Journal of Urban Economics 153 (2026) 103857

Contents lists available at ScienceDirect

Journal of Urban Economics

journal homepage: www.elsevier.com/locate/jue

The effects of comprehensive pay reform on achievement in urban schools

Eric Hanushek[a], Jin Luo[b], Andrew Morgan[c], Minh Nguyen[d], Ben Ost[e,*], Steven Rivkin[e], Ayman Shakeel[f]

[a] Stanford University, 450 Serra Mall, Stanford, CA 94305, USA
[b] Sichuan University, No. 24 South Section 1, Yihuan Road, Chengdu, Sichuan 610065, China
[c] South Dakota State University, 1015 Campanile Ave, Brookings, SD 57007, USA
[d] Ball State University, 2000 W University Ave, Muncie, IN 47306, USA
[e] University of Illinois at Chicago, 1200 W Harrison St, Chicago, IL 60607, USA
[f] Abt Global, United States

ARTICLE INFO

Keywords:
Economics of education
Urban education
Teacher accountability
Pay for performance

Academic performance of disadvantaged students has been stubbornly hard to improve, particularly in urban schools. The Dallas Independent School District (Dallas ISD) addressed this problem in 2015 with a radical change in the structure of teacher compensation, basing it entirely on classroom effectiveness as determined by an evaluation system incorporating achievement growth, classroom observations and student surveys. We evaluate this institutional change with synthetic control methods that compare math and reading achievement in Dallas ISD schools with achievement for schools in other high-poverty Texas districts. We find large and significant positive effects on math achievement that increase steadily over time. For reading, there is no clear evidence of improvement in Dallas ISD relative to the synthetic control. A mechanism analysis shows that changes in teacher composition account for a majority of the math achievement increase.

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 District (DISD). Individuals wishing to use these data must apply for access to DISD. The conclusions of this research do not necessarily reflect the opinions or official position of the Texas Education Agency or Dallas ISD. We benefitted from helpful comments from anonymous reviewers.

1. Introduction

Since the Coleman Report in 1966, there has been an undercurrent of thought that schools could not do much to alter the achievement patterns established by family circumstances (Coleman et al. (1966)). Supporting this possibility, a half century of effort to close achievement gaps between poor and nonpoor students has made little progress (Hanushek et al. (2022)). This pessimistic overhang is particularly strong in urban school districts, where high rates of crime and joblessness, poorly funded public services and challenging political conditions likely contribute to substantial inertia in the quality of public education. We evaluate a unique restructuring of the personnel systems in the Dallas Independent School District (DISD) that suggests that it is possible to improve student outcomes at scale.

A growing body of evidence finds positive effects of teacher pay for performance incentives (e.g. Lavy (2020); Pham, Nguyen, and Springer (2021)), but there is strong inertia in school personnel systems. Meaningful challenges to the status quo of strict educator salary schedules that vary little with effectiveness, of inflated teacher evaluations, and of strong job security in urban districts have generally been rare. One exception in the late 2000s was Michelle Rhee's introduction of the IMPACT personnel reform in the Washington D.C. school district. Although the standard salary structure based on the experience and graduate education of the teacher remained largely intact, the reform used rigorous teacher evaluations by outside experts and student achievement as the basis for involuntary removal of very low-performing teachers and for substantial pay increases for highly effective teachers in high-poverty schools. The results were encouraging. Using regression discontinuity design methods, Dee and Wyckoff (2015) finds that an elevated threat of dismissal increased the probability of exit of poor performers and raised the effectiveness of low-performing teachers who remained in the Washington, D.C. district. Adnot, Dee, Katz, and Wyckoff (2017) additionally find that higher turnover of teachers rated less effective appeared to increase

[*] Corresponding author.
E-mail address: bost@uic.edu (B. Ost).

https://doi.org/10.1016/j.jue.2026.103857
Received 19 November 2024; Received in revised form 10 March 2026;
Available online 24 March 2026

grade-average math and reading achievement in the subsequent year. Evaluating the aggregate effect of the DC personnel reform is complicated, however, because several other reforms were implemented simultaneously. A Mathematica report on D.C. outcomes finds that the combined effect of the personnel reform, expansion of the charter school sector and intensified school choice improved NAEP scores relative to a matched control group (Dotter, Chaplin, and Bartlett (2021)).

Building on No Child Left Behind and the promising IMPACT findings, the Obama administration introduced the Race to the Top (RTT) federal legislation that incentivized states to introduce high-stakes achievement-based educator accountability. Using federal legislation to alter local behavior is a challenging enterprise, and Bleiberg et al. (2025) show that the reforms typically lacked the strong dismissal threat of IMPACT, often did not incentivize school administrators, and included quite modest levels of performance pay. Not surprisingly, they find that RTT induced reforms had little if any effect on achievement. A related paper suggests that adverse effects on teacher labor supply may have also dampened the benefits of the reforms (Kraft, Brunner, Dougherty, and Schwegman (2020)).

This general policy failure provides the context in which Dallas ISD introduced by far the most dramatic personnel reforms in American public education. The Teacher Excellence Initiative (TEI) was designed to align total compensation much more closely with effectiveness, to strengthen incentives for current teachers, and to alter teacher composition towards more effective educators. In a radical departure from the rigid single-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 evaluates 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 evaluation inflation, TEI fixes the distributions of teachers across rating categories. The additional inclusion of school average achievement as a determinant of teacher evaluations recognizes the importance of teamwork.

Dallas ISD also adopted a parallel personnel reform for principals that recognized that TEI would require principals to evaluate teachers rigorously, support teacher improvement, and make difficult personnel decisions. The Principal Excellence Initiative (PEI) constituted a more modest departure from the status quo as principal salaries were already determined relatively flexibly and principals already had limited job security. Nonetheless, it is important to recognize that the TEI reforms were implemented in a context where principals were incentivized to implement the reform with fidelity. A large proportion of a principal's evaluation is based on student achievement and the effectiveness of TEI implementation, including the alignment between the principal's subjective evaluations and teacher value added to achievement.

We evaluate the effect of TEI on elementary school math and reading achievement for students in grades three to five using the synthetic control method to construct counterfactual achievement trends.[1] From a donor pool of schools from all Texas districts with at least the state average of 60 percent low-income students, we construct a synthetic control school for each school in Dallas ISD based on a comparison of achievement during the pre-treatment period. PEI became effective in 2013 and TEI in 2015. Recognizing the possibilities that PEI may have had direct effects on achievement and that anticipation of TEI may have affected teacher behavior, we omit 2013 and 2014 and use achievement in the years 2004 to 2012 to construct the synthetic control district.

Positive and significant effects on math achievement emerge in the year following TEI implementation and increase over time until they exceed 0.1 standard deviations in 2019. For reading, there is no clear evidence of improvement in Dallas relative to the synthetic control. The fact that math improvement occurs with a one-year delay is not surprising because teachers do not receive evaluation scores, do not have ratings connected with salary level, and do not have detailed information on performance until after the first year of implementation.

To examine the mechanisms for the impacts of TEI, we first describe the attrition of teachers based on their effectiveness. The close 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 that educator composition may play a role in overall district outcomes. 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 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 teacher- and nonteacher- related factors other than composition including stronger performance incentives and enhanced professional development. We find that models that control for teacher fixed effects and experience sharply reduce the estimated improvement in Dallas relative to the other high-poverty districts, suggesting that teacher composition plays an important role. A Gelbach (2016) decomposition reveals that it is fixed differences in teacher effectiveness rather than experience that drives these changes.

Our findings highlight the possibility that achievement in large cities with many children from poor and disadvantaged families can be improved by structural changes in personnel evaluation and compensation – institutional changes long proposed in the economics literature (Kershaw and McKean (1962)).

2. Literature on performance pay

Researchers have studied the effect of performance pay on productivity in both education and the broader labor market. Our analysis contributes to these literatures, as it provides one of the few evaluations of a large-scale, permanent, performance-pay policy.

