0088)	(0.0083)	(0.0084)
Relative Years * Treated	-0.0088	-0.0159***	-0.0161***
	(0.0054)	(0.0052)	(0.0053)
Control variables	N	Y	Y
Policy controls	N	N	Y
Grade FE	Y	Y	Y
Year FE	Y	Y	Y
District FE	Y	Y	Y
R-squared	0.8634	0.8644	0.8644
N	328,215	328,215	328,215

* Note: The table presents the estimates from equation (1) using the SEDA. Grade, Year, District FE are included in all specifications. Column (2) uses the control variables including student charactriscits of the districts such as gender, race/ethnicity, special education status, limited english learners, and free lunch status. Column (3) adds the policy control such as RTTT winner states and percentage of teachers receiving performance pay in the districts. Standard errors clustered at district level are shown in parentheses. Statistically significant at ***1%, **5%, and *10%

Table 2: The heterogeneous effects of releasing SGMs to teachers on student math scores

	(1)	(2)	(3)
DID	-0.0161***	0.0178	-0.0283***
	(0.0053)	(0.0116)	(0.0050)
DID*High		-0.0461***	
		(0.0123)	
DID*Low			0.0461***
			(0.0123)
Control variables	Y	Y	Y
Policy controls	Y	Y	Y
Grade FE	Y	Y	Y
Year FE	Y	Y	Y
District FE	Y	Y	Y
R-squared	0.8644	0.8660	0.8660
N	328,215	328,215	328,215

* Note: The table presents the estimates from equation (3) using the SEDA. Grade, Year, District FE as well as control variables are included in all specifications. Column (1) shows the estimates of high-performing districts and Column (2) shows the estimates of low-performing districts. Standard errors clustered at district level are shown in parentheses. Statistically significant at ***1%, **5%, and *10%

Table 3: The overall effects of releasing VA to teachers on student achievement

	Math		Reading	
	(1)	(2)	(3)	(4)
Pilot	-0.014	-0.018	0.0023	-0.0014
	(0.013)	(0.011)	(0.0086)	(0.0047)
Statewide	-0.020	-0.019	-0.0061	-0.0076
	(0.014)	(0.012)	(0.012)	(0.0092)
Control variables	N	Y	N	Y
Grade FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
School FE	Y	Y	Y	Y
Observations	48,769	48,769	48,769	48,769
R-squared	0.785	0.751	0.807	0.784

* Note: The table presents the estimates from equation (3) using Ohio data. Grade, Year, District FE are included in all specifications. Column (2) and Column (4) use the control variables including student charactriscits of the schools such as gender, race/ethnicity, limited english learners, and economically disadvantage status. Standard errors clustered at district level are shown in parentheses. Statistically significant at ***1%, **5%, and *10%

Table 4: The distributional effects of VA on student achievment

VARIABLES	(1) advanced	(2) accelerated	(3) proficient	(4) basic	(5) limited
Panel A. Math					
pilot	-0.016**	-0.0089**	0.0072**	0.0071**	0.011
	(0.0074)	(0.0039)	(0.0034)	(0.0035)	(0.0082)
statewide	-0.0090*	-0.012*	0.0020	0.0030	0.016
	(0.0053)	(0.0066)	(0.0059)	(0.0040)	(0.013)
R-squared	0.711	0.517	0.504	0.575	0.685
Panel B. Reading					
pilot	-0.0053**	-0.0048	0.012**	0.0028	-0.0051
	(0.0024)	(0.0046)	(0.0052)	(0.0030)	(0.0081)
statewide	-0.0056	0.0046	-0.0051	0.0058*	0.00030
	(0.0052)	(0.0044)	(0.0051)	(0.0034)	(0.014)
R-squared	0.587	0.634	0.599	0.645	0.673
Control variables	Y	Y	Y	Y	Y
Grade FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
School FE	Y	Y	Y	Y	Y
N	48,769	48,769	48,769	48,769	48,769

* Note: The table presents the estimates from equation (4) using Ohio data. Grade, Year, District FE, and control variables are included in all specifications. Standard errors are adjusted with SUR and shown in parentheses. Statistically significant at ***1%, **5%, and *10%

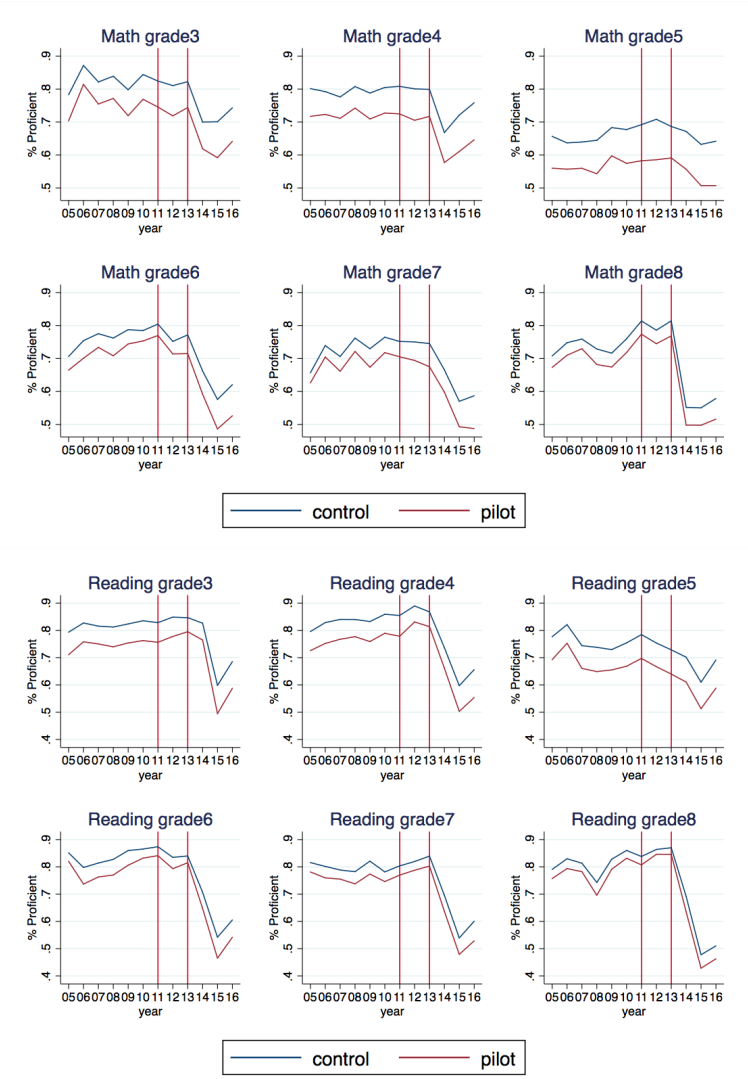
Table 5: The heterogeneous effects of VA on student achievment

