A BACKGROUND PAPER FOR THE HOOVER EDUCATION SUCCESS INITIATIVE

A Consideration of Educator Evaluation and Compensation Reform

STEVEN RIVKIN, BEN OST, ANDREW MORGAN, AND MINH NGUYEN

The divergence between evidence on educator quality and labor markets on the one hand and the structure of most compensation and systems on the other raises the possibility that evaluation and compensation reform could substantially raise the quality of instruction, particularly in disadvantaged schools. Most districts offer extra teacher pay for the possession of an MA degree despite extensive evidence that it is not associated with higher-quality instruction; most districts raise teacher pay with experience even beyond those early career years in which increases in the quality of instruction are concentrated; few districts offer higher pay to teachers working in chronically low-performing schools despite difficulties attracting and retaining educators; rigorous evaluation based on supervisor observations and student outcomes has not been the norm despite evidence of the value of feedback for teacher growth.[1]

So why does the salary scale based on education and experience persist in most districts? Some point to politics and the strength of unions and education associations, while others focus on the difficulties inherent in getting the incentives right.[2] Whatever the reason, it is challenging to design an effective personnel system that supports student growth in ways that raise longer-term education, labor market, and social outcomes. In a nutshell, educator and program effectiveness must be measured in real time despite the primacy of outcomes in the longer term. This highlights the importance of a thoughtful approach to reform that incorporates multiple measures of educator performance and mitigates unproductive behavior, including narrow teaching to a test and cheating. But it in no way justifies the status quo.

Opposition to higher pay in low-performing, disadvantaged schools that struggle to attract and retain teachers is similarly puzzling, particularly as those against differentiated compensation often rail against large achievement gaps. Perhaps this simply reflects opposition to any deviation from the traditional salary schedule. Regardless, in designing a targeted pay program to support children in low-performing schools, it is important to understand the impediments to attracting and retaining more effective educators. These include not just poverty or racial composition, as evidenced by substantial variation in school performance across schools in high-poverty areas.

Fortunately, there is much to learn from the recent evaluation and compensation reforms implemented by the Washington, DC, and Dallas Independent School Districts (Dallas ISD), and these will be the focus of much of this paper. They take seriously the incentive

and measurement issues in program design and implement modifications as evidence on program effectiveness emerges. Given the richness of the personnel systems and key role for educator judgment, elevating the quality of school leadership is of primary importance.

Prior to discussing the IMPACT and Dallas ISD programs, we present selective evidence on evaluation, incentive pay, and targeted pay for schools serving disadvantaged children. The studies highlight both the potential gains from reform and pitfalls that may arise, and provide the evidence necessary for improvements in program design and operations. The final section of the paper summarizes the findings and presents some key issues to consider.

Selected Evidence on Evaluation and Pay Reform

Most public school districts use teacher salary schedules based largely on experience and highest degree, largely ignoring differences in effectiveness and variation in the pool of available teachers by subject and school characteristics. Salary compression may discourage more productive teachers from working in the schools and may provide few incentives to educators (Hoxby and Leigh 2004; Leigh 2013). Educator evaluation appears not to have been very robust in many districts, particularly those in which performance does not influence salary (Weisberg et al. 2009). The absence of rigorous evaluation may adversely affect the quality of instruction regardless of the compensation structure, as evidence indicates that rigorous evaluation based on supervisor observations elevates the quality of instruction as measured by student achievement. Taylor and Tyler (2012) find that observation-based evaluation and support of midcareer teachers significantly raised math achievement in the year of the evaluation and even more so in subsequent years. Steinberg and Sartain (2015) find that observation-based evaluation using the Danielson Framework increased mathematics and reading achievement in a randomized controlled pilot study in the Chicago Public Schools; although positive, the estimated effects on math achievement were not significant. These findings contrast with the far less positive evidence on the benefits of conventional professional development programs.[3]

