The Effects of Comprehensive Educator Evaluation and Pay Reform on Achievement

Eric Hanushek*, Jin Luo**, Andrew Morgan***, Minh Nguyen****, Ben Ost*****, Steven Rivkin****** and Ayman Shakeel*******

August 2023

A fundamental question for education policy is whether outcomes-based accountability including comprehensive educator evaluations and a closer relationship between effectiveness and compensation improves the quality of instruction and raises achievement. We use synthetic control methods to study the comprehensive teacher evaluation and compensation system introduced in the Dallas Independent School District (Dallas ISD) in 2015. Under this far-reaching reform, evaluations that are used to support improvement and determine salary depend on a combination of supervisor evaluations, student achievement, and student survey responses. The reform replaced salary scales based on experience and educational attainment with those based on evaluation scores, a radical departure from decades of rigid salary schedules. The synthetic control estimates reveal positive and significant effects of the reforms on math and reading achievement that increase over time. From 2015 through 2019, the average math achievement for the synthetic control district fluctuates narrowly between -0.27 s.d. and -0.3 s.d., while the Dallas ISD average increases steadily from -0.28 s.d. in 2015 to -0.08 s.d. in 2019, the final year of the sample. Though the increase for reading is roughly half as large, it is also highly significant.

This research was supported by grants from the Arnold Foundation and from the CALDER Research Network. The analysis uses confidential data supplied by the Dallas Independent School Districts (DISD). Individuals wishing to use these data must apply for access to DISD.

* Stanford University, University of Texas at Dallas, and NBER; **University of Texas at Dallas; ***University College London; ****Ball State University; *****University of Illinois at Chicago; ******University of Illinois at Chicago, University of Texas at Dallas, and NBER; *******University of Illinois at Chicago

1. Introduction

The 2001 No Child Left Behind legislation (NCLB) under the Bush Administration required states to use student outcomes including achievement in measuring school performance. Race to the Top (RTT) legislation under the Obama administration extended the accountability of NCLB by providing large incentives for states to reform teacher evaluation and compensation systems. Over this period new teacher evaluation and accountability policies were instituted by the vast majority of states (National Council on Teacher Quality (2017)). Yet, a fundamental question that remains is whether rigorous educator evaluation that affects compensation improves student outcomes.

Bleiberg et al. (2023) documents "the massive effort to introduce new high-stakes teacher evaluation systems," but they find that state reforms, even those classified as more rigorous, failed to significantly increase the quality of instruction and achievement. On the one hand, this evidence raises doubts about the potential for changes in personnel practices to elevate the quality of instruction. On the other hand, less than one percent of teachers were rated unsatisfactory following the reforms, and the compensation linked with higher performance was typically small. These factors, also highlighted by Bleiberg et al. (2023), suggest the incentives that were introduced may have been too weak to induce meaningful changes. The evidence from Washington DC, for example, suggests that stronger reforms may have a more positive effect in the quality of instruction. Evaluation of the Washington DC IMPACT program, using regression discontinuity design (RDD) estimates, show that an elevated threat of dismissal both increases the probability of exit of poor performers and raises performance of those who remain in the district (Dee and Wyckoff (2015)). In addition, Adnot, Dee, Katz, and Wyckoff (2017) find that higher turnover of teachers rated less effective raised grade-average math and reading achievement in the subsequent year. Yet a personnel system with expanded sanctions and greater salary uncertainty may adversely affect educator supply and potentially dampen the gains from collaboration.[1] Therefore, it is crucial to measure the total effect of a far-reaching reform on student outcomes.

We study the unprecedented reforms introduced by Dallas Independent School District (Dallas ISD) to understand the effects of a comprehensive overhaul of teacher and principal evaluation and compensation systems. The Teacher Excellence Initiative (TEI), implemented in 2015, introduced multiple-measure evaluation systems that align compensation with effectiveness. In a radical departure from decades of the rigid teacher salary schedules commonly found across the country, TEI replaced salary scales based on experience and educational attainment with those based on evaluation scores. The district rates teachers on their contributions to student achievement, supervisor observations and student feedback and uses the aggregate evaluation scores to place educators into ratings categories that are the primary determinant of salary. To protect the budget from tendencies toward evaluation inflation and further deter the arbitrary treatment of teachers, TEI fixes the distributions of teachers across rating categories. In addition, the inclusion of school average achievement as a determinant of teacher evaluations recognizes the importance of teamwork. Finally, district assessments have been developed to measure outcomes in grades and subjects that lack a state standardized test. The system has been modified in the years since adoption, but the foundational principles remain in place.

[1] Kraft, Brunner, Dougherty, and Schwegman (2020) find evidence that state reforms adversely affected the supply of teachers, particularly in hard-to-staff schools. This is consistent with the notion that reforms tended not to account adequately for factors outside educator control and that the size of pay increases to high-performing educators did not offset the additional risk and possible other dis-amenities associated with the reforms (Rothstein (2015)).

This paper first investigates the reform effects on achievement and then considers the contribution of changes in the composition of teachers to any improvement. Attribution of any achievement gains to the reforms requires more than simple examination of trends, since state policies or underlying demographic trends could, for example, improve outcomes for all urban or high poverty districts. Also, it is important to allow the treatment effects to evolve over multiple years. Short-term disruptions including extensive educator turnover across the experience distribution accompanied this reform, and these can mask longer-term benefits in the initial treatment years.

We use synthetic control methods to construct counterfactual achievement trends based on a donor pool of schools from the largest 20 Texas districts with at least 60 percent low-income students.[2] The weight assigned to each comparison school is chosen to minimize the distance between each school in Dallas ISD and average achievement in the synthetic control district in the pre-treatment period. The teaching evaluation reform (TEI) begins in 2015, but we only use the 2004-2012 pre-period for matching purposes because of a related principal evaluation reform (Principal Excellence Initiative) that begins in 2013.[3] Importantly, we see that Dallas and the synthetic control have similar achievement in the period between PEI and TEI, suggesting that PEI alone is insufficient to generate immediate improvement. The similar achievement between Dallas and the synthetic control in the two years before TEI is reassuring from an identification perspective because these years are not used in constructing the synthetic control.

[2] Including more districts yields similar results and we show robustness to including the largest 50 low-income districts instead of just the largest 20. Limiting the donor pool likely worsens pre-treatment fit, but we make this restriction in order to reduce the likelihood of interpolation bias. Abadie (2021) notes that interpolation bias is most likely when the donor pool includes many units that are very different from the treated unit.

[3] Dallas ISD implemented the Principal Excellence Initiative (PEI) in 2013 to strengthen incentives for administrators to engage in rigorous evaluation and support of teachers. Given that TEI depends strongly on teacher evaluation scores, PEI provides an important foundation for the implementation of TEI.