VARIABLES	(1) overall	(2) advanced	(3) accelerated	(4) proficient	(5) basic	(6) limited
Panel A: Math						
DID	-0.023	-0.017*	-0.010*	0.0052	0.013***	0.0094
	(0.017)	(0.0097)	(0.0063)	(0.0052)	(0.0047)	(0.014)
DID*High	0.014	-0.0013	0.0087	0.0069	-0.0055	-0.0087
	(0.017)	(0.0082)	(0.0083)	(0.0073)	(0.0051)	(0.017)
R-squared	0.752	0.712	0.519	0.505	0.578	0.688
Panel B: Reading						
DID	0.00056	-0.0055	-0.0063	0.012**	0.0051	-0.0056
	(0.0091)	(0.0035)	(0.0063)	(0.0048)	(0.0039)	(0.0078)
DID*High	0.0021	0.0065	-0.00076	-0.0036	-0.0038	0.0016
	(0.012)	(0.0060)	(0.0060)	(0.0057)	(0.0039)	(0.011)
R-squared	0.784	0.588	0.635	0.600	0.645	0.674
Control variables	Y	Y	Y	Y	Y	Y
Grade FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
School FE	Y	Y	Y	Y	Y	Y
N	48,769	48,769	48,769	48,769	48,769	48,769

* Note: The table presents the heterogeneous effects in Ohio. Grade, Year, District FE, and control variables are included in all specifications. Column (1) shows the impact on overall proficiency level of the school. Column (2)-(3) present the heterogeneous effects. Standard errors are adjusted with SUR and shown in parentheses. Statistically significant at ***1%, **5%, and *10%