In education, there is an extensive literature on performance pay for teachers, but the results are quite heterogeneous across programs and studies. For example, Fryer (2013); Goodman and Turner (2013), Fryer, Levitt, List, and Sadoff (2022), Glazerman and Seifullah (2012), Springer et al. (2010) and Sojourner, Mykerezi, and West (2014) find small or null effects, whereas Dee and Wyckoff (2015) and Lavy (2002, 2009) find evidence of significantly improved outcomes from performance pay. In a meta-analysis, Pham, Nguyen, and Springer (2021) estimate an average effect of 0.043 SD.[2] Pham, Nguyen, and Springer (2021) argue that the literature is too thin to provide strong evidence on which components of performance pay drive efficacy, but some patterns emerge such as larger effects in contexts that include professional development, larger effects in elementary schools and larger effects from individual, rather than group incentives. Relative to common studies in this literature, we consider a much more extensive reform because it is a district-wide, permanent change to the entire salary schedule, rather than a short-duration program that provides bonuses to a subset of

[1] The reform also affects middle schools, but we focus on evaluating the effect on elementary school students because there was a simultaneous policy that accelerated the middle school math curriculum and shifted many students to take middle school math exams one grade level above their current grade. Note that 3rd grade is the first grade tested in the state accountability system.

[2] There is also a literature on pay for performance in developing contexts, and these studies tend to find much larger effects, possibly because of the low baseline levels of effort (e.g. Muralidharan and Sundararaman (2011); Duflo, Hanna, and Ryan (2012).

teachers. Because of the scale and complexity of the DISD reforms, it is uncertain ex ante whether to expect larger effects than those in prior contexts.

Outside of education, there is a robust literature on performance pay with the broad takeaway that performance pay is very effective in jobs where effort maps clearly to well-measured unidimensional output. For example, Lazear (2000) finds a 44 % improvement in productivity of auto glass workers when switching to a piece-rate salary schedule. In standard deviation units, this is a dramatic 0.84 SD improvement. Shearer (2004) finds similarly large effects in the context of tree planting, where performance pay increases productivity by more than the control group standard deviation. When tasks are multidimensional or output is difficult to measure, the effect of performance pay is less clear. For example, a literature in psychology relying mainly on lab-based experiments finds that simple tasks are substantially improved from performance incentives, but performance on complex tasks is sometimes worsened by performance incentives (Weibel, Rost, and Osterloh (2010)).

In a recent meta-analysis on performance pay for civil servants, George and van der Wal (2023) find a positive, albeit small improvement in productivity from performance pay. The most closely connected literature to education is studies of performance pay in health care, and similar to education, the results are mixed. Most studies in healthcare are focused on improving narrow outcomes such as vaccine take-up (Kouides et al. (1998)) or cancer screening rates (Hillman et al. (1998)). A broad meta-analysis on performance pay by Hasnain, Manning, and Pierskalla (2014) also suggests a mix of results: studies including Hillman et al. (1998); Hillman et al. (1999) and Grady, Lemkau, Lee, and Caddell (1997) find no effect, whereas studies including Fairbrother, Hanson, Friedman, and Butts (1999); Fairbrother et al. (2001), Roski et al. (2003), and Kouides et al. (1998) find large increases.

3. Dallas ISD evaluation and compensation reforms

The personnel reforms involve a complicated and integrated system of evaluations and rewards for educational effectiveness. 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 determination.[3]

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 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 a principal's compensation and continued employment to student 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.[4]

The integrated multi-measure evaluation system and accompanying effectiveness-based compensation structure are designed to support teacher growth, strengthen incentives that improve instruction, and attract strong educators to Dallas ISD.[5] 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. TEI combines the scores on the three components into a single evaluation score, recognizing that information details vary by grade and subject taught. The evaluation score constitutes the primary determinant of salary, and supervisors also use the information from all three components to support teacher improvement and growth.[6]

a. Teacher Evaluation

The multi-measure structure of TEI places the largest weight on supervisor evaluations derived mainly from classroom observations but also includes assessments of student performance and student survey responses for most teachers. We will focus on the effect of TEI on state standardized test scores, but, importantly, Shakeel (2023) shows that teacher effectiveness based on the Dallas ISD metrics is 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. Appendix Table A1 lists the domains and indicators within each domain that comprise the teacher performance rubric; teachers receive 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 through spot, extended and informal observation. TEI specifies ten, 10–15-minute spot observations and one 45-minute extended observation per year. The observations focus on

[3] There were some exceptions including educators in their first year in the district and some protections against salary decreases.

[4] 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 (2022) shows substantial evaluation inflation despite these efforts. Nevertheless, he also finds little change over time in the correlation between subjective and objective performance measures.

[5] 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

[6] Dallas ISD categorizes the three interrelated components of TEI as 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.

instructional practice and classroom structure (Domains 2 and 3). 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 surveys are summarized by a single score for each teacher who has 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 each 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 consist 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 scores relative to the scores of a designated peer group of schools based on prior achievement, although 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. Appendix Table A2 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.

b. Supporting Teacher Development

Evidence including that by Taylor and Tyler (2012) and Steinberg and Sartain (2015) highlight the value of teacher observations and feedback for professional growth, and the DISD reforms emphasize teacher feedback based on observations and outcomes along with the principal's role as an instructional leader. Each of the three components of the teacher 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 and fostering self-learning among teachers.

c. Performance Pay

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 1, conditional on the constraint that a teacher cannot move up or down more than one effectiveness level per year. This excludes early career teachers from the higher categories, as 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. Finally, the Master level requires a teacher to be Exemplary II for at least two years and is only possible at specific, hard-to-staff schools. To maintain budget stability and deter evaluation inflation, the category boundaries of evaluation scores are determined by a target distribution (see Appendix Fig. A1).

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.

In addition to being unique among performance pay for teachers, TEI is fairly unique with regards to performance pay in the broader labor market. Though performance incentives are common outside of education, a Bureau of Labor Statistics report (Bureau of Labor Statistics (2022)) indicates that incentive pay in the private sector generally takes the form of bonuses rather than affecting base salary and is often based on group, rather than individual performance.

4. 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 (Hanushek et al. 2026). 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.[7] 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 state 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 and STAAR is introduced during the pre-treatment period, it is important that the synthetic control analysis minimizes achievement differences in a pre-period that spans both 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 enable the description of educator movements in and out of schools and districts including Dallas ISD. Student-teacher matches become available in 2012, and starting in 2013, we are able to calculate teacher classroom average achievement and value added on the STAAR tests.

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

[7] 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.

Table 1
Compensation tied with teacher effectiveness levels in the initial year of TEI.

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

Source: Teacher Guidebook p36.

evaluation and effectiveness ratings and compensation. These data also contain identifiers that enable us to link the TEI and PEI information with student and staff longitudinal data.

5. Empirical model

We estimate the effect of the Dallas reforms on elementary school math and reading scores using the synthetic control method (SCM) developed by Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010). To allow for a more flexible fit and accommodate the many differences between Dallas ISD and other Texas districts, we set schools rather than districts as the focal unit. Conceptually, this approach constructs a synthetic control for each Dallas ISD school based on a weighted average of potential control schools throughout the state where the weights are chosen to minimize the pre-treatment difference in outcomes between each Dallas ISD school and its synthetic control school for the years before 2013. The synthetic controls for all Dallas ISD schools are then aggregated across all Dallas ISD elementary schools to construct the synthetic control district for Dallas ISD.[8] Even though TEI does not begin until 2015, we exclude the years 2013 and 2014 from the construction of the control schools to avoid possible anticipation effects because PEI is implemented in 2013 and TEI is publicly being discussed during these years.

The baseline donor pool for constructing the synthetic control schools is all elementary schools in the set of Texas districts with at least 60 percent students from poverty households. We subsequently investigate the robustness of the estimates by progressively restricting the donor pool to schools in increasingly larger districts in order to be more similar overall to the large urban district of Dallas ISD.

