# Attracting and Retaining Highly Effective Educators in Hard-to-Staff Schools

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

June 2025

***Abstract***

*Attracting and retaining effective teachers in high-poverty schools remains a persistent challenge. We evaluate the impact of an initiative that offered substantial performance-based compensation to highly effective educators in the lowest-achieving schools. We find that the program led to dramatic improvements in achievement, nearly closing the gap between the targeted schools and the district average. When the program was later reversed for some campuses, effective teachers departed at high rates, resulting in an immediate decline in student achievement. The results underscore the potential for effectiveness-based compensating differentials to attract and retain effective teachers in disadvantaged schools.*

---

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

*Chicago Public Schools; **Ball State University; ***Stanford University, University of Texas at Dallas, and NBER; ****University of Illinois at Chicago; *****University of Illinois at Chicago, University of Texas at Dallas, and NBER

## 1. *Introduction*

Elevating achievement in the lowest-performing public schools is among the most pressing challenges for education policy, but to date most efforts to turn around low-achieving schools have had limited success. Though teacher quality is known to be critical for student success (Chetty, Friedman Rockoff 2014; Rivkin, Hanushek and Kain 2005), attracting high-quality teachers to high-poverty schools remains a persistent challenge. The fundamental economic issue is that teachers prefer working at more advantaged schools (Bonhomme et al. 2016) and districts have been reluctant to provide compensating differentials for working at disadvantaged schools. Moreover, most districts lack detailed information on educator performance and strong administrator incentives to hire and retain the most effective teachers, two factors highlighted by Bates et al. (2025) as key to leveraging compensating pay to raise the quality of instruction. Consistent with the notion that policies that expand the supply of teachers will have limited benefits if administrators lack the ability and incentive to hire the best candidates, Clotfelter, Ladd, and Vigdor (2011) find that paying a premium at high-poverty schools successfully improves teacher retention but had little effect on teacher quality. In this paper, we provide the first empirical evidence on a policy that combines a generous compensating differential for working at low-achieving schools, with the information and incentives for principals to hire the best applicants.

In 2016, the Dallas Independent School District (DISD) implemented the Accelerating Campus Excellence (ACE) program to turn around its most disadvantaged schools.[1] Though multidimensional, a central component of the ACE intervention is providing large effectiveness-based compensating differentials to attract and retain the most effective educators.[2] This program built on district-wide personnel reforms that required rigorous educator evaluations and replaced salary schedules based on experience and post-graduate education with compensation determined by the evaluations. Moreover, educator effectiveness in the prior year determined the magnitude of the compensatory pay for working in the most disadvantaged schools, meaning that the Dallas

[1] We use the spring year to reference academic years, e.g., 2019 for the 2018-2019 year. The ACE program included elementary and middle schools, and we focus on elementary schools to enable analyses of immediate effects and effects on achievement at the next schooling level. Beginning in 8th grade, math testing varies by the grade in which a student takes algebra, complicating the analysis of middle school effects on high school achievement.

[2] Educators include teachers, counselors, coaches and school leadership. To simplify exposition, we sometimes refer to just "teachers", but it is important to keep in mind that a broad group of ACE educators receive the stipends.

ISD intervention met the conditions laid out in Bates et al, 2025. To expedite the transformation of the teacher workforce at ACE schools, incumbent educators at the time of implementation had to undergo a rigorous screening process to retain a position and in practice, less than 20 percent of existing teachers were retained. We find that this intervention substantially improves teacher quality and test scores, raising performance of the lowest achieving schools to nearly the district-wide average.

This path-breaking intervention emerges directly from consideration of well-known labor-market fundamentals combined with extensive evidence on the variation in teacher effectiveness among schools and districts.[3] Districts have not used large salary inducements to staff entire schools with highly effective teachers largely because of the lack of comprehensive performance measures, union resistance to differentiated pay, and an unwillingness to increase compensation enough to offset labor-market disadvantages of low-achieving, high-poverty schools.[4] Dallas ISD overcame these impediments, enabling it to focus its turnaround efforts on the elevation of educator effectiveness.

We evaluate the ACE intervention using a difference-in-differences design where other low-performing Dallas schools serve as a control group. ACE schools were specifically selected using test scores two years before implementation. To mimic this selection process, we form a group of control schools based on school-level test scores two years prior to ACE implementation.[5] This selection process guards against common negative shocks during the period between ACE school selection and program implementation. Although our control group is less negatively selected than the actual ACE schools, they follow remarkably similar pre-treatment trends both before program selection and in the run up to program implementation.

[3] The issues of rigid salary schedules unrelated to effectiveness have been discussed frequently and over a long period; see Kershaw and McKean (1962) and Hanushek and Rivkin (2004). An extensive literature including Rivkin, Hanushek, and Kain (2005) documents the substantial variation in teacher effectiveness; see also Hanushek and Rivkin (2010, (2012). There is also substantial evidence on the attraction to teachers of both higher pay and higher achieving students (Hanushek, Kain, and Rivkin (2004)).
[4] The closest policy is IMPACT, which pays larger performance stipends to high quality teachers in disadvantaged Washington DC schools (Dee and Wyckoff (2015)). Unlike ACE, IMPACT did not include the transformational element of removing a majority of educators in the first year of the program.
[5] Forming the control group based on t-2 test scores has a similar logic to estimating synthetic difference-in-differences, but our approach has two important advantages. First, our selection of controls is based on institutional knowledge of the actual selection process and is based on a single pre-policy year, limiting concerns of overfitting. Second, because we only use t-2 test scores to identify the control group, we are able to provide more meaningful evidence on pretrends on either side of t-2 than would be possible if we matched on all pre-period outcomes.

Decisions about the exact size and composition of the control group have little influence on the results because the improvements at ACE schools prove so dramatic.

We find that the ACE program leads to large, immediate improvements in academic achievement, and there is suggestive evidence that program impacts continue into middle school ($6^{th}$ grade) for those with two or more years of treatment. The effects on $6^{th}$ grade scores increase as the dosage increases from two to three years in an ACE elementary school; the estimated treatment effects on $6^{th}$ grade scores approach 0.3 standard deviations in reading and exceed 0.4 standard deviations in math. Although the effects are imprecisely estimated, they are consistent with the notion that the multi-measure evaluation protocol identified effective teachers who raised the acquisition of skills that persisted into later grades.

Two subsequent policy changes provide further evidence on the efficacy of the ACE intervention. First, two years after the initial ACE implementation, the district expanded the program, allowing us to evaluate the effect of ACE for a second wave (dubbed ACE 2). We find very similar effects for this second wave of schools. Second, in 2019, the district eliminated most elements of the program for all but one school in the original 2016 ACE cohort because the achievement growth made them no longer eligible for the program. Following the elimination of ACE incentives for the first cohort of ACE schools, there was an exodus of many highly effective teachers and an immediate reversal of much of the achievement gains. This reversal highlights the central importance of the targeted compensating differentials and is further evidence of the efficacy of the ACE program. We show that the teachers who left ACE 1 schools following the removal of stipends specifically sought out the relatively small number of schools that still provided ACE stipends. In the control schools, 9% of exiting teachers moved to a stipend-providing school, whereas 44% of teachers exiting ACE 1 in 2019 moved to a stipend-providing school. This pattern reinforces our conclusion that stipends are important for attracting and retaining highly effective teachers.

We assess whether changes in student composition at ACE schools contributed to the program effects. Because of limited public discussion prior to ACE implementation and because of the broader history of transitory reforms targeting low-achievement schools, student selection based on ACE is unlikely in the year of adoption, and we find no evidence of changes in student characteristics from 2015 to 2016 or from 2017 to 2018 at ACE 1 or ACE 2 schools. We assess the influence of compositional changes in later years by examining changes in student characteristics following ACE adoption and show that results are robust to intent-to-treat specifications defined by school enrollment in the first year of the program.

Though educator stipends constituted roughly 85 percent of the program budget, the ACE intervention incorporated other components including an increase in instructional time, an extended after-school program, a requirement to adopt data-driven instruction, funds for school uniforms and enhanced professional development. We emphasize the central role of educator quality as this was the focus of the program, but we cannot rule out that part of the improvement at ACE schools comes from other program elements. That said, estimates from the literature on the effect of other components suggests that non-educator related components of ACE likely account for less than one quarter of the improvement.[6]

## 2. *Prior evidence on school turnarounds and compensating differentials for educators*

ACE is a school turnaround intervention whose central component is the payment of effectiveness-based compensating differentials to attract and retain highly effective educators in some of the lowest performing schools. This section describes relevant research on turnarounds and compensating pay for educators in turn.

### *2.a. School turnarounds*

Existing research on school turnarounds reports mixed results, with some studies finding null or negative effects (Heissel and Ladd (2018); Dougherty and Weiner (2019); Pham et al. 2025) and other studies finding significant gains from turnarounds (Dee (2012); Papay (2015); Strunk et al. (2016); Schueler, Goodman, and Deming (2017); Zimmer, Henry, and Kho (2017); Fryer (2014)). Though the specifics of each turnaround differ, a common element is that school turnarounds tend to be multifaceted. For example, Schueler, Goodman, and Deming (2017)

[6] We discuss the literature on each non-personnel component in the conceptual framework section. One particularly useful reference point is Fryer (2014), which evaluates the effect of a bundle of best practices including increased instruction time, daily high dosage tutoring of three students per adult, data-driven instruction, replacement of principals, replacement of teachers with low value-added or negative classroom observations, staff evaluation and feedback, and establishment of a culture of high expectations on elementary school achievement including contracts signed by parents. Relative to ACE, the Fryer (2014) bundle includes much more extensive non-educator components, but is more limited with regards to personnel changes. Fryer (2014) finds improvement math score improvements that are approximately one third as large as the ACE effects, and little or no significant effects on reading.

examine the effect of a district-wide turnaround in Lawrence, Massachusetts that included changes in school autonomy, school accountability, data-driven instruction, replacement of principals, replacement of teachers, and extensive tutoring. They find improvements in math of 0.2-0.3 SD and up to 0.1 SD in reading. Though they cannot separately identify the effect of each component, they find that an intensive tutoring program for struggling students during school vacations was an important driver of the improvement. The school turnaround evaluated in Fryer (2014) is similarly multifaceted with increased instruction time, daily high dosage tutoring, data-driven instruction, replacement of principals, replacement of teachers, staff evaluation, and establishment of a culture of high expectations. He finds that this bundle raised math achievement by roughly 0.15 standard deviations but had little or no significant effect on reading achievement. Relative to the interventions considered in Schueler, et al. (2017) and Fryer (2014) the ACE intervention is structured to have smaller non-personnel components and larger personnel components. Most importantly, in both Schueler et al. (2017) and Fryer (2014) there is no compensating differential to attract high-quality educators, and the dismissal rates of existing teachers are far lower. In Schueler et al. (2017), 8% of teachers and 56% of principals are removed as part of the turnaround. In Fryer (2014), nearly all principals were replaced and the teacher turnover rate at treated elementary schools rises by approximately 10 percentage points relative to control schools. In contrast, the ACE intervention increases the teacher turnover rate by nearly 50 percentage points relative to control schools.