States and school districts have adopted a variety of evaluation and compensation reforms designed to strengthen the relationship between performance, typically measured by student achievement, and compensation. No Child Left Behind (NCLB) required states to develop longitudinal data sets that could be used to measure educator or school effectiveness, and the expanded availability of and access to achievement data facilitate the use of test-score-based compensation programs. The expanded data availability mitigates concerns highlighted by Murnane and Cohen (1986) about the lack of reliable measures of performance, though Podgursky and Springer (2007) point out that concerns remain over the attribution of achievement outcomes to educators or schools. Critics of pay-for-performance programs continue to raise doubts about their appropriateness in education due largely to the multidimensional nature of teaching. Reward structures may introduce perverse incentives, including a focus on particular students or subjects at the expense of others, teaching to the test in a way that compromises deeper learning, an overemphasis

on achievement at the expense of the production of behavioral or socio-emotional skills, cheating, or the exclusion of low-performing students from the tested sample.[4]

The evidence on perverse incentives illuminates the importance of test-instrument quality and the related reward structure. A large benefit to memorization can induce poor education practices in terms of deeper learning. In addition, an incentive system that focuses on the crossing of a single threshold, such as pass rate, can distort the allocation of resources and effort through an overemphasis on students near the cutoff.

A growing body of literature examines the effects of performance pay, where performance is measured by student achievement. In their 2007 review, Podgursky and Springer conclude that the few studies available generally find positive effects on test scores, but evidence also suggests that strategic behavior may dampen the benefits of such programs. They write that "education policymakers need to be careful in designing such programs, and must expect to continually refine the programs as they learn about behavioral responses." More recent work produces generally less favorable results for the efficacy of incentives. Fryer (2013) finds little evidence of effects of relatively small group incentives on either achievement or teacher behavior; Marsh et al. (2011) produce similar findings. Springer et al. (2011) report little or no evidence that a group incentive program raises achievement or alters practice, and Goldhaber and Walch (2012) find little or no evidence that teachers participating in the Denver PROCOMP incentive program produced larger test-score growth than nonparticipants. Similar results emerge for individual incentives. In an experimental study of Nashville middle school mathematics teachers, Springer et al. (2010) find only limited evidence that payments for teachers who produce large average growth relative to historic standards raise achievement; the hypothesis of no significant achievement effect is not rejected in most grades and years. In contrast to these results, Lavy (2009) finds significant effects of performance incentives structured as a tournament in high school matriculation scores in Israel. Evidence indicates that changes in teaching methods, enhanced after-school teaching, and increased responsiveness to students' needs all contributed to the improvement.

Much of the cited evidence comes from incentive pay experiments rather than district-wide interventions, which are more difficult to evaluate because of the absence of randomly assigned treatment and control schools or teachers. Importantly, the experimental programs tend to be limited in scope and duration, focusing mostly on raising the effort of existing teachers, while broader, longer-lasting reforms would be expected to stimulate teacher development and alter the composition of teachers by making the district more attractive to more effective teachers. The comprehensive district evaluation and compensation reforms undertaken by both Washington, DC (IMPACT) and Dallas ISD (TEI and PEI) would be expected to affect the quality of instruction through all of these channels. IMPACT incorporates a much more developed multiple-measure and high-stakes evaluation structure than virtually all other teacher incentive programs, and the Dallas ISD evaluation and compensation reforms, Principal Excellence Initiative (PEI) and Teacher Excellence

Initiative (TEI) go even farther than IMPACT in integrating the evaluation of teachers and administrators and assessing teachers on the basis of student work.

Dallas ISD also introduced a program that offers substantial stipends to effective educators willing to work in chronically low-performing schools. Prior evidence suggests that additional pay can increase retention, but supplemental pay for working in disadvantaged schools (which does not typically depend upon effectiveness) is unlikely to have a major impact on the quality of instruction.[5] Clotfelter, Ladd, and Vigdor (2011) argue that these types of pay premia are unlikely to equalize teacher quality across advantaged and disadvantaged schools because they differentially attract teachers with worse credentials. Conversely, a recent randomized controlled trial provides evidence that a program which paid effective educators $10,000 per year for two years succeeded in attracting small numbers of high-value-added teachers to designated schools and modestly raised achievement (Glazerman et al. 2013). This intervention led to a situation in which recruits receive much higher pay than existing teachers who might be equally effective. Although it asked the important question of whether highly effective teachers would continue to be effective after a move to a lower-performing, higher-poverty environment, it was far less ambitious than an intervention that used stipends to overhaul school leadership and the teaching force in a coordinated effort. Given evidence on the benefits of having strong peer teachers, the hiring of effective teachers would be expected to raise performance across the school and potentially reduce the difficulties of hiring and retaining effective teachers (Jackson and Bruegmann 2009). Whether financial inducements to qualified educators could be implemented at scale and transform low-performing schools is a fundamental question, and the Washington, DC, and Dallas ISD reforms constitute just such transformative interventions.

