Getting Effective Educators in Hard-to-Staff Schools

Andrew Morgan*, Minh Nguyen*, Eric Hanushek***, Ben Ost*, and Steven Rivkin**

February 2019

PRELIMINARY AND INCOMPLETE
PLEASE DO NOT CITE

Abstract

Financial incentives to attract and retain educators in difficult to staff schools have typically not differentiated by educator performance, and this substantially weakens efforts to elevate the quality of instruction and raise achievement. The Dallas Independent School District developed a program that provided educators substantial additional compensation to teachers and principals willing to work in very low-achievement schools. Crucially, the amount of the additional pay depended upon a teacher or principal's effectiveness in previous years as measured by comprehensive systems of evaluation. Moreover, the district committed to staffing these schools with effective educators. Difference-in-differences estimates reveal dramatic achievement increases in reading and especially mathematics that substantially narrow the gap between students in these schools and the district average, highlighting the potential for targeted compensation programs linked with educator effectiveness to reduce the gap in school quality and achievement

*University of Illinois at Chicago; **Stanford University, University of Texas at Dallas, and NBER, University of Illinois at Chicago, University of Texas at Dallas, and NBER

1. Introduction

Attracting and retaining effective educators constitutes a primary challenge to raising the quality of instruction, achievement and social mobility in schools that are chronically low achieving. Evidence suggests that additional pay can increase retention, but supplemental pay for working in disadvantaged schools does not typically depend upon effectiveness and therefore is highly unlikely to have a major impact on the quality of instruction.[1] Clotfelter et al. (2011) argues that these types of pay premia are unlikely to equalize teacher quality across advantaged and disadvantaged schools because they differentially attract teachers with worse credentials. A recent randomized controlled trial provides evidence that a program that paid effective educators $10,000 per year for two years succeeded in attracting small numbers of high value-added teachers to designated schools and modestly raising achievement (Glazerman et al, 2013). Whether financial inducements to qualified educators could be implemented at scale and transform low-performing schools is a fundamental question for those seeking to raise the quality of instruction in disadvantaged urban and rural schools. The identification of effective teachers and school leaders requires a system that accurately measures educator quality, a challenging problem given the difficulties of measuring educator effectiveness.

In an effort to raise the quality of teaching and school leadership the Dallas Independent School District (DISD) introduced comprehensive evaluation and compensation reform that rated

[1] Research that investigates the effects of programs designed to attract educators to hard-to-staff schools includes Clotfelter, et al. (2008); Clotfelter, Ladd and Vigdor (2011); Steele et al. (2010); Springer, Swain and Rodriguez (2015); Cowan and Goldhaber, (2015); Springer et al. (2010); Springer, Swain and Rodriguez (2016); and Glazerman et al (2013)

educators on the basis of supervisor observations, student achievement, and student or family surveys and established compensation systems based on these evaluations. Concerns about persistent low performance and the possibility that the reforms could exacerbate staffing difficulties in low-performing schools, DISD formulated and implemented a compensatory pay policy that offered bonuses and substantial salary increases to educators willing to work in seven of the persistently lowest-achievement schools in the district. Importantly, Dallas ISD structured the program such that the pay premia depend upon the teacher's evaluation during the previous year. They also made a commitment to transform these schools immediately; over 60 percent of the teachers and all seven principals in ACE schools were replaced prior to the 2015-16 academic year.

The ratings produced by the comprehensive evaluation reform provided the necessary information on educator quality, and a sizeable pool of educators applied to work in an ACE school. Note that roughly 60 percent of DISD teachers are rated effective, and among ACE teachers roughly one third were in their first three years and thus not eligible for a proficient rating, another 28 percent were rated proficient, and less than 40 percent were in the higher, distinguished category. This is clearly not an intervention that is reliant on stars.

Our empirical analysis uses a difference-in-differences design based on the fact that the central administration in Dallas identified a set of 25 schools that would potentially receive the intervention, but only 7 schools were initially selected into ACE; the remaining were referred to as "near-ACE" and provide a potential comparison group. ACE grew by six additional schools three years later in the second wave of the program, three coming from the "near-ACE" group and the remainder from other Dallas ISD schools. The first wave of ACE includes four primary

schools and three middle schools, while the second wave of ACE includes five primary schools and one middle school.

We show that at baseline, schools in the first wave of ACE adoption are systematically more disadvantaged than near-ACE schools, but the two groups are on nearly identical achievement trends prior to policy implementation. These trends diverge sharply following ACE adoption, and ACE schools catch up with and overtake near-ACE schools within one year. Schools in the second wave of ACE are on a systematically downward trend in the years leading up to ACE adoption, reflective of district efforts to target struggling schools. This trend reverses immediately following ACE adoption in 2017-2018. By embedding salary inducements to teach in low-performing schools into comprehensive systems of evaluation linked with compensation, DISD was remarkably effective at their goal of transforming the district's worst performing schools. In a single year, the test-score gap between the targeted, low-performing schools, and the average DISD school closed by more than 50 percent and these gains continue in subsequent years. This dramatic success is consistent with the notion that a strong personnel system combined with the commitment of resources necessary to attract effective educators can dramatically improve school quality.

Though the central analysis estimates the overall effect of ACE, we also explore potential mechanisms through which the reform affects the quality of achievement. ACE not only overhauls the composition of teachers and replaces the principal, but it also amplifies the incentives for teachers and principals to elevate their performance and achievement. A more effective principal would be expected to enhance teacher development, and the evidence suggests that teachers benefit from working with highly effective peers.[2] Finally, the structure of

[2] Jackson and Bruegmann (2009)

the teacher pay premia strengthened incentives because the supplement amount depends on the evaluation score. Each ACE educator receives a $2,000 annual bonus, and teacher salary supplements increase from $6,000 for those in their first three years or not rated proficient, $8,000 for those rated proficient, and $10,000 for high-scoring teachers who successfully complete the distinguished teacher process. Therefore, the supplements increase the salary differentials under the teacher compensation reform, elevating the financial incentives to raise the quality of instruction.[3] The pay-for-performance literature has mixed findings with some studies finding positive effects and many studies finding null effects.[4] A recent recent meta-analysis (Pham, Nguyen and Springer 2018) finds that on average, pay-for-performance improves outcomes slightly (0.03 standard deviations for studies of the United States). Regardless, given that the district wide reforms provide performance incentives for all Dallas ISD educators, we do not expect incentives to be a primary channel through which ACE raises achievement.

Although ACE focuses solely on personnel policies and educator quality, the dramatic changes in the composition of educators has elements of a school turnaround intervention. The literature on school turnarounds reports mixed results, with several studies finding null or negative effects (Heissel and Ladd, 2016); Dougherty and Weiner, 2017) and other studies finding large gains from turnarounds (Dee, 2012; Strunk et al., 2016; Papay, 2015; Schueler et al (2016); Zimmer, Henry, & Kho, 2016). Schueler et al (2016) investigated a district-led turnover effort in Lawrence, Massachusetts that generated mathematics and reading achievement

[3] Principals and educators in other positions received fixed bonuses that did not vary by prior rating.

[4] See (Fryer, 2013; Goldhaber and Walch, 2012; Balch and Springer, 2015; Figlio and Kenny, 2007; Fryer, Levitt, List and Sadoff, 2012; Dee and Wyckoff, 2015; Brehm, Imberman and Lovenheim, 2015; and Glazerman and Seifullah, 2012).

increases that approach those produced by ACE. In contrast to the ACE reform that focused on raising the quality of school instruction, however, the Lawrence, Massachusetts turnaround incorporated a number of components including expansion of the school year by an average of 200 hours for all elementary and middle school students and enrichment activities after school at a per-student cost that far exceeds the value of the salary supplements paid to ACE educators.