We find positive and significant effects on math and reading achievement that emerge in the year following TEI implementation and increase over time. Finding no immediate improvement in the first implementation year of 2015 is not surprising because teachers do not receive evaluation scores, ratings connected with salary level, or detailed information on performance until after the 2015 school year. Furthermore, it is plausible that any immediate effects in response to the strengthened performance incentives and midyear teacher meetings to discuss classroom observations could be offset by the disruptive nature of implementing a new program.

From 2015 through 2019 (the last year in our data), the average achievement for the synthetic control district fluctuates narrowly between -0.27 s.d. and -0.3 s.d., while the Dallas ISD average increases steadily from -0.28 s.d. in 2015 to -0.1 s.d. in 2018 and -0.08 s.d. in 2019, the final year of the sample.[4] Although the increase for reading is roughly half as large, it follows the same time pattern and is also highly significant beginning in 2018 based on permutation test p-values. The closer relationship between pay and effectiveness would be expected to increase educator effort and strengthen the relationship between educator persistence in the district and effectiveness. Consistent with this, we find that educators who exit the district have substantially lower evaluation scores on average than those who remain despite the absence of explicit removal triggers from the reforms.

The selective nature of teacher turnover suggests thateducator composition may play a role; however, this is only suggestive as we do not have direct measures of the effectiveness of new entrants prior to their arrival in Dallas ISD. We deduce the contribution of fixed differences

[4] It is possible that the principal reform implemented in 2013 contributes to these improvements in Dallas if its effects occur only with a multi-year delay. As such, one interpretation of our results is that they are the effect of the overall reform to educator evaluation – not just the teacher evaluation component.

in teacher effectiveness and those related to experience by comparing overall changes over time in average achievement with estimates of average changes over time within teachers, controlling for experience. The within-teacher changes capture the influences of all factors other than composition including stronger performance incentives and enhanced professional development. The differences between overall and within-teacher achievement changes then provide estimates of the contributions of teacher composition. This analysis finds that composition accounts for approximately one third of the growth in math achievement, or roughly 0.08 standard deviations.[5] The remaining channels, including the strengthened incentives for teachers and more effective teacher support, account for the majority of the change, but their contributions cannot be disentangled from one another.

2. Dallas ISD Evaluation and Compensation Reforms

After three years of discussion and development, the Teacher Excellence Initiative (TEI) was approved by the Dallas ISD Board of Trustees in May 2014. It replaced the evaluation and salary system (Dallas Professional Development and Appraisal System) that had been in place for 22 years and that used years of service and post-graduate schooling as the primary salary determinants. TEI dramatically alters the evaluation and compensation structures by requiring schools to collect far more information about teachers and to use the information for assessment, for professional development, and for salary.[6]

Dallas ISD established the foundation for the successful implementation of TEI by first introducing PEI and offering extensive principal training in teacher evaluation and support both

[5] The widespread use of reading support programs outside of the regular classroom led us to focus on math.
[6] There were some exceptions including educators in their first year in the district and some protections against salary decreases.

prior to and following its introduction. As a comprehensive evaluation and compensation reform, PEI shares many characteristics with TEI. Perhaps most important from the perspective of successful implementation of TEI, it provides strong incentives for principals to raise the quality of instruction in their schools by tying compensation and continued employment to achievement and teacher development. This discourages the arbitrary treatment of teachers, as does a component of PEI that penalizes principals for a divergence between their subjective teacher evaluations and the objective measure of teacher effectiveness based on achievement. [7]

TEI contains a student achievement component, a performance component based largely on supervisor observations of teaching, and a survey component based on feedback from students. The integrated multi-measure evaluation systems and accompanying effectiveness-based compensation structure are designed to support teacher growth, strengthen incentives to improve instruction, and attract strong educators to Dallas ISD. These are the primary channels through which the reforms are expected to raise the quality of instruction and consequently lead to higher test scores and improvements in educational attainment and labor-market outcomes.[8]

[7] PEI places substantial weight on effectiveness as an instructional leader. Almost 20 percent of the PEI performance component focuses directly on improving teacher effectiveness and congruence between teacher performance and student achievement. Thus, the principal is rated on their work in support of teachers and the alignment between the subjective teacher evaluation and teacher effectiveness at raising achievement. Morgan (2021) shows substantial evaluation inflation despite these efforts. Nevertheless, it also finds little change over time in the correlation between subjective and objective performance measures.

[8] Sources for the discussion of TEI include TEI Presentation (2015); TEI Rulebook (2015). "Rules and Procedures for Calculating TEI Evaluation Scores and Effectiveness Lev; TEI SLO Rubric (2014); TEI Student Achievement Templates (2015); TEI Teacher Performance Rubric (2014); Weerasinghe, D. (2008). How to compute school and classroom effectiveness indices: The value-added model implemented in Dallas Independent School District (retrieved at 4/20/2015). Sources for the discussion of PEI include Final 2014-2015 DISD Principal Handbook Sept; DISD 2014-2015 Salary Handbook; Principal Professional Development-Dec 2012; Principal Evaluation Rubric-General-Dec 2012; Principal Evaluation-Concept Paper-17 Jan 2013; Professional Development Hours – 18 Mar 2013; Miles M. (2013) Superintendent's Principal Evaluation System Report to the Board and Community. http://www.dallasisd.org/site/default.aspx?PageType=3&DomainID=7954&ModuleInstanceID=24529&ViewID=047E6BE3-6D87-4130-8424-D8E4E9ED6C2A&RenderLoc=0&FlexDataID=22163&PageID=20637

TEI activities can be categorized in three components - Defining Excellence, Supporting Excellence and Rewarding Excellence - each plays an important role in achieving the district goals. Defining Excellence describes the vision of effective teaching and teaching evaluation. Supporting Excellence refers to evidence-based professional development efforts based on the information generated by TEI. Finally, Rewarding Excellence refers to the connection between evaluation score and salary level.

a. Defining Excellence

Recognizing the possibility of strategic behavior in response to the rewards for raising end-of-year achievement, the multi-measure structure of TEI places the largest weight on supervisor evaluations based largely on classroom observations and also includes student survey responses for most teachers. Since only Dallas ISD teachers use this evaluation system, it is not possible to compare changes over time for Dallas ISD teachers with those in other Texas districts along all these dimensions. We therefore focus on the effect of TEI on state standardized test scores. Shakeel (2022) shows that teacher effectiveness based on the Dallas ISD metrics are significantly related to achievement in subsequent grades, suggesting that more effective educators based on TEI metrics produce lasting increases in human capital and not just increases in the high-stakes tests directly related to their compensation.