Relative to the literature on school turnarounds, our paper is the first example of a district initiative that uses effectiveness-based compensating differentials to transform the quality of a school's educators. In addition to examining a novel reform, ours is the first analysis to assess whether effects on achievement persist once students are no longer treated and transition into the next schooling level.[7]

*2.b. Compensatory pay*

A wide range of programs designed to improve the quality of instruction at traditional public schools for disadvantaged students have been implemented, but the overall record in the United

[7] Gilraine and Pope (2021) contrasts teacher effects on contemporaneous and future test scores to isolate permanent influences, and Dinnerstein and Opper (2022) demonstrate that educators respond to policies that elevate the importance of end-of-year test scores.

States has not been very successful. Research that investigates the effects of programs designed to attract educators to hard-to-staff schools includes Clotfelter et al. (2008); Springer et al. (2010); Steele, Murnane, and Willett (2010); Clotfelter, Ladd, and Vigdor (2011); Glazerman et al. (2013); Springer, Swain, and Rodriguez (2016); and Cowan and Goldhaber (2018).

Clotfelter, Ladd, and Vigdor (2011) found that paying a premium at high-poverty schools successfully improves teacher retention but had little effect on teacher quality because the pay premia differentially led to the hiring of teachers with worse credentials. This outcome is consistent with the simulations presented in Bates et al. (2025) which illustrate the value of tying stipends to teacher effectiveness and having strong incentives to hire effective teachers.

The targeting of academically strong college students through alternative certification programs including Teach for America (TFA) has shown modest success (Boyd et al. (2006), Kane, Rockoff, and Staiger (2008)).[8] Washington, D.C. School District has offered sizeable rewards to teachers who succeeded in generating substantial achievement increases in the most disadvantaged schools through the IMPACT reform. Although this strengthened incentives and made those schools more attractive, it was not implemented in a manner that enabled the identification of the targeted pay effects on achievement or the quality of instruction (Dee and Wyckoff (2015)).

Glazerman et al. (2013) report on a randomized controlled trial that offered salary stipends for high value-added teachers to transfer to designated high-poverty schools. They find that $10,000 per year payments to teachers succeeded in attracting high-value added teachers, and that classrooms targeted to receive the stipends had achievement that was 0.1 to 0.25 standard deviations higher. Unlike the ACE intervention, only a small number of teachers in each school qualify for the stipends, so it is not intended to be a school-wide transformative intervention. Interestingly, the ACE intervention improves school-wide outcomes by more than the targeted classrooms in Glazerman et al. (2013). There are a variety of possible explanations for the larger effects in ACE compared to Glazerman et al. (2013). First, peer effects or input complementarity could lead whole-school transformations to have larger effects than more limited interventions (Jackson and Bruegmann 2009). Second, the $10,000 stipends in Glazerman et al. (2013) were understood to be temporary payments, whereas the ACE program

[8] Glazerman, Mayer, and Decker (2006) find that TFA teachers have higher value-added for math (0.15 student standard deviations) and are equally effective for reading.

was presented as a permanent increase. Finally, selection in the ACE program was based on a multi-measure evaluation system rather than just achievement value added, making it less noisy, less susceptible to strategic behavior including teaching to the test, and more politically acceptable.

Because the magnitude of the ACE stipend is increasing with teacher effectiveness, our study also relates to a literature on teacher merit pay summarized by Pham, Nguyen, and Springer (2021). Their meta-analysis finds that on average merit pay increases student outcomes by 0.043 student standard deviations, with larger effects in elementary school and smaller effects for smaller performance incentives. As we discuss in more detail below, while the compensating differential for working in an ACE school is large, the incremental increase in the stipend for better performance is relatively modest.

## 3. Dallas ISD's personnel reforms

Dallas ISD is a large urban school district in north Texas comprised of roughly 160,000 students in 230 schools. After a lengthy development process, the district dramatically changed its personnel systems beginning in 2013 including the replacement of its traditional teacher salary schedule with a compensation system based on performance-based evaluations. After performance-based compensation systems for administrators and teachers were fully implemented, the district introduced the evaluation-based incentives for working in disadvantaged schools (ACE) that are analyzed here.

### *a. District-wide evaluation and compensation reforms*

The personnel reforms first focused on principals with the introduction of the Principal Excellence Initiative (PEI) in 2013, and two years later, the reforms extended to teachers under the Teacher Excellence Initiative (TEI), described in more detail in Appendix A. Both PEI and TEI moved from a traditional salary schedule based largely on years of experience and academic degrees to compensation systems more closely related to performance. For teachers, evaluation incorporated test-based measures of value-added and achievement relative to others teaching comparable students, an extensive observational component using prescribed rubrics for evaluation, and student surveys. Specifics varied by subject taught, grade and other factors. For principals, supervisor evaluations of work product including their effectiveness at supporting teacher improvement replaced classroom observations, and the achievement component included a measure of success at closing achievement gaps. Importantly, the separate evaluation components are aggregated to a single evaluation score that is used to rank teachers (principals) and place the educators into bins that are the primary determinant of salaries.[9] TEI divides teachers into four major groupings with fixed proportions in each category: Exemplary, (1 category), Proficient (3 categories), Progressing (2 categories), and Unsatisfactory (1 category). PEI uses a similar method to allocate principals. Appendix Table A1 shows how the evaluation categories map to teacher salaries (top row). After an intensive informational campaign and extensive administrator training, these reforms were implemented uniformly across the entire district. Hanushek et al. (2023) investigate the overall reform effects on math and reading achievement using synthetic control methods and find that average math achievement in Dallas ISD increased substantially following implementation of the reforms.

*b. Accelerating Campus Excellence (ACE)*

In academic year 2015-2016, one year following TEI adoption, Dallas ISD introduced the Accelerating Campus Excellence (ACE) program that was designed to raise the quality of instruction and achievement in chronically low-performing schools. Our description below focuses on ACE implementation in elementary schools.

Although this intervention incorporates several additional components, attracting and retaining effective educators is the centerpiece of the program. Educators who applied and were selected to work at ACE campuses received signing bonuses of $2,000 and annual stipends that depended upon position and, for teachers, on TEI effectiveness rating for the prior year. Stipend amounts equaled $13,000 for a principal, $11,500 for an assistant principal, $8,000 for a counselor, $6,000 for an instructional coach. For teachers, stipends were based on effectiveness level and ranged from $6,000 to $10,000. Appendix Table A1 shows the stipend associated with each effectiveness level. The stipends are best thought of as salary increases that are contingent on working in an ACE school. Because principal compensation in PEI depends in part on student achievement, principals have a direct incentive to hire teachers who are most likely to increase achievement.

[9] The linkage to salaries followed a hold-harmless period used to move from the traditional salary system to the new system. In addition, placement in the highest rating categories requires participation in the distinguished teacher review process. Finally, early career teachers are not eligible for the higher ratings categories.

Based on the pre-specified distribution of ratings underlying the TEI reforms, approximately 20 percent of Dallas ISD teachers would qualify for the $10,000 pay premium, 40 percent of teachers would qualify for an $8,000 pay premium by obtaining a proficient rating, and 37 percent would qualify for a $6,000 premium by receiving a progressing rating due either to being new to the district, to having only one or two years of prior experience, or to failing to reach proficiency. In actuality, 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 in the first year of the ACE program, illustrating the district's success at recruiting teachers from the upper part of the effectiveness distribution.

The program required all existing teachers in ACE schools to re-apply for positions and to be interviewed along with the other applicants. Approximately 80 percent of teachers and all principals in a school newly designated as ACE were different from the teachers and principals who had been in the school the previous year. Such turnover by itself would be expected to adversely affect the quality of instruction, as teachers adjust to different schools and in many cases different grades.[10]

The ACE program was initially rolled out to four elementary schools in 2016 and then an additional five elementary schools in 2018. We refer to the first set of schools as ACE 1 and the second set of schools as ACE 2.[11] In one of the ACE 1 schools (Umphrey Lee), the district found evidence of cheating on the 2013 state standardized test, meaning that the school does not have reliable test records for 2012 and 2013. Because this complicates the examination of pretreatment trends, our baseline analysis excludes Umphrey Lee. However, we show that the inclusion of Umphrey Lee yields very similar estimated effects.[12] Our preferred analysis includes three ACE elementary schools in the first wave and three in the second wave, so it is important to highlight that the evidence we provide is based on a limited number of schools. Furthermore,

[10]Hanushek, Rivkin, and Schiman (2016) find that much of the negative effect of turnover can be explained by lost general and grade-specific human capital. Ost (2014) identifies the substantial returns to recent experience teaching the same grade.