The next section describes the administrative and program data, and Section 3 discusses the Dallas ISD evaluation and pay reforms. Section 4 presents the differences-in-differences empirical model and estimates, highlighting potential threats to identification and relevant evidence. Section 5 explores the contributions of potential channels to the achievement increase. The final section summarizes the findings and considers implications for personnel policies.

2. Data

The data for analysis are merged from several sources. Data on student and staff characteristics come from Dallas ISD administrative data as submitted to the Texas Education Agency. We use student standardized math and reading test scores from the State of Texas Assessments of Academic Readiness (STAAR), as measures of student achievement, where the tests are standardized within Dallas ISD separately by subject and year. Other student information includes courses, grades and schools attended, race, gender, indicator for students qualifying for programs such as free or reduced lunch, gifted, special education, and limited English proficiency. Staff information also contains demographics and courses, grades and schools taught.

We also have access to unique data that include scores and sub-metrics for all the components used in the evaluation rating process, such as teacher performance as measured by

rubric-based observations, and student or family perception as measured by surveys from students and parents. We construct a panel that links teachers, students, and schools together from the 2011-2012 to 2015-2016 school years to study Wave 1. To date we lack the crosswalk necessary to match students and teachers in Wave 2, but we expect to receive the file within the week. Therefore, we restrict some of the analysis to Wave 1.

Table 1 shows descriptive statistics for ACE schools, near ACE schools, Tier 1 schools, and non-Tier 1 schools before the policy. ACE and Near ACE schools are lower performing and have a substantially higher percentage of African American students and much lower percentages of Hispanic and LEP students than other Dallas ISD schools. There are smaller gaps between ACE and near-ACE schools in achievement and the black enrollment share; the achievement deficits for ACE schools approach 0.15 standard deviations in both math and reading. ACE teachers also have substantially lower math and reading value-added, and average teacher experience is twice as high in near-ACE schools as in ACE schools.

3. Institutional Background:

Dallas is a large urban school district in north Texas comprised of roughly 160,000 students and 230 schools. Several years prior to the adoption of ACE, DISD undertook a process that fundamentally alters the evaluation and compensation of district educators. Prior to describing the ACE reform, we outline some of the main features of the principal and teacher evaluation and compensation reforms.

2.a. TEI and PEI

The district introduced the Principal Excellence Initiative (PEI) during the 2012-2013 academic year and the Teacher Excellence Initiative (TEI) during the 2014-2015 academic year. Though they differ in many details, the two reforms share a similar structure. Each contains an

achievement component based on standardized assessments, a performance component based largely on supervisor observations and judgements, and a survey component based on feedback from students or families. There are target distributions for the overall distributions into ratings categories and the components of TEI and PEI to limit evaluation inflation and retain control over the personnel budget. The current-year composite evaluation score determines the evaluation rating category, and the two-year average score determines the salary bin (referred to as effectiveness rating) with some qualifications. PEI and TEI delineate in great deal the requirements of the initiatives, points awarded for each criteria, and educator responsibilities for carrying them out. We now highlight some main features of each and relevant implementation details.

The PEI evaluation component is determined by both overall achievement and success at reducing the achievement gap. The district developed numerous assessments to measure achievement in subjects and grades lacking a state-standardized test. Initially three separate achievement scores were calculated, and the number of points assigned was the highest from three alternatives: Status (percentage of tests with scores at a specified standard); a value-added measure; and achievement score relative to the scores of a designated peer group of schools based on prior achievement. Subsequently, the status alternative was dropped. The number of achievement points depends on success at reducing achievement gaps by race and ethnicity. This codifies the objective of equity and support for students in demographic groups that have lower average achievement in the district and state.

PEI places substantial weight on effectiveness as an instructional leader. Almost 20 percent of the 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. The congruence component of the evaluation is designed to mitigate the tendency to inflate subject evaluations and to deter arbitrary judgements of teachers based on factors other than the quality of teaching. Unlike the case for TEI, attendance and enrollment also contribute to the performance score for principals.

TEI has a similar structure as PEI, but naturally there are important differences between teacher and principal evaluation systems. Supervisor classroom observations constitute the primary source of evidence for the performance score. TEI specifies ten, 10- to 15-minute spot observations of each teacher and one 45-minute extended observation per year by the designated supervisor, typically the principal or assistant principal. The supervisor is required to provide written feedback following all observations and conference with the teacher following the extended observation.

Student Perception, is obtained through a student survey conducted in the second week of April. Most students in grades 3-12 complete two surveys, one online and one in paper. Results from the survey will be summarized by a statistic for teachers with sufficient number of student responses. 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.

The achievement score is based on the results for a teacher's students (when available) and the outcomes for the entire school. This is intended to foster collaboration and a common mission, but it likely also handicaps teachers who work in schools with a high fraction of ineffective educators. This may exacerbate difficulties of attracting and retaining teachers in low-performing schools, the problem ACE was designed to remedy.

Differences in grade, subject, and role lead to cases in which a teacher may not have a measure of achievement for her own students or student survey results. TEI assigns different weights to the three evaluation components depending upon the availability of these two sources of information.

Though teacher pay is determined largely by the evaluation score, there are other consideration including, teachers must apply to and pass a distinguished teacher review process in order to place into the highest evaluation and pay categories; experience and education determine the salary for teachers new to Dallas ISD; teachers with fewer than three years of experience are limited to the maximum rating and compensation they can receive; and district teachers who taught in Dallas ISD prior to the TEI reform cannot have their nominal pay lowered below their pre-reform level.

2.b. Accelerating Campus Excellence

In academic year 2015-2016, one year after TEI adoption, Dallas ISD implemented the Accelerating Campus Excellence (ACE) program that pays educators a transfer bonus of $2,000 per year and a salary stipend to work in designated low-performing schools. ACE principals receive an additional $13,000 salary stipend to lead an ACE school, and the district worked to put leaders in place that they expected to be successful. In contrast to the case for principals, the teacher pay premium depends on the rating in the previous year, where the lowest rated teachers get no premium and "distinguished" teachers get $10,000 per year. Based on the target distribution of ratings, approximately 20% of Dallas ISD teachers qualify for the $10,000 pay premium by having passed distinguished teacher review, 40% of teachers qualify for an $8,000 pay premium by obtaining a proficient rating, and 37% qualify for a $6,000 premium by receiving a progressive rating due either to being inexperienced or to failing to reach proficiency.

In the first year of the ACE program, 40 percent of ACE teachers qualified for a $10,000 stipend, 28 percent for an $8,000 stipend and 32 percent for a $6,000 stipend. Out of the 32 percent in the progressing category FILL IN

In addition to providing incentives to transfer to an ACE school, the stipends amplify the TEI pay for performance incentive for teachers in an ACE school by increasing the differential between ratings categories. For example, at a non-ACE school, moving from the level just below distinguished up to the first rung in the distinguished category raises salary by $5,000.[5] At an ACE school, the same rating change raises salary by $7,000.

Dallas identified a total of 25 low-performing schools that they considered for the ACE intervention, but ultimately designated seven as ACE schools in 2015-2016 based on persistent low-achievement. Another six schools were selected for the second wave of ACE in 2017-2018. Over 60 percent of teachers and all seven principals were replaced in the first ACE wave, and the same held for the second wave. IS THIS ACCURATE?? Such turnover would be expected to affect adversely the quality of instruction, as teachers adjust to different schools and in many cases different grades. Consequently, effect of ACE may increase over time as the teaching force stabilizes.