Let $Y_{it}^{D=1}$ be the potential outcome at school i when the policy is in effect and let $Y_{it}^{D=0}$ be the potential outcome at school i when no policy is in effect. The indicator D is 1 for each Dallas school and zero otherwise. For each year in the post-period, we know the realized outcomes at Dallas schools and need to estimate $Y_{it}^{D=0}$. The synthetic control method estimates this counterfactual by taking a weighted average of potential control school outcomes in each year, where these weights are constrained to be constant over time. Specifically, the counterfactual outcome for year t is

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

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-treatment period (defined as $t < 0$) shown in the equation below.[9]

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

The synthetic control estimator of the impact of TEI is simply the average difference between Dallas schools and the synthetic control schools in each year following TEI's introduction.

Following the approach in Abadie, Diamond, and Hainmueller (2010), we conduct inference using a permutation test that compares the estimated effect for Dallas ISD to a distribution of placebo estimated effects. As discussed in Cavallo, Galiani, Noy, and Pantano (2013), because the main estimate is based on the average of many treated units, it is important that the placebo estimates are also based on averages across many treated units. In our case, the number of placebo units assigned to treatment in each simulation should be equal to the number of Dallas ISD schools. It is not possible to simulate all possible permutations of placebo treatment since there are far too many potential ways to draw groups of schools from the donor pool. Instead, we randomly sample from the distribution of possible permutations 1000,000 times with replacement (See Galiani and Quistorff (2017) for details on this procedure).

6. The impact of TEI

The main synthetic control analysis uses a donor pool of all elementary schools in high-poverty districts, and we then illustrate the sensitivity of the estimates to the restriction of the donor pool to schools from the largest 50, 20 and 10 districts. Figs. 1 and 2 present plots of math and reading achievement in Dallas and the synthetic control both before and after the introduction of TEI in 2015, and Table 2 presents the exact estimated effects and p-values. Appendix Table A3 shows that results are very similar if we weight estimates by school enrollment.

Fig. 1 provides compelling evidence of improved math outcomes in Dallas following the adoption of TEI. The magnitude of the TEI effect rises to approximately 0.1 standard deviations by 2019. Importantly, Fig. 1 shows that Dallas and the synthetic control district not only have similar math scores from 2004–2012 but also continue to have similar outcomes in 2013 and 2014, years that are not used in the matching algorithm. This provides evidence of common pretreatment trends in the Dallas and control schools. The treatment effects shown in column 1 of Table 2 increase in magnitude from 2016 to 2019 and are statistically significant at the 5 % level in 2016 and at the 1 % level for 2017 to 2019.

For reading, Fig. 2 shows that there is little evidence of long-term improvement in Dallas relative to the synthetic control. Both Dallas and the synthetic control show rising test scores from 2004 to 2011, generally falling test scores from 2011 to 2015 and rising test scores from 2016 to 2019. With the exception of 2013 and 2015, Dallas and the synthetic control have similar reading scores throughout the 2004–2019 period. Though the 2013 divergence between the synthetic control and Dallas suggests some caution in interpreting the reading results, the broad similarity between the synthetic control and Dallas in the post-

[8] Synthetic control estimation requires a strongly balanced panel, and we drop the small number of Dallas ISD schools that do not serve a tested grade throughout the period.

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

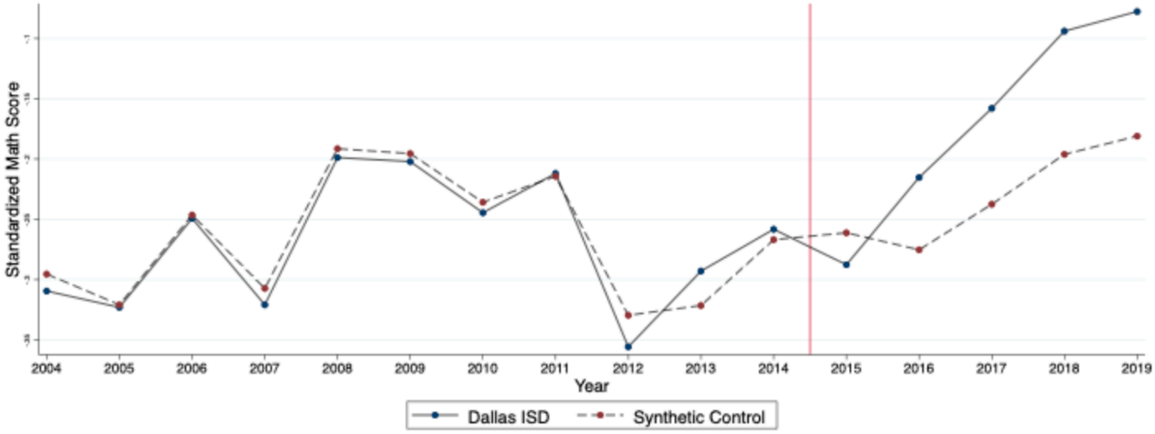


Fig. 1. Synthetic control analysis of math achievement.
Notes: The figure plots average math achievement in Dallas ISD and the synthetic control over time.

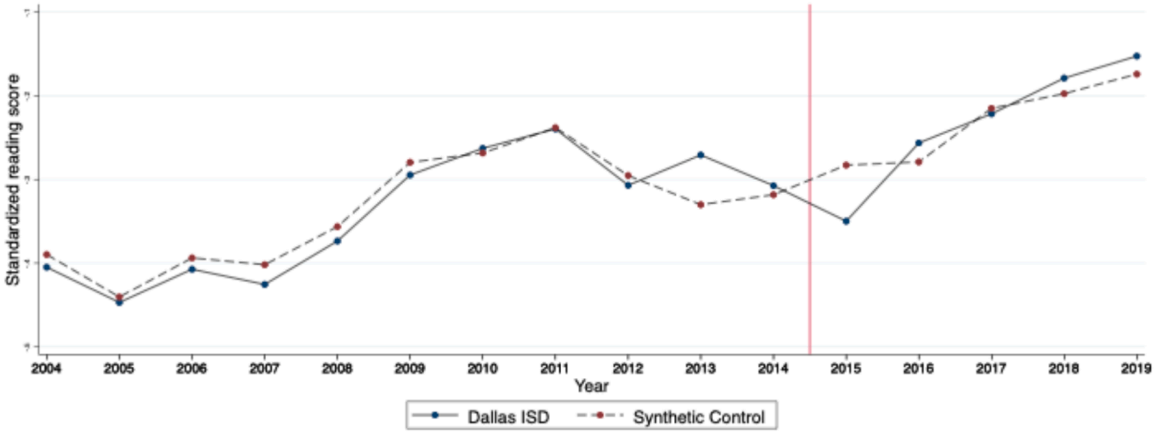


Fig. 2. Synthetic control analysis of reading achievement.
Notes: The figure plots average reading achievement in Dallas ISD and the synthetic control over time.

period suggests that TEI does not improve reading scores in Dallas relative to the synthetic control.[10]

Importantly, Fig. 1 and Fig. 2 show that achievement in both Dallas ISD and the synthetic control district increase steadily following 2015, indicating small achievement gains in high-poverty Texas districts relative to the state as a whole. This underscores the importance of developing adequate control schools because simple comparisons of Dallas ISD with the remainder of the state would lead to the overestimation of TEI effects.

The gradual, rather than immediate, increase in math achievement aligns with expectations for the impacts of major personnel reforms. Significant changes in evaluations and pay likely engender substantial initial disruptions, and the outcomes would be expected to evolve over time through individual teacher improvements and changes in the composition of teachers. In 2015, the first year of TEI, there is no evidence of improved math outcomes in Dallas relative to the synthetic control. The lack of immediate improvement might reflect that in 2015, teachers had not yet received evaluation scores so detailed information on performance did not yet inform professional development or personnel decisions. In 2016, math scores in Dallas and the synthetic control diverge, and the positive math achievement gap between Dallas ISD and the synthetic control district grows noticeably in the following

Table 2
Synthetic control estimates and p-values of the effects on math and reading scores.