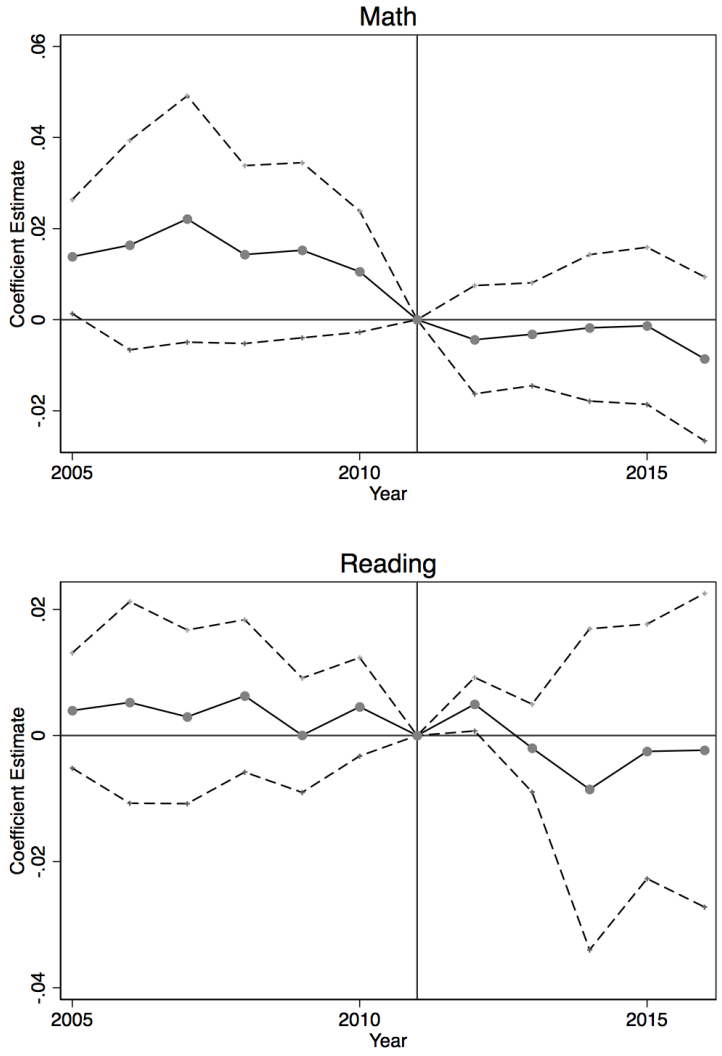
A Appendix A

Figure A.1: Proficiency level of schools in Ohio



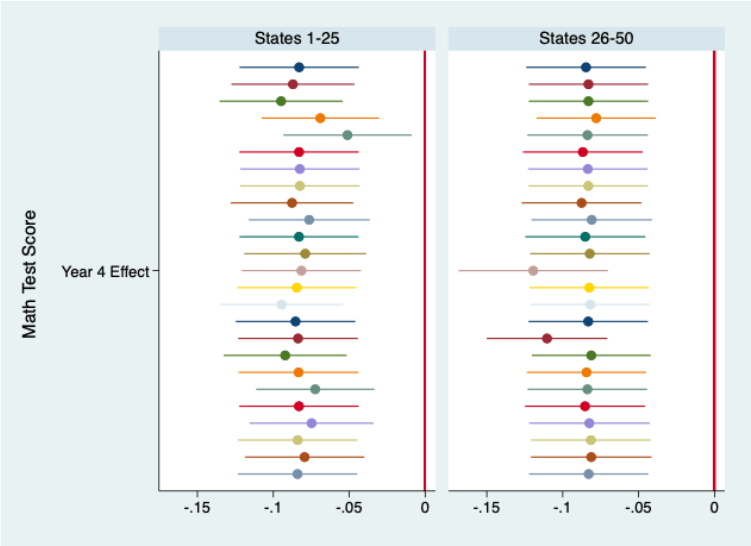
Notes: The graphs show the percentage of students who are proficienct by control and treatment groups. The teachers in pilot schools recieved VA scores in 2011. Ohio implemented the VA measures statewide in 2013. The significant drop in proficiency level since 2014 is due to the change in test system in Ohio.

Figure A.2: Event study results for Ohio



Notes: The graphs present the result of event study model for Ohio. The point estimates show the effect of providing VA scores to teachers on school proficiency level relative to year of implementation.

Figure A.3: Robustness check: sensitivity to outliers



Notes: The graphs present the result of robustness check from estimating equation (1) as described in the text using the SEDA. Each point represents a point estimate excluding a given state from the regression. The 95% confidence intervals are calculated using standard errors that are clustered at the district level.

Table A1: State policy detail

	AL	AK	AZ	AR	CO	CT	DE	FL	GA
Pilot									
Year of Partial Pilot			2012-2014	2012-2013	2011-2012	2012-2013			2011-2014
Year of Full Pilot	2009-2010	2015-2016		2013-2014			2011-2012		
Student Test Score Used	N	Y	Y	N	Y	Y	Y		Y
Summative Rating Provided	Y	N	Y	N	N	N	N		N
Percentage			20						
Statewide									
New Teacher Eval. System									
Year of VA provided								2011	
Year of the New Teacher Eval. System	2015	2016	2013	2014	2013	2014	2012	2012	2014
Student Test Score Used	N	N	Y	N	Y	Y	Y	Y	Y
Type of Measure			student growth		student growth	SLO	SLO	VA	SGP
Student Growth in Eval.									
Year of Inclusion in Summative Rating			2013		2013	2014	2012	2012	2014
Percentage			33		50	45	50	50	50

	HI	ID	IL	IN	KS	KY	LA	ME	MD
Pilot									
Year of Partial Pilot	2011-2013			2011-2012	2011-2012	2012-2013	2009-2012		2011-2013
Year of Full Pilot			2015-2016		2013-2014	2013-2014		2014-2015	
Student Test Score Used	N		Y	Y	N	Y	Y	Y	Y
Summative Rating Provided	N		N	Y	N	N	Y	Y	Y
Percentage				25-50			50		50
Statewide									
New Teacher Eval. System									
Year of VA provided			2012				2011		
Year of the New Teacher Eval. System	2013	2014	2016	2012	2014	2014	2012	2015	2013
Student Test Score Used	Y	Y	Y	Y	Y	Y	Y	Y	Y
Type of Measure	SGP	student growth	SLO	SGP	student growth	SLO	VA	SLO	SLO
Student Growth in Eval.									
Year of Inclusion in Summative Rating	2013	2014	2016	2016	2014	2014	2012	2015	2013
Percentage	25	33	30	no info	no info	no info	50	20	50

	MA	MI	MN	MS	MO	NV	NH	NJ	NM
Pilot									
Year of Partial Pilot			2013-2014	2011-2012	2012-2013		2012-2013	2011-2013	2012-2013
Year of Full Pilot		2011-2013				2013-2014			
Student Test Score Used		Y	Y	Y	Y	N	Y	Y	Y
Summative Rating Provided		Y	Y	N	N	N	N	Y	Y
Percentage			35				0-25	0-10	
Statewide									
New Teacher Eval. System									
Year of VA provided									
Year of the New Teacher Eval. System	2013	2015	2014	2014	2014	2015		2013	2013
Student Test Score Used	Y	Y	Y	Y	Y	Y		Y	Y
Type of Measure	SGP	SGP	VA	SGP	SLO	student growth		SGP	VA
Student Growth in Eval.									
Year of Inclusion in Summative Rating	2014 for level 4 2015 for all	2015	2014	2014	2014	2016		2013	2013
Percentage	no info	25	35	30	no info	20		45	50

	NY	ND	NC	OH	OK	OR	PA	RI	SC
Pilot									
Year of Partial Pilot			2008-2009	2011		2012-2013			2013-2014
Year of Full Pilot			2009-2012		2012-2013	2013-2014			
Student Test Score Used			Y		Y	Y			Y
Summative Rating Provided			N		Y	N			N
Percentage					35				30
Statewide									
New Teacher Eval. System									
Year of VA provided					2012		2006		
Year of the New Teacher Eval. System	2011	2015	2011	2013	2013	2014	2013	2012	2015
Student Test Score Used	Y	Y	Y	Y	Y	Y	Y	Y	Y
Type of Measure	SGP	student growth	VA	VA	VA	SLO	student growth	SGP	VA
Student Growth in Eval.									
Year of Inclusion in Summative Rating	2011	2015	2011	2013	2015	2014	2015	2013	2016
Percentage	20	no info	15	50	35	no info	15	30	20