Figure 1 illustrates the dramatic increases in educator quality in Wave 1 ACE schools as measured by PEI and TEI. The lower panel shows that the share of principals rated effective in ACE schools rose from 0 in the year prior to ACE to more than 70 percent in the first year of the program. This share also increased more modestly in near-ACE schools but fell by almost 10 percentage points in the remaining schools that constitute the vast majority in Dallas. This

[5] In the terminology used by the district, this is a movement from proficient 1 to proficient 2.

decline results largely from the entry of many educators who had not previously been principals into the principal position.

The upper panel of Figure 1 illustrates a similar pattern for teachers. The share of teachers in Wave 1 ACE schools that had at least a proficient rating in the 2014-15 academic year rose from 29 percent in 2014-15 to 72 percent in 2015-16; the share declined a small amount in near-ACE schools and only slightly in the rest of the district. Wave 2 schools show a similar pattern, with an almost 50 percent increase in proficient share, far greater than the small increases in the near-ACE and other Dallas ISD schools. Clearly ACE substantially increased the shares of educators rated effective in ACE relative to near-ACE schools, and following a description of the data we investigate whether this upgrading of educators is associated with higher achievement.

4. Estimation of ACE effects

This section describes the empirical approach used to identify the effects of ACE on mathematics and reading achievement and presents estimates results on two waves of ACE schools. The conceptual discussion highlights potential threats to identification, and the results include difference-in-difference estimates and evidence that pertains to these threats.

4.a. Empirical approach

Identification of the effects of ACE on mathematics and reading achievement requires a valid counterfactual, and a difference-in-differences specification that uses the 18 schools designated as near-ACE as the control group seems like a promising approach. This approach requires the satisfaction of the common trends assumption, and we describe primary threats to this assumption and relevant evidence on each. Following this discussion, we present the basic

difference-in-differences results using the near-ACE schools as the comparison group and illustrate the annual achievement changes by school classification: ACE, near-ACE, and the remaining schools. Comparisons of trends prior to the policy period provide information on the validity of the common trends assumption, and achievement changes during the policy period provides suggestive evidence on the importance of channels other than fixed changes in educator composition such as greater stability and teacher growth.

Our primary analysis is based on a simple difference-in-differences design that includes demographic controls in some specifications. This design is facilitated by the fact that DISD identified both ACE schools (that received the intervention) and "near-ACE" schools that were considered for, but did not receive the ACE designation in either 2016 or 2018. Our baseline estimating equation restricts the analysis to just ACE and near-ACE schools. Because there are relatively few treated schools, clustering at the school level has the potential to over or under state standard errors (CITE). In future drafts, we will consider using p-values based on wild-bootstrapping, though recent work argues that this also may provide misleading inference. For now, we cluster standard errors at the school-by-year level.

There are several potential threats to credibly identifying the effect of the ACE program on student and school outcomes. First, as with any difference-in-differences approach, it is possible that the near-ACE schools would have trended differently in the post-period even without program implementation. Based on discussion with administrators, there were no policies targeted towards near-ACE schools that did not also apply to ACE schools. Combined with the fact that the achievement for Wave 1 ACE schools trends similarly to that for near-ACE schools performance prior to the policy, we believe that the available evidence supports the common trends assumption for Wave 1. In contrast, schools assigned to Wave 2 trend far more

negatively in the two years prior to their designation. This raises the possibility that a temporary negative shock precipitated assignment to Wave 2. However, given the extensive information Dallas ISD has on each school, we believe a more likely explanation is that classification resulted from information indicating serious problems in these schools. Nonetheless, the divergent Wave 2 achievement trends raises questions about the interpretation of Wave 2 ACE effects, and we focus attention on Wave 1.

A second threat to identification is the possibility that students (or their parents) respond to ACE designations by transferring their students into or out of the ACE schools. If enrollment changes, this could alter the composition of ACE schools and thereby improve school average test scores without affecting any individual student's outcomes. We investigate this possibility using a falsification where we test for whether there is an "effect" of the ACE school designation on the pre-policy scores and demographics of students in ACE schools.

A third threat to identification is the possibility that the near-ACE schools are also affected by the policy. For example, ACE may adversely affect the quality of educators in near-ACE schools through the loss of teachers to ACE schools or greater difficulty attracting and retaining effective teachers and principals. This concern is mitigated by the fact that Wave 1 ACE schools represent only 7 out of 234 Dallas schools and thus spillover effects in the teacher market are likely to be relatively small. We show that in practice, fewer than 2% of near-ACE teachers moved to ACE schools. In fact the modal number of teachers lost to Wave 1 ACE schools is one among all schools in the district.

An issue separate from identification considerations but relevant for policy interpretation is the effect of ACE on the distribution of teacher quality. Although ACE schools are too small a share to substantially affect the aggregate teacher market, it remains the case that the policy

caused many effective teachers and principals to move to ACE schools. If teachers who left ACE schools following the adoption of the policy are simply redistributed among other schools, it is possible that the policy has no aggregate effect on student achievement in Dallas.[6] If this is not the case, then the effect depends upon the quality difference between teachers who left an ACE school and exited the district and entrants into non-ACE schools. Importantly, the long equilibrium of a system that pays substantial stipends for educators in low-performing schools is likely to differ substantially the short-term changes following policy adoption. We intend to explore these issues in future work.

4.b. Baseline results

Our primary goal is to assess how the ACE designation affects students' outcomes. As a first step towards this goal, we plot math and reading scores in ACE wave 1, ACE wave 2, near-ACE and other Dallas schools from 2011-2012 to 2017-2018. Academic years are indicated based on the spring, so the first ACE wave is treated in 2016 and the second wave is treated in 2018. Figure 1 shows that both waves of ACE and the near ACE schools were on downward trajectories in the four years before the first wave of ACE (2012 to 2015). Schools that were selected for ACE during the first wave were lower performing than near-ACE schools, but they were on very similar trends. When the first wave of ACE adoption occurs in 2016, average math achievement in ACE schools increases by approximately 0.3 standard deviations, surpassing near-ACE schools and closing the gap between ACE and the other Dallas ISD schools by over 50%. Near-ACE schools also make improvements in 2016, though these are far smaller than those made in the ACE schools.

[6] Even without mean improvement, reallocation may be a desirable outcome as it reduces inequality with no apparent efficiency loss.

Schools that were selected for ACE in 2018 (wave 2) were similar to near-ACE schools from 2012 to 2015 but experienced sharp declines in math and reading achievement between 2015 and 2017. This is consistent with the district selecting schools that were trending downward for the second wave of ACE treatment. Schools in the second wave of ACE increased math scores by approximately 0.4 standard deviations in 2018.

In addition to previewing our main results, Figure 1 highlights the magnitude of the baseline deficiency of ACE schools. In the year just before adoption, both waves of ACE schools were approximately 0.5 standard deviations below other Dallas ISD schools. By 2018, both waves of ACE schools were within 0.1 standard deviations of other Dallas ISD schools, while near-ACE schools still lagged by approximately 0.3 standard deviations.

In Figure 2, we show the analogous results for reading scores. The pattern is similar to that of math scores, but the ACE improvement is less dramatic. As with math scores, wave 1 ACE schools and near-ACE schools trended very similarly up until 2015. In 2016, both wave 1 ACE and near-ACE schools improve, but the improvement is twice as large among ACE schools. ACE wave 2 schools trend downward until 2017 and then improve substantially in 2018.