Year	Math (1)	Reading (2)
2013	0.029 [0.122]	0.059 [0.000]
2014	0.009 [0.716]	0.011 [0.536]
2015	-0.027 [0.385]	-0.067 [0.002]
2016	0.06 [0.014]	0.023 [0.353]
2017	0.08 [0.003]	-0.007 [0.830]
2018	0.102 [0.000]	0.018 [0.579]
2019	0.103 [0.001]	0.022 [0.571]

Notes: This table provides exact estimates and p-values (in brackets) corresponding to Figs. 1 and 2. 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.

[10] Prior research consistently finds that schools have less impact on reading achievement compared to math achievement; see, for example, Hanushek and Rivkin (2010), Koedel, Mihaly, and Rockoff (2015),Bacher-Hicks and Koedel (2023). This finding is often attributed to parents having a greater impact on reading than on math achievement.

years. By 2019 (the last year of our data), the gap exceeds 0.1 standard deviations.

Columns 1 and 2 of Table 3 describe the weights used in forming the synthetic controls for the math and reading specifications respectively, focusing on the five districts whose schools get the largest weights. Column 1 shows that Houston schools contribute 20 % of the weight, none of the remaining districts contribute more than 10 percent, and the majority (60 %) is divided among schools from many districts; Column 2 shows similar weights for reading scores. This mitigates concerns that a policy reform in another district is driving the estimates. The substantial weight on Houston ISD schools is not surprising given that it is the largest Texas district and most similar to Dallas ISD. On aggregate, almost 80 percent of the weight comes from districts that contribute at most 5 percent.

The sizeable contribution of smaller districts could, however, be problematic if schools in these districts experience systematically different economic, social and policy shocks than those in large Texas districts including Dallas ISD. In Table 4, we therefore assess the robustness of our estimates to successively restricting the donor pool to the 50 (Column 1), 20 (Column 2) and 10 (Column 3) largest high-poverty districts. The estimates for math reported in the top panel support the findings of a substantial TEI treatment effect that increases gradually in magnitude over time, revealing the same qualitative pattern as the baseline specification but increasing in magnitude as the donor pool becomes restricted to schools from larger districts. The 2019 coefficient increases from 0.1 with the unrestricted donor pool of schools from high-poverty districts to 0.18 with a donor pool of schools from the 10 largest high-poverty districts. When using only the largest districts in the donor pool, some specifications show statistically significant effects for reading. However, the consistently smaller estimates than those for math and the insignificant estimates in Table 2 lead us to be cautious in making any claim regarding reading improvement.

To understand better the impacts of TEI and the test score improvements observed in Dallas, we use the same synthetic control approach to describe changes in proportion special education, proportion ELL, and proportion retained in grade in Dallas ISD relative to synthetic control districts determined separately for each outcome. Each of these elements has the potential to affect achievement through impacts on learning or performance on tests, holding knowledge fixed. We also consider whether the reform affected the probability a student exits Dallas ISD. This outcome is of direct interest since it provides suggestive evidence on student preferences for schooling under the reform; it also provides suggestive evidence on the potential contribution of sample selection to the math achievement growth.

Table 5 shows generally small and statistically insignificant effects on these outcomes for most years. We find no significant effect on the probability of grade retention in any year. There is a significant increase in the probability of exit from Dallas and a significant decrease in the probability of ELL classification in 2015, but these are temporary and unlikely to explain the test score improvement that is concentrated in the period 2017–2019.

Table 3
Weights used to construct synthetic control.

	Math	Reading
Houston	0.197	0.264
Laredo	0.081	0.085
Fort Worth	0.049	0.05
Galveston	0.036	0.071
Mullin	0.036	
Port Arthur		0.035
All other districts	0.601	0.495

Notes: The synthetic control approach assigns a weight to each school, and this table describes the aggregate amount of weight received by each district, highlighting the top five districts in terms of aggregate weight. The complete list of districts is available from the authors upon request, but it is too long to present in a table given that no other district gets more than a 0.035 wt and many districts receive very low weight. Mullin and Port Arthur get a positive weight for both the math and reading synthetic control groups, but we only report their weight when they are in the top 5.

Table 4
Synthetic control estimates and p-values of the effects of TEI on math and reading scores for alternative donor pools.

Donor Pool Math	largest 50 districts	largest 20 districts	largest 10 districts
2013	0.009	0.015	0.024
	[0.606]	[0.403]	[0.315]
2014	0.002	0.001	0.028
	[0.944]	[0.975]	[0.421]
2015	-0.028	-0.025	0.017
	[0.202]	[0.261]	[0.533]
2016	0.04	0.066	0.115
	[0.098]	[0.012]	[0.000]
2017	0.054	0.068	0.134
	[0.027]	[0.008]	[0.000]
2018	0.092	0.104	0.171
	[0.001]	[0.001]	[0.000]
2019	0.107	0.111	0.184
	[0.000]	[0.001]	[0.000]
Reading			
2013	0.046	0.043	0.049
	[0.002]	[0.004]	[0.000]
2014	0.009	0.007	0.035
	[0.623]	[0.668]	[0.041]
2015	-0.046	-0.048	-0.011
	[0.012]	[0.018]	[0.658]
2016	0.043	0.047	0.08
	[0.024]	[0.042]	[0.000]
2017	0.028	0.031	0.086
	[0.169]	[0.177]	[0.000]
2018	0.047	0.049	0.112
	[0.046]	[0.038]	[0.000]
2019	0.056	0.047	0.129
	[0.021]	[0.094]	[0.000]

Note: Only schools in districts with a poverty rate of at least 60 percent are included in the potential donor pool. The estimated effects in this table are the gap between Dallas and the synthetic control and the p-values (in brackets) are based on the permutation test described in the text.

The most noteworthy result is that the special education share increases by approximately 2 percentage points in Dallas ISD relative to the synthetic control in both 2018 and 2019. Placing more students in special education can potentially contribute to the improved test scores during this time period, but quantitatively, the contribution is likely to

Table 5
Synthetic control estimates and p-values of the effects of TEI on the probabilities of classification as special education or limited English proficient, being retained in grade, and exiting the district.

	Share Special Education	Share Limited English Proficient	Share Retained in Grade	Share that Leave District
2013	0.008	-0.015	-0.002	0.023
	(0.534)	(0.680)	(0.312)	(0.132)
2014	0.003	0.009	-0.001	0.017
	(0.992)	(0.991)	(0.903)	(0.042)
2015	0.001	-0.051	-0.001	0.034
	(0.997)	(0.002)	(0.573)	(0.001)
2016	0.001	0.015	0.000	-0.022
	(0.999)	(0.875)	(0.922)	(0.663)
2017	0.005	-0.012	-0.004	-0.006
	(0.950)	(0.962)	(0.475)	(0.860)
2018	0.022	0.011	-0.003	0.023
	(0.003)	(0.990)	(0.355)	(0.175)
2019	0.024	-0.030	0.002	na
	(0.000)	(0.147)	(0.261)	na

Notes: The estimated effects in this table are the gap between Dallas and the synthetic control and the p-values (in brackets) are based on the permutation test described in the text.

be small for two reasons. First, nearly all special education students still take the state standardized tests, so placing a student in special education does not alter the composition of test takers. In 2019, 98 % of special education students in Dallas took the exam, and the 2 % who did not are likely to be severe cases whose classification status is unlikely to be changed by TEI.[11] Second, even if special education services were to increase test scores by better meeting student needs, the aggregate performance gains would almost certainly be small given the small fraction of students induced into special education and evidence on the achievement effects of special education programs. The high end of estimates on the effect of special education classification based on analysis that accounts for selection is approximately 0.1 standard deviations (Schwartz, Hopkins, and Stiefel 2021), though they note that effects are considerably smaller among Black and Hispanic students that are the majority of the Dallas sample. Even using the estimate of 0.1 SD, the aggregate effect of placing an additional 2 % of students into special education would only be 0.002 SD.