Performance, achievement and perception comprise the three components of the evaluation system. Table 1 lists the domains and indicators within each domain that comprise the teacher performance rubric; teacher receives scores for their performance on each. Every teacher is assigned a primary evaluator who is typically the principal or assistant principal. The evaluator monitors and collects evidence to assess performance mainly through spot, extended and informal observation. TEI specifies ten, 10- to 15-minute spot observations and one 45-minute extended observation per year. The observations focus on Domains 2 and 3, instructional practice and classroom structure. The supervisor is required to provide written feedback following all observations and to meet with the teacher following the extended observation. Artifacts and informal observations also contribute to the performance score, as these constitute the evidence of performance on the first and fourth domains.

Student perception is based on a survey conducted in the second week of April. Most students in grades 3-12 complete two surveys, one online and one on paper. Results from the survey are summarized by a single score for each teacher with at least a minimum number of responses; student surveys do not contribute to the evaluation score of some teachers including those in grade 2 or below. Points are assigned based on the target distribution at grade-level to assure equity because early grade-level students tend to provide more positive responses.

Both school average achievement and classroom achievement contribute to the achievement component except for teachers whose role is not associated with a student assessment. All school-level achievement measures are based on the state standardized test results. Teacher-level measures consists of Student Learning Objective (SLO) and Standardized Teacher-level Student Achievement Measures. SLO is a measure of student improvement during the year based on assessments that are not standardized tests; SLO contributes to the evaluation scores of all teachers, while classroom achievement contributes to the evaluation scores of teachers whose students take a standardized test. The district computes multiple measures of school and classroom achievement, and the highest metric for a teacher is used to determine their number of achievement points. Initially the alternatives included status (percentage of tests with scores that met a specified standard); value added; and achievement score relative to the scores of a designated peer group of schools based on prior achievement. Subsequently, the district eliminated the status alternative. The district uses target distributions to assign points for the school and teacher achievement components based on the standardized tests.

The evaluation score equals a weighted sum of points earned on the three components, where the weights depend on the role and grade level. Table 2 describes the four categories of teachers and differences among the weights for the three components. Category is determined primarily by the availability of student survey responses and results of a state or district assessment.

Teachers are divided into ratings categories based on scores and whether an application for recognition as a distinguished teacher is approved, a requirement for a rating of proficient II or higher. Table 3 lists the nine evaluation categories.

b. Supporting Excellence

Evidence including Taylor and Tyler (2012) highlight the value of teacher observations and feedback for professional growth, and the reforms emphasize the importance of teacher feedback based on observations and outcomes and the principal’s role as an instructional leader. Each of the three components of the evaluation system provides information used in teacher support and professional development. In addition to the written feedback and conferences following observations, achievement data are collected and analyzed to help improve instruction. An online resource bank of videos and modules was developed to support school leaders and instructional coaches in generating a clear and common vision of the TEI program in the system and foster self-learning among teachers**.**

c. Rewarding Excellence

Except for a teacher in her first or second year in Dallas ISD, salary is based on the average of evaluation points earned in the most recent two years; for teachers in their second year, it is based on evaluation points in the previous year only. The average score divides teachers into the nine effectiveness levels listed in Table 3, conditional on certain constraints: a teacher cannot move up or down more than one effectiveness level per year; completion of three years of service as a classroom teacher is a necessary condition to be considered for the Proficient I level; the Proficient II level and above requires teachers to go through the Distinguished Teacher Review (DTR) process, and to be at Exemplary II, teachers need to have at least one year qualifying as an Exemplary teacher; And Master level has additional requirements. To maintain budget stability and deter evaluation inflation, the category boundaries are determined by a target distribution (see Figure 1).

The system also includes safeguards to protect against downside risk: 1) It takes three consecutive years in a lower ratings category for teacher salary to go down by one level; 2) a salary will not fall below the teacher's salary in 2014-15 for those employed in that year; 3) a teacher starting after 2014-15 will not receive a salary lower than their entry-level salary; and 4) the compensation scale will be adjusted at least once per three years to keep salary levels competitive with other districts.

3. Administrative and Program Data

We use both Texas state administrative data housed at the University of Texas at Dallas Education Research Center (ERC) and administrative and program data provided by Dallas ISD. The Public Education Information Management System (PEIMS), TEA's statewide educational database, reports key demographic data including race, ethnicity, and gender for students and school personnel as well as program characteristics including subsidized or free lunch eligibility. PEIMS also contains detailed annual information on teacher and administrator role, experience, salary, education, class size, grade, population served, and subject taught. Beginning in 1993, the Texas Assessment of Academic Skills (TAAS) was administered each spring to eligible students enrolled in grades three through eight.[9] In 2003 the state substituted the TAKS in place of the TAAS, and in 2012 STAAR replaced the TAKS. We focus on the years 2004 to 2019, (year refers to spring of the academic year), which covers parts of the TAKS and STAAR test regimes. We transform all test results into standardized scores with a mean of zero and variance equal to one for each subject, grade, and year, meaning that our achievement measures describe students by their relative position in the overall state performance distributions. Because TAKS and STAAR differ, it is important to account for changes associated with the test-regime change. The synthetic control analysis minimizes achievement differences in a pre-period that spans the two test regimes.

The longitudinal data contain unique student and educator identifiers that enable us to follow students and educators across districts and schools as long as they remain in a Texas public school. These linkages permit the estimation of value added, and they also enable the description of educator movements in and out of schools and districts including Dallas ISD. We merge educator and student data by campus, grade, and year for the entire period and additionally by teacher, grade and year beginning in 2013.

The Dallas ISD administrative data include demographic and program information contained in the state data system, achievement data, and the disaggregated TEI and PEI components used to determine evaluation and effectiveness ratings and compensation. These

[9] Many special education and limited English proficient students are exempted from the tests. In each year roughly 15 percent of students do not take the tests, either because of an exemption or because of repeated absences on testing days.

data also contain identifiers that enable us to link the TEI and PEI information with student and staff longitudinal data.

4. Estimating the Impact of Personnel Reform

The lack of a natural comparison group led us to create a synthetic control district to serve as the counterfactual for Dallas ISD. This control district is created from elementary and middle schools in large, high-poverty districts. In the main specification the donor pool includes all schools from the largest 20 high-poverty districts (other than Dallas ISD), where high-poverty districts have at least 60 percent of the students qualify for a subsidized or free lunch. Schools in the synthetic control district are selected from the donor pool and weighted to minimize the pre-period average achievement gaps between the synthetic control district and each school in Dallas ISD. The selection of schools rather than districts as the focal unit dampens the impacts of reform efforts and shocks in other districts. We subsequently investigate the robustness of the estimates by examining the sensitivity of the estimates to expanding the donor pool to include the largest 50 high-poverty districts.

We estimate the effect of the Dallas reforms using the synthetic control method (SCM) developed by Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010). Conceptually, rather than comparing Dallas schools to a specific set of control schools, this approach forms a synthetic control group, which is a weighted average of potential control schools throughout the state. The weights are chosen to minimize the pre-treatment difference in outcomes between Dallas schools and synthetic control schools for the years prior to the introduction of PEI. We exclude the years 2013 and 2014 from the pre-treatment period because

PEI is implemented in 2013 and TEI is publicly being discussed during these years, raising the possibility of anticipation effects.