We view Figure 1 and 2 as indicating a large ACE effect, but it is worth noting that the wave 2 ACE schools trend quite differently than near-ACE schools leading up to ACE adoption in 2018. If these diverging trends continued, wave 2 ACE schools would have fallen even further behind near-ACE schools suggesting that our DID estimates may be downward biased. That said, it is also possible that wave 2 ACE schools would have recovered in the absence of ACE. More generally, the diverge in pre-trends introduces uncertainty into the estimates of program effects on Wave 2 schools. To provide a more credible counterfactual group, in a future

draft we plan to use a synthetic control approach to study the second wave of ACE. For the first wave of ACE, the synthetic control approach is unnecessary since near-ACE schools appear to trend very similarly to wave 1 ACE schools prior to the policy.

Table 2 presents difference-in-difference estimates for the effects of ACE on achievement for Wave 1 and Wave 2 separately. All coefficients are highly significant based on clustering the standard errors by school and year, though we expect the results to remain significant once the standard errors are corrected given the results. Estimated effects on math achievement are larger than those for reading in both periods, equaling 0.24 standard deviations in Wave 1 and 0.30 standard deviations in Wave 2. Reading coefficients from specifications with student demographic controls equal 0.11 in Wave 1 and 0.20 in Wave 2. The fact that the estimates are largely insensitive to the inclusion of the controls provides additional support for the difference-in-differences approach.

4.c. Specification checks

The difference-in-differences approach relies on the assumptions that changes in outcomes in the control group provides a valid counterfactual for changes in outcomes in the treatment group. As discussed above, the presence of parallel pre-trends in Figures 1 and 2 for Wave 1 but not for Wave 2 schools leads us to focus on our Wave 1 results for now, though the differential trends between near-ACE and Wave 2 schools possibly driven by temporary negative shocks to Wave 2 schools are unlikely to explain the large improvement at Wave 2 schools following ACE adoption. Importantly, the parallel trends in the pre-period do not bear on other possible confounding factors including changes in student composition in response to ACE and spillover effects on the composition of educators in near-ACE schools.

Though common trends in the pre-period for wave 1 is reassuring, there are several threats to internal validity that the common trend test has no power to detect. First, the ACE designation has the potential to alter parents' perceptions of school quality such that ACE schools attract a different type of student post-policy. If this occurs, we would still correctly identify the causal effect of ACE at the school level, but the mechanism would be changes in student composition as opposed to improvement in school quality. Table 3 reports changes in demographic and program characteristics for Wave 1 ACE schools, near-ACE schools and remaining Dallas ISD schools following ACE implementation and reveals only very small changes in any of the characteristics. Share low income fell by 1 percent in ACE and near-ACE schools, and share black fell by 3 percent in near-ACE schools while remaining constant in ACE schools. Thus there is little or no evidence that student composition drives or even contributes to the achievement increases in ACE schools.

A second threat to internal validity is the possibility that the labor market for teachers is sufficiently thin so that the incentive for high quality teachers to move to ACE schools and departure of many who previously taught in an ACE school adversely affected the quality of instruction in near-ACE schools. This would introduce spillover effects on near-ACE schools that inflate the estimates. Table 4 presents the teacher transition matrix for the years surrounding the two ACE waves, and the quite small numbers of teachers who transition between ACE and near-ACE schools mitigate concerns about this potential problem. For Wave 1, only 13 out of 633 near-ACE teachers in 2016 just arrived from an ACE school, and only 8 teachers transitioned from a near-ACE to an ACE school during this time. The numbers are somewhat higher for Wave 2 but still small: 13 out of 545 near-ACE teachers just arrived from an ACE school and 16 ACE teachers came from a near-ACE school. Moreover, ACE schools hired a total

of 156 teachers in their first year of implementation, less than 2% of the total Dallas teacher labor market. We cannot rule out the possibility that ACE schools hired some high quality teachers who otherwise would have ended up at near-ACE schools, but the negligible direct movement suggests that ACE is unlikely to substantially affect the stock of teachers at near-ACE schools.

A finding that much of the ACE effect comes from a decline in the quality of instruction at other schools would indicate the presence of substantial spillovers to non-ACE schools. Although the TEI and PEI reforms affected all schools and thus achievement trends in the district as a whole do not reflect the aggregate effects of ACE, it is informative to compare achievement trends in Dallas ISD with those in the rest of Texas. Figures 3 and 4 taken from Hanushek et al (2019) illustrate trends over time in the differences between Dallas ISD achievement at the 25th, 50th, and 75th percentiles and the corresponding achievement percentiles for all other Texas public schools, where achievement is standardized for the state as a whole. The figures cover the period from 2005 to 2017, but the discontinuous jumps in 2012, the year of a change in the test regime, raise questions about prior years. In all three percentiles in both the mathematics and reading distributions achievement increases in Dallas ISD relative to the rest of Texas by roughly 0.15 standard deviations in math and 0.1 standard deviations in reading between 2015 and 2017, the period of Wave1. Given that ACE comprised only seven schools in Wave 1, the figures indicate that improvement in Dallas ISD was widespread and that ACE achievement increased substantially relative to the rest of the state.

5. Contributions of various channels

The introduction of additional pay linked with teacher effectiveness for those working in ACE schools can affect achievement through a number of channels. We consider 5 potential mechanisms: 1) fixed differences in teacher effectiveness resulting from the influx of effective teachers and departure of less effective teachers. We expect this to be the primary channel through which ACE raises the quality of instruction, but it is not necessarily the case; 2) improvement in the quality of school leadership; 3) improvements in teacher peer quality; 4) strengthened incentives to improve practice and raise achievement due to the positive association between the size of the ACE salary stipend and the evaluation rating; and 5) adverse effects of disruption following the implementation of ACE. Evidence shows that a transition to a new school or being assigned a new grade lowers the quality of instruction, and a disproportionate number of teachers were in this position in the initial years of ACE implementation.[7] This may offset some of the positive effects of other channels, and lower turbulence would be associated with a higher program benefit.

We are not seeking to decompose the total effect into all its component parts, as some of the factors discussed above are conceptually intertwined and there are no natural control groups for each mechanism. For example, an effective principal may hire more effective teachers or help existing teachers improve. Relatedly, the various components of ACE may not be additively separable (e.g. the effect of disruption may depend on principal quality). However, the rich information produced by PEI and TEI includes evaluations of specific practices such as a principal's effectiveness at supporting and evaluating teaching that can be related to teacher growth as measured by subjective evaluations and statistical estimates of value added. We will also use the longitudinal data to estimate the effects of school and grade switching on the quality

[7] Hanushek et al (2017) and Ranfield et al (2013) show that the share of teachers new to the school is negatively related to achievement, and Ost (2014) finds that teacher grade switching adversely affects the quality of instruction.

of instruction and apply the estimates to the transition rates in different years and school types to learn more about the likely contribution of the transitions. As a check on our estimates, we will compare them to estimates of transition and teacher peer effects in the literature and use existing estimates as an alternative approach to simulating the contributions of some of the channels. However, this draft focuses on changes in teacher effectiveness as measured by TEI.

Given the importance of changes in the composition of teachers, we will also attempt to decompose the ACE effects into the contributions of teacher composition and all other factors. Simply adding teacher fixed effects to the baseline differences in differences model and measuring the change in the treatment effect provides an estimate of the role of improvements in fixed teacher quality, but this approach requires the satisfaction of strong assumptions and will be pursued once the requisite links between teachers and students for all years are available. The current draft does illustrate the sensitivity of the ACE effect to the inclusion of controls for teacher experience and teacher fixed effects.