Washington, DC: IMPACT Reform

Beginning in the 2009–10 academic year, District of Columbia Public Schools (DCPS) undertook an ambitious program to reform the way educators are evaluated, retained, and compensated. DCPS's IMPACT program evaluates and rewards educators based on multiple measures of performance. Teachers are assigned points and ratings based on classroom observations, adherence to DCPS policy, and their own student-achievement growth and survey responses. IMPACT also serves to provide feedback and assist teachers in maintaining better professional practices and growth.

Since its outset, IMPACT had the mission of retaining and rewarding teachers who were determined to be the most effective and dismissing teachers who were found to be least effective. This initiative, unsurprisingly, came with significant political costs. The chancellor who instituted the policy, Michelle Rhee, resigned from her post shortly after the mayor who appointed her, Adrian Fenty, lost his reelection bid in what was widely seen as a referendum on DCPS policy. There was also substantial pushback from the Washington Teachers' Union, particularly after the initial round of termination for teachers rated ineffective (Lewin 2010). Allegations of widespread teacher cheating as a result of IMPACT also arose during the program's early years, though independent investigators hired by DCPS found only

a very limited number of suspicious incidents (Brown, Strauss, and Stein 2018). Despite the polarizing nature of the program and formal opposition by the Washington Teachers' Union, IMPACT remains in DC public schools through the 2019–20 school year.

While the program has seen considerable changes during its tenure, the core components, consisting of multiple classroom observations for teachers during the school year and metrics for student achievement growth, have remained the same. The crux of the policy hinges on the threat of dismissal for the lowest-performing teachers, as well as considerable bonus pay for the highest-performing teachers. Teachers are rated on a scale ranging from one hundred to four hundred points, with differing weights on each IMPACT component based on subject and grade taught. The first main component for all teachers is a series of classroom observations throughout the school year, almost always done now by an administrator at the school. Evaluators use a set rubric to assign points from observations, which evaluate practices such as how well teachers do in keeping students engaged and how well teachers design content that is both rigorous and clearly conveyed. The second main component is teachers' value added to standardized student test scores and is only available for teachers in grades 4 through 8 who teach math or reading. Teachers are also assigned points based on student surveys and how much they contribute to their school's environment and keep professional practices.

Points are then totaled and teachers are assigned one of five ratings based on the number of points. Each rating carries with it either a threat of dismissal or a change in salary based on union affiliation and prior rating history. Teachers rated in the lowest category, "ineffective," are subject to termination at the end of the school year in which they receive that rating, regardless of union affiliation. Teachers who are "minimally effective" or "developing" receive no salary change and are subject to termination if they receive these ratings for two and three consecutive years, respectively. "Effective" teachers have traditional salary assignment and growth based on education and experience and are not subject to automatic termination, and members of the Washington Teachers' Union may be eligible for additional pay if they work in a high-poverty school. Teachers in the highest category, "highly effective," also proceed normally along the pay scale, and union members who get this rating may receive bonus pay, depending on the poverty level and student achievement level of the school in which they work.

For highly effective teachers, additional pay can be substantial. All highly effective teachers receive minimum bonus pay of $2,000 in the 2019–20 school year. Bonus pay then increases for teachers in high-poverty schools, in schools the district deems lowest performing, and for those teaching subjects that count toward standardized testing metrics. If a teacher meets all of these criteria, bonus pay can max out at $25,000 per year. Additionally, teachers meeting highly effective status for multiple years can make the additional pay permanent in the form of applied extra service credit or movement to a higher education tier.

