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

Hwanoong Lee*

JOB MARKET PAPER

This version: October 2018

The most recent version and the Online Appendix are available on my [website](#).

Abstract

This paper 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. ([JEL Codes: H75, I28, J08, J45](#)).

*Ph.D. Candidate, Department of Economics, Michigan State University, 486 W. Circle Drive, 110 Marshall-Adams Hall, East Lansing, MI 48824-1038 USA. Email: lee2hwan@msu.edu. Words can barely express how grateful I am to Todd Elder and Scott Imberman for their guidance, time and encouragement throughout the process of this project and throughout my MSU years. I am also indebted to Michael Bates, Joshua Cowen, Seth Gershenson, Steven Haider, Lesley Turner, Stephen Woodbury, and Ben Zou for their insights and suggestions, and participants at the 2018 Annual Meeting for Education Finance and Policy, and the 2017 Annual Meeting for Missouri Valley Economic Association for their helpful comments and discussion. Thanks also goes to Kara Bonneau and the North Carolina Education Research Data Center (NCERDC) as well as representatives of Guilford County Schools, and Winston-Salem/Forsyth Community Schools. All errors are my own.
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 (Doherty 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 (Glazerman et al., 2013). 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 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 et al., 2017; Imberman and Lovenheim, 2015; Dee and Wyckoff, 2015; Glazerman 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, 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 a set of empirical strategies 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.

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 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 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

\(^2\)See Ballou et al. (2004) for a detailed description of the model
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 (Bénabou 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 et al., 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 fi-
nally, 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.

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. 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. 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. 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

---

3In 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”

4I 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.
of the districts 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.

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.

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 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 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 Online Appendix Table C.1, 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. 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 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

---

5I 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.
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:

\[
y_{it} = \sum_{g=3}^{5} (Y_{it-1} \times Grade_g)\gamma_{1g} + X_{it}\gamma_{2} + \delta_{d} \times \delta_{t} + TG_{g} \times \delta_{t} + \beta_{0}TD_{d} \times TG_{g} + \beta_{1}TD_{d} \times TG_{g} \times Post_{t} + \varepsilon_{it} \tag{1}
\]

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 \((\delta_{d})\), year \((\delta_{t})\), 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 \(\beta_{1}\) 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 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. 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:

$$W^*(V) = \arg\min_w (X_0 - \sum_{d=1}^{D} w_d \cdot X_d)' V (X_0 - \sum_{d=1}^{D} w_d \cdot X_d)$$  \hspace{1cm} (2)$$

---

6In Online Appendix Figure B.1, 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.
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:

\[
\hat{\alpha}_t = y_t^0 - \sum_{d \in D} w^*_{d}(V)y_t^d
\]

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.

\[
DD_{ws} = \left( \frac{1}{T - T_0} \sum_{t \geq 2008} \hat{\alpha}_t \right) - \left( \frac{1}{T_0} \sum_{t < 2008} \hat{\alpha}_t \right)
\]

The remaining challenges to implementing the synthetic control is selecting the pre-treatment outcomes and predetermined variable as predictors. When the number of pre-treatment 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

\(^7\)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 (see, https://web.stanford.edu/~jhain/synthpage.html).
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 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 Online Appendix A. 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.\textsuperscript{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.

\textsuperscript{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 2, suggesting that prior test scores are more important than other demographic variables in predicting future test scores.
4.2 Results

4.2.1 Guilford

Figure B.2 in Online Appendix shows the estimates using the event study model with co-
variates, 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 achieve-
ment effects on student achievement in math. In column (1) of Table 3, I estimate the simple
DID 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 3, there is evidence that providing VA in-
formation 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, $\beta_1$ 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 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.\footnote{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.} In Online Appendix Figure B.3, 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 hypothesis 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 3 to evaluate how the failure of this assumption would affect the permutation inference. Figure 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 3). The vertical lines indicate the estimates reported in Table 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 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.

4.2.2 Winston-Salem

Online Appendix Figure B.4 shows the location of donor districts underlying the Synthetic Winston-Salem with a preferred predictor set.\(^\text{10}\) 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.

Figure 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

\(^{10}\)For interested readers, I report the combination of districts and weights underlying the Synthetic Winston-Salem for math and reading in Online Appendix Table C.2.
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 Online Appendix Figure B.5, 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 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 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 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.