Table 5 shows average teacher performance scores and the share of students who respond that the teacher is very supportive of learning in the year first year a school is in the ACE program for both waves and the year immediately prior for three sets of teachers: 1) teachers in ACE schools in the prior year who leave prior to the initial ACE year; teachers not in an ACE school in the prior year who enter an ACE school in the first year of the program; and teachers who remain in ACE following the start of the program. Because a large fraction of teachers who leave ACE schools prior to the start of the program exit the district, we do not report the scores of those who remain during the first program year. Nonetheless, the table presents a consistent pattern of large differences between the averages for teachers who leave a school about to begin the ACE program and the new entrants and relatively stable performance scores and survey

responses regardless of whether a teacher is entering an ACE school or remaining in an ACE school during the transformation. The average performance-score difference between ACE entrants and those who exit the school after the prior year equals 0.73 standard deviations for Wave 1 (0.48-0.25) and 1.2 standard deviations for Wave 2. The share of students who strongly agree that the teacher supports learning is also higher for entrants.

In contrast to the large differences between entrants and leavers, changes within teachers tend to be small regardless of whether then move into or remain in an ACE school. Despite the fact that teachers entering an ACE school have a different principal and student body, entrants experience an average decline of less than 0.1 standard deviations in Wave 1 and an average increase of less than half that large in Wave 2. This stability does not support the concern that working in a low-performing school could make it more difficult to obtain a higher rating, though it is important to recognize that teachers enter a transformed school with an effective principal and highly rated group of peer teachers. Those who remain in an ACE school during the ACE implementation also have a different principal and higher quality peers. Nevertheless, there is virtually change in their average performance score in Wave 1, though there is an increase of roughly 0.3 standard deviations for Wave 2. Overall, the pattern in Table suggests a primary role for compositional change and a lesser role for improvements in the quality of instruction.

Finally, Table 6 illustrates the sensitivity of ACE treatment effects to the inclusion of controls for teacher experience and ultimately teacher fixed effects. This is illustrative, as it includes only one year of achievement under ACE and requires additional steps to identify the contribution of fixed differences in teacher quality to the ACE effect. With teacher fixed effects included, the ACE coefficient is identified by two types of teachers. First, there are retained

teachers who were at ACE schools prior to ACE implementation and stayed on afterwards. Second, there are transitioning teachers who began at non-ACE schools and moved to ACE. Importantly, the change in performance among transitioning teachers is based on an across-school comparison of their performance at the ACE school compared to their performance at their more advantaged previous school. Furthermore, these teachers experience substantial disruption given that they are changing schools and are likely to change course assignments as well.

The estimates in Table 6 are quite similar for math and reading despite being less precise than the baseline difference in difference estimates based on more than one year in the post-period. Inclusion of the teacher experience controls has virtually no effect on the estimates, while inclusion of teacher fixed effects eliminates the ACE effect. This pattern is consistent with changes in teacher composition serving as the main channel through which ACE operates, but as stated above these estimates are only suggestive. Being able to match teachers with students for the later years along with the estimation of additional specifications provide a richer picture of the contributions of different mechanisms to the dramatic improvement of ACE schools.

6. Summary and Policy Considerations

We document the remarkable success of a program implemented by Dallas ISD that embeds increased compensation for working in a low-performing school into a comprehensive evaluation and pay structure in which educator effectiveness serves as the primary determinant of the level of compensation. The evaluation systems generate information that can be used to identify effective teachers and support the growth of all teachers, and the district salary structure establishes the practice of differentiated pay by the quality of instruction and leadership. ACE

amplifies the salary differences by level of effectiveness provides the incentives for educators to take the risk and transition to a low-performing school.

Several factors make our empirical findings credible. First, there is a well-defined control group that was considered for, but did not ultimately receive the ACE designation. This group of schools was on a very similar trend to Wave 1 ACE schools in the years prior to the ACE intervention. Second, we see little change in student composition at ACE schools relative to near-ACE schools suggesting that altered student composition does not explain the test score improvement. Finally, there are two waves of ACE and the timing of test score improvement perfectly lines up with the timing of ACE implementation.

The existing data limitations that will be addressed imminently limit the investigation of the channels through which the personnel reform raises school quality and achievement. Nonetheless, descriptive evidence is consistent with the notion that changes in the composition of teachers accounts for a substantial share of the ACE effect, while teacher improvement due to better support from the principal or peers or enhanced incentives seems modest. Yet this result is preliminary and may change with additional data and analysis.

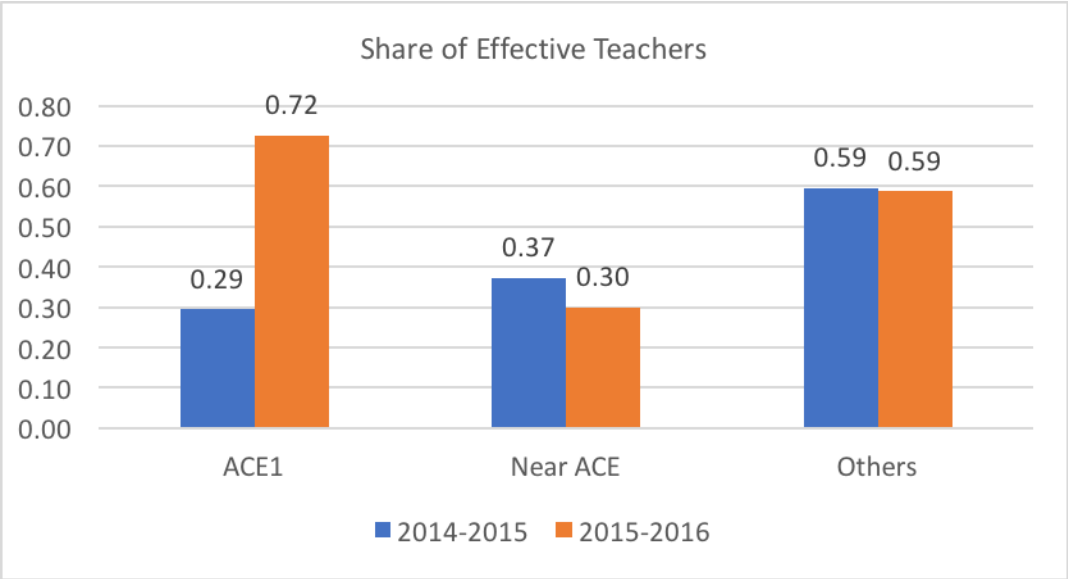
The results demonstrate the potential to elevate the quality of instruction in even the lowest-performing urban schools if the district is willing to devote adequate resources and has the capacity to identify effective educators. An important policy question concerns the optimal structure for such a targeted compensation program, and this depends crucially on the dynamic forces that lead some schools to be very low achieving. One approach would be to target the highest poverty or most racially isolated schools, but the evidence suggests that the situation is much more complicated than this. Figure 5 plots school-average achievement taken over two different two year periods against predicted achievement based on student demographic

characteristics for the top and bottom quartiles of the distribution based on predicted achievement. The figure shows that there is substantial variation around the regression line. In fact predicted achievement explains roughly half the variation in actual achievement. Of course, this is a limited set of variables, and random shocks and test-measurement error reduce the R^2. Yet these are large samples of tested students. Regardless, even averaged over a two-year period low-achievement is difficult to predict. This is consistent with the large ACE effects, as a group of effective educators can be successful even in high poverty schools.