	SD	TN	TX	UT	VA	WA	WV	WI	WY
Pilot									
Year of Partial Pilot	2013-2014	2010-2011	2014-2016	2012-2013	2011-2012	2010-2012	2011-2012	2012-2013	2014-2015
Year of Full Pilot									
Student Test Score Used	Y	N	Y	N	Y	N	N	Y	N
Summative Rating Provided	N	N	Y	N	Y	N	N	N	N
Percentage					40		20	50	
Statewide									
New Teacher Eval. System									
Year of VA provided		1996							
Year of the New Teacher Evaluation System	2014	2011	2016	2015	2012	2015	2015	2014	2016
Student Test Score Used	Y	Y	Y	Y	Y	N	N	Y	Y
Type of Measure	SLO	VA	VA	SGP	SGP	SLO		SLO	SGP
Student Growth in Eval.									
Year of Inclusion in Summative Rating	2015	2011	2017	2015	2012	2015	2015	2014	2016
Percentage	no info	35	20	20	40	no info		50	no info

Table A2: Summary Statistics of SEDA

	Adopted		Non-adopted	
	Mean	SD	Mean	SD
Math Score	256.2807	20.93127	261.2401	20.34998
ELL	0.069	0.129	0.033	0.064
Free Lunch	0.399	0.213	0.381	0.210
Special Ed	0.119	0.058	0.142	0.043
Asian	0.029	0.070	0.019	0.041
Black	0.066	0.160	0.090	0.178
Hispanic	0.170	0.248	0.109	0.177
White	0.716	0.302	0.755	0.265
RTTT	0.000	0.000	0.450	0.503
Performance Pay	0.500	0.527	0.725	0.452
N	57,460		272,983	

* Notes: The table presents the descriptive statistics for the sample used in nationwide analysis using the SEDA.

Table A3: Summary Statistics Ohio

	Pilot (Treated)		Control	
	Mean	SD	Mean	SD
unit of observations: school-year (2005-2010)				
Female	0.485	(0.036)	0.485	(0.021)
Limited English	0.042	(0.082)	0.026	(0.080)
Econ Disadvantaged	0.545	(0.270)	0.401	(0.258)
Asian	0.025	(0.060)	0.018	(0.039)
Black	0.248	(0.307)	0.127	(0.241)
Hispanic	0.046	(0.080)	0.033	(0.075)
White	0.651	(0.330)	0.791	(0.269)
N	3582		9066	

* Notes: The table presents the descriptive statistics for the sample from ODE

Table A4: Event study model

Event Study Model	(1)	(2)	(3)
Year -7	0.061**	0.034	0.036
	(0.026)	(0.027)	(0.028)
Year -6	0.073***	0.052**	0.054**
	(0.021)	(0.023)	(0.023)
Year -5	0.036*	0.021	0.024
	(0.020)	(0.021)	(0.021)
Year -4	0.009	-0.007	-0.005
	(0.017)	(0.017)	(0.018)
Year -3	0.034***	0.023*	0.025*
	(0.013)	(0.013)	(0.014)
Year -2	0.013	0.002	0.003
	(0.010)	(0.010)	(0.010)
Year -1	0.006	0.003	0.003
	(0.007)	(0.006)	(0.006)
Year 1	0.018**	0.012	0.011
	(0.008)	(0.008)	(0.008)
Year 2	0.014	-0.000	-0.001
	(0.014)	(0.014)	(0.015)
Year 3	-0.006	-0.023	-0.024
	(0.021)	(0.020)	(0.021)
Year 4	-0.061***	-0.083***	-0.083***
	(0.023)	(0.023)	(0.024)
Year 5	-0.085***	-0.109***	-0.112***
	(0.026)	(0.026)	(0.029)
Year 6	-0.080***	-0.097***	-0.102***
	(0.030)	(0.033)	(0.036)
Year 7	-0.150***	-0.169***	-0.175***
	(0.039)	(0.042)	(0.045)
Control variables	N	Y	Y
Policy controls	N	N	Y
Grade FE	Y	Y	Y
Year FE	Y	Y	Y
District FE	Y	Y	Y
R-squared	0.864	0.865	0.865
N	328,215	328,215	328,215

* Notes: The table presents the estimates from equation (2) using the SEDA. Grade, Year, District FE are included in all specifications. Column (2) uses the control variables including student charactriscits of the districts such as gender, race/ethnicity, special education status, limited english learners, and free lunch status. Column (3) adds the policy control such as RTTT winner states and percentage of teachers receiving performance pay in the districts. Standard errors clustered at district level are shown in parentheses. Statistically significant at ***1%, **5%, and *10%