7. Contributions of educator selection

The bundling of many reform components precludes direct estimation of specific mechanisms underlying the overall TEI impact. We cannot separate the contributions of strengthened incentives, enhanced professional development, and other channels to the overall treatment effects. But we can separate the contribution of overall teacher composition from those of the other channels. 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 TEI raises district quality.[12]

We begin by describing the evaluation scores of stayers and leavers during the post-policy period. We focus on selection out rather than selection into Dallas ISD because of the absence of comparable prior measures of effectiveness for most entrants into Dallas ISD. The existing literature on selective attrition yields mixed findings. For example, Goldhaber, Gross, and Player (2011) find that higher quality teachers are more likely to persist in teaching, whereas West and Chingos (2009) and Nguyen, Qureshi and Ost (2025) find that higher quality teachers leave the public school system at similar rates to less effective ones. In a comprehensive meta-analysis, Nguyen, Pham, Crouch, and Springer (2020) states “across seven studies, we find that increases in teacher effectiveness scores are not associated with increased odds of attrition.” They note that there is suggestive evidence that higher quality teachers are less likely to leave teaching, but that result is not statistically significant.

In addition to examining exit from the district, we also describe transitions by whether a teacher remains in a tested grade and whether they leave their school but remain in the district. Fig. 3 shows that there is pronounced negative selection out of tested grades that is largest for teachers who leave the district. The average evaluation scores of teachers who remain in their tested grade exceed those who leave the district following the school year by more than 0.5 standard deviations. The lower two panels of Fig. 3 show that this strong negative selection holds for both the performance and achievement components. Finally, all 3 teacher quality measures show smaller negative selection for those who move to a non-tested grade but remain in Dallas ISD; the effectiveness of teachers moving to non-tested grades is similar regardless of whether they stay in the same school.

Although the negative selection of exits in the 2015–2019 period is suggestive,[13] we cannot directly assess the contribution of selective attrition to the impacts of TEI. First, we cannot calculate value added and do not have the TEI performance metrics in the pre-policy period.[14] And second, there is no performance information for teachers entering Dallas ISD from other districts. Luo (2023) does show that even 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. Nonetheless, we are unable to identify the overall effect of TEI on positive selection.

Instead of estimating the effect of TEI on selection patterns, we separate the contribution of teacher composition from those of the other channels. The sample includes all elementary school math teachers in high-poverty districts so that we can see how Dallas achievement patterns differ from other high-poverty districts. We compare estimates of achievement changes over time from a regression of achievement on a set of year dummies with estimates from a regression that also includes teacher fixed effects and a full set of experience dummies for years 1 to 10 and 11 plus. This latter regression shows the time path of achievement based solely on within-teacher achievement changes. We focus the decomposition on math achievement because it is the subject with the clearest evidence of improvement. To assess how the role of teacher composition differs in Dallas compared to other high-poverty Texas districts, we interact the year effects with a Dallas indicator.[15] The parameter of interest is how these Dallas-by-year effects change when teacher composition is controlled for.

Eq. (1) models achievement for student i in year t with teacher j as a function of a Dallas indicator (D), year dummy variables (Y), a set of experience dummies *exp*, a teacher fixed effect (η_j) and a random error:

$$A_{ijt} = \alpha + \omega D_i + \sum_{t=2013}^{2019} \gamma_t Y_t + \sum_{t=2013}^{2019} \delta_t D_i Y_t + \sum_{x=1}^{10+} \lambda_x exp_x + \eta_j + \varepsilon_{ijt} \qquad (1)$$

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 Dallas-by-year dummies $\left(\widehat{\delta_t^{no\ fe}}\right)$ capture the influences of all factors including teacher composition that contribute to the achievement difference between Dallas ISD and other districts in year t relative to the omitted baseline year (2014). The inclusion of teacher fixed effects and the experience dummies shuts the teacher composition channel by considering just within-teacher variation not related to experience, 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 of the contribution of teacher composition to changes over time in Dallas relative to other districts.

If all of the improvement in Dallas ISD schools comes from replacing

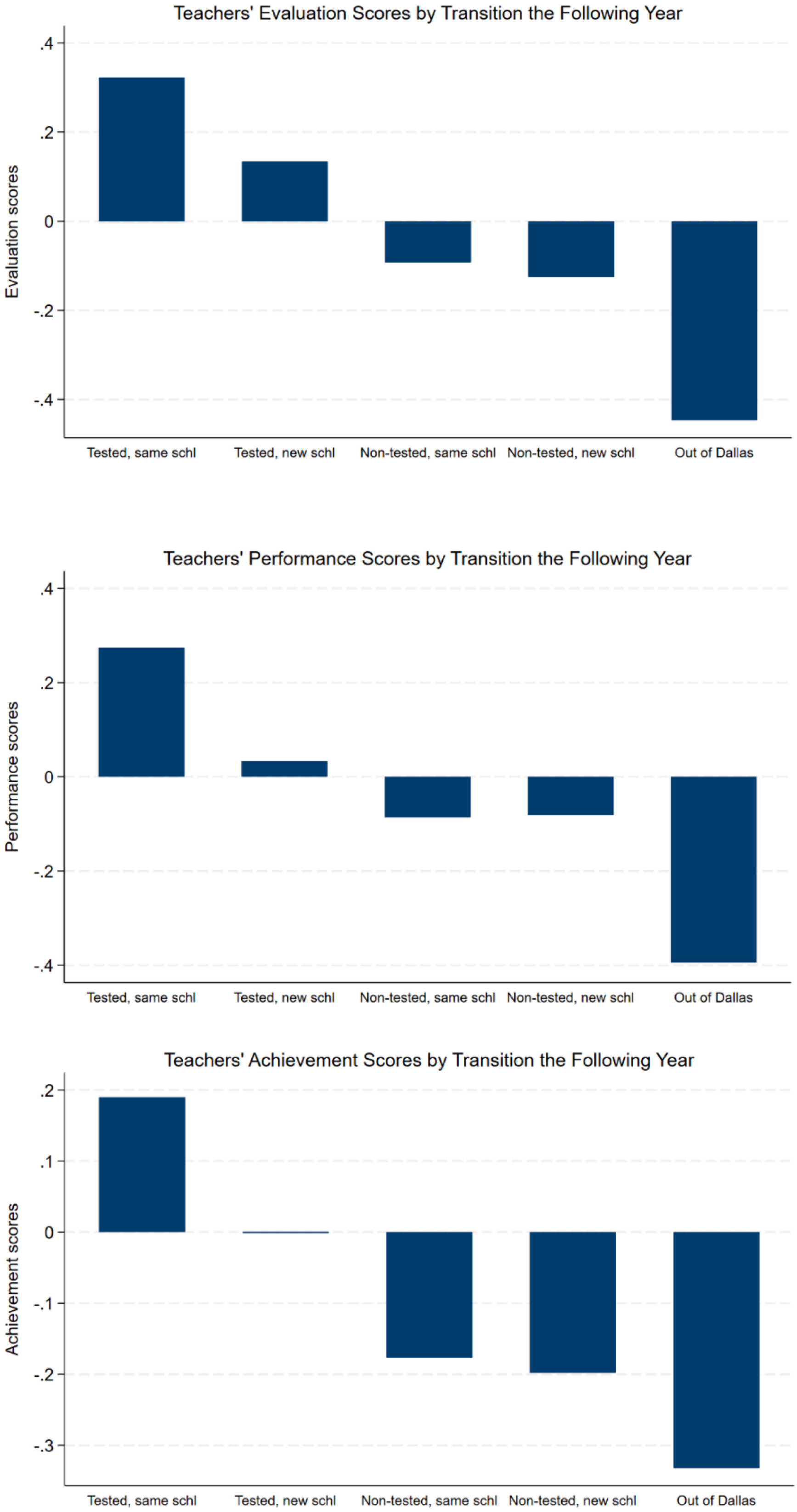
[11] We have also verified directly that there is no TEI effect in any year on the fraction of students who take the exam.

[12] Note that compositional effects were previously identified as a central element of the IMPACT program in Washington, DC (Dee and Wyckoff (2015)).

[13] The exit of less effective teachers from the district should unambiguously improve outcomes. The movement of less effective teachers to non-tested grades will have a more complicated effect on student outcomes as it depends on the grade-specific match of teachers moving to non-tested grades and whether there is a persistent effect of having a lower quality teacher in a non-tested early grade.