Figure 6 plots the difference between actual and predicted achievement in the later period against the difference in the early period, and there is substantial movement between periods, particularly in the lowest predicted achievement quartile. Although this work is quite preliminary, the larger fluctuations among lower predicted achievement schools is consistent with wider swings in quality. Whether this is indicative of greater susceptibility to negative shocks is something we intend to investigate. We are particularly interested in the possibility that such schools have a difficult time escaping from persistent poor performance and are more prone to rapid deterioration such as that experienced by ACE 2 schools. The extensive information on educator effectiveness produced by TEI and PEI can facilitate a more in-depth understanding of the dynamics of school quality fluctuations and the forces that contribute to concentrations of ineffective educators in some schools. This would provide information that could contribute to the design of compensation policies designed to foster equity and successful interventions in cases of deteriorating or persistently low school quality.

Bibliography (incomplete)

Figure 1. Shares of principals and teachers rated effective in the first year of ACE and the previous year, by school type



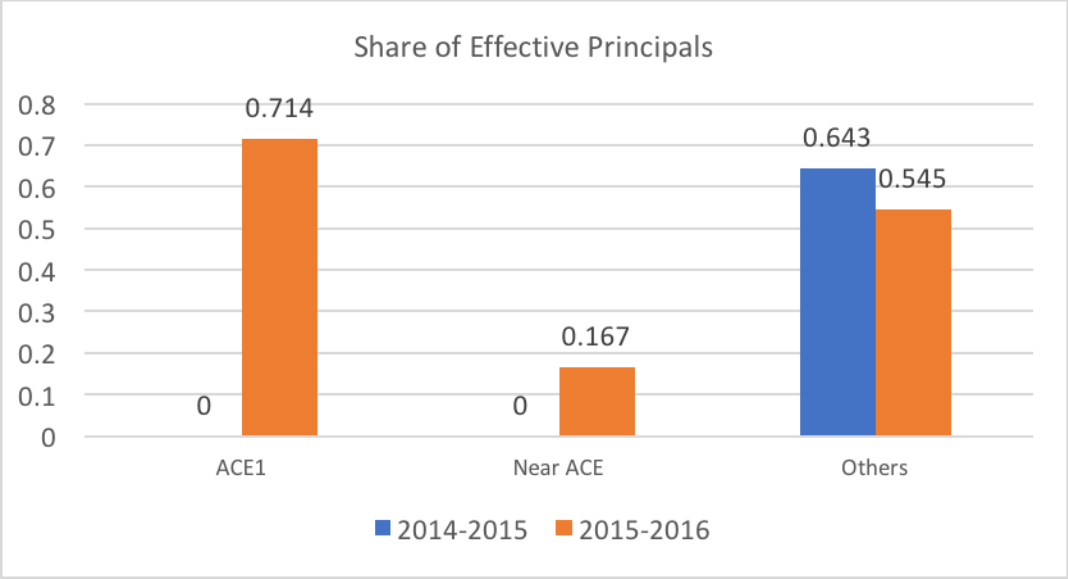
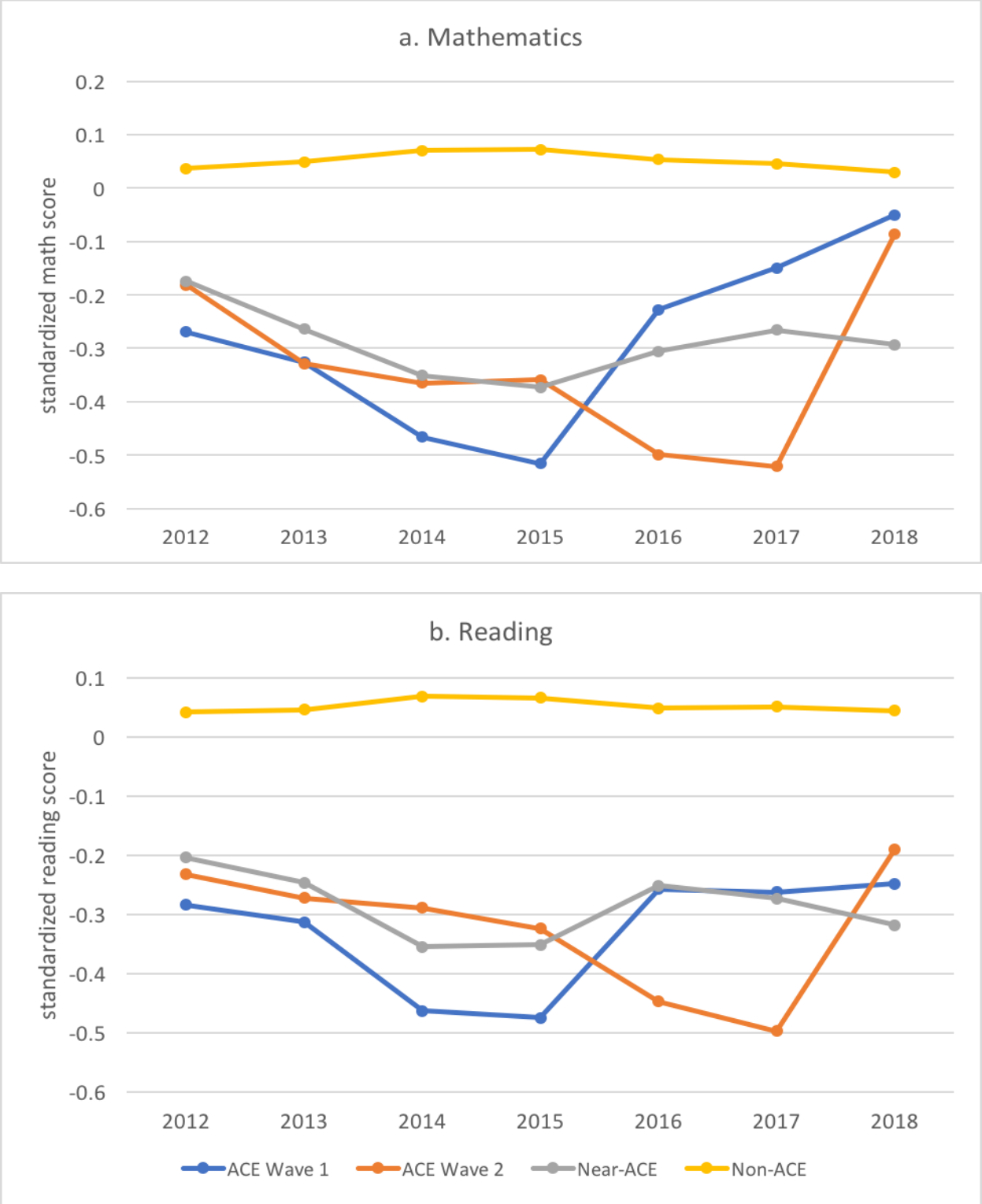


Figure 2. Trends in mathematics and reading achievement, by school category: 2012-2018



nonace refers to all schools other than ACE and near-ACE

Figure 3. Differences in Math Achievement at the 25th, 50th, and 75th percentiles between Dallas ISD and other Texas public schools, 2005 to 2017