Table A5: Event study model with hetereogeneous effects

	Event Study Estimator (1)		(2)		Event Study Estimator * High (3)		(4)
Year -7	0.113***	Year 1	0.029**	Year -7	-0.161***	Year 1	-0.038***
	(0.037)		(0.012)		(0.030)		(0.013)
Year -6	0.112***	Year 2	0.038*	Year -6	-0.144***	Year 2	-0.067***
	(0.028)		(0.020)		(0.021)		(0.023)
Year -5	0.047*	Year 3	0.039	Year -5	-0.072***	Year 3	-0.115***
	(0.027)		(0.029)		(0.023)		(0.031)
Year -4	0.022	Year 4	-0.031	Year -4	-0.074***	Year 4	-0.080**
	(0.023)		(0.039)		(0.020)		(0.041)
Year -3	0.034*	Year 5	-0.085**	Year -3	-0.035**	Year 5	-0.037
	(0.017)		(0.043)		(0.016)		(0.041)
Year -2	-0.006	Year 6	-0.108**	Year -2	0.008	Year 6	0.009
	(0.013)		(0.055)		(0.013)		(0.053)
Year -1	-0.023***	Year 7	-0.227***	Year -1	0.044***	Year 7	0.085
	(0.009)		(0.067)		(0.010)		(0.065)
Policy controls	Y						
Grade FE	Y						
Year FE	Y						
District FE	Y						
R-squared	0.865						
N	328,215						

* Notes: The table presents the estimates from equation (2) using the SEDA. Grade, Year, District FE are included in all specifications. Column (2) uses the control variables including student charactriscits of the districts such as gender, race/ethnicity, special education status, limited english learners, and free lunch status. Column (3) adds the policy control such as RTTT winner states and percentage of teachers receiving performance pay in the districts. Standard errors clustered at district level are shown in parentheses. Statistically significant at ***1%, **5%, and *10%

Table A6: Robustness check: using statewide adoption year

	(1)	(2)	(3)
Relative Years to Policy Adoption	-0.0101**	-0.0070*	-0.0075*
	(0.0040)	(0.0041)	(0.0042)
Treated	0.0234**	0.0230**	0.0224**
	(0.0103)	(0.0092)	(0.0094)
Relative Years * Treated	-0.0083	-0.0155***	-0.0158***
	(0.0056)	(0.0054)	(0.0055)
Control variables	N	Y	Y
Policy controls	N	N	Y
Grade FE	Y	Y	Y
Year FE	Y	Y	Y
District FE	Y	Y	Y
R-squared	0.8634	0.8644	0.8644
N	328,215	328,215	328,215

* Notes: The table presents the estimates from equation (1) using the statewide adoption year as a treatment. Standard errors clustered at state level are shown in parentheses. Statistically significant at ***1%, **5%, and *10%

Table A7: Robustness check: using statewide adoption year

	(1)	(2)	(3)
DID	-0.0158***	0.0089	-0.0374***
	(0.0055)	(0.0120)	(0.0047)
DID*High		-0.0463***	
		(0.0127)	
DID*Low			0.0463***
			(0.0127)
Control variables	Y	Y	Y
Policy controls	Y	Y	Y
Grade FE	Y	Y	Y
Year FE	Y	Y	Y
District FE	Y	Y	Y
R-squared	0.8644	0.8649	0.8649
N	328,215	328,215	328,215

* Notes: The table presents the estimates from equation (2) using the statewide adoption year as a treatment. Standard errors clustered at state level are shown in parentheses. Statistically significant at ***1%, **5%, and *10%

B Appendix B: North Carolina

B.1 Background and Data

Before the statewide implementation of individual teacher VA into the teacher evaluation system in 2013, two school districts started releasing VA information to teachers and principals: Guilford in 2000 and Winston-Salem in 2008. Same as the case from Ohio, teacher effectiveness of each teacher in this report is estimated using the EVAAS. SAS calculates the VA measure for each given academic year, and this information is presented in the EVAAS teacher report.

The VA information is estimated for teachers who teach subjects for which the state of North Carolina requires end-of-grade (EOG) assessments for math and reading for 4th to 8th graders. SAS did not provide VA measures for 3rd grader teachers. Once the VA score is estimated using the student test scores, teachers and principals can access the VA information. Also, the EVAAS teacher report presents the teacher effects at different student achievement levels: low, middle, and high. Teachers having access to this information might give incentives for teachers to strategically react to the VA score disclosure by focusing their efforts on students at different levels of initial ability. Thus, I exploit two school districts adoption of individual VA measures ahead of the state adoption to examine the differential effects on students.

In order to further explore the heterogeneous effect of providing VA scores to teachers on student performance, I use the student records covering the period from 1997-2011 of NCERDC. The data from the NCERDC contains the EOG math/reading test scores of 3rd to 8th graders and a rich set of students, school, teacher characteristics. The student characteristics include grade, gender, race, and exceptional status, including the academically gifted.

One main advantage of using the data from the NCERDC is the ability to define students in the same classroom. The primary objective of this part is to determine whether and how

differentially students learning is influenced when a teacher receives VA information. Thus, it is important to identify the students that correspond to the test scores form each classroom. The data from NCERDC contains the identifiers that attempt to link students in the same classroom, which allows me to identify a students performance relative to the entire test score distribution in that classroom. I used the student's first attempt test score, thus excluded the re-attempt record for the analysis. I also excluded charters schools from the sample. In the end, there are 116 school districts in the sample.

Guildford started receiving the teacher VA score from 2000, and Winston-Salem began in 2008. Summary statistics of certain key variables for North Carolina, including Guilford and Winston-Salem, are shown in Table B1. Table B1 compares the means and standard deviation of Guildford (Winston-Salem) and the rest of the districts for the pre-treatment periods, respectively. The two districts that adopt VA do have a higher percentage of black students than the rest of the state. However, the averages in achievement in these districts are not different from the state average.