[14] 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.

[15] We do not use the synthetic control weights for this exercise because of complications arising from including teacher fixed effects when teachers might move between schools with positive and zero weight. As such, we do not view the Dallas-by-year interactions as estimates of the treatment effect since other Texas schools may not perfectly capture counterfactual trends for Dallas. For the purpose of assessing the role of teacher composition, however, this need not confound conclusions because if there are Dallas-specific shocks, these would affect both the model with and without teacher composition controls.



Teachers' Evaluation Scores by Transition the Following Year

Teachers' Performance Scores by Transition the Following Year

Teachers' Achievement Scores by Transition the Following Year

Fig. 3. Mean teacher overall evaluation and component scores, by annual transition status.

less effective teachers with more effective teachers and from changes in the experience distribution, then we would expect to find small and insignificant Dallas-by-year 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 Dallas-by-year coefficients to be insensitive to the inclusion of teacher fixed effects and experience. When both teacher composition and other factors contribute to the overall treatment effects, the difference between the Dallas-by-year effects with and without the teacher fixed effects and experience controls provides an estimate of the contribution of teacher composition. To statistically test the change in the Dallas-by-year effects, we estimate Eq. (1) both with and without the teacher composition controls in a single regression and test the marginal change in these effects when teacher composition is accounted for $(\widehat{\delta_t^{fe}} - \widehat{\delta_t^{no\ fe}})$.

Specifically, we follow Oberfichtner and Tauchmann (2021) and duplicate each observation to construct two identical stacked samples. We use the first sample to estimate the $\widehat{\delta_t^{no\ fe}}$ effects without the teacher controls, and we use the second sample to estimate the $\widehat{\delta_t^{fe}}$ effects controlling for the teacher controls. Stacking the two samples allows us to simultaneously estimate these two models and include an interaction term that tests whether the $\widehat{\delta_t}$ effects differ across the two models. The exact estimating equation is

$$\begin{aligned} A_{ijt} &= \alpha + \kappa S_{ijt} + \omega D_i + \omega^S D_i S_{ijt} + \sum_{t=2013}^{2019} \left(\gamma_t Y_t + \gamma_t^S Y_t S_{ijt}\right) \\ &\quad + \sum_{t=2013}^{2019} \left(\delta_t D_i Y_t + \delta_t^S D_i Y_t S_{ijt}\right) + S_{ijt}\eta_j + S_{ijt} \sum_{x=1}^{10+} \lambda_x \exp_x + \varepsilon_{ijt} \end{aligned}$$

where S_{ijt} is an indicator for the second stacked sample and all other variables are as in Eq. (1). Interacting S_{ijt} with the teacher FE and experience indicators (but leaving out the main effects) means that these controls are a constant in the $S_{ijt} = 0$ sample. The parameters of interest are the δ_t^S, which capture the change in the Dallas-specific event study indicators when the teacher FE and experience indicators are added to the model. Standard errors are clustered at the teacher level accounting both for correlation within a teacher and for the correlation across the identical samples.

Table 6 reports the set of year and Dallas-by-year dummy coefficients for regressions with no teacher composition controls (Column 1) and both teacher fixed effects and experience controls (Column 2).[16] The similarity of the year effects on Columns 1 and 2 shows that teacher composition explains little of the achievement changes over time in the control districts. In contrast, the addition of the teacher composition variables substantially reduces the magnitudes of the Dallas ISD/year interactions for the years 2016 to 2019. Column 3 shows that these changes are significant at the 1 percent level in 2018 and 2019. By 2019, composition is estimated to account for 0.086 standard deviations, slightly more than half of the 2019 achievement difference between Dallas ISD and the other high-poverty district schools in Column 1.

Table 6
Year dummy coefficients for Dallas and all other high-poverty districts, by inclusion of teacher composition controls.

Includes teacher composition controls	no	Yes	p values for test of differences in coefficients
2013	-0.015 (0.006)	-0.015 (0.005)	0.996
2014	ref.	ref.	
2015	-0.037 (0.006)	-0.04 (0.005)	0.437
2016	-0.032 (0.007)	-0.03 (0.006)	0.883
2017	-0.003 (0.007)	0.008 (0.007)	0.117
2018	0.029 0.007	0.028 (0.008)	0.805
2019	0.03 (0.007)	0.028 (0.009)	0.727
Dallas x 2013	-0.06 (0.021)	-0.076 (0.019)	0.41
Dallas x 2014	ref.	ref.	
Dallas x 2015	-0.027 (0.020)	-0.016 (0.017)	0.412
Dallas x 2016	0.041 (0.022)	0.016 (0.019)	0.261
Dallas x 2017	0.062 (0.023)	0.025 (0.021)	0.11
Dallas x 2018	0.124 (0.023)	0.053 (0.022)	0.002
Dallas x 2019	0.153 (0.024)	0.067 (0.024)	0.000
	1907,113	1907,113	

Notes: The table shows how the year and Dallas-by-year interaction terms change when teacher composition controls are added. Standard errors, clustered at the teacher level are shown in parentheses. Column 3 shows p-values testing the equality of coefficients across columns 1 and 2.

To assess whether the teacher fixed effects or teacher experience is a more important driver of the compositional change, we use the Gelbach (2016) decomposition.[17] Though the Gelbach decomposition has the advantage of not depending on the order in which controls are added, a challenge for our context is that it requires explicitly estimating models for each covariate which is computationally infeasible with very high-dimensional fixed effects. However, if we focus on just Dallas ISD, the Gelbach decomposition is estimable. We thus conduct the Gelbach decomposition on just the Dallas sample and estimate how controlling for teacher experience and/or teacher fixed effects alters the estimated year effects. Given the Table 6 results that show little or no effects of teacher composition on the year effects for the control schools, we believe the focus on Dallas ISD schools yields informative findings. Column 1 of Table 7 shows the total change in the year fixed effects from including the teacher controls, and columns 2 and 3 show the relative importance of the teacher FE and experience, respectively. The results of the Gelbach decomposition demonstrate that the teacher fixed effects drive the teacher composition contribution to the achievement increase; the changes due to experience are small, concentrated around TEI implementation, and go in the opposite direction. Importantly, the sizeable contribution of teacher composition highlights the potential for personnel reforms to alter teacher composition in ways that raise the quality of instruction and achievement.

As discussed previously, 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, of peer

[16] Note that roughly 5 percent of students are not matched with a single math teacher and consequently dropped from the sample. There are only small differences between the average achievement and low-income share of included and excluded students, and the achievement growth between 2015 and 2019 is modestly higher for this sample than for synthetic control sample that includes the observations that are not matched with a single teacher.

[17] The Gelbach (2016) decomposition is based on a direct application of the omitted variable bias formula where the full model is used to estimate the relationship between each covariate and the outcome, and a series of regressions estimate the relationship between the treatment variable of interest and each covariate. See Gelbach (2016) for further details.

Table 7
Gelbach decomposition of changes in year effects due to teacher fixed effects and experience.

	total contribution of teacher composition	contribution of teacher fixed effects	contribution of teacher experience
2013	0.0165	0.009	0.007
	(0.007)	(0.008)	(0.002)
2014	ref.	ref.	ref.
2015	-0.015	-0.007	-0.008
	(0.006)	(0.007)	(0.003)
2016	0.014	0.02	-0.007
	(0.008)	(0.009)	(0.004)
2017	0.013	0.019	-0.006
	(0.009)	(0.011)	(0.004)
2018	0.059	0.058	0.001
	(0.011)	(0.013)	(0.004)
2019	0.075	0.074	0.001
	(0.013)	(0.015)	(0.004)

Notes: Using the Gelbach decomposition, the table presents how the year effects change when controlling for teacher composition and how much of that change comes from teacher fixed effects versus teacher experience. The sample is restricted to just Dallas ISD for computational feasibility of the Gelbach decomposition.

teacher effects, or of improvements in school leadership. Nonetheless, their contributions and those of other factors including improvements in academic support and school climate account for less of the math achievement gain than teacher composition.