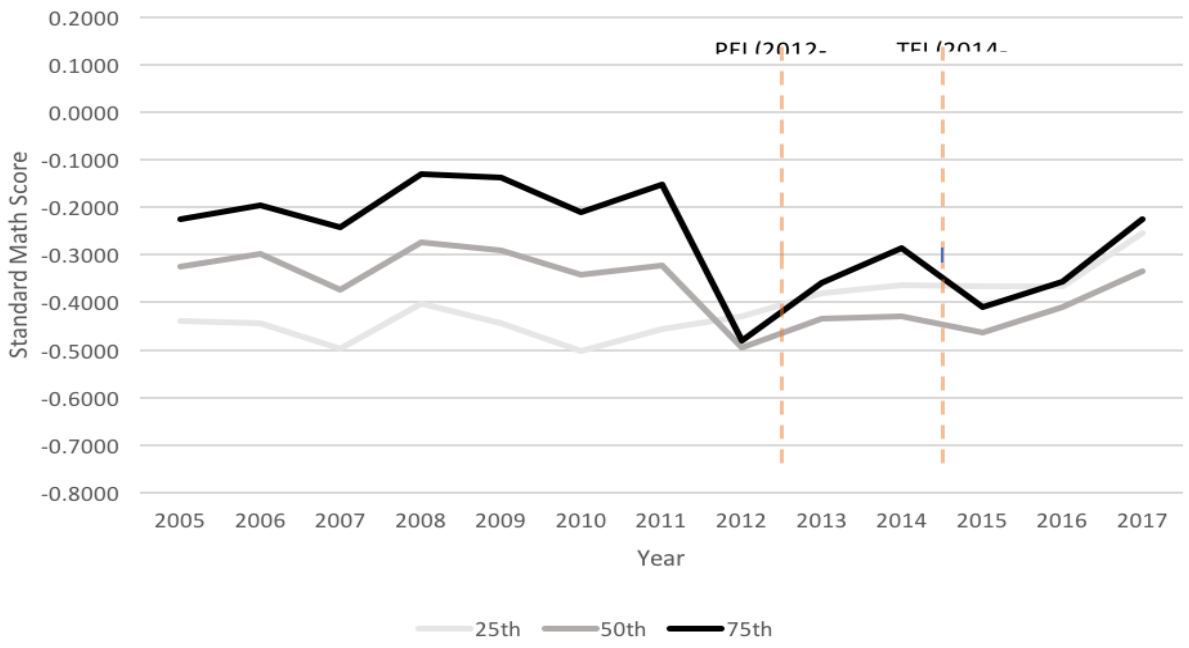


Figure 4. Differences in Reading Achievement at the 25th, 50th, and 75th percentiles between Dallas ISD and other Texas public schools, 2005 to 2017

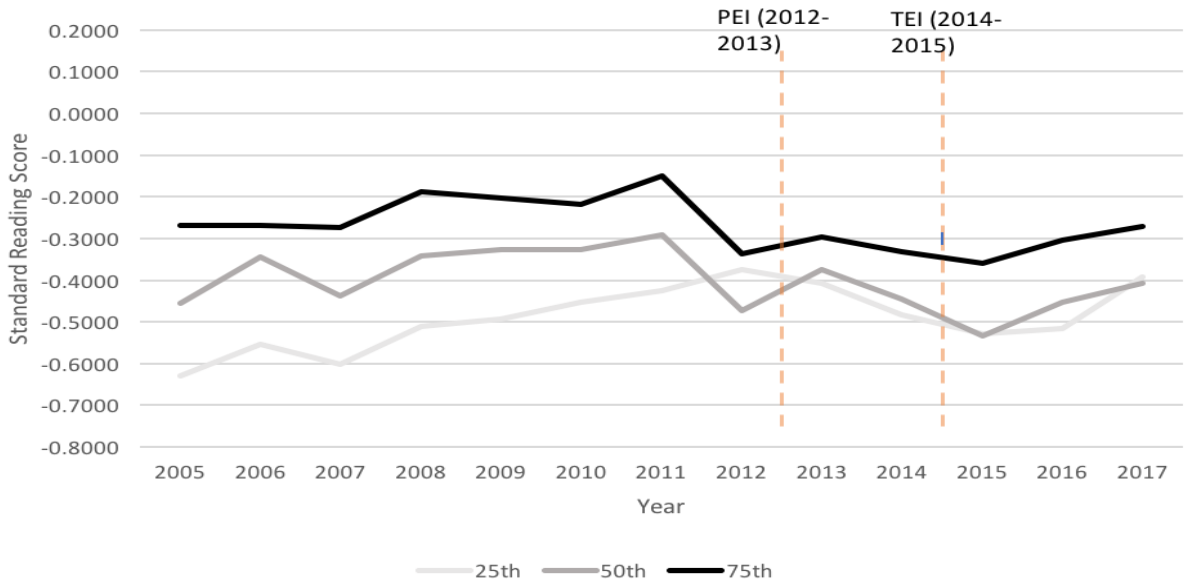
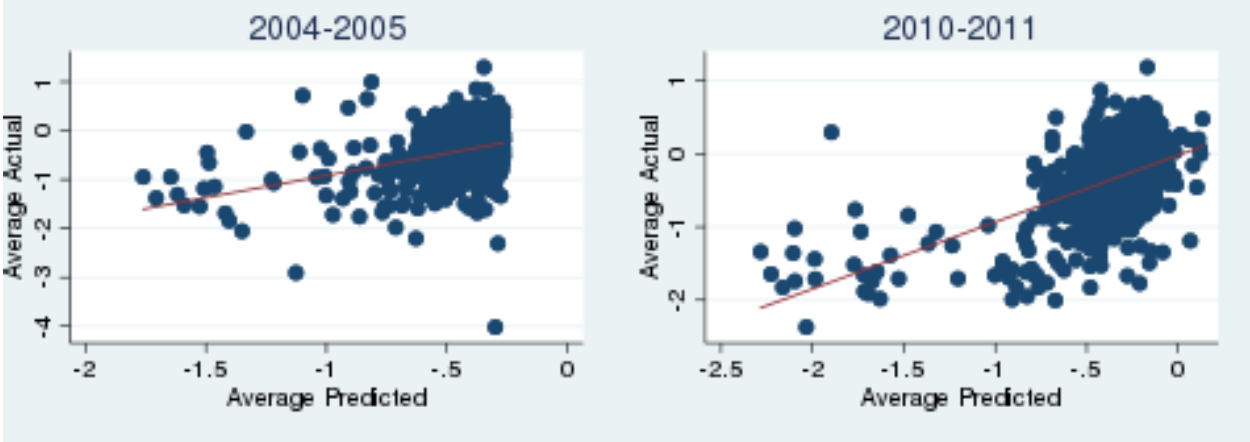


Figure 5. School-average achievement over a two year period plotted against predicted achievement based on school demographics for schools in the lowest and highest predicted achievement quartiles: 2004/05 and 2010/11

Lowest predicted achievement quartile



Highest predicted achievement quartile

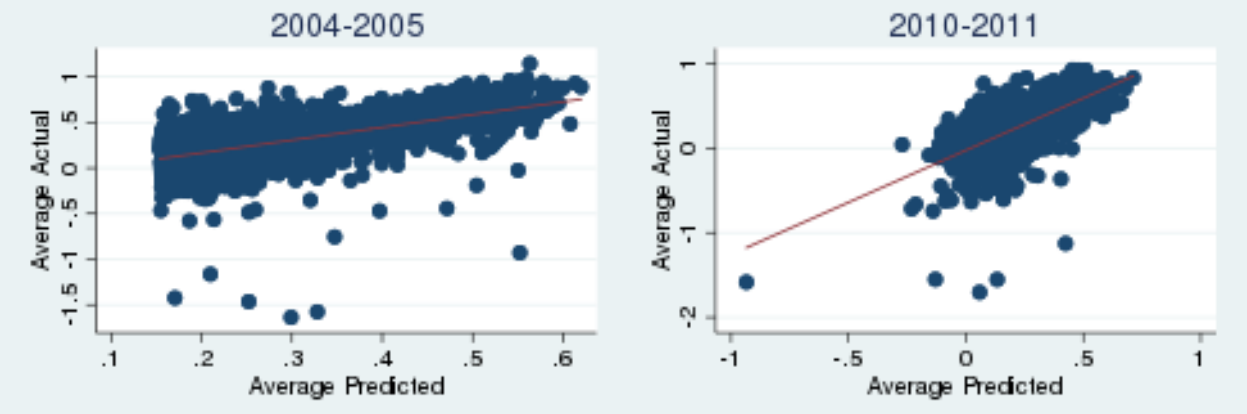
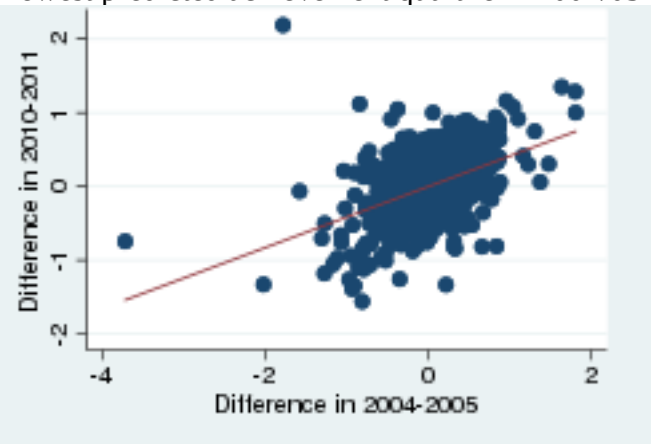