B.2 Empirical Strategy

I examine the overall effect of VA policy on student achievment by implementing difference-in-differences model and event study model. First, I use the event study model as presented above to examine whether policy has a mean effect on student performance.

$$Y_{ist} = \beta_o + \alpha_s + \lambda_t + \theta_g + \sum_{\tau=-k}^{K} \beta_\tau D_{ist}^{\tau} + X_{ist}'\gamma + \epsilon_{ist} \tag{B.1}$$

With the availability of the individual-level data from NCERDC, Y_{ist} becomes the achievement in of student i in district s in year t. The test scores are normalized with the pre-treatment periods. The variable $D_{ist}^{\tau} = I\left(t - t_{0i} = \tau\right)$ is an indicator equal to one for being τ time periods relative to is initial treatment (t_{0i}). τ ranges from -3 to 3 for Guildford and

Winston-Salem with $\tau = 0$ being the year of initial treatment. The control districts are randomly divided into two groups and attached to each of the treated districts for the analysis. Thus, the sample is the stack of two groups: 1997-2003 years for Guilford and 2005-2011 years for Winston-Salem, and corresponding control districts.

There are two reasons why I define the estimation sample in this way. First, I need to have a student's records in pre-treatment periods and after-treatment periods, as I want to examine the distributional effect of the policy. Second, Winston-Salem displays the negative trend in 2002-2004 years (shown in Figure B1), so including these years in the analysis will spuriously estimate the positive effect for Winston-Salem. I also control for the student characteristics, such as gender, race/ethnicity, and gifted status. The standard errors are clustered at school district level. The parameters of interest in equation (B.1) are β_1 to β_3, which show the time-varying effects of the policy among students who are first exposed to this policy in relative years 1 to 3. I show a full set of beta estimates in the Figure B2 and Table B2. The β_{-3} to β_{-1} estimate in equation (B.1) serves as a test of the parallel trend assumption.

I also examine the effects with difference-in-differences model.

$$Y_{ist} = \beta_o + \alpha_s + \lambda_t + \theta_g + \beta_1 D_{ist} + X'_{ist}\gamma + \epsilon_{ist} \tag{B.2}$$

Only difference in this equation is that D_{ist} is an indicator for whether a district currently release the VA to teachers. Here two treatment districts are Guilford and Winston-Salem.

I further investigate the heterogeneous effect found in Ohio by using the student-level test scores from North Carolina.

$$Y_{ist} = \beta_o + \alpha_s + \lambda_t + \theta_g + \beta_1 D_{ist} + \beta_2 D_{ist} \times I(H_{ist-1}) + X'_{ist}\gamma + \epsilon_{ist} \tag{B.3}$$

Where $I(H_{ist-1})$ is indicator of initial district achievement; and all other variables are defined as in equation (B.1). In order to investigate how the effect of VA varies across

the distribution of test scores, I generate the indicator of students previous achievement status in two ways. First, to see where a student stands relative to the states test score distribution, I compare the students prior achievement to the state median. Second, to see whether effect differs on students achievement status within the classroom, I compare the student's performance to the class average. In addition to the heterogeneity analysis using difference-in-differences method, I also present the results using an event study model. These are estimated using the same equation except the indicator of previous achievement is interacted with the event study indicators.

B.3 Results

Table B2 shows the results of the event study model for math score. I only report the results for the math scores due to the pre-trends in reading scores. As shown in the Figure B3, reading scores in pre-policy periods show positive trends. Thus, the result of negative estimates in the event study model might be representing the test scores reverting to the mean.

First, the estimates in column (2) shows the evidence of VA positively affecting average test scores in math. The math scores immediately increase by 0.052 SD, and the size of the estimate grows to 0.154 SD after two years and 0.183 SD after three years. Column (1) in Table B3 shows the results from DID model. Similar to the results from event study model, students attending schools in districts with VA policy have 0.075 SD gain in math score. This result is similar (although slightly small in magnitude) to findings from (Lee, 2019): he found a 0.096 SD increase in math in Guildford.

However, the estimates from the heterogeneous analysis show that the VA is detrimental to the top half of the distributin, even with a positive impact on the average test scores. Column (3) and (4) from Table B2 and column (2) from Table B3 shows the negative impact on high-performing students. These results indicate that VA harms the previously high-performing students, which is one possible explanation as to why my results differ from

those found in the literature.

However, these results should be interpreted with caution since the statistical inference might be biased. Clustering standard errors with only a small number of clusters can underestimate the true standard errors. To address this issue, I conduct a randomization test. I randomly draw two treatment districts from the sample for 500 times and generate the distribution of t-statistics from the same regression. Comparing with true t-statistics from the results shows that there is possibility of overrejection. The results from randomization test is presented in Figure B4.