[11]Two of the five ACE 2 schools, Onesimo Hernandez Elementary School and J. W. Ray Learning Center, were consolidated/closed in 2019 complicating outcome analysis for these schools. Our analysis of the second wave of ACE is thus focused on the three schools that exist in both 2018 and 2019. If we instead include the consolidated schools, we cannot examine 2019 outcomes since they no longer exist in 2019, but we find similar 2018 effects.

[12] The exception is that if we include Umphrey Lee in the pretrend assessment, the ACE schools show a large test score drop from 2013 to 2014 that is not observed in the control. This is because 2013 was the last year that Umphrey Lee's test scores were deceptively high through cheating. Importantly, even with Umphrey Lee included, the treatment and control groups show very similar trends from 2014 to 2015, (t-2 to t-1).

the small number of schools raises the possibility that we might have insufficient power to detect moderate effects and presents a challenge for inference. In practice, the ACE effects are large enough that we can statistically reject zero effects at conventional significance levels using permutation test p-values that account for the small number of treated schools.

In 2019, Dallas ISD scaled back the intervention for three of the four ACE 1 elementary schools; the fourth was assigned to a new ACE cohort.[13] The reduced program eliminated the after-school program component, and importantly, salary stipends for most teachers. The teachers who continued to receive stipends took on the role of education leaders and worked additional hours in support of professional development for teachers in the school. In other words, the program was no longer an effectiveness-based compensating differential but instead reverted to the more common extra-pay-for-extra-work model. Teacher transitions following 2018 therefore provide evidence on the role of effectiveness-based stipends in teacher retention.

The first wave of ACE schools was selected in 2015 based primarily on 2014 test scores. This implementation poses a challenge for a standard difference-in-differences identification strategy, because either idiosyncratic 2014 shocks or downward trends in performance could lead to an ACE designation. On the one hand, a purely transitory negative shock in 2014 could lead to these schools making unusually large recovery-induced gains in 2015 that continue into 2016, the first year of treatment. On the other hand, a downward trend in performance might both lead to the initial ACE designation and continue into 2016. To address these concerns, we construct a control group based explicitly on low performance in 2014. Since our control group is also selected based on low 2014 performance, it also may have experienced negative 2014 shocks or persistent negative trends. Our main analysis defines the control as the lowest 15 percent of non-ACE 1 schools in terms of 2014 average test scores, but we show robustness to varying this definition of low performance. ACE 2 schools were selected primarily based on low-achievement in 2016 (t-2), and the control group is similarly selected based on 2016 test scores.

## 4. *Channels of Potential Impact*

The ACE intervention likely affects learning and achievement through multiple channels. Personnel reforms are clearly the centerpiece of ACE, representing approximately 85 percent of

[13] The 2019 ACE cohort was given a less intensive version of the ACE intervention with more limited stipends. All four ACE 1 schools no longer received the original ACE intervention in 2019.

program expenditures, but the remaining channels may also contribute to the program effects. The ACE intervention caused the following policy differences: (1) All ACE teachers have to re-apply for their jobs and only 20% of teachers were rehired; (2) Teachers who choose to work at an ACE school receive a one-time bonus of $2,000 and an annual salary stipend of $6,000 to $10,000 depending on the evaluation rating in the previous year; (3) Changes in the evaluation rating while at an ACE school may cause larger salary shifts than identical evaluation rating changes at a control school because in some cases the change moves the teacher to another ACE stipend level; (4) ACE principals, vice principals, counselors and coaches receive a stipend, and all existing principals were replaced; (5) a 1-hour increase in the master schedule; (6) ACE students had access to extended after-school programming that ran until 7:00 pm as opposed to the 6:00 pm finish of after-school programming accessible to students at control schools; (7) a greater emphasis on and support for data-driven instruction; funding for uniforms; and some enhancement of professional development.

The simultaneous implementation of all ACE components precludes the decomposition of the overall impact into the separate channels. Below, we describe the various channels through which the ACE intervention may affect outcomes and, where possible, reference existing literature on the likely magnitude of each component.

### *4a. Personnel reforms*

The effectiveness-determined stipends would be expected to expand the supply of teachers to ACE schools and amplify the performance incentives imbedded in the TEI salary structure. In the absence of compensating differentials, the most disadvantaged schools have been beset historically by high levels of teacher vacancies, significant teacher turnover, and the disproportionate reliance on new teachers. The ACE stipends were designed to incentivize teachers to work in these schools with an effectiveness-based compensating differential introduced to make the incentives strongest for more productive teachers (Brown and Andrabi 2020). In combination with the large performance incentives for principals embedded in PEI, the large stipend for principals to move to an ACE school, and strong district commitment to ACE, the personnel reforms reflect the district's thoughtful approach to raising the quality of instruction at highly disadvantaged schools, satisfying the criteria described in Bates et al. (2025). Moreover, the requirement that all existing teachers at ACE schools must reapply for an ACE teaching position expedites any changes in teacher composition. If ACE stipends attract higher-quality teachers, teacher peer effects can lead the increase in the quality of instruction to exceed the changes in average teacher effectiveness based on past performance.[14] The unique structure of the Dallas ISD personnel system, combined with the large scale of the ACE compensating differential, makes it difficult to use existing literature to predict the expected magnitude of changes in teacher quality.

In addition to expanding the teacher supply, the connection between the ACE stipend value and the evaluation rating amplifies the performance incentives. Appendix Table A1 compares the Dallas ISD salary scale with the salary scale at ACE schools. The ACE stipends strengthen the performance incentives for the roughly 60 percent of ACE teachers at the progressing and proficient 1 effectiveness levels by raising the reward from moving to a higher effectiveness category. For example, moving from progressing 2 to proficient 1 increases salary by $3,000 at most Dallas schools, but by $5,000 at ACE schools. For a teacher considering a wider range of possible future salaries, increasing from progressing 1 to master increases salary by $41,000 at most Dallas schools, but by $45,000 at ACE schools.[15] ACE stipends do not affect the financial reward from moving up one category for the 40 percent of ACE teachers already at a rating of Proficient 2 or above. Overall, the ACE performance incentives are a modest intensification of an existing performance incentive. The literature on pay for performance in the United States finds that even large performance incentives have modest to null effects on achievement (Pham et al. 2021), so we do not expect intensifying the existing performance incentive to constitute a primary channel through which ACE affects the quality of instruction and achievement.

Finally, ACE stipends are also paid to personnel in other roles including counselors, assistant principals and principals. We do not aim to isolate the effect of various personnel, both because it is empirically intractable and because some effects are conceptually inextricable (e.g. a principal's efficacy may stem from hiring effective teachers). In addition to directly affecting outcomes, improvements in administration and other personnel may interact with improvements in teacher quality if there are complementarities between various ACE components.[16]

[14] Jackson and Bruegmann (2009) find positive teacher peer effects.
[15] This calculation excludes the "unsatisfactory" category as only 3% of teachers in Dallas are unsatisfactory and none of them work at ACE schools.
[16] Though we do not aim to isolate the effect of administrator quality improvements, it is unlikely that those improvements in isolation could drive a large portion of the ACE effect as Branch et al (2022) find that a one standard deviation change in principal effectiveness is associated with a roughly 0.05 standard deviation change in achievement, and Bartenen et al (2024) finds an even smaller effect that is closer to 0.03 standard deviations.

*4b. Nonpersonnel channels*

The non-personnel components include uniforms, extended school time, after-school programming, data-driven instruction, and professional development. Estimates from prior literature suggest that these components are unlikely to drive large increases in student achievement. We now consider these components in turn. There is limited evidence on the effects of uniforms in elementary schools, though Gentile and Imberman (2009) find no evidence of a positive achievement effect.

It is difficult to quantify the likely effect of the extended school time because we do not know what portion of this time was used for instruction or for each subject. To gauge the potential impact of extended time on test score improvement, we assume that 100% of the extra time was devoted to instruction, split equally between math and reading. Using a student fixed effect approach first developed by Lavy (2015) and extended in Rivkin and Schiman (2015), Bietenbeck and Collins (2020) estimate the effect of instructional time using multiple waves of PISA and find that each additional hour of weekly instructional time increases test scores by approximately 0.02 standard deviations. If each subject saw 2.5 extra hours, the increase in instructional time might increase outcomes by 0.05 SD.

The after-school extension is unlikely to substantially raise achievement since only 10% of students participated and the increase in time relative to the control group was modest.[17] Moreover, in a review of after-school programs, Zief, Lauver and Maynard (2006) finds no evidence that after-school programs increase test scores, and more recently Drange and Sandsør (2024) find no effect of a universal after-school program.

The effect of professional development is expected to be context specific, but most evidence suggests that general professional development is ineffective. Two large-scale randomized

[17] The ACE after-school programs were provided by community partners such as Big Brothers, Big Sisters and focused on social and emotional support and mentorship. Participation in the after-school component was not required, and while there are no records of student attendance, district administrators estimate that approximately 10% of students participated. Students at control schools also have access to after-school programs, but programming at ACE schools goes through 7:00 pm whereas programming at control schools goes through 6:00 pm. We lack data on participation rates at control schools or in the pre-policy period, but we know that after-school programs have 75 available slots per school, which is more than 10% of the typical elementary school. Because we the lack data on what fraction of ACE and control students participate in after-school programs, or what fraction of participating ACE students stayed through 7:00 pm, the marginal increase in after-school program participation is not known.

control trials find no evidence of a positive effect, and this finding is consistent across a variety of settings (Garet et al. 2008; Randel et al. 2011). In the ACE context, we expect the formal professional development to have minimal effect, but it is certainly possible that the general ACE environment that includes high-quality peers, coaches and principals contributes to teacher efficacy. Furthermore, ACE includes a teacher coaching component, and recent evidence suggests that coaching is more effective than generalized professional development (Kraft, Blazar and Hogan 2018).