Figure 6. Relationship between the difference between actual and predicted achievement in the two periods, by poverty quartile

Lowest predicted achievement quartile in 2004/05



Highest predicted achievement quartile in 2004/05

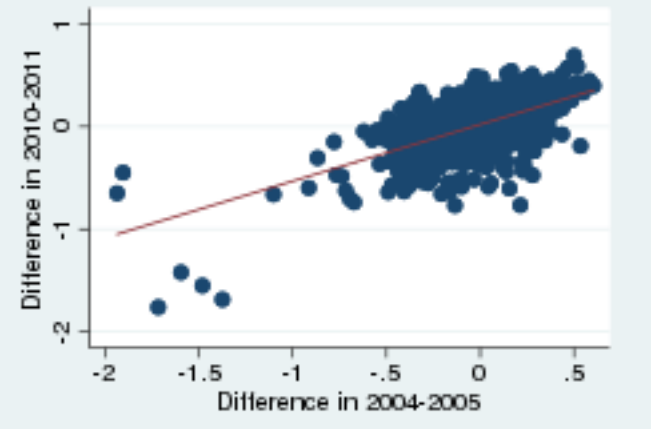


Table 1. Student demographic and program characteristics and teacher evaluations and characteristics in 2014-2015, by school type

school type	ACE Wave 1	ACE Wave 2	Near ACE	Other Schools
Student variables				
Math Score	-0.516	-0.359	-0.373	0.0722
Reading Score	-0.495	-0.364	-0.368	0.0329
Free or Red. Price Lunch	0.965	0.981	0.967	0.926
White	0.00684	0.0123	0.0106	0.0458
Afr. American	0.594	0.474	0.461	0.180
Hispanic	0.384	0.485	0.508	0.744
Native American	0.00190	0.00130	0.00252	0.00377
Asian	0.000760	0.00973	0.00162	0.0129
Special Educ.	0.125	0.0857	0.0955	0.0831
LEP	0.283	0.385	0.392	0.553
Male	0.543	0.517	0.523	0.516
Teacher variables				
student perception score	0.657	0.798	0.667	0.706
standardized observation score	-0.290	-0.233	-0.191	0.0775
reading value added	-0.155	0.0317	-0.0288	0.0340
math value added	-0.0888	-0.0329	0.00470	0.0783
years of experience	4.875	8.688	9.811	10.77

Table 2. Difference in Differences Estimates of ACE Effects on Math and Reading Achievement, by Wave of ACE

	ACE Wave 1				ACE Wave 2			
	Math		Reading		Math		Reading	
ACEXPost	0.250***	0.246***	0.120**	0.114*	0.284***	0.310***	0.182*	0.203**
	(0.0754)	(0.0666)	(0.0567)	(0.0575)	(0.107)	(0.0923)	(0.100)	(0.0819)
ACE	-0.109***	-0.0634	-0.0991***	-0.0654*	-0.0812	-0.0874**	-0.0583*	-0.0618**
	(0.0408)	(0.0387)	(0.0305)	(0.0335)	(0.0532)	(0.0408)	(0.0334)	(0.0281)
Post	0.000704	-0.00752	0.00895	0.00101	-0.00463	-0.0201	-0.0389	-0.0563
	(0.0377)	(0.0358)	(0.0302)	(0.0312)	(0.0653)	(0.0592)	(0.0451)	(0.0456)
Controls	N	Y	N	Y	N	Y	N	Y
Observations	54,751	54,751	55,267	55,267	48,639	48,639	48,984	48,984

Notes: Robust standard errors clustered by school-year are in parentheses; * p<0.1; ** p<0.05; *** p<0.01;
The sample is limited to ACE and near-ACE schools. Controls include indicators for student race-ethnicity, gender, special education status, free or reduced price lunch eligibility, and classification as limited English proficient.

Table 3. Changes in selected student demographic and program characteristics following the implementation of ACE

Comparison periods	2016 to 2018 minus 2012 to 2015		2018 minus 2017	
	ACE Wave 1	near-ACE	ACE Wave 2	Near-ACE
Share low income	0.01	-0.01	0.04	-0.02
Share white	0.00	0.00	0.00	0.00
Share black	0.00	-0.03	0.03	0.00
Share Hispanic	0.01	0.03	-0.03	0.00
Special special education	-0.02	-0.01	0.01	0.00
Share limited English proficient (LEP)	0.03	0.04	-0.02	0.01
Observations	3,424	7,890	1,596	4,579

Table 4. Teacher transitions by origin and destination school type and year

Origin	Destination			
	ACE	near-ACE	other Dallas ISD schools	Out of district
2015 to 2016				
Wave 1 ACE	58	13	98	121
near-ACE	8	407	46	190
other Dallas ISD schools	156	33	7,212	1,903
Out of district	66	180	1,944	n.a.
2017 to 2018				
Wave 2 ACE	18	13	95	67
near-ACE	16	404	42	112
other Dallas ISD schools	126	24	7,256	1,695
Out of district	39	94	1,507	n.a.

Table 5. Average performance score and student perception of teachers in Wave 1 and Wave 2 ACE Schools, by timing of entry and exit

teacher transition status	exit school immediately prior to ACE implementation		enter ACE school following program implementation		remain in ACE school following program implementation	
	performance score	share of students who strongly agree that the teacher supports learning	performance score	share of students who strongly agree that the teacher supports learning	performance score	share of students who strongly agree that the teacher supports learning
Wave1						
Before ACE	-0.25	0.72	0.48	0.83	0.14	0.78
During ACE	n.a.	n.a.	0.39	0.82	0.13	0.72
Wave 2						
Before ACE	-0.93	0.64	0.27	0.80	-0.30	0.79
During ACE	n.a.	n.a.	0.29	0.84	0.03	0.76

Table 6. Difference in Differences Estimates of Wave 1 ACE Effects on Math and Reading Achievement based on only the first year of ACE, by controls for teacher effectiveness

	Math			Reading		
ACEXPost	0.213**	0.214**	-0.0106	0.0974	0.0984	0.00789
	(0.0923)	(0.0911)	(0.116)	(0.0734)	(0.0733)	(0.0568)
Observations	37,475	37,475	37,475	37,959	37,959	37,959
Full set of teacher experience dummies	N	Y	N	N	Y	N
Teacher fixed effects	N	N	Y	N	N	Y

Note: See notes to Table 2.