8. Conclusions

Previous inability to raise substantially the achievement of disadvantaged students in large urban districts has led to pessimism about the prospects for developing policies that could improve the effectiveness of urban schools. But the general conclusions about the overwhelming challenges of urban school systems have generally rested on accepting the overall structure of teacher incentive and compensation systems. The positive experiences of the IMPACT reforms in Washington, D.C. offered the promise of significant impacts of rigorous evaluation accompanied by performance pay, exactly the reform structure adopted and extended in Dallas ISD. The comprehensive Dallas reforms replaced the dependence of teacher salary on experience and post-graduate degrees in a system with strong incentives for classroom effectiveness. This radical deviation from the personnel systems commonly used in US school districts aligns with long-standing recommendations of economists that called for a closer connection between pay and performance (Kershaw and McKean (1962)).

Using a synthetic control approach, we find that Dallas elementary students improved in math following the reform. Effect sizes of 0.1 standard deviations are large, particularly in comparison to much more costly interventions such as large reductions in class size. Relative to the pay-for-performance interventions reviewed in Pham, Nguyen, and Springer (2021), the Dallas reform is much more impactful than the average effect of 0.043 SD.

Teacher composition appears to be an important factor driving the reform impacts. The extensive principal training in instructional leadership and strong incentives for principals to elevate the quality of instruction and to faithfully evaluate teacher effectiveness may be important mediating factors in the successes of TEI.

Our estimates are best interpreted as the effect of TEI when a single district adopts the policy and may not scale were all districts to adopt it. When a single district adopts TEI, the policy can affect teacher composition by attracting high-quality teachers from nearby districts or encouraging low-performing teachers to exit to other districts. However, if all state districts were to adopt the policy simultaneously, this cross-district-movement channel would be shut down, reducing the size of the policy effect. Importantly, wide-spread adoption could strengthen other margins of response, particularly in the long run. A system-wide shift to performance-linked compensation would have larger effects on aggregate entry into the profession since an individual considering whether to become a teacher may not strongly consider the pay structure in a single district. Furthermore, system-wide adoption may alter expectations about the permanence of such reforms, reshape professional norms, and influence teacher training in ways that improve effectiveness and future compensation.

The policy changes in Dallas ISD and previously implemented reforms in Washington, D.C. demonstrate that radical changes that strengthen the links between teacher effectiveness and labor market outcomes can be instituted and sustained in large urban districts. As the failure of the RTT legislation illustrated, however, instituting such fundamental changes are difficult to enact through broad policy reforms. Nevertheless, based on the Dallas ISD experience with TEI and PEI, the Texas legislature moved in 2019 to provide incentives for other Texas districts to develop similar programs of teacher evaluation and pay. By the end of 2023, over 250 districts were participating in the Teacher Incentive Allotment program that required the use of achievement and observations of teaching in pay for performance systems. This response led the legislature to expand the underlying grant program significantly in 2025 with the Teacher Retention Allotment (TRA). Whether the statewide benefits will match those produced in Dallas ISD remains to be seen.

CRediT authorship contribution statement

Eric Hanushek: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Jin Luo:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Andrew Morgan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Minh Nguyen:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis. **Ben Ost:** Writing – review & editing, Writing – original draft, Supervision, Methodology. **Steven Rivkin:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Ayman Shakeel:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation.

Appendix A

Table A1
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 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	Artifacts and informal observations	15
Domain 2: Instructional Practice	2.1. Establishes clear, aligned standards-based lesson objective(s) (3x) 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)	Spot, extended and informal observations	48
Domain 3: Classroom culture	3.1. Maximizes instructional time (spot) (3x) 3.2. Maintains high student motivation (2x) 3.3. Maintains a welcoming environment that promotes learning and positive interactions (2x)	Spot, extended and informal observations	21
Domain 4: Professionalism and Collaboration	4.1. Models good attendance for students 4.2. Follows policies and procedures, and maintains accurate student records 4.3. Engages in professional development	Artifacts and informal observations	15

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

Table A2
Teacher categories and evaluation templates.

Teacher Category	Teacher Performance	Student Achievement	Student Perception
Category A: Most grade 3–12 teachers whose students take an Assessment of Course Performance (ACP), The State of Texas Assessments of Academic Readiness (STAAR), or Advanced Placement (AP) exam, including most K-5 special teachers	50	35	15
Category B: Most K-2 teachers whose students take an ACP or Iowa Test of Basic Skills (ITBS)/Logramos	65	35	0
Category C: Most grade 3–12 teachers whose students do not take an ACP, STAAR, or AP assessment but who are able to complete a student survey (e.g. Career and Technical Education (CTE) teachers)	65	20	15
Category D: Any teachers whose students do not take an ACP, STAAR, or AP assessment nor are eligible to complete a student survey (e.g. pre-K teachers, Teachers not-of-record such as special education inclusion teachers, talented and gifted teachers)	80	20	0

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

Table A3
Synthetic control estimates, by use of enrollment weights.

	Math		Reading	
	not weighted	weighted	not weighted	Weighted
2013	0.029	0.029	0.059	0.056
2014	0.009	0.006	0.011	0.006
2015	-0.027	-0.034	-0.067	-0.066
2016	0.06	0.05	0.023	0.014
2017	0.08	0.075	-0.007	-0.005
2018	0.102	0.103	0.018	0.024
2019	0.103	0.113	0.022	0.028

Notes: The table shows how estimates differ depending on whether we use enrollment weights.

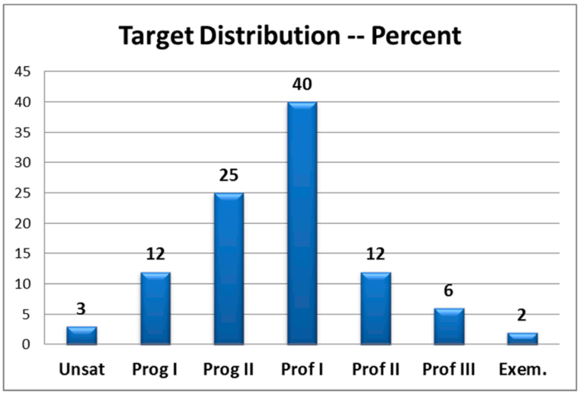


Fig. A1. Target distribution of teacher effectiveness scales.
Source: TEI Rulebook v4.1 (DISD (2017)).

References

Abadie, Alberto, Diamond, Alexis, Hainmueller, Jens, 2010. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. J. Am. Stat. Assoc. 105 (490), 493–505.

Abadie, Alberto, Gardeazabal, Javier, 2003. The economic costs of conflict: a case study of the Basque Country. Am. Econ. Rev. 93 (1), 113–132.

Adnot, Melinda, Dee, Thomas, Katz, Veronica, Wyckoff, James, 2017. Teacher turnover, Teacher quality, and student achievement in DCPS. Educ. Eval. Policy Anal. 39 (1), 54–76.

edited by Bacher-Hicks, Andrew, Koedel, Cory, 2023. Estimation and interpretation of teacher value added in research applications. In: Hanushek, Eric A., Machin, Stephen, Woessmann, Ludger (Eds.), Handbook of the Economics of Education, Handbook of the Economics of Education, 6. Elsevier, pp. 93–134. edited by.

Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M.A., Springer, M.G., 2025. Taking teacher evaluation to scale: the effect of state reforms on achievement and attainment. J. Political Econ. Microecon. 3 (3), 000-000.

Cavallo, Eduardo, Galiani, Sebastian, Noy, Ilan, Pantano, Juan, 2013. Catastrophic natural disasters and economic growth. Rev. Econ. Stat. 95 (5), 1549–1561 (December).

Coleman, James S., Campbell, Ernest Q., Hobson, Carol J., McPartland, James, Mood, Alexander M., Weinfeld, Frederic D., York, Robert L., 1966. Equality of Educational Opportunity. U.S. Government Printing Office, Washington, D.C.

Dee, Thomas S., Wyckoff, James, 2015. Incentives, selection, and teacher performance: evidence from IMPACT. J. Policy Anal. Manage. 34 (2), 267–297. Spring.