Figure B1: Math Score Trend

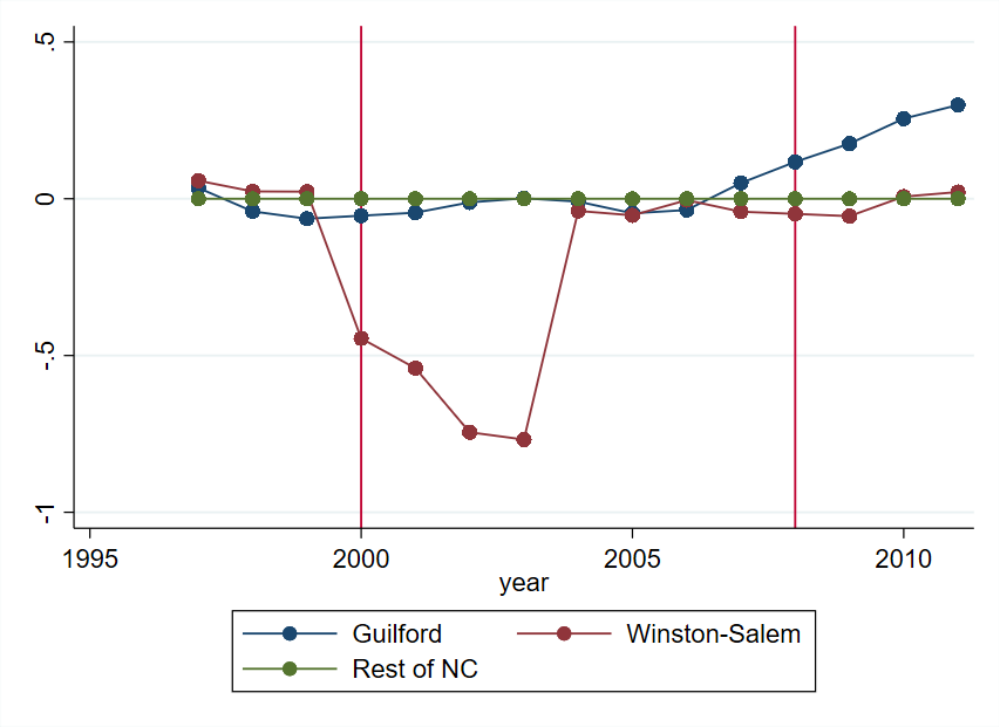


Figure B2: North Carolina: Event Study- Math

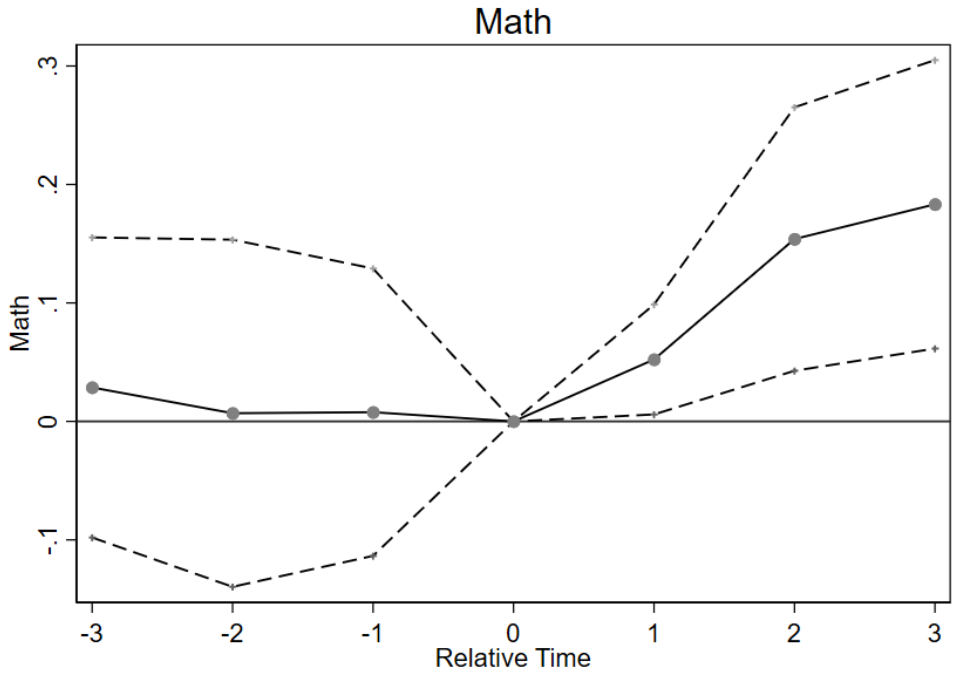


Figure B3: Reading Score Trend

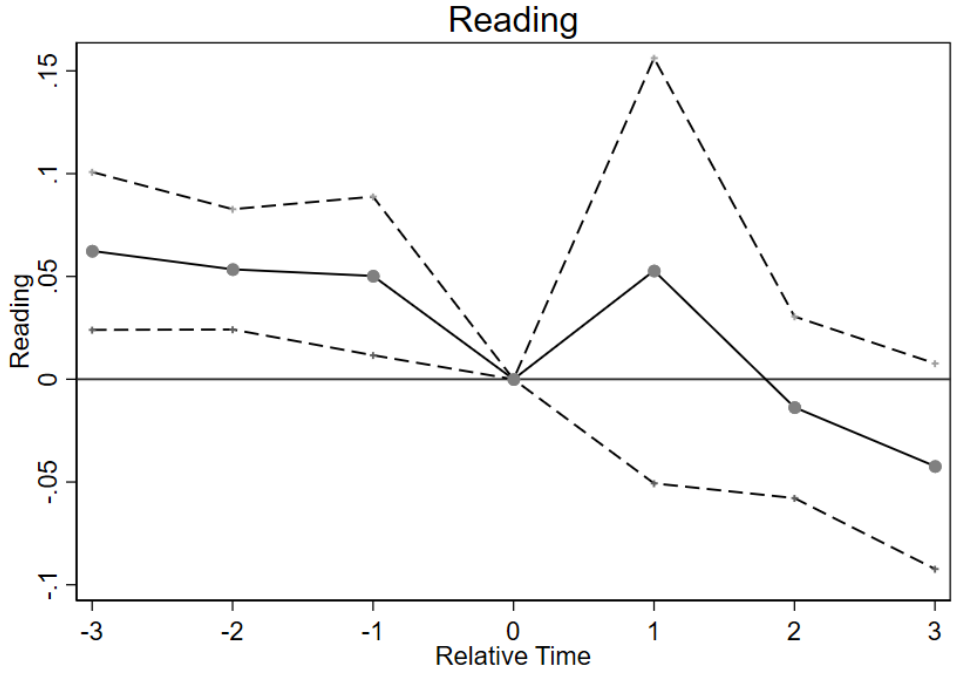


Figure B4: Robustness Check: Randomization Test

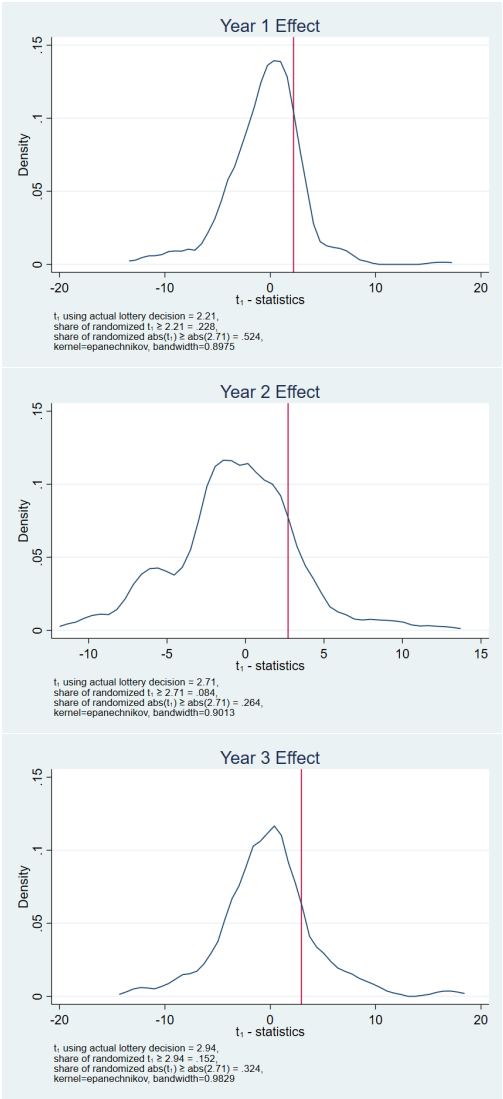


Table B1: Sample Summary Statistics

	Guilford		Rest of NC		Winston-Salem		Rest of NC	
	1997-1999				2005-2007			
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Black	0.34	0.22	0.27	0.24	0.36	0.26	0.24	0.24
Hispanic	0.03	0.07	0.02	0.04	0.07	0.08	0.08	0.10
White	0.62	0.23	0.68	0.26	0.47	0.29	0.61	0.28
Asian	0.01	0.03	0.01	0.04	0.05	0.07	0.02	0.04
Gifted	0.19	0.00	0.14	0.05	0.26	0.01	0.15	0.05
Female	0.51	0.12	0.51	0.11	0.50	0.11	0.50	0.12
N	36161		1273999		64660		1309826	

* Notes: The table presents the summary statistics from the NCERDC.

Table B2: Event Study Results

	(1)	(2)	(3)	(4)
Year -3	-0.00498	0.0286	0.0186	0.0437
	(0.0186)	(0.0646)	(0.0204)	(0.0714)
Year -2	-0.0278	0.00689	-0.0191	0.00186
	(0.0433)	(0.0747)	(0.0337)	(0.0800)
Year -1	-0.0298	0.00777	-0.127***	-0.109**
	(0.0291)	(0.0619)	(0.00964)	(0.0467)
Year 1	-0.00994	0.0523**	0.0327	0.0607***
	(0.0220)	(0.0237)	(0.0554)	(0.0144)
Year 2	0.0291	0.154***	0.199***	0.181***
	(0.0255)	(0.0567)	(0.0156)	(0.0206)
Year 3	0.0140	0.183***	0.247***	0.232***
	(0.0535)	(0.0621)	(0.0157)	(0.0210)
Year -3 * High			-0.0135	-0.0690
			(0.0220)	(0.0914)
Year -2 * High			0.0225	-0.0170
			(0.0168)	(0.0814)
Year -1 * High			0.202***	0.166***
			(0.0338)	(0.0513)