ACE schools were required to use data-driven instruction, but the marginal program increase is likely to be modest since all schools in the district use data extensively. The literature on data-driven instruction is wide-ranging and most studies include data-driven instruction as a component of a broader initiative. Carlson, Borman and Robinson (2011) evaluate the effect of a data-driven reform where consultants from John Hopkins worked with districts to establish quarterly student benchmark assessments and provide extensive training. Relative to a control group that did not use data-driven instruction, the reform increased math test scores by approximately 0.06 SD, although this estimate is sensitive to whether the analysis accounts for baseline scores.[18]

Overall, we expect the ACE non-personnel components combined likely have a modest effect on achievement. Even under the assumptions that 100% of the extra school time is devoted to instruction and that none of the control schools use any data-driven instruction, the existing research suggests an upper bound of approximately 0.11 SD of math improvement at ACE schools.

## 5. *Data and Descriptive Statistics*

The analytical database is constructed from several sources. Data on student and staff characteristics come from Dallas ISD administrative data as submitted to the Texas Education Agency. We use the math and reading test scores from the State of Texas Assessments of Academic Readiness (STAAR), which we standardize within Dallas ISD separately by subject,

[18] Theoretically, the RCT does not require controlling for baseline scores, but models that control for baseline scores show a 0.06 improvement and models that omit this control show an insignificant 0.002 effect. The effect on reading test scores was statistically insignificant in all models but is negative in models that omit the baseline score control and positive in models that control for baseline scores.

grade, and year to have mean zero and standard deviation one.[19] Other student information includes race, gender, and indicators for students qualifying for programs such as free or reduced-price lunch, gifted, special education, and limited English proficiency. Staff information includes role, experience, subject, grade, and school.

We also have access to unique data that include scores and sub-metrics for all the inputs into the TEI evaluations and rating categories along with salaries and stipend amounts. The evaluation data include scores for teacher performance as measured by rubric-based observations, student perceptions reported in surveys of students above the second grade, and achievement. As we describe below, the information used to produce teacher evaluation scores differs by grade and subject taught. The educator stipend data contain information for all ACE-school educators in each year a school participated in the program. We combine the data sets and construct a panel that links teachers, students, and schools together from the 2011-2012 to 2018-2019 school years.

Table 1 shows descriptive statistics for the two waves of ACE schools and their respective control groups for the year prior to program implementation for each wave. Compared to the controls, ACE schools are lower performing and have a higher percentage of Black students and lower percentages of Hispanic and LEP students. Most students in Dallas are eligible for subsidized lunch, and there is little difference in this between treatment and control. Financial resources are generally standardized across schools, and we observe similar student-teacher ratios and teacher-aid to teacher ratios at ACE and control schools. There are 3 ACE schools and 22 control schools in each wave.

The requirement that existing teachers re-apply to continue working at an ACE school is expected to increase teacher turnover in the first year of implementation. In the top and bottom panels of Figure 1 we show the turnover rate by year for ACE 1, ACE 2, and their respective control groups. Turnover is 80% or higher during program implementation for both waves of ACE, reflecting the fact that relatively few teachers were rehired.[20]

---

[19] Test scores used in the teacher value-added analysis are standardized by grade, subject and year at the state rather than the district level.

[20] In addition to the spike in teacher turnover, non-teaching personnel also turned over at higher rates in the year of ACE adoption. Principals were all involuntarily replaced, so their turnover was 100%. Other non-teaching personnel had a turnover rate of 45%, which while considerably higher than their baseline turnover rate of 13%, is far lower than the teacher turnover rate.

ACE may increase teacher evaluation scores either because ACE attracts higher quality teachers or because the program raises scores through its effects on achievement or teacher performance based on supervisor observations. To isolate the role of teacher composition, we describe all teachers in terms of their t-1 evaluation rating.[21] Figures 2 and 3 illustrate the rightward shifts in the ratings distributions for ACE schools and the absence of such changes in control schools. For both ACE 1 (Figure 2) and ACE 2 (Figure 3), the shift in teacher evaluation ratings is transformational: Before ACE, the vast majority of teachers were rated in the bottom 3 categories whereas after ACE, far fewer teachers fall in this range. The increase in the share of ACE 1 teachers rated Proficient II or above exceeded 50 percentage points. For ACE 2, a 25 percentage-point increase in the share rated Proficient 1 accompanied a 35 percentage-point increase in the share rated Proficient II. Remarkably, prior to the program, none of the teachers in ACE 2 schools had a rating above Proficient I, highlighting the difficulties that urban districts often have in staffing schools serving highly disadvantaged populations.

Large reductions in the share of ACE 2 and particularly ACE 1 teachers with no prior experience contribute to their rightward shifts in the ratings distributions. This share dropped by 10 percentage points following the implementation of ACE 2 and a whopping 30 percentage points following the implementation of ACE 1. In contrast, the share of teachers with no prior experience fell by less than 4 percentage points in the ACE 1 and ACE 2 controls. While the fraction of novice teachers fell substantially, the overall change in average experience was modest as the average experience level of exiting teachers was approximately 7.5 years and the average experience level of replacement teachers was approximately 9.75 years. That said, the fraction novice is likely more central than the average level of experience as evidence documents diminishing returns to experience (Papay and Kraft 2015).

The extent to which the evaluation scores and ratings capture differences in the quality of mathematics and reading instruction is a key determinant of program impacts on achievement. Table 2 reports the average t-1 mathematics and reading value added estimates for the subset of teachers who enter or exit an ACE school, teach in a tested grade in year t and have a value-

[21] Describing teachers in terms of their t-1 evaluation ratings helps avoid conflating school-wide ACE effects with compositional effects, but it is important to note that t-1 evaluation scores are based on performance at a different school and may not fully predict performance at the new school. The information content of t-1 evaluation scores for the ACE program depends on whether there is a school component that drives evaluation scores and whether there is an important match component.

added estimate in t-1.[22] The entrant-exit differences are remarkably similar across the cohorts and reveal much larger changes in mathematics value added than in reading value added. Average mathematics value added of entrants exceeds that of exits by 0.26 standard deviations for both ACE cohorts; the corresponding difference for reading value added equals 0.14 standard deviations in 2015 and 0.075 standard deviations in 2017.

The larger differences in math than in reading suggest a stronger connection between the evaluation scores and effectiveness at math instruction than effectiveness at reading instruction. Appendix Table A2 reports the correlations between the value added and both the overall evaluation score and the performance component for all Dallas ISD teachers with value added estimates in 2015 and 2016 and in 2017 and 2018, weighted by the number of test-takers. For both ACE cohorts the correlation with the total evaluation score is roughly one third larger in math than in reading: 0.62 versus 0.45 in 2015/16 and 0.55 versus 0.39 in 2017/18. The fact that even larger relative differences emerge for the performance component of the score which is based primarily on classroom observations suggests that the district evaluation rubric better captures the skills and practices that determine effectiveness at mathematics instruction.

In Appendix Figure A1, we show the distribution of math and reading teacher value-added before and after ACE adoption compared to the overall distribution of teacher value-added at other Dallas schools.[23] This description is not intended to isolate compositional change since teacher value-added includes a school component. For both math and reading, there is clear evidence of a rightward shift in value-added following ACE adoption, though the magnitude is more dramatic for math than for reading. Comparing post-intervention ACE value-added to other Dallas schools shows that not only has ACE caught up with other Dallas schools, but it has surpassed them.

Finally, to assess whether ACE simply attracts higher quality teachers or if the teachers who transfer also perform better at ACE, we compare the value-added of ACE entrants at their first year at ACE (t+0) to their value-added during the last year at their previous school (t-1). We find that at ACE entrants have 0.16 higher VA in math and 0.14 higher VA in reading compared to their VA in their previous school. This within-teacher improvement could be driven by any of

[22] Because the vast majority of entrants to control schools are new to Dallas ISD and do not have value-added estimates in t-1, Table 2 focuses on ACE teachers.
[23] Other Dallas schools includes all non-ACE schools, not just control schools. That distribution is provided as a reference point, not as a counterfactual.

the ACE components or by fixed differences between the ACE school effects and the school effects at previous schools.

## 6. *Effects of ACE on achievement*

We investigate the efficacy of ACE by comparing test scores at ACE and control schools. We first present the raw test-score data. Subsequently we describe the empirical models used to identify the program effects and to address potential confounding factors. Finally, we present regression results including extensive sensitivity analysis.

*a. A simple description of achievement trends*

The top and bottom panels of Figure 4 offer strong evidence that the ACE intervention led to dramatic growth in math and reading achievement. Not surprisingly, compared to both the district average and the control schools, average math achievement in ACE 1 schools began very low. Performance in 2014 dipped, consistent with a transitory negative 2014 shock at the already low achieving ACE 1 schools, and we observe a small recovery from 2014 to 2015. Importantly, although we use only the 2014 test scores to select the control schools, they follow a remarkably similar trend to the ACE schools over the entire pretreatment period. Control schools, which by design are not quite as negatively selected, also exhibit low performance in 2014 and a 2014 to 2015 recovery. For ACE 2, the control and treatment groups do not match quite as closely, but it remains the case that they are on very similar overall trends: the gap between ACE 2 and the control is virtually identical in 2012 and 2017 (t-6 and t-1).

The ACE schools show an immediate and very large increase in achievement upon program implementation. The math score increases for both ACE waves exceeded 0.4 standard deviations. Achievement barely changed in either control group. The achievement decline following the removal of salary stipends in 2019 at all but one ACE 1 school reinforces the interpretation of the achievement pattern as being driven by the teacher incentives. Scores in ACE 1 schools fall below those seen in the prior three years. The bottom panel of Figure 4 shows a quite similar pattern for reading scores. As generally found, schools and programs have smaller effects on reading than math, and the impacts of ACE are not an exception. Nonetheless, the improvement pattern following ACE implementation is similar to that for math, including the decline in 2019.

There are no obvious alternative explanations for the time pattern of performance after the implementation of ACE. Nevertheless, the possibility remains that other factors including changes in student characteristics contribute to the achievement growth. We therefore undertake a more comprehensive analysis of ACE effects on both contemporaneous and future achievement.