More formally, let Y_{it}^{T} be the potential outcome at school *i* when the policy is in effect and let Y_{it}^{C} be the potential outcome at school *i* in the pre-period, defined as the years prior to 2013. For each year in the post-period, we know the realized outcomes at Dallas schools and need to estimate Y_{it}^{C}. The synthetic control method estimates this counterfactual by taking a weighted average of control school outcomes in each year, where these weights are constrained to be constant over time. Specifically, defining an indicator D_i that is 1 for each Dallas school and zero otherwise, the counterfactual outcome for year t is

$$\sum_{D_i=0} w_i^* Y_{it}^{C}$$

where the weights are chosen to minimize a specific objective function. Because we match on all pre-treatment outcomes, the nested optimization component of the synthetic control approach greatly simplifies and all pre-period years receive equal weight (Kaul, Klößner, Pfeifer, and Schieler (2022). As such, in our case the synthetic control approach simply chooses weights, w_i^*, to minimize the sum-of-squared differences between each Dallas school and synthetic control schools in the pre-period (defined as $t < 0$) shown in the equation below.[10]

$$\sum_{t<0} \left(Y_{it}^{D=1} - w_i^* Y_{it}^{D=0}\right)^2$$

The synthetic control estimator is simply the average difference between Dallas schools and the synthetic control schools. Following the approach in Abadie, Diamond, and Hainmueller (2010), we conduct inference using a permutation test that compares the estimated effect for Dallas to a distribution of placebo estimated effects. Because there are many treated schools, the

[10] This is implemented using the user-written *synth_runner* routine for Stata, described in Galiani and Quistorff (2017).

distribution of placebo estimates is based on averages where the number of placebo units used in each average is the number of Dallas schools. This approach is described in more detail inCavallo, Galiani, Noy, and Pantano (2013). With many Dallas schools and many potential controls there are a large number of possible averages, and we sample from this distribution 1,000,000 times with replacement (See Galiani and Quistorff (2017) for details on this procedure).

5. The Impact of TEI

We begin with synthetic control estimates based on a donor pool of the largest 20 high-poverty districts before illustrating the sensitivity of the estimates to expansions of the donor pool.

a. Main Estimates

Figures 2 and 3 present plots of math and reading achievement in Dallas and the synthetic control both before and after the introduction of TEI in 2015; Table 4 presents the exact estimated effects and p-values. While perhaps unsurprisingly because of the construction of the control group, achievement patterns in Dallas mimic those of the synthetic control group before 2013. More importantly, there is a very close match between Dallas ISD and the synthetic control district achievement during the years 2013 and 2014 that are not used in the construction of the synthetic control weights, supporting the credibility of the synthetic control group methodology in this case.

In 2015, the first year of TEI, we see no evidence of improved outcomes in Dallas relative to the synthetic control. Although teachers did face strengthened performance incentives and have midyear meetings considering classroom observations, teachers did not receive evaluation scores, ratings connected with salary level, or detailed information on performance until after the 2015 school year.

In 2016, outcomes in Dallas and the synthetic control diverge, and the positive gap between Dallas ISD and the synthetic control district grows noticeably in the following years. By 2019 (the last year of our data), the gap exceeds 0.2 standard deviations (Figure 2). This gap reflects an achievement increase in Dallas ISD and not an achievement decline in the synthetic control district, supporting the conclusion that the reforms succeeded in elevating the quality of instruction and achievement. The gradual, rather than immediate, increase in achievement is not surprising given that many program benefits are expected to take time to be fully realized and the initial implementation of TEI may have been disruptive.

Column 1 of Table 4 shows the exact estimated effects in the post-policy period, along with p-values based on the permutation methods described above. It shows that the reform effect is significant at the 5 percent level in 2016 and at the 1 percent level in the subsequent years. None of the estimates prior to 2016 are significant at any conventional level.

b. **Figure 3 and Column 2 of Table 4 show that as with math, reading scores improve substantially from 2016 to 2019. A key difference is that reading scores fell below the synthetic control in Dallas in 2015 and so despite the improvements from 2016 to 2019, Dallas does not statistically exceed the synthetic control for reading until 2018. Furthermore, the reading gains are roughly half as large as those in math. Finding a larger effect in math than reading is a common result for many educational policies and is often attributed to the large role that families play in reading development.** **Sensitivity analysis**

Figures 4 and 5 present synthetic control plots for an expanded donor pool that includes schools from the largest 50 low-income districts. Columns 3 and 4 of Table 4 show exact estimates and p-values. Math achievement in Dallas ISD continues to track the synthetic control district closely until diverging in 2016. By 2019 the gap approaches 0.2 standard deviations, only slightly smaller than the differential observed in the main specification. The differences for 2017 to 2019 are significant at the 1 percent level.

As is the case for math, Figure 5 shows that the expansion of the donor pool introduces only minor changes to the reading treatment effect estimates. Although they are slightly smaller in 2018 and 2019 with the expanded donor pool, they remain significant at the 1 percent level. All in all, the insensitivity of the estimates to a substantial expansion of the donor pool provides additional support to the finding that the reforms significantly increased achievement in Dallas ISD.

6. Contributions of Educator Selection

TEI involves simultaneous changes in the strength of incentives, the information available for mentoring and professional development, the nature of retention decisions, and the composition of applicants and new hires, thus complicating efforts to disentangle the contributions of each. If the much closer alignment between effectiveness and salary alters the composition of entrants to and exits from Dallas ISD, educator composition could emerge as an important channel through which the TEI raises district quality. A first order issue, therefore, is understanding the impact of the reforms on teacher selection. We focus on selection out rather than selection into Dallas ISD due to the absence of comparable prior measures of effectiveness for most entrants into Dallas ISD. No other Texas district uses a similar evaluation system, and estimates of teacher value added are available only for the small fraction of entrants who previously taught in a tested grade in another district. Even for these teachers, estimates of value added in their previous schools would conflate teacher, school and district effects, just as would be the case if we were to measure effectiveness for entrants based on value added following their arrival to Dallas ISD. In addition, the fixed distributions of ratings designed to mitigate evaluation inflation and limit budget growth mean that the ratings distributions do not capture aggregate improvements or declines in educator effectiveness over time.