Year 1 * High			-0.0426	-0.0800***
			(0.0627)	(0.0223)
Year 2 * High			-0.170***	-0.122***
			(0.0137)	(0.0187)
Year 3 * High			-0.203***	-0.164***
			(0.0135)	(0.0188)
Control variables	N	Y	Y	Y
Grade FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
District FE	Y	Y	Y	Y
R-squared	0.028	0.309	0.578	0.532
N	1,732,693	1,732,693	1,732,693	1,732,693

* Notes: The table presents the results from event study model using the NCERDC. All specifications include Grade, District, Year FE. Column (1) includes no control variables. Column (3) and (4) show the heterogeneous effects. Column (3) use student achievment relative to the state performance, and column (4) use student achievment relative to the classroom in generating high-performing indicator. Standard errors clustered at district level are shown in parentheses. Statistically significant at ***1%, **5%, and *10%

Table B3: The effect of VA on student achievement

	(1)	(2)
DID	0.0752***	0.0987***
	(0.0242)	(0.0174)
DID*High		-0.0576***
		(0.0075)
Control variables	Y	Y
Grade FE	Y	Y
Year FE	Y	Y
District FE	Y	Y
R-squared	1,732,693	1,732,693
N	0.3089	0.5778

* Notes: The table presents the results from DID model using the NCERDC. All specifications include Grade, District, Year FE and control variables. Column (1) shows the overall effect of VA on student math score. Column (2) shows the heterogeneous effect using student achievment relative to the state performance.. Standard errors clustered at district level are shown in parentheses. Statistically significant at ***1%, **5%, and *10%