### b. Analytical framework

Our test-score analysis uses a student-level event study framework to identify the effects of the ACE program. Throughout, we use the control groups discussed above and exclude other Dallas schools from the analysis. Our baseline specification stacks ACE 1 and ACE 2 treatment and control schools and estimates an event study following the Calloway and Sant'Anna (2021) approach with no covariates, conducting inference using their suggested wild bootstrap method. In a standard application with no covariates, one would take a simple 2-way difference-in-difference between each treatment cohort and the control group in period t + j relative to the omitted year of t-1. The control group is defined as either units that are never treated or units that are not yet treated in period t + j. One complication for our setting is that there are institutional reasons to expect that the year effects differ across treatment waves. In particular, ACE 1 (and its control group) are expected to have unusually low achievement in 2014 and ACE 2 (and its control group) are expected to have unusually low achievement in 2016. A simple application of the Calloway and Sant'Anna (2021) approach would assume a single set of year effects and both control groups would be used as the counterfactual for both waves of ACE. This is particularly problematic when comparing ACE 1 (treated in 2016) to the ACE 2 control group since the latter is specifically selected based on having low 2016 scores. To avoid this concern, we modify the Calloway and Sant'Anna (2021) approach so that only control group 1 is used as the counterfactual for ACE 1 and only control group 2 is used as the counterfactual for ACE 2. The resulting estimates are equivalent to averaging the results from the estimation of separate models for ACE 1 and ACE 2.

In addition to presenting the Calloway and Sant'Anna (2021) overall estimates of the effect of ACE, we also analyze the two waves separately, removing complications related to staggered treatment effects. Equation (1) describes the event-study specifications used in this analysis:

$$A_{it} = \alpha_0 + \gamma_0 ACE_i + \rho_t + \sum_{k \neq -1} \delta_k D_{it}^k + \epsilon_{it} \tag{1}$$

where $A_{it}$ denotes math or reading test score for student i in year t, $ACE_i$ is an indicator for schools in the relevant ACE wave, $\rho_t$ is a year fixed effect and $D_{it}^k$ are event-time indicators that are 1 if the school is k years from treatment and zero otherwise. The event-time indicators omit $-1$ and span from -4 to +3 for ACE 1 and from -6 to +1 for ACE 2. The parameters of interest are $\delta_k$, which capture the divergence between ACE and the control group in year t+k relative to the omitted year. The time effects span 2012 to 2019, with the year before treatment taken as the omitted group (2015 for ACE 1 and 2017 for ACE 2). Because there are only 3 treated schools per wave, we conduct permutation-based randomized inference and report p-values rather than basing inference on clustered standard errors that are known to perform poorly in contexts with few treated clusters (MacKinnon and Webb 2017). Our permutation test follows the approach described in (Heß 2017) for difference-in-difference setups where data are clustered at the school level. The procedure tests the sharp null hypothesis that the treatment had no effect on any unit by comparing the observed DID estimate to a reference distribution generated by reassigning the treatment indicator across clusters. The p-value is a 2-sided test that is the proportion of 500 placebo estimates that are larger in absolute value than the actual estimate.

To assess dosage effects, we investigate the effect of the number of years of treatment (one, two or three years for ACE 1 and one or two years for ACE 2), running separate regressions for each dosage. The estimating equation is identical to equation (1), but we restrict the sample to students who have been at an ACE or control school for at least n years by the end of year t. In the absence of the sample restriction, some students in an ACE 1 school in 2017 had two years of treatment while others who moved to the school in 2017 had only one year of treatment. This means that estimates of $\delta_2$, for example, do not identify the effects of three years of treatment in the absence of sample restrictions. Below we discuss steps taken to account for any endogenous selection in school moves.

Importantly, only some of the event-time coefficients are informative of the relevant dosage effect. For example, for n=3 (3 years of ACE or control attendance), the $\delta_0$ coefficient captures the effect of attending ACE for two pre-treatment years and one treatment year and is consequently not informative of the effect of 3 years of exposure. As such, for the dosage analysis, we are primarily interested in the coefficient $\delta_{n-1}$, which corresponds to the student having exactly n years of exposure following program start. This is measured relative to the omitted relative-time indicator, $k = -1$, corresponding to the last fully untreated cohort with n years of exposure. Importantly, even though only some event-time indicators are of interest, we include the entire set of indicators in the model to ensure that only $k = -1$ is the reference year. Since tests for elementary school students are first administered in grade 3, the one-year of exposure specification include students in grades 3-5 in year t=0, the two-year specification samples include students who were in grades 2-4 in year t=0, and the three-year specification samples include students who were in grades 1-3 in year t=0.[24]

To assess the effect of ACE 1 elementary school attendance on test scores at the next schooling level (6th grade), we modify equation (1) so that the outcome is 6th grade rather than contemporaneous test scores. In this case, we are primarily interested in the coefficient $\delta_{n-1}$, which corresponds to the student having exactly n years of elementary-school exposure following program start. For example, when studying the effect of 3 years of exposure, the $\delta_2$ coefficient captures the effect of attending a treated school for grades 3-5 on 6th grade test scores, relative to the last fully-untreated cohort. It is the 6th grade students in 2019 who were treated for three years, and their achievement is compared with achievement for the 6th grade students in 2016, the last untreated cohort.[25]

Restriction of a sample to students with n successive years at a treatment or control school potentially introduces selection bias, and we therefore also report intent-to-treat estimates based on samples that include all students with n years of potential exposure. In other words, we assign treatment or control status based on school attended n-1 years earlier regardless of which Dallas ISD elementary school the student attended in the subsequent years.[26]

To assess the role of student composition, we use our same event-study model but make the outcome a series of student characteristics rather than actual achievement. To summarize the change in characteristics, we also estimate models where the outcome is predicted, rather than actual achievement. We generate predicted achievement by regressing math (reading)

[24] Longer exposure corresponds to exposure in earlier grades, raising the possibility that heterogeneous treatment effects could contribute to differences by exposure length. Preliminary estimates (not reported) revealed little variation by grade, leading us to combine students across grades.

[25] Some students who attend an ACE elementary school also attend an ACE middle school, but this cannot drive estimates as ACE elementary attendance does not increase the odds of ACE middle school attendance and the rates of overlap in ACE attendance across schooling levels is low. For example, only 2% of those with 3 years of potential ACE elementary attendance attend an ACE middle school.

[26] An alternative to the intent-to-treat approach is instrumenting actual years of exposure with potential years of exposure, but this relies on a linearity assumption (e.g. three years of exposure is exactly three times as beneficial as one year of exposure) that is likely to be violated.

achievement on student characteristics separately by grade in the year prior to the beginning of treatment for each ACE wave. Then, we use the coefficients to predict achievement for ACE and control students in grades 3-5 with valid test scores. As a further check on bias from compositional change, we also estimate models that include student fixed effects to compare outcomes for the same student prior to and following ACE implementation.

Another potential threat to identification is that the ACE treatment could affect control schools. For example, ACE may adversely affect the quality of educators in control 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 teachers employed in an ACE elementary school in 2016 represent less than 2 percent of Dallas teachers, implying that spillover effects on any given control school are likely to be small. Among all schools in the district, the modal number of teachers lost to ACE schools is zero. For control schools specifically, less than 3% of control-school teachers in 2015 transferred to an ACE school in 2016. To assess whether spillovers to control schools represents a first-order source of bias, we show robustness to excluding any school that lost a teacher to an ACE school from the control group.

Though we expect that spillovers are unlikely to substantially affect any control school, it is important to acknowledge that part of the benefit of ACE may come from drawing high quality teachers away from other schools. In the worst-case scenario, ACE could simply transfer high-quality teachers across schools, with no improvement for the district as a whole.[27] Even in this case, the policy would improve equity as schools that lose teachers to ACE schools are substantially higher performing, with 0.5 SD higher average test scores and 0.85 SD higher average teacher evaluation scores. Furthermore, there is extensive evidence that more advantaged, higher-achieving schools have an easier time attracting and retaining teachers, so the improvements at ACE schools need not be symmetric to the losses at schools losing a teacher to an ACE school. The higher salaries in high-poverty schools would expand the teacher applicant pool if some prospective teachers are willing to work at an advantaged school but will only work at a disadvantaged school if they receive a stipend.

[27] Evaluating district-wide performance during this time, Hanushek et al (2023) finds that test scores in Dallas as a whole are rising from 2015-2019 relative to a synthetic control group. This likely reflects the effect of the overall TEI program rather than the effect of ACE given that TEI affects the entire district, while ACE only directly affects a small fraction of schools.

*c. Estimates of contemporaneous effects*

Table 3 shows the results of the effects of ACE, estimated by stacking ACE 1 and ACE 2 and using the Calloway and Sant'Anna (2021) approach. To maintain a balanced panel, we look at effects in the 4 years prior to implementation and the 2 years following implementation. The reference group is t-1. For both math and reading, none of the pre-period estimates are statistically significant, and the estimates are generally small in magnitude. In the first year of ACE implementation, we estimate an ACE effect of 0.46 standard deviations for math scores and 0.22 for reading scores, both of which are highly significant. In the second year of ACE these estimates are 0.54 for math and 0.34 for reading. The fact that the effect on t+2 scores is not double the magnitude of the effect on t+1 scores suggests that the ACE effect on test scores is not additive.[28]

To examine a longer pre-trend for ACE 2 and a longer post-trend for ACE 1, we examine event study plots separately by ACE wave. Plots of the math and reading event-study coefficients for both waves of ACE shown in Figure 5 reveal little or no evidence of differential trends prior to ACE implementation, and there are sharp increases in test scores in the year of implementation. The confidence intervals shown in Figure 5 may be unreliable given the small number of clusters, so we rely on permutation tests for inference. Table 4 shows the point estimates and permutation test p-values. For math, the estimated effect is approximately 0.5 standard deviations in both waves, whereas reading estimates are closer to 0.3 standard deviations. Importantly, the much smaller t+3 coefficient for ACE wave 1 corresponds to 2019, the year in which most components of the ACE treatment were removed in these schools.