We use the student-teacher matches to identify the contribution of changes in teacher composition to the overall increase in math achievement. The presence of both stayers and leavers enables the separation of the contributions of teacher composition from those of all other channels through which TEI affected learning and achievement.

a. Teacher selection

Theoretically, TEI may promote more positive selection since higher quality teachers benefit from the pay differentiation. Though we lack data from before the reform to assess how selection patterns change, Figure 6 describes teacher evaluation scores by annual transition status. There is pronounced negative selection out of the district, as the average evaluation scores of teachers who remain in Dallas ISD exceed those who leave following the school year by more than 0.5 standard deviations. The lower two panels show that this strong negative selection holds for both the performance and achievement components. Though the negative selection out of Dallas in the 2015-2019 period is encouraging, we cannot directly assess the role that composition plays in the efficacy of TEI because we lack pre-policy data on teacher quality. Even in the post-period, we have no information on the efficacy of entrants prior to their entry into Dallas ISD. That said, Luo (2022) shows that though a low TEI rating does not trigger dismissal, it increases the probability of leaving Dallas ISD, suggesting that the stronger connection between effectiveness and salary may have contributed to the positive selection of stayers.

b. The contribution of teacher composition

The bundling of many components precludes the direct estimation of the contributions of strengthened incentives, enhanced professional development, better school leadership and other channels to the overall treatment effects. We can, however, separate the contribution of teacher composition from those of the other channels. We focus on math achievement because of the more extensive contributions of educators other than the classroom teacher of record to reading and language arts instruction. We compare estimates of the achievement changes over time from a regression of math achievement on a set of year dummies with the same regression that adds teacher fixed effects and a full set of experience dummies for years 0 to 10 and 11 plus. This latter regression considers the time path of achievement based on within-teacher achievement changes over time.

Equation 2 models achievement for student i in year t with teacher j as a function of a set of year dummy variables (*D*), a set of experience dummies *exp*, a teacher fixed effect (η_j) and a random error:

$$A_{ijt} = \alpha + \sum_{t=2016}^{2019} \delta_t D_t + \sum_{x=1}^{10+} \lambda_x exp_x + \eta_j + \varepsilon_{ijt} \tag{2}$$

In the absence of teacher fixed effects and experience controls, the teacher fixed effect (η_j) and the experience effects become part of the error, and the coefficients on the year dummies, $\widehat{\delta_t^{no\ fe}}$capture the influences of all factors including teacher composition that contribute to the difference between achievement in year t and achievement in 2015, the omitted baseline year. The inclusion of teacher fixed effects and the experience dummies shuts the teacher composition channel by considering just within-teacher variations, and the estimate $\widehat{\delta_t^{fe}}$ captures the influences of the other factors only. Therefore, the difference between $\widehat{\delta_t^{no\,fe}}$ and $\widehat{\delta_t^{fe}}$ provides estimates for the contribution of fixed and experience related differences in teacher composition between 2015 and year t to the achievement change over that period. Importantly, this analysis does not use the synthetic control or the pre-policy period since we are not trying to evaluate the overall effect of TEI with this exercise.[11] The purpose of the synthetic control group is to account for possible global shocks, but for this exercise, if there are global shocks that are not accounted for in equation (2), these will affect both the model with and without teacher FE and so the difference between the two models remains interpretable.

If all of the improvement in Dallas ISD schools comes from replacing worse teachers with better teachers and changes in the experience distribution, then we would expect to find small and insignificant year-dummy coefficients for the teacher fixed effect specifications. In the diametrically opposite case, if teacher composition accounts for none of the reform effects, we would expect the year-dummy coefficients to be insensitive to the inclusion of teacher fixed effects and experience. If, however, both teacher composition and other factors contribute to the overall treatment effects, the difference between the interaction term coefficients with and without the teacher fixed effects and experience controls will provide an estimate of the contribution of teacher composition.

Table 5 reports the set of year dummy coefficients (2015 is the baseline year) for regressions of achievement on year dummies with no teacher composition controls (Column 1) and both teacher fixed effects and experience controls (Column 2). Column 1 shows that the 2019 dummy variable coefficient of 0.24 is similar to the findings of the synthetic control analysis for math

[11] Student-teacher matches are not available in the pre-policy period. We drop unmatched students.

achievement shown in Figures 2 and 4.[12] Column 2 shows that the addition of both teacher fixed effects and experience controls has little effect on the 2016 dummy but it reduces the 2019 dummy variable coefficient from 0.24 to 0.16 standard deviations.[13] This 0.08 standard deviation decline suggests that by 2019 teacher composition accounts for roughly one third of the achievement gain following the implementation of the reforms. A 0.08 standard deviation compositional change is a meaningful shift given the evidence that a one standard deviation difference in the math teacher effectiveness distribution equals approximately 0.12 standard deviations in Texas (Rivkin, Hanushek, and Kain (2005)).

Importantly, teacher composition accounts for only one of the channels through which the reforms could have increased the quality of instruction. We are not able to identify the contributions of increases in effort in response to the strengthened incentives, peer teacher effects, or improvements in school leadership. Their contributions and those of other factors including improvements in academic support and school climate account for the majority of the math achievement gain but cannot be separately identified. This does suggest, however, that piecemeal reforms as opposed to the comprehensive reforms involving principals might have quite different results.

7. Conclusions

The comprehensive reforms introduced in Dallas ISD in 2015 eliminated much of the dependence of salary on experience and post-graduate degrees, radically altering the system

[12] Note that a small fraction of students are not matched with teachers, so the sample and changes in achievement differ slightly from the main analysis.

[13] Appendix Table a1 reports the single year of experience coefficients, where 0 years is the omitted category. The estimates suggest that the vast majority of the return to experience for Dallas Teachers occurs in the first three years of teaching, similar to other research on Texas (Hanushek et al, 2005).

commonly used in US districts. System details reflect careful consideration of the potential for unintended consequences including evaluation inflation, the arbitrary treatment of teachers, and strategic responses including teaching to the test. Aligning the relationship between educator effectiveness and pay dramatically strengthened performance incentives, while the development of a multiple-measure evaluation system based on student outcomes, supervisor observations and student or family feedback recognized the pitfalls of a singular reliance on achievement or subjective evaluations by supervisors. Importantly, value-added and achievement relative to comparable students rather than pass rates or absolute achievement levels in absolute terms were introduced to insulate evaluations from factors including family circumstances that were outside of educator control.

The synthetic control analysis shows that the reforms succeeded in markedly raising math and to a lesser extent reading achievement. Effect sizes exceeding 0.2 standard deviations for math and 0.1 standard deviations for reading are large, particularly in comparison to such much more costly interventions as a large reduction in class size. The teacher fixed effects analysis further shows that changes in teacher composition accounted for roughly one third of the math achievement increase. Other channels including instructional improvements driven by the strengthened incentives and enhanced support for teachers, extensive principal training in instructional leadership, and strong incentives for principals to elevate the quality of instruction are likely candidates for the remaining improvement.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program." *Journal of the American Statistical Association* 105, no. 490: 493-505.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The economic costs of conflict: A case study of the Basque Country." *American Economic Review* 93, no. 1: 113-132.

Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff. 2017. "Teacher Turnover, Teacher Quality, and Student Achievement in DCPS." *Educational Evaluation and Policy Analysis* 39, no. 1: 54-76.