Although overall student achievement in Washington, DC, has been increasing since IMPACT's inception in the 2009–10 school year, there is little causal evidence on the overall effect of the program. Notably, however, there has been work looking into how

IMPACT affected teacher turnover in the district and how teacher turnover affects student achievement. Dee and Wyckoff (2015) and Adnot et al. (2017) present evidence that differential quality in teacher entrances and exits as a result of the program raised student achievement. Using a regression discontinuity design, Dee and Wyckoff (2015) show that exits of lower-quality teachers and retention and improvement of more effective teachers increased teacher quality in the district. Teachers who scored just below the effective rating, and were thus subject to a higher automatic dismissal potential, were 11 percentage points less likely to be retained in DCPS relative to those just above. Additionally, teachers who scored just above the highly effective threshold were more likely to be retained by an insignificant 3 percentage points. Teachers who remained in DCPS around both thresholds also had higher IMPACT scores in the following school year by about a quarter of a standard deviation of current-year IMPACT scores. Although these effects suggest the potential for improvement, they only hold for the teachers who are around the threshold. Additionally, it is unclear how overall student achievement is affected by the higher teacher turnover rate. Adnot et al. (2017) attempt to answer this by employing a difference-in-differences design, comparing changes in achievement in a school-grade cell where teacher turnover occurred to school-grade cells where no turnover occurred. Teacher turnover is defined separately by effectiveness rating, where high turnover induced by ineffective or minimally effective ratings is believed to raise teacher quality and achievement. Conversely, higher turnover of highly effective teachers would suggest lower teacher quality and student achievement. Their estimates suggest that teachers who teach math and exit are replaced by teachers whose IMPACT scores are about 0.33 standard deviations higher on average, and those who teach reading and exit are replaced by teachers whose IMPACT scores are roughly 0.08 standard deviations higher on average. Student achievement increases by an insignificant 0.08 standard deviations in math and 0.05 standard deviations in reading. Their estimates also imply that higher rates of turnover among highly effective teachers insignificantly decrease student achievement in both math and reading. Higher exit rates for less effective teachers seem to increase student achievement by about 0.2 standard deviations in math and 0.14 standard deviations in reading.

Principal and Teacher Excellence Initiatives in Dallas ISD

IMPACT incorporates a much more developed multiple-measure and high-stakes evaluation structure than virtually all other pay-for-performance programs, and the Dallas ISD evaluation and compensation reforms, Principal Excellence Initiative (PEI) and Teacher Excellence Initiative (TEI), go even farther than DC in integrating the evaluation of teachers and administrators and assessing teachers on the basis of student work. Dallas developed numerous assessments for teachers in subjects and grades not included in the state-standardized testing protocol. Moreover, these programs take additional steps to mitigate potential adverse effects of high-stakes evaluation systems, incorporate effects on achievement gaps explicitly into principal evaluation, and completely eliminate the salary steps based on experience and education for administrators and teachers. Following descriptions of TEI and PEI, I report some preliminary evidence on test-score trends following the reform.

TEI and PEI Overview

Dallas ISD introduced the Principal Excellence Initiative (PEI) during the 2012–13 academic year and the Teacher Excellence Initiative (TEI) during the 2014–15 academic year. Though they differ in many details, the two reforms share a similar structure. Each contains an achievement component based on student outcomes, a performance component based largely on supervisor observations, and a survey component based on feedback from students or families; the multiple dimensions, including feedback from students and families, broaden the scope of the evaluation far beyond narrow achievement outcomes. With some qualifications, the two-year average evaluation score determines the salary bin for all educators except those in their first year in the district. We discuss TEI first and then highlight differences for PEI.[6]

After three years of discussion and development, the Teacher Excellence Initiative was approved by the DISD Board of Trustees in May 2014 for evaluating and compensating more than eleven thousand teachers in Dallas between 2014 and 2015. It replaced the Dallas Professional Development and Appraisal System, which used years of service and college degrees as the main determinants of salary and had been in place for twenty-two years. TEI dramatically alters the evaluation and compensation structures by requiring schools to collect far more information about teachers, to use the information for both assessment and professional development, and with the exception of the first year in the district, to determine compensation on the basis of the evaluations.