We investigate the robustness of the main results to a variety of alternative specifications. For compactness, we focus on the immediate effects of ACE 1 and ACE 2 rather than presenting robustness of the entire event study for every alternative specification.[29] First, we examine the sensitivity of the estimates to the inclusion of school fixed effects or school by grade fixed effects. Given that the baseline model already accounts for average differences between ACE and control schools, it is not surprising that neither the school or school-by-grade fixed effects have a meaningful effect on estimates for either ACE 1 or ACE 2 (see Appendix Table A3).

[28] A non-additive effect is consistent with the rapid fadeout of teacher effects documented in Lefgren and Sims (2010). Importantly, this need not imply that students are forgetting material learned as tested content changes across grade levels, particularly for math.

[29] Estimates that are not reported find similar robustness for later treatment years.

None of the math or reading coefficients change by more than 0.02 standard deviations. Second, we examine the sensitivity of the estimates to the inclusion of student fixed effects that control for compositional changes following the implementation of the ACE treatment. All estimates remain highly significant, increasing slightly in the ACE 1 specifications and decreasing somewhat in the ACE 2 specifications (see Appendix Table A3). Importantly, the student fixed effects change the parameter of interest to be the effect of ACE for relatively stable students.

To assess whether student composition changes following ACE adoption, we estimate placebo event-study models that substitute student characteristics and predicted math and reading achievement based on these characteristics in place of the achievement outcomes.[30] Appendix Table A4 shows estimates for ACE 1 and Appendix Table A5 shows estimates for ACE 2, and these reveal little or no evidence of the type of endogenous selection that would improve outcomes at ACE schools relative to controls. In fact, the only statistically significant estimate shows a slight increase in the share of low-income students at ACE 2 schools in 2019, a change that if real would not be expected to improve outcomes. Focusing on the predicted outcomes, none of the coefficients in either the treatment or pre-treatment period are significant at conventional levels, and none are larger in magnitude than 0.04. Moreover, all of the coefficients in the first year of treatment are negative and very small, indicating the absence of any enrollment response in anticipation of the program and therefore any selection bias in the intent-to-treat estimates.

Next, we examine the robustness of our estimates to alternative definitions of the control and treatment groups (see Appendix Table A6). The primary analysis uses the bottom 15% of schools in terms of achievement in t-2 to form the control group, and we investigate the sensitivity of the estimates if we instead use the bottom 5%, 10%, 20% or 25% of schools to form the control group. For both math and reading and for both ACE 1 and ACE 2, the results are quite similar across all these control groups. The estimates are similarly insensitive to the addition of ACE 1 and ACE 2 schools excluded from the preferred specifications. The inclusion of both Umphrey Lee to the ACE 1 specifications and the two ACE 2 schools that closed in 2019

[30] We do not use lagged test scores to characterize the composition of ACE students as lagged test scores would be affected by ACE for the second cohort and later, but we have verified that there is no change in lagged test score from the year before ACE to the first year of ACE.

to the ACE 2 specifications leads to slight increases in the estimated effects on both math and reading achievement in both ACE waves.

Though each wave of ACE appears to trend very similarly to the control group, ex-ante, there are concerns that a phenomenon similar to an Ashenfelter dip could lead to overstating the efficacy of ACE. In particular, because ACE schools are selected based on poor performance in t-2, there is the possibility that ACE schools would recover more than control schools, even in the absence of the ACE intervention. If an Ashenfelter dip were driving our estimates, we might expect to see that improvement between t-2 and t-1 at ACE schools is much larger than improvement between t-2 and t-1 at control schools. Table 4 showed no evidence of this given that the t-2 coefficient is small and statistically insignificant. To provide further evidence on whether Ashenfelter dips drive our estimates, we conduct a falsification test where we pretend that ACE determination was based on 2012 scores and compare the very lowest performing schools to a control group consisting of schools in the bottom 15% of 2012 performance. We then replicate our event study to examine whether the placebo "treated" schools diverge from the control schools in 2014 and 2015. This falsification exercise helps us assess whether the lowest performing schools generally catch up with the bottom 15% of schools in the next few years. Appendix Table A7 shows no evidence of rapid improvement in 2014 and 2015 for the placebo treatment relative to the control. None of the coefficients are positive and significant at conventional levels, with the largest taking a value of 0.089 with a p-value of 0.28.

If ACE schools attract high-quality teachers from control schools, this may lead to overstating the effect of ACE, since control groups outcomes would be falling relative to a pure control condition. Ex-ante, we do not expect these spillovers to be large because control schools lost relatively few teachers to ACE. That said, to provide suggestive empirical evidence, in Table A8 we show how results differ if we modify the control group to exclude schools that lose any teachers to an ACE school. This only provides suggestive evidence since control schools may be affected by the ACE program, even if they do not directly lose a teacher to an ACE school. For example, a high-quality teacher may apply to an ACE school rather than a control school. Nevertheless, the fact that none of the coefficients from the modified samples are smaller than those that include all control schools suggests that spillover effects on control schools are unlikely to drive our estimates.

Though we view the earlier permutation-based inference as most reliable, Appendix Table A9 shows that our conclusions are generally similar if we calculate p-values using the cluster-wild bootstrap approach as implemented via Roodman et al (2019). For math, the 2016 and 2017 coefficients are statistically significant at the 5% level for ACE 1 and the 2018 and 2019 coefficients are statistically significant at the 5% level for ACE 2. The 2018 coefficient for math for ACE 1 is statistically significant at the 10% level. For reading, both ACE 1 and ACE 2 show effects that are significant at the 10% level, though 2018 is just shy of significance for ACE 2 (p-value 0.11). For 2019 (the year ACE 1 is no longer in effect), the ACE 1 coefficients are not statistically significant.

*d. Treatment dosage effects*

Because of entry to and exit from ACE schools during the treatment periods, ACE 1 and ACE 2 schools include students treated for one, two or three years following the initial program year. Consequently, we turn now to specifications that estimate the effects of specific treatment dosages. In Table 5, we show the estimated effect of 1, 2 or 3 years of exposure to the ACE program on math achievement (top panel) and reading achievement (bottom panel), separately by ACE wave. Columns 1,2 and 4 report the effects of actual exposure, and Columns 3 and 5 report ITT estimates of the effects of potential exposure. Since ACE 2 begins in 2018, we observe at most 2 years of exposure, meaning that Columns 4 and 5 are blank for this wave. The exposure effects are estimated as described earlier and are interpreted as the change in outcomes in ACE vs control schools for students who attend for *n* years.

All estimates of ACE effects on math achievement are large and highly significant based on permutation test p-values for both ACE waves. The 1-year-of-exposure coefficients exceed 0.4 standard deviations, and the coefficients on two years of exposure are larger in magnitude than the estimated one-year effect, but the differences are modest and not significant for both waves. The estimated effect of three years of exposure is slightly smaller than that for two years for ACE 1, but the difference is not statistically significant. As expected, the ITT estimates in Columns 3 and 5 are smaller than the corresponding actual exposure coefficients. Nevertheless, they all exceed 0.35 standard deviations with p-values below 0.05.

The estimates for reading in the lower panel also reveal little variation by ACE wave, though they are smaller than those for math. Nevertheless, the ITT estimates are all significant at the 0.05 level based on the permutation tests and at least 0.2 standard deviations in magnitude, strong evidence of sizeable ACE treatment effects on reading achievement.

*e. Effects on 6th grade test scores*

To investigate whether ACE augments the acquisition of skills that persist after students leave elementary school, we change the outcome to be 6th grade math and reading scores. Columns 1, 2 and 4 in Table 6 report coefficients from specifications that measure actual years of exposure, and columns 3 and 5 report ITT estimates from specifications that measure two or three years of potential exposure. In contrast to the large effect of 1 year of exposure on contemporaneous math test scores, we see little evidence of improved 6th grade scores for students with 1 year of ACE exposure. However, students with 2 or 3 years of exposure to an ACE elementary school show point estimates suggesting lasting improvements of approximately 0.3-0.4 standard deviations in math achievement that increase with dosage years. The permutation test p-values equal 0.098 and 0.150 for the estimates of effects of two and three years of exposure on math scores. As expected, the ITT estimates are again smaller than the estimates for actual exposure. Although the magnitudes are large and consistent with positive dosage effects, the lower precision suggests these results should be interpreted cautiously.

For reading, Column 4 of the lower panel of Table 6 shows that the estimated effects of three years of elementary school ACE exposure are substantial though smaller than estimates for math and not statistically significant. In addition, the estimated effects of two years of exposure are much smaller, as are the ITT estimates for both two and three years of exposure.

Panel A of Appendix Figure A2 illustrates kernel density estimates of the distributions of 6th grade math achievement for ACE and control students in 2016, the last pre-treatment cohort, (left diagram), and 2019, the fully treated cohort (right diagram). The contrast illustrates the rightward shift across the math achievement distribution for ACE students in comparison to students in the control schools between 2016 and 2019: the ACE distribution lies slightly to the left of the control distribution in 2016 but is substantially to the right of the control distribution in 2019. Panel B of Appendix Figure A2 shows that the distribution of reading scores does not shift as clearly as math.

Importantly, the divergence between the patterns of contemporaneous and future effects of attendance at an ACE elementary school highlights the importance of a careful consideration of treatment dynamics. The evidence strongly supports the existence of large, immediate ACE effects on contemporaneous elementary school math and reading achievement in both waves. But, the evidence also suggests that only multiple years of exposure to an ACE elementary school establish the learning patterns that raise middle school math achievement. Therefore, evidence based on a single year of treatment may not be adequate to understand the potential for an education intervention to affect the acquisition of valued skills that persist into the future.