Bleiberg, Joshua, Eric Brunner, Erica Harbatkin, Matthew A. Kraft, and Matthew G. Springer. 2023. "Taking Teacher Evaluation to Scale: The Effect of State Reforms on Achievement and Attainment." NBER Working Paper Series No. 30995. Cambridge, MA: National Bureau of Economic Research (March).

Cavallo, Eduardo, Sebastian Galiani, Ilan Noy, and Juan Pantano. 2013. "Catastrophic Natural Disasters and Economic Growth." *The Review of Economics and Statistics* 95, no. 5 (December): 1549-1561.

Dee, Thomas S., and James Wyckoff. 2015. "Incentives, selection, and teacher performance: Evidence from IMPACT." *Journal of Policy Analysis and Management* 34, no. 2 (Spring): 267-297.

Galiani, Sebastian, and Brian Quistorff. 2017. "The Synth_Runner Package: Utilities to Automate Synthetic Control Estimation Using Synth." *The Stata Journal* 17, no. 4: 834-849.

Kaul, Ashok, Stefan Klößner, Gregor Pfeifer, and Manuel Schieler. 2022. "Standard Synthetic Control Methods: The Case of Using All Preintervention Outcomes Together With Covariates." *Journal of Business & Economic Statistics* 40, no. 3 (July): 1362-1376.

Kraft, Matthew A., Eric J. Brunner, Shaun M. Dougherty, and David J. Schwegman. 2020. "Teacher accountability reforms and the supply and quality of new teachers." *Journal of Public Economics* 188(August): 104212.

Luo, Jin. 2023. Teachers' Responsiveness to Performance-Based Pay: Evidence from a Large Urban School District in Texas. *Mimeo*

National Council on Teacher Quality. 2017. *State teacher policy yearbook, 2017*. Washington: National Council on Teacher Quality.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, schools, and academic achievement." *Econometrica* 73, no. 2 (March): 417-458.

Rothstein, Jesse. 2015. "Teacher Quality Policy When Supply Matters." *American Economic Review* 105, no. 1: 100-130.

Shakeel, Ayman. 2023. "High-Stakes Objective and Subjective Teacher Evaluation Measures and Student Skill Development". *Mimeo*

Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance." *American Economic Review* 102, no. 7 (Decenber): 3628-51.

Table 1: Teacher performance rubric.

Domain	Indicator of teacher practice	Evidence used	Max. points
Domain 1: Planning and Preparation	1.1. Demonstrate knowledge of content, concepts, and skills	Artifacts and informal observations	15
	1.2. Demonstrates knowledge of students		
	1.3. Plans or selects aligned formative and summative assessments		
	1.4. Integrates monitoring of student data into instruction		
	1.5. Develops standards-based unit and lesson plans		
Domain 2: Instructional Practice	2.1. Establishes clear, aligned standards-based lesson objective(s) (3x)	Spot, extended and informal observations	48
	2.2. Measures student mastery through a demonstration of learning (DOL) (spot) (3x)		
	2.3. Clearly presents instructional content (spot) (3x)		
	2.4. Checks for academic understanding (2x)		
	2.5. Engages students at all learning levels in rigorous work (3x)		
	2.6. Activates higher-order thinking skills (2x)		
Domain 3: Classroom culture	3.1. Maximizes instructional time (spot) (3x)	Spot, extended and informal observations	21
	3.2. Maintains high student motivation (2x)		
	3.3. Maintains a welcoming environment that promotes learning and positive interactions (2x)		
Domain 4: Professionalism and Collaboration	4.1. Models good attendance for students	Artifacts and informal observations	15
	4.2. Follows policies and procedures, and maintains accurate student records		
	4.3. Engages in professional development		

Source: compiled from TEI Teacher Performance Rubric and the TEI Presentation

Table 2: Teacher categories and evaluation templates

Teacher Category	**Teacher Performance**	**Student Achievement**	**Student Perception**
Category A: Most grade 3-12 teachers whose students take an ACP, STARR, or AP exam, including most K-5 special teachers	50	35	15
Category B: Most K-2 teachers whose students take an ACP or ITBS/Logramos	65	35	0
Category C: Most grade 3-12 teachers whose students do not take an ACP, STARR, or AP assessment but who are able to complete a student survey (e.g. CTE teachers)	65	20	15
Category D: Any teachers whose students do not take an ACP, STARR, or AP assessment nor are eligible to complete a student survey (e.g. pre-K teachers. Teachers not-of-record such as SPED inclusion teachers, TAG teachers)	80	20	0

Source: Compiled from TEI Teacher Guidebook p.6 and TEI Rulebook p.9

Table 3: Compensation tied with teacher effectiveness levels in the initial year of TEI

Unsatisfied	Progressing		Proficient			Exemplary		Master
	I	II	I	II	III	I	II	
$45K	$49K	$51K	$54K	$59K	$65K	$74K	$82K	$90K

Source: Teacher Guidebook p36.

Table 4: Synthetic control estimates and p-values of the effects on math and reading scores

	Donor pool includes largest 20 districts		Donor pool includes largest 50 districts	
Year	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)
2013	0.017	0.029	0.001	0.002
	[0.258]	[0.058]	[0.974]	[0.998]
2014	0.010	-0.017	0.015	-0.035
	[0.626]	[0.342]	[0.458]	[0.408]
2015	0.012	-0.055	-0.002	-0.065
	[0.751]	[0.003]	[0.922]	[0.001]
2016	0.074	0.028	0.040	-0.003
	[0.030]	[0.518]	[0.126]	[0.896]
2017	0.112	0.029	0.077	0.003
	[0.000]	[0.522]	[0.001]	[0.861]
2018	0.177	0.064	0.164	0.052
	[0.000]	[0.044]	[0.000]	[0.017]
2019	0.212	0.093	0.186	0.078
	[0.000]	[0.035]	[0.000]	[0.004]

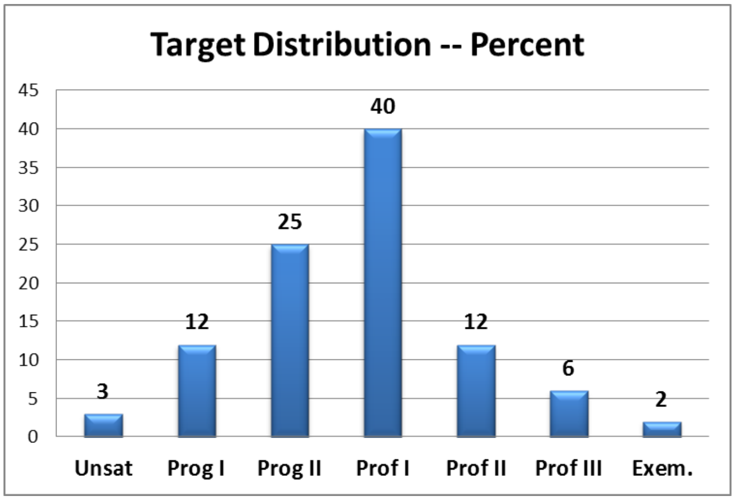
Notes: This table provides exact estimates and p-values (in brackets) corresponding figures 2-5. The estimated effects in this table are the gap between Dallas and the synthetic control and the p-values are based on the permutation test described in the text.