Dotter, Dallas, Chaplin, Duncan, Bartlett, Maria, 2021. Impacts of School Reforms in Washington, DC on Student Achievement. Mathematica, Washington, DC. August 12.

Duflo, Esther, Hanna, Rema, Ryan, Stephen P., 2012. Incentives work: getting teachers to come to school. Am. Econ. Rev. 102 (4), 1241–1278.

Fairbrother, G., Hanson, K L, Friedman, S., Butts, G C, 1999. The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates. Am J Public Health 89 (2), 171–175.

Fairbrother, Gerry, Siegel, Michele J., Friedman, Stephen, Kory, Pierre D., Butts, Gary C., 2001. Impact of financial incentives on documented immunization rates in the inner City: results of a randomized controlled trial. Ambulatory Pediatrics 1 (4), 206–212 (2001/07/01/).

Fryer, Roland G., 2013. Teacher incentives and student achievement: evidence from New York City public schools. J Labor Econ 31 (2), 373–427.

Galiani, Sebastian, Quistorff, Brian, 2017. The Synth_Runner package: utilities to automate synthetic control estimation using synth. Stata J 17 (4), 834–849.

Gelbach, Jonah B., 2016. When do covariates matter? And which ones, and how much? J Labor Econ 34 (2), 509–543. April.

George, Bert, van der Wal, Zeger, 2023. Does Performance-related-pay work? Recommendations for practice based on a meta-analysis. Policy. Des. Pract. 6 (3), 299–312. April.

Glazerman, Steven, Seifullah, Allison, 2012. An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) After Four Years. Mathematica Policy Research. March 7.

Goldhaber, Dan, Gross, Betheny, Player, Daniel, 2011. Teacher career paths, teacher quality, and persistence in the classroom: are public schools keeping their best? J. Policy Anal. Manage. 30 (1), 57–87. Winter.

Goodman, Sarena F., Turner, Lesley J., 2013. The design of teacher incentive pay and educational outcomes: evidence from the New York City bonus Program. J Labor Econ 31 (2), 409–420. April.

Hanushek, Eric A., Light, Jacob D., Peterson, Paul E., Talpey, Laura M., Woessmann, Ludger, 2022. Long-run trends in the U.S. SES-achievement gap. Educ Finance Policy 17 (4), 608–640.

Hanushek, Eric A., Rivkin, Steven G., 2010. Generalizations about using value-added measures of teacher quality. Am. Econ. Rev. 100 (2), 267–271. May.

Hanushek, Eric, Luo, Jin, Morgan, Andrew, Nguyen, Minh, Ost, Ben, Rivkin, Steve, Shakeel, Ayman, 2026. The Effects of Comprehensive Pay Reform on Achievement in Urban Schools. Inter-university Consortium for Political and Social Research [distributor], Ann Arbor, MI. https://doi.org/10.3886/E246745V1, 2026-03-10.

Hasnain, Zahid, Manning, Nick, Pierskalla, Jan Henryk, 2014. The promise of performance pay? Reasons for caution in policy prescriptions in the core civil service. World Bank Res Obs 29 (2), 235–264. August.

Hillman, A L, Ripley, K., Goldfarb, N., Nuamah, I., Weiner, J., Lusk, E., 1998. Physician financial incentives and feedback: failure to increase cancer screening in Medicaid managed care. Am J Public Health 88 (11), 1699–1701.

Hillman, Alan L., Ripley, Kimberly, Goldfarb, Neil, Weiner, Janet, Nuamah, Isaac, Lusk, Edward, 1999. The use of physician financial incentives and feedback to improve pediatric preventive care in medicaid managed care. Pediatrics. 104 (4), 931–935.

Kaul, Ashok, Klößner, Stefan, Pfeifer, Gregor, Schieler, Manuel, 2022. Standard synthetic control methods: the case of using all preintervention outcomes together with covariates. J. Business Econ. Stat. 40 (3), 1362–1376. July.

Kershaw, Joseph A., McKean, Roland N., 1962. Teacher Shortages and Salary Schedules. McGraw-Hill, NY.

Koedel, Cory, Mihaly, Kata, Rockoff, Jonah E., 2015. Value-added modeling: a review. Econ Educ Rev 47, 180–195.

Kouides, Ruth W., Nancy, M., Bennett, Bonnie, Lewis, Joseph D., Cappuccio, William H., Barker, F., 1998. Performance-based physician reimbursement and influenza immunization rates in the elderly. Am J Prev Med 14 (2), 89–95, 1998/02/01/.

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

Lavy, Victor., 2002. Evaluating the effect of teachers' group performance incentives on pupil achievement. J. Polit. Econ. 110 (6), 1286–1317. December.

Lavy, Victor., 2020. Teachers' Pay for performance in the long-run: the dynamic pattern of treatment effects on students' Educational and labour market outcomes in adulthood. Rev. Econ. Stud. 87 (5), 2322–2355.

Lazear, Edward P., 2000. Performance pay and productivity. American Economic Review 90 (5), 1346–1361. December.

Luo, Jin., 2023. Teachers' Responsiveness to Performance-Based Pay: Evidence from a Large Urban School District in Texas. University of Illinois Chicago (mimeo).

Muralidharan, Karthik, Sundararaman, Venkatesh, 2011. Teacher performance pay: experimental evidence from India. J. Polit. Econ. 119 (1), 39–77. February.

Nguyen, Tuan D., Pham, Lam D., Crouch, Michael, Springer, Matthew G., 2020. The correlates of teacher turnover: an updated and expanded meta-analysis of the literature. Educ. Res. Rev. 31 (November), 100355.

Oberfichtner, M., Tauchmann, H., 2021. Stacked linear regression analysis to facilitate testing of hypotheses across OLS regressions. Stata J 21 (2), 411–429.

Pham, Lam D., Nguyen, Tuan D., Springer, Matthew G., 2021. Teacher Merit pay: a meta-analysis. Am. Educ. Res. J. 58 (3), 527–566.

Roski, Joachim, Jeddeloh, Robert, An, Larry, Lando, Harry, Hannan, Peter, Hall, Carmen, Zhu, Shu-Hong, 2003. The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines. Prev. Med 36 (3), 291–299, 2003/03/01/.

Schwartz, A.E., Hopkins, B.G., Stiefel, L., 2021. The effects of special education on the academic performance of students with learning disabilities. J. Policy Anal. Manage. 40 (2), 480–520.

Shakeel, Ayman., 2023. High-Stakes Objective and Subjective Teacher Evaluation Measures and Student Skill Development. University of Illinois Chicago, Chicago. August 10.

Shearer, Bruce., 2004. Piece rates, fixed wages and incentives: evidence from a field experiment. Rev. Econ. Stud. 71 (2), 513–534. April.

Sojourner, Aaron J., Mykerezi, Elton, West, Kristine L., 2014. Teacher pay reform and productivity: panel data evidence from adoptions of Q- comp in Minnesota. J. Human Resources 49 (4), 945–981.

Springer, Matthew G., Ballou, Dale, Hamilton, Laura, Le, Vi-Nhuan, Lockwood, J.R., McCaffrey, Daniel F., Pepper, Matthew, Stecher, Brian M., 2010. Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. National Center on Performance Incentives, Vanderbilt University, Nashville, TN.

Steinberg, Matthew P., Sartain, Lauren, 2015. Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. Educ Finance Policy 10 (4), 535–572. Fall.

Taylor, Eric S., Tyler, John H., 2012. The effect of evaluation on teacher performance. Am. Econ. Rev. 102 (7), 3628–3651. December.

Weibel, A., Rost, K., Osterloh, M., 2010. Pay for performance in the public sector—Benefits and (hidden) costs. J. Public Administration Res. Theory 20 (2), 387–412.

edited by West, Martin R., Chingos, Matthew M., 2009. Teacher effectiveness, mobility, and attrition in Florida. In: Springer, Matthew (Ed.), Performance incentives: Their growing Impact on American K12 Education. Brookings Institution Press, Washington, DC, pp. 251–272. edited by.