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.
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.\footnote{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.} 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.\footnote{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.}

\[ y_{it} = Y_{ij-1} \beta_0 + X_{it} \beta_1 + C_{ijt} \beta_2 + S_{ist} \beta_3 + \mu_t + \mu_g + \mu_j + \epsilon_{it} \] (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 $\mu_t$ is a set of year effects to control for the year-specific shocks, $\mu_g$ is a grade fixed effect, and $\mu_j$ 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.\footnote{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 Online Appendix Table C.3, I reports the main results with alternative VA specifications.}

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, $\mu_j$, are normalized, and then linked to teachers in the 1999-2003 data for Guilford and 2006-2011 data for Winston-Salem.\footnote{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.} 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 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 generalizing 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:

\[
y_{it} = \alpha + \sum_{k=-2}^{2} \beta_k year_{T+k} \times VA_j + \sum_{k=-1}^{2} year_{T+k} + Y_{it-1} \gamma_1 + X_{it} \gamma_2 + C_{it} \gamma_3 + S_{ist} \gamma_4 + \theta_s + \epsilon_{it} \tag{6}
\]

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 \(\theta_s\) is a school fixed effect and \(\epsilon_{it}\) 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 \(\beta_k\), 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.\textsuperscript{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:

\[
y_{it} = \beta_0 + \beta_1 V A_j + \beta_2 T D_d \times V A_j + \beta_3 V A_j \times I(\text{year}_t = T - 1) + \beta_4 V A_j \times I(\text{year}_t \geq T) + \sum_{k=2}^{k=-2,k\neq-1} \delta_k \text{year}_t \times T D_d \times V A_j + Y_{it-1} \gamma_1 + \gamma_2 + C_{it} \gamma_3 + S_{ist} \gamma_4 + \delta_d \times \delta_t + \varepsilon_{it} \quad (7)
\]

The model includes an indicator variable, \( I(\text{year}_t = T - 1) \), which is equal to 1 when the year \( t \) is the adopting year, and \( I(\text{year}_t \geq T) \), 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 \( \beta_4 \) picks up the impact of other policies that are common across 15 large districts, and the set of parameters \( (\delta_k) \) from the three-way interaction terms demonstrate how the impact of one-standard-deviation-higher VA teachers

\textsuperscript{15}In Figure B.6 of the Online Appendix, I also show the estimates using between school variation as well as within-teacher variation. The results are similar regardless of the model specifications.
on student achievement gains were evolved differently in the treated districts.\textsuperscript{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.\textsuperscript{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.

5.2 Results

Figure 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

\textsuperscript{16}Since these coefficients are normalized to one year prior to the year when teachers first received the VA information, the $T \neq 1$ year triple interaction term ($year_{T-1} \times TD_d \times VA_j$) is the omitted variable. However, for Winston-Salem, the $T \neq 2$ 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 $T \neq 1$ year.

\textsuperscript{17}Chingos and West (2011) reported that high-VA teachers are less likely to be assigned to a low-stakes teaching position.
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 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.\footnote{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.}

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 Online Appendix Figure B.7, I display the estimated coefficients of the triple interaction terms in Eq.(7). Consistently with the results in Figure 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 7 presents 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 Online Appendix Figure B.8, 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.

5.3 Robustness Checks

To evaluate whether the baseline results are robust to how I construct teacher VA measures, Online Appendix Table C.3 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 \( (\delta_k) \) in Eq.(7), I estimate the more parametric model that the postpolicy period indicator is interacted with the two-way interaction term \( (TD_d \times VA_j) \).

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

\[19\] Specifically, I estimate the following model.
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 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. 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.

$$y_{it} = \beta_0 + \beta_1 VA_j + \beta_2 TD_d \times VA_j + \beta_3 VA_j \times I(year_t = T - 1) + \beta_4 VA_j \times I(year_t \geq T) + \delta_{1} TD_d \times VA_j \times I(year_t = T - 1) + \delta_{2} TD_d \times VA_j \times I(year_t \geq T) + y_{it-1} \gamma_1 + X_{it} \gamma_2 + C_{it} \gamma_3 + S_{ist} \gamma_4 + \delta_{d} \times \delta_{t} + \epsilon_{it}$$

The estimated parameter reported in Tables B.3 and B.4 is $\delta_2$. All notations are the same as in Eq.(7).