## 7. *How important are stipends to the retention of ACE teachers?*

A critical question concerns the necessity of performance-based stipends in the ACE schools over the longer-term. Is just the initial introduction of stipends and public commitment to hiring high-quality teachers sufficient to move the ACE schools from a bad to a much better equilibrium in terms of educator effectiveness? Or is the continuation of stipends necessary to retain high quality teachers at these previously low-performing schools? If effective teachers value collaboration with high-quality peers, they may not leave if the stipends are eliminated, even if they would not have moved to ACE schools initially in the absence of the program. Alternatively, if time-invariant factors such as location or a high poverty rate among students hinder efforts to attract and retain educators, elimination of the stipends could lead to an increase in turnover that comes disproportionately from highly effective teachers due to the larger decline in compensation and their better alternative opportunities. The elimination of stipends and other ACE 1 program components at 3 of the 4 ACE 1 schools provides an opportunity to investigate these issues.

Table 7 provides a detailed description of how teacher retention changes following the elimination of effectiveness-determined stipends. In describing the destinations, we differentiate between leaving Dallas ISD, moving to a non-ACE Dallas school and moving to a new school that provides the ACE stipends. Only ACE 1 schools lose the program in 2019, so this last outcome refers to teachers moving to either ACE 2 schools or a new wave of ACE 3 schools.

Panel A shows that the end of ACE 1 leads to a large increase in turnover, with 44% of teachers leaving ACE 1 after 2018 compared to 22% leaving after 2017. The detailed analysis of destinations demonstrates that this increase is entirely driven by school-to-school transitions, as exit from the district is actually lower in 2018 compared to 2017. Perhaps most striking, the second row of Panel A shows that a disproportionate share of teachers leaving ACE 1 schools in 2018 move to a school that is still providing ACE stipends. In the control schools, only 3% of teachers leave in 2018 to go to a school that still has ACE stipends, reflecting the fact that relatively few schools have these stipends. But at ACE 1 schools, 20% of teachers leave ACE 1 to go to a new school that still provides ACE stipends, consistent with teachers following the stipends. As a percent of exiting teachers, 44% of exits from ACE 1 schools and only 9% of exits from control schools move to a school offering stipends.

Panel B shows clear evidence that the increase in attrition at ACE schools is driven by highly rated teachers. Attrition actually declines for teachers in the bottom two effectiveness categories (low level), from 25% to 10%. In contrast, attrition increases substantially for teachers in the highest level, rising from 13% to 45%. There are relatively few teachers leaving from the lowest rating level, but those that do all move to a school that provides stipends. For the higher two rating groupings, the proportion who move to a stipend-providing school is similar at 20% for the proficient group and 18% for the high-level group.

Panel C examines transitions separately by experience. Teachers exiting from ACE 1 schools in 2019 disproportionately move to stipend-providing schools at every experience level, but the pattern is particularly striking for highly experienced teachers. Following the elimination of stipends at ACE 1 schools, 28% of highly experienced ACE 1 teachers move to a stipend-providing school, whereas only 3% of highly experienced control teachers move to a stipend providing school. 68% of highly experienced teachers leave ACE 1, whereas less than 10% of teachers who leave control schools.

The stark differences in exit patterns across effectiveness levels provides clear evidence that high quality teachers disproportionately leave following the elimination of ACE 1 stipends. If the exiting teachers were replaced with similarly high-quality teachers, this need not result in a net decline in quality, but descriptive statistics on the replacement teachers suggests that they are lower quality. Among teachers moving to ACE 1 schools in 2019, only 6% were rated highly at their previous school, 24% were rated proficient or lower at their previous school and 70% were new to the district and therefore have no past rating. This is in sharp contrast to the teachers exiting ACE 1, a majority of whom were highly rated in 2018. Though we do not have baseline quality measures for those new to the district, we know that the replacement teachers are considerably less experienced than the exiting teachers; 68% of replacement teachers having fewer than 5 years of experience, roughly double the rate for exiting teachers. One of the most consistent findings of prior research is that new teachers on average significantly improve their effectiveness over the first few years of teaching (Hanushek and Rivkin 2012; Jackson, Rockoff, and Staiger 2014; Bacher-Hicks and Koedel 2023). This reversion to new teachers in the ACE 1 schools mirrors a common situation for low achieving, high poverty urban schools (e.g., Lankford, Loeb, and Wyckoff 2002).

## *8. Cost-benefit analysis*

To gauge whether the improvements in test scores justify the cost of the ACE program, we conduct a back-of-the-envelope cost-benefit analysis. This is meant to be approximate because measuring the full general equilibrium costs and benefits is beyond the scope of the analysis. In addition to the cost-benefit analysis, we also examine the cost efficacy of ACE relative to a class-size reduction policy and relative to a tutoring policy. The cost-efficacy analysis has the advantage of requiring fewer assumptions since it does not require assigning a monetary value to the improved test scores.

The cost-benefit analysis is simplified along a number of important dimensions, some of which are expected to exaggerate cost efficacy and some of which are expected to understate cost efficacy. First, in calculating costs, we measure only direct costs, ignoring potential secondary costs such as negative spillover effects on schools that lose teachers to an ACE school. Second, in measuring benefits, we focus only on test score benefits, ignoring potential improvements to non-cognitive skills. Third, we consider the costs and benefits of a single year of implementation rather than the cumulative effect of many years of implementation. Fourth, rather than attempting to separately value the math and reading score improvements, we simplify and count the ACE benefit as the average effect across math and reading. Finally, in estimating the economic value of test score improvements, we only estimate the expected increase in earnings, ignoring other benefits such as reduced crime (Deming 2011).

Dallas conducted a cost analysis of the ACE program in 2016 that estimates the total cost of implementing ACE to be $1,197 per student, with $1,019 going directly to staff stipends. We estimate that the program results in approximately 0.5 SD higher math scores and 0.25 SD higher reading scores, so the average test score improvement is 0.375 SD. Per thousand dollars in per-student spending, the ACE intervention therefore increases average tests scores by 0.313 SD. In order to conduct cost-benefit analysis, it is necessary to place a monetary value on the 0.313 SD improvement caused by a $1,000 per student increase in spending. Chetty, Friedman and

Rockoff (2014) estimate that a 0.144 SD improvement in test scores (induced by a 1 SD improvement in teacher quality) results in an approximate $7,000 increase in the lifetime present value of earnings ($7,692 in 2016 dollars). Assuming that test score improvements induced by ACE have a similar long-term effect as test score improvements related to variation in teacher value added, this implies that the discounted lifetime earnings improvement caused by $1,000 spent on ACE is approximately $16,719. This earnings benefit is much larger than the $1,000 cost, but we emphasize that this cost-benefit analysis is only suggestive given the various simplifications and limitations discussed above.

To compare the cost efficacy of the ACE intervention to the cost efficacy of class-size reductions, we rely on cost estimates from Harris (2009), adjusting for inflation. Harris (2009) estimates that a reduction from 22 students per class to 15 students per class costs $1,859 per student in 2016 dollars. Baseline class sizes in Dallas are approximately 22 students in early grades, so a class size reduction of this type is a relevant counterfactual policy. Estimates of the benefit of reducing class size vary across studies and we consider both estimates from the STAR experiment as estimated by Schanzenbach (2006) and estimates from Rivkin, Hanushek and Kain (2005). The Rivkin et al. (2005) estimates vary by grade level, but based on their average estimate, a 7-student reduction in class size (22 to 15) would be expected to improve elementary student test scores in by approximately 0.05 standard deviations. STAR experiment estimates from Tennessee suggest that reducing class size from 22 to 15 improves test scores by approximately 0.15 SD (Schanzenbach 2006). Using the Rivkin et al. (2005) class size benefit estimates, a thousand dollar increase in per-student spending increases test scores by approximately 0.027 SD and using the STAR experiment estimates, a thousand dollar increase in per-student spending increases test scores by approximately 0.08 SD. Regardless of which estimates we use, the ACE intervention appears to be much more cost effective than class size reductions as it increases test scores by 0.313 SD per $1,000.

Another useful benchmark for cost efficacy is high-dosage tutoring, which is known to be among the most effective interventions for improving academic achievement. Guryan et al. (2023) provides estimates of both costs and benefits for a tutoring intervention aimed at high school students in Chicago. Their intervention costs approximately $3,500 per student and generates math score improvement ranging from 0.18 to 0.4 SD. Even using the upper range of efficacy, the cost efficacy of ACE is well above the cost efficacy of tutoring (0.313 SD per

$1000 for ACE versus 0.11 per $1000 for tutoring). Again, this is only intended as suggestive since it does not account for the variety of other benefits that ACE and tutoring likely provide, nor does it account for the fact that the SD units are likely different across studies.

## *9. Summary and Conclusions*

Improving achievement in low-performing urban schools has been identified as a high priority for education policy. Yet, a range of tactics have been employed over the past half century with limited overall success. Dallas ISD addressed this challenge by using the information produced by its evaluation and compensation reforms as the basis for effectiveness-adjusted payments that provided the compensating differentials to attract and retain effective teachers in the lowest achievement schools. By adopting a rigorous screening process included information produced by the Dallas ISD multi-measure evaluation system, the district leveraged the supply increase induced by the targeted salary increases. Roughly 80 percent of the teachers and all principals were replaced at the schools in each ACE wave, and the treated schools saw dramatic achievement increases.

Given its success, a critical question is whether the ACE program can be scaled. One form of scaling would be to increase the proportion of ACE schools within a district, and we expect that there are limits to how much the program can be expanded on this dimension. Under the plausible assumption that the supply of high-quality teachers is more elastic to specific schools than to the district as a whole, increasing the proportion of ACE schools should eventually dampen the benefits of the program. A second form of scaling would be to implement the ACE program (along with the complementary teacher evaluation system) in additional districts. Though we have no direct evidence on whether the program will be as successful in other districts, scaling on this dimension need not lead to lower program efficacy assuming a relatively small fraction of each district receives ACE-like programming and limited spillovers across districts.