Such a comprehensive system has the potential to raise the quality of leadership and instruction through more favorable selection into and out of the district, provide more effective support for improvement and professional development, and strengthen work incentives. Development of assessments for most subjects and grades and incorporation of components to mitigate pressures on principals to inflate teacher evaluations should produce more accurate and informative evaluations that treat educators similarly. Explicit incentives to reduce achievement gaps embed equity into the reform, and the inclusion of student or family survey responses broadens the scope of the evaluation and empowers stakeholders.

TEI activities can be categorized in three components—Defining Excellence, Supporting Excellence, and Rewarding Excellence—and each plays an important role in achieving the DISD goals.

Defining Excellence

Defining Excellence refers to the district's vision for effective teaching and teaching evaluation, in which the principal conveys the school goals to teachers as part of goal setting. In September, the principal or another designated supervisor meets with each teacher to establish student-learning objectives and a professional development plan. Performance, achievement, and perception constitute the three components of the evaluation system. Table 1 lists the

domains and indicators within each domain that compose the teacher performance rubric. Teachers are scored for their performance on each indicator. Each teacher is assigned a primary evaluator, usually a principal or assistant principal. The evaluator monitors and collects evidence to assess the teacher performance, mainly through spot, extended, and informal observation. TEI specifies ten ten- to fifteen-minute spot observations of each teacher and one forty-five-minute extended observation per year. The observations focus on domains 2 and 3, instructional practice and classroom culture. The supervisor is required to provide written feedback following all observations and conference with the teacher following the extended observation. Artifacts and informal observations also contribute to the performance score, as these constitute the evidence of performance for the first and fourth domains.

The student perception score comes from a student survey conducted in the second week of April. Most students in grades 3–12 complete two surveys, one online and one on paper. Results from the survey are summarized by a statistic for teachers with sufficient student responses. Points are assigned based on the target distribution at grade level to ensure equity, because early grade-level students tend to provide more positive responses.

Table 1. Teacher performance rubric

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

Source: Dallas ISD 2014a and Miles 2015.

The achievement component includes both school average achievement and classroom achievement, except for teachers whose role is not associated with a TEI assessment. All school-level student achievement measures are based on state standardized test results. Teacher-level measures consist of Student Learning Objective (SLO) and Standardized Teacher-Level Student Achievement Measures. SLO is a measure of student improvement during the year based on assessments that are not standardized tests; it is a part of all teachers' templates. Teachers with students who take a standardized test are scored based on classroom achievement. In the cases of both school average and classroom achievement, alternative measures are generated based on the test results, and the highest metric is used to determine the number of achievement points. Initially the alternatives included status (percentage of tests with scores at a specified standard); a value-added measure; and an achievement score relative to the scores of a designated peer group of schools based on prior achievement. There are target distributions for the school and teacher achievement components based on the standardized tests.

The evaluation score equals a weighted sum of points earned on the three components, where the weights depend on the role and grade level. The achievement component receives much lower weight for those whose role is not associated with a district or state assessment, and surveys are not administered to students in second grade or below. Table 2 describes the four categories of teachers and differences among the weights for the three components.

Teachers are divided into ratings categories based on scores and whether an application for recognition as a distinguished teacher is approved, which is a requirement for a rating of

Table 2. Teacher categories and evaluation templates

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

Source: Dallas ISD 2014a.

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

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

Source: Dallas ISD 2014a.

proficient II or higher. Table 3 lists the nine evaluation categories of the initial framework; there have subsequently been some modifications.

Supporting Excellence

Each of the three components of the evaluation system provides information used in teacher support and professional development, and the district works to ensure that the system has sufficient elements to support excellence. In addition to written feedback and conferences following observations, data on student achievement are collected and analyzed to help teachers improve instruction. The district also provides significant support and professional development opportunities for administrators and teachers. Finally, the district developed an aligned curriculum for the whole district and trained teachers in this curriculum.