20See Kane and Staiger (2008) who outline the procedure to compute the EB estimates.
To evaluate how teacher attrition affects my results, in Table C.4 of the Online Appendix, 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 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.\textsuperscript{21}

\section{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.

\textsuperscript{21}Panel A of Online Appendix Table C.5 shows the corresponding p-values of the baseline estimates in Online Appendix Table C.4 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.
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 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.\textsuperscript{22,23}

Table 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

\textsuperscript{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.

\textsuperscript{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.
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 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 C.5 in Online Appendix. 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.

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 teach-
ers 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 mobil-
ity 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.

Table 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

24For example, the post-treatment period of my analysis for Guilford covers 2001 to 2003, whereas the
period for Bates spans from 2001 to 2011.
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 average test scores in treated districts.\(^27\)

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

\(^25\) 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.

\(^27\) 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. Unfortunately, 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.
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 information 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.
References


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


Figures and Tables

Figure 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

<table>
<thead>
<tr>
<th>Year</th>
<th>N</th>
<th>Mean Student Score</th>
<th>Mean Pred. Score</th>
<th>Pred. Score %tile</th>
<th>Pred. Score %tile</th>
<th>Teacher Effect</th>
<th>Std Error</th>
<th>Teacher vs Comparison Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>21</td>
<td>349.8</td>
<td>351.9</td>
<td>32</td>
<td>-1.9</td>
<td>0.5</td>
<td>Below</td>
<td></td>
</tr>
</tbody>
</table>

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 SAS 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.

Copyright © 2007 SAS Institute Inc., Cary, NC, USA. All Rights Reserved

Source: Author's reproduction using a copy of a Guilford County Schools’ Value-Added Report.
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 3: 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 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 4: Distribution of Placebo Estimates of Math Score by Number of School Districts

A. Use 109 control districts

B. Use 60 control districts

C. Use 30 control districts

D. Use 15 control 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 3 and the horizontal lines show the corresponding p-values.
Figure 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.
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 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 post-policy 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.
Table 1: Summary Statistics for Guilford, Winston-Salem, and the Rest of Districts

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Guilford</td>
<td>Rest of NC</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Female</td>
<td>0.507</td>
<td>0.500</td>
</tr>
<tr>
<td>White</td>
<td>0.493</td>
<td>0.500</td>
</tr>
<tr>
<td>Black</td>
<td>0.408</td>
<td>0.491</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.030</td>
<td>0.171</td>
</tr>
<tr>
<td>Gifted</td>
<td>0.188</td>
<td>0.390</td>
</tr>
<tr>
<td>Special Edu.</td>
<td>0.113</td>
<td>0.317</td>
</tr>
<tr>
<td>Parent Edu (less than HS)</td>
<td>0.468</td>
<td>0.499</td>
</tr>
<tr>
<td>Math score</td>
<td>0.070</td>
<td>1.017</td>
</tr>
<tr>
<td>Reading score</td>
<td>0.057</td>
<td>1.007</td>
</tr>
<tr>
<td>Observations</td>
<td>96,414</td>
<td>1,569,084</td>
</tr>
</tbody>
</table>