The program as implemented brought the average achievement in the previously lowest performing schools close to the district average. But doing so required extra funds to pay the compensating differentials that appear to be key to attract and retain highly effective teachers at the ACE schools. When the stipends paid to ACE 1 educators were largely removed following the achievement increases, turnover jumped among the most effective teachers and test scores fell substantially. The full cost of a targeted salary program such as ACE depends upon the responsiveness of effective teachers to targeted salary increases in treated schools and the salary increases necessary to restore teacher quality to its pre-program level in nontreated schools. Regardless, it remains the case that the ACE program achieves objectives that have been key policy goals for decades. Specifically, it manages to attract high quality educators to the lowest quality schools and in doing so, it transforms educational outcomes at those schools. This has positive distributional effects, and it demonstrates that it is possible to create meaningful rapid improvements in high-poverty urban schools within the context of the traditional public school system.

**References**


Bacher-Hicks, Andrew, Cory Koedel (2023). Estimation and Interpretation of Teacher Value Added in Research Applications. In *Handbook of the Economics of Education, Vol 6*, edited by Eric A. Hanushek, Stephen Machin, Ludger Woessmann: Elsevier: 93-134.

Bartanen, Brendan & Husain, Aliza N. & Liebowitz, David D. (2024). Rethinking principal effects on student outcomes," *Journal of Public Economics*, 234(C),

Bates, M., Dinerstein, M., Johnston, A. C., & Sorkin, I. (2025). Teacher labor market policy and the theory of the second best. *The Quarterly Journal of Economics*, 140(2), 1417-1469.

Bietenbeck, Jan, Matthew Collins (2020). New Evidence on the Importance of Instruction Time for Student Achievement on International Assessments Working Paper No. 2020:18. Department of Economics: Lund University (September).

Boyd, Don, Pam Grossman, Hamilton Lankford, Susanna Loeb, James Wyckoff (2006). How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement. *Education Finance and Policy* 1 (2): 176-216.

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and policy analysis*, *33*(3), 378-398.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American economic review*, *104*(9), 2633-2679.

Clotfelter, Charles T., Elizabeth Glennie, Helen F. Ladd, Jacob L. Vigdor (2008). Would Higher Salaries Keep Teachers in High-Poverty Schools? Evidence from a Policy Intervention in North Carolina. *Journal of Public Economics* 92 (5-6): 1352-1370.

Clotfelter, Charles T., Helen F. Ladd, Jacob L. Vigdor (2011). Teacher Mobility, School Segregation, and Pay-Based Policies to Level the Playing Field. *Education Finance and Policy* 6 (3): 399-438.

Cowan, James, Dan Goldhaber (2018). Do Bonuses Affect Teacher Staffing and Student Achievement in High Poverty Schools? Evidence from an Incentive for National Board Certified Teachers in Washington State. *Economics of Education Review* 65: 138-152.

Dee, Thomas (2012). School Turnarounds: Evidence from the 2009 Stimulus. NBER Working Paper No. 17990. Cambrdige, MA: National Bureau of Economic Research (April).

Dee, Thomas S., James Wyckoff (2015). Incentives, Selection, and Teacher Performance: Evidence from Impact. Journal of Policy Analysis and Management 34 (2): 267-297.

Dougherty, Shaun M., Jennie M. Weiner (2019). The Rhode to Turnaround: The Impact of Waivers to No Child Left Behind on School Performance. Educational Policy 33 (4): 555-586.

Drange, N., & Sandsør, A. M. J. (2024). The effects of a free universal after-school program on child academic outcomes. *Economics of Education Review*, *98*, 102504.

Figlio, David, Susanna Loeb (2011). School Accountability. In Handbook of the Economics of Education, Vol. 3, edited by Eric A. Hanushek, Stephen Machin, Ludger Woessmann. Amsterdam: North Holland: 383-421.

Fryer, Roland G., Jr. (2014). Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments. Quarterly Journal of Economics 129 (3): 1355–1407.

Glazerman, Steven, Daniel Mayer, Paul Decker (2006). Alternative Routes to Teaching: The Impacts of Teach for America on Student Achievement and Other Outcomes. Journal of Policy Analysis and Management 25 (1): 75-96.

Glazerman, Steven, Ali Protik, Bing-ru Teh, Julie Bruch, Jeffrey Max (2013). Transfer Incentives for Highperforming Teachers: Final Results from a Multisite Randomized Experiment, NCEE 2014-4003. Washington, DC: U.S. Department of Education (November).

Goodman-Bacon, Andrew (2021). Difference-in-Differences with Variation in Treatment Timing. Journal of Econometrics 225 (2): 254-277.

Guryan, J., Ludwig, J., Bhatt, M.P., Cook, P.J., Davis, J.M., Dodge, K., Farkas, G., Fryer Jr, R.G., Mayer, S., Pollack, H. and Steinberg, L. (2023). Not Too Late: Improving Academic Outcomes among Adolescents. *American Economic Review*, *113*(3), 738-765.

Hanushek, Eric A, Jin Luo, Andrew Morgan, Minh Nguyen, Ben Ost, Steven G. Rivkin, Ayman Shakeel (2023). The Effects of Comprehensive Educator Evaluation and Pay Reform on Achievement. Unpublished manuscript  (March).

Hanushek, Eric A., John F. Kain, Steve G. Rivkin (2004). Why Public Schools Lose Teachers. Journal of Human Resources 39 (2): 326-354.

Hanushek, Eric A., Steven G. Rivkin (2004). How to Improve the Supply of High Quality Teachers. In Brookings Papers on Education Policy 2004, edited by Diane Ravitch. Washington, DC: Brookings Institution Press: 7-25.

Hanushek, Eric A., Steven G. Rivkin (2010). Generalizations About Using Value-Added Measures of Teacher Quality. American Economic Review 100 (2): 267-271.

Hanushek, Eric A., Steven G. Rivkin (2012). The Distribution of Teacher Quality and Implications for Policy. Annual Review of Economics 4: 131-157.

Hanushek, Eric A., Steven G. Rivkin, Jeffrey C. Schiman (2016). Dynamic Effects of Teacher Turnover on the Quality of Instruction. Economics of Education Review 55: 132-148.

Harris, D. N. (2009). Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. Educational Evaluation and Policy Analysis, 31(1), 3-29.

Heß, S. (2017). Randomization inference with Stata: A guide and software. *The Stata Journal*, *17*(3), 630-651.

Heissel, Jennifer A., Helen F. Ladd (2018). School Turnaround in North Carolina: A Regression Discontinuity Analysis. Economics of Education Review 62: 302-320.

Jackson, K. and E. Bruegmann. 2009. Teaching students and teaching each other: The importance of peer learning for teachers. American Economic Journal: Applied Economics, 1(4), 85–108.

Jackson, C. Kirabo, Jonah E. Rockoff, Douglas O. Staiger (2014). Teacher Effects and Teacher Related Policies. *Annual Review of Economics* 6: 801-825.

Kane, Thomas J., Jonah E. Rockoff, Douglas O. Staiger (2008). What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City. Economics of Education Review 27 (6): 615-631.

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

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of educational research*, *88*(4), 547-588.

Lankford, Hamilton, Susanna Loeb, James Wyckoff (2002). Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis. *Educational Evaluation and Policy Analysis* 24 (1): 37-62.

Lavy, Victor (2015). Long Run Effects of Free School Choice: College Attainment, Employment, Earnings, and Social Outcomes at Adulthood. NBER Working Paper 20843. Cambridge, MA: National Bureau of Economic Research

Papay, John P. (2015 of Conference). The Effects of School Turnaround Strategies in Massachusetts. Paper presented at APPAM Fall Research Conference, November, at Miami.

Pham, Lam D., Sean P. Corcoran, Gary T. Henry, Ron Zimmer (2025). Do the Effects Persist? An Examination of Long-Term Effects after Students Leave Turnaround Schools. *American Educational Research Journal* 62 (1): 180-213.

Pham, L. D., Nguyen, T. D., & Springer, M. G. (2021). Teacher merit pay: A meta-analysis. American Educational Research Journal, 58(3), 527-566. ChicagoRivkin, Steven G., Eric A. Hanushek, John F. Kain (2005). Teachers, Schools, and Academic Achievement. Econometrica 73 (2): 417-458.

Rivkin, Steven G., Jeffrey C. Schiman (2015). Instruction Time, Classroom Quality, and Academic Achievement. Economic Journal 125 (588): F425-F448.

Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., & Webb, M. D. (2019). Fast and wild: Bootstrap inference in Stata using boottest. *The Stata Journal*, *19*(1), 4-60.

Schanzenbach, D. W. (2006). What have researchers learned from Project STAR?. Brookings papers on education policy, (9), 205-228.

Schueler, Beth E., Joshua S. Goodman, David J. Deming (2017). Can States Take over and Turn around School Districts? Evidence from Lawrence, Massachusetts. Educational Evaluation and Policy Analysis 39 (2): 311-332.

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

Springer, Matthew G., Walker A. Swain, Luis A. Rodriguez (2016). Effective Teacher Retention Bonuses: Evidence from Tennessee. Educational Evaluation and Policy Analysis 38 (2): 199-221.

Steele, Jennifer L., Richard J. Murnane, John B. Willett (2010). Do Financial Incentives Help Low-Performing Schools Attract and Keep Academically Talented Teachers? Evidence from California. Journal of Policy Analysis and Management 29 (3): 451-478.

Strunk, Katharine O., Julie A. Marsh, Ayesha K. Hashim, Susan Bush-Mecenas, Tracey Weinstein (2016). The Impact of Turnaround Reform on Student Outcomes: Evidence and Insights from the Los Angeles Unified School District. Education Finance and Policy 11 (3): 251-282.

Zief, S. G., Lauver, S., & Maynard, R. A. (2006). Impacts of after-school programs on student outcomes. Campbell Systematic Reviews, 2(1), 1-51.

Zimmer, Ron, Gary T. Henry, Adam Kho (2017). The Effects of School Turnaround in Tennessee's Achievement School District and Innovation Zones. Educational Evaluation and Policy Analysis 39 (4): 670-696.