Table 5. Year dummy coefficients from regressions of math achievement on year indicators, by inclusion of teacher fixed effects and experience indicator variables (2015 is the omitted year; standard errors clustered by teacher in parenthesis)

teacher fixed effects	no	yes
dummy variables for single years of experience from 1 to 10 and an indicator for 11 or more years of experience	no	yes
2016	0.047	0.048
	(0.007)	(0.008)
2017	0.138	0.110
	(0.007)	(0.010)
2018	0.182	0.140
	(0.007)	(0.011)
2019	0.237	0.162
	(0.007)	(0.014)

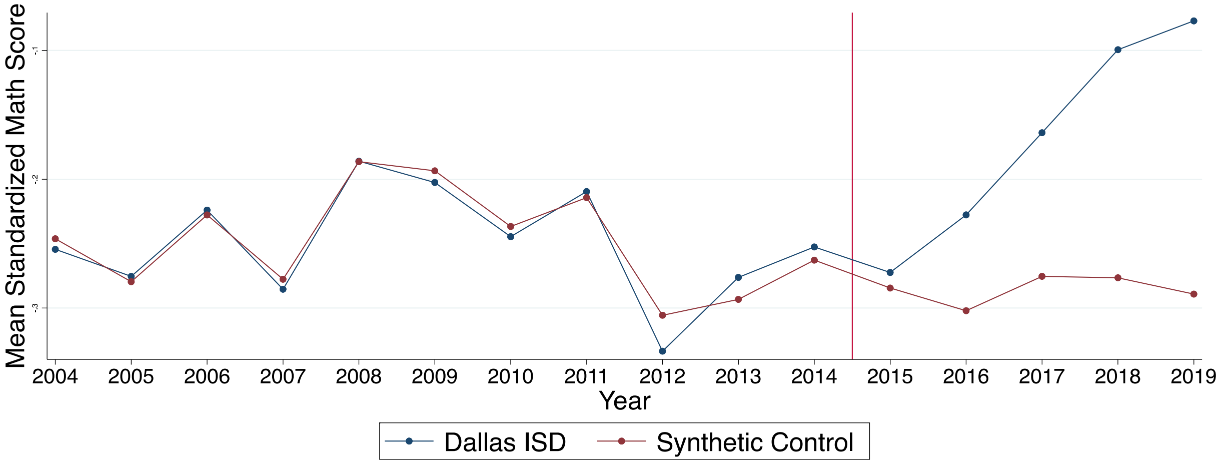
Notes: The coefficients in the left column come from a regression of math achievement on a full set of year dummies (2015 is the excluded year), and the coefficients in the right column come from a teacher fixed effect regression on a full set of year dummies and dummy variables for single years of experience from 1 to 9 and an indicator for 10 years of experience or more (0 years of experience is the omitted category).

Figure 1: Target distribution of teacher effectiveness scales



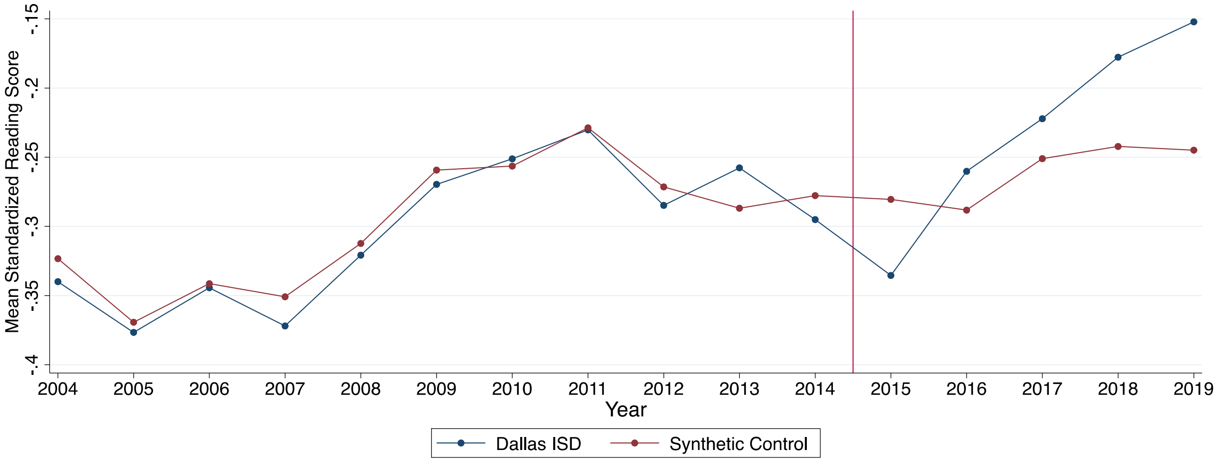
Source: TEI Rulebook v4.1 (DISD (2017)).

Figure 2. Synthetic control analysis of math achievement using the 20 largest high poverty districts



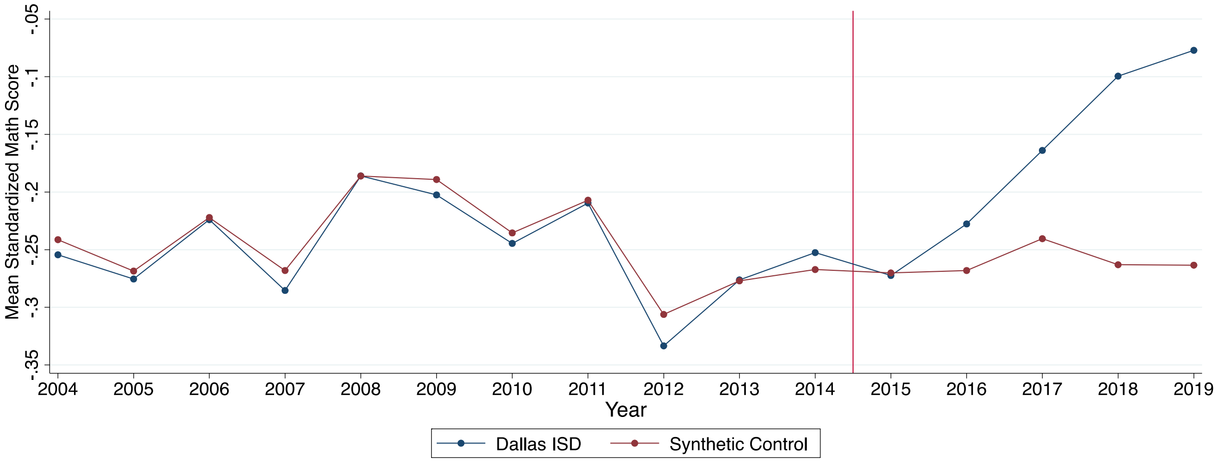
Notes: The figure plots average math achievement in Dallas ISD and the synthetic control over time. The synthetic control is constructed using schools from the 20 largest high-poverty districts as the donor pool.