Unit of observations: Student-year

<table>
<thead>
<tr>
<th></th>
<th>Guilford</th>
<th>Rest of NC</th>
<th>Winston-Salem</th>
<th>Rest of NC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>0-3 years’ experience</td>
<td>0.226</td>
<td>0.418</td>
<td>0.223</td>
<td>0.416</td>
</tr>
<tr>
<td>4-10 years’ experience</td>
<td>0.261</td>
<td>0.439</td>
<td>0.261</td>
<td>0.439</td>
</tr>
<tr>
<td>11+ years’ experience</td>
<td>0.513</td>
<td>0.500</td>
<td>0.515</td>
<td>0.500</td>
</tr>
<tr>
<td>White</td>
<td>0.743</td>
<td>0.435</td>
<td>0.850</td>
<td>0.357</td>
</tr>
<tr>
<td>Black</td>
<td>0.236</td>
<td>0.425</td>
<td>0.137</td>
<td>0.344</td>
</tr>
<tr>
<td>Advanced Degree</td>
<td>0.277</td>
<td>0.448</td>
<td>0.277</td>
<td>0.448</td>
</tr>
<tr>
<td>Observations</td>
<td>4,871</td>
<td>89,122</td>
<td>5,809</td>
<td>143,968</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>RMSPE</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A. Math Test Score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-treatment</td>
<td>0.033</td>
<td>0.037</td>
<td>0.035</td>
<td>0.036</td>
</tr>
<tr>
<td>Post-treatment</td>
<td>0.105</td>
<td>0.112</td>
<td>0.110</td>
<td>0.107</td>
</tr>
<tr>
<td><strong>Panel B. Reading Test Score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-treatment</td>
<td>0.029</td>
<td>0.034</td>
<td>0.033</td>
<td>0.032</td>
</tr>
<tr>
<td>Post-treatment</td>
<td>0.090</td>
<td>0.095</td>
<td>0.096</td>
<td>0.093</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predictors</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Annul outcomes</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biannual outcomes</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both biannual outcomes</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other controls</td>
<td></td>
<td></td>
<td></td>
<td>Y</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DID strategy</th>
<th>DDD strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Across Districts</td>
<td>Within Guilford</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td><strong>Panel A. Math Score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TD_d \times Post_t$</td>
<td>0.108***</td>
<td>0.091***</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>$TG_g \times TD_d \times Post_t$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Panel B. Reading Score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TD_d \times Post_t$</td>
<td>0.014</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>$TG_g \times TD_d \times Post_t$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Comparison</strong></td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td><strong>District</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Comparison</strong></td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Grade</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>1,075,784</td>
<td>96,414</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Synthetic Control Methods</th>
<th>DID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) (2) (3) (4) (5)</td>
<td></td>
</tr>
<tr>
<td><strong>Panel A. Math Score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DID estimates</td>
<td>-0.005</td>
<td>0.015</td>
</tr>
<tr>
<td>P-value from two-tailed test</td>
<td>0.905</td>
<td>0.810</td>
</tr>
<tr>
<td><strong>Panel B. Reading Score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DID estimates</td>
<td>0.020</td>
<td>0.023</td>
</tr>
<tr>
<td>P-value from two-tailed test</td>
<td>0.823</td>
<td>0.790</td>
</tr>
</tbody>
</table>

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 5: Teacher Characteristics

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

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 6: Estimates of the Relationship between Classroom Composition and Teacher VA

<table>
<thead>
<tr>
<th></th>
<th>Guiford</th>
<th>Winston-Salem</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Math VA</td>
<td>Reading VA</td>
</tr>
<tr>
<td></td>
<td>DID DDD</td>
<td>DID DDD</td>
</tr>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>DID Post</td>
<td>-0.395</td>
<td>0.086</td>
</tr>
<tr>
<td></td>
<td>(0.319)</td>
<td>(0.163)</td>
</tr>
<tr>
<td>DID Post</td>
<td>-0.534</td>
<td>-0.524</td>
</tr>
<tr>
<td></td>
<td>(0.360)</td>
<td>(0.339)</td>
</tr>
</tbody>
</table>
| Panel A. Classroom Size
|                  | -0.030  | -0.010  | 0.029   | -0.008   |
|                  | (0.040) | (0.014) | (0.032) | (0.017)  |
| Panel B. Lagged average math score
|                  | 0.040   | 0.035   | -0.002  | -0.012   |
|                  | (0.044) | (0.036) | (0.040) | (0.065)  |
| Panel C. Lagged average reading score
|                  | 0.036   | -0.004  | 0.044   | -0.012   |
|                  | (0.032) | (0.013) | (0.030) | (0.015)  |
| Panel D. Parents’ education
|                  | 0.147** | -0.039  | 0.133*  | -0.025   |
|                  | (0.059) | (0.035) | (0.073) | (0.037)  |
| observations     | 847     | 6810    | 847     | 6810     |
| observations     | 811     | 7298    | 811     | 7298     |

Notes: The table reports 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 7: Estimates of the Relationship between Teacher Turnover and Teacher VA

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

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.