Figure 3. Synthetic control analysis of reading achievement using the 20 largest high poverty districts



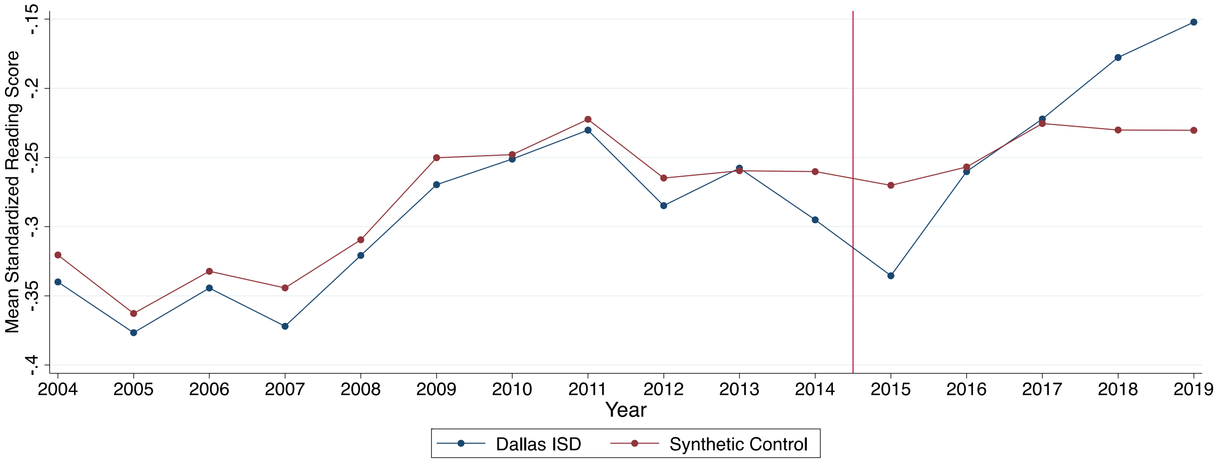
Notes: The figure plots average reading achievement in Dallas ISD and the synthetic control over time. The synthetic control is constructed using schools from the 20 largest high-poverty districts as the donor pool.

Figure 4. Synthetic control analysis of math achievement using the 50 largest high poverty districts



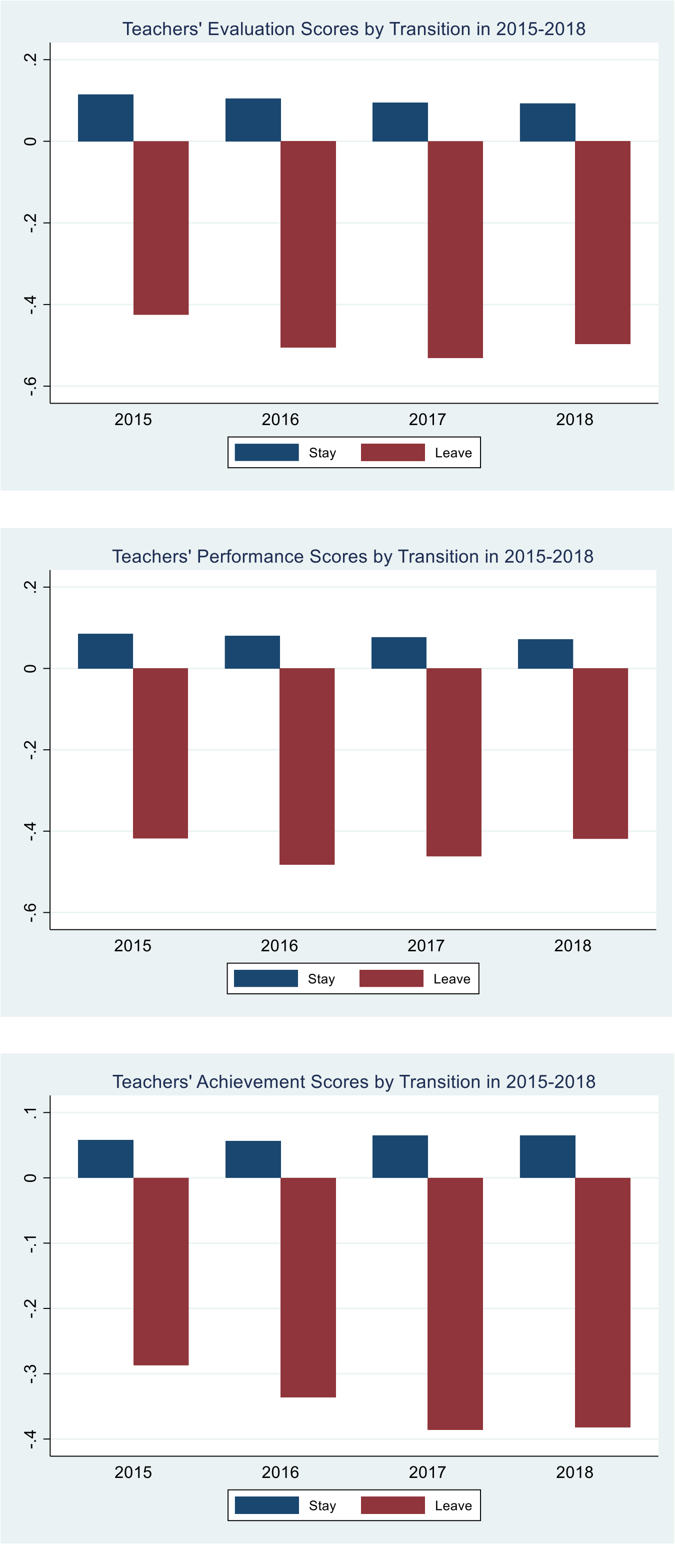
Notes: The figure plots average math achievement in Dallas ISD and the synthetic control over time. The synthetic control is constructed using schools from the 50 largest high-poverty districts as the donor pool.

Figure 5. Synthetic control analysis of reading achievement using the 50 largest high poverty districts



Notes: The figure plots average reading achievement in Dallas ISD and the synthetic control over time. The synthetic control is constructed using schools from the 50 largest high-poverty districts as the donor pool.

Figure 6. Mean teacher overall evaluation and component scores, by annual transition status



Teachers' Evaluation Scores by Transition in 2015-2018

Teachers' Performance Scores by Transition in 2015-2018

Teachers' Achievement Scores by Transition in 2015-2018