School leaders and teachers have many ways to access district support. An online bank of videos and modules was developed to facilitate school leaders and instructional coaches in generating a clear and common vision of the TEI program in the system. This resource also allowed teachers to learn about the TEI program and to improve their instruction. There are professional development modules so that teachers can learn how to improve their effectiveness with small teams or in whole groups. Workshops were designed to strengthen campus and content expertise in areas of need. The district also created a mechanism that pairs highly rated teachers with progressing teachers in teaching summer school to raise instructional capacity. With experts and university partnerships, the district also developed a set of academies to support teachers at all levels to grow.

Rewarding Excellence

Except for a teacher in her first year in Dallas, salary is based on the average of evaluation points earned in the most recent two years. This average is divided into the nine effectiveness levels, listed in table 3, conditional on certain constraints. Teachers cannot move up or down more than one effectiveness level per year. Completion of three years of service as a classroom teacher is a necessary condition to be considered for the proficient I level. The proficient II level and above require teachers to go through the Distinguished Teacher Review (DTR) process, and to be at exemplary II, teachers need to have at least one year qualifying as an exemplary teacher. And the master level entails additional requirements. To maintain budget stability and deter evaluation inflation, the category boundaries are

determined by a target distribution that fixes within a narrow range the share of teachers who fall in each category.

The system also includes safeguards to protect teachers against downside risk: 1) it takes three consecutive years in a lower rating category for teacher salary to go down by one level; 2) salaries never go below the 2014–15 level; 3) teachers starting after 2014–15 never receive salaries lower than their entry levels; and 4) the compensation scale is adjusted at least once every three years to keep salary levels competitive.

Principal Excellence Initiative

PEI went into effect for the 2012–13 academic year and is quite similar to TEI. The evaluation includes both performance and achievement components, where the performance component contains information obtained from a family survey as well as other sources. The district devotes substantial resources to build the skills and capacity of principals: principals went through 135 hours of professional development in the school year 2011–12 and 175 hours in 2012–13. As is the case for teachers, principal compensation is determined by the earned effectiveness level except for those in their first year.

A fundamental contrast between IMPACT in Washington, DC, and the Dallas ISD reforms is the use of outside evaluators of teaching in Washington but not in Dallas ISD. On the one hand, outside experts provide alternative perspectives based on observations of teachers across many schools, and such an approach mitigates concerns that it is challenging for school administrators to provide critical feedback given their relationships with teachers and their desire to maintain a positive climate in the school. On the other hand, honest feedback and willingness to receive constructive criticism are important elements of an effective learning community. Finally, the use of outside evaluators may increase costs. Thus, there are pros and cons to each approach.

PEI places substantial weight on effectiveness as an instructional leader. Table 4 lists the metrics used in principal evaluation. Almost 20 percent of the performance component focuses directly on improving teacher effectiveness and congruence between teacher performance and student achievement. Thus, principals are rated on their work in support of teachers and on the alignment between the subjective teacher evaluation and teacher effectiveness at raising achievement. The congruence component of the evaluation is designed to mitigate the tendency to inflate subject evaluations and to deter arbitrary judgments of teachers based on factors other than the quality of teaching. Unlike the case for TEI, attendance and enrollment also contribute to the performance score for principals.

Not surprisingly, the achievement component also differs from that used in TEI, particularly with respect to the tests included and concerns about inequality. Over 10 percent of the achievement score depends on success in reducing achievement gaps by race and ethnicity. This

Table 4. Measuring principal effectiveness: the metrics

	Area	*Points*
Performance (60%)	Performance rubric	30
	System review	10
	Improving teacher effectiveness	5
	Congruence between teacher performance and student achievement	5
	Student enrollment or student attendance	5
	Parent climate survey	5
Achievement (40%)	School STAAR results	10
	Feeder group STAAR results	3
	District common assessments	7
	School achievement gap	5
	College ready rate (HS); 7th-grade writing (MS); 4th-grade writing (ES)	10
	Career ready rate (HS); 8th-grade reading and math (MS); 5th-grade reading and math (ES)	5

Source: Dallas ISD 2013a.

codifies the objective of equity and support for students in demographic groups that have lower average achievement in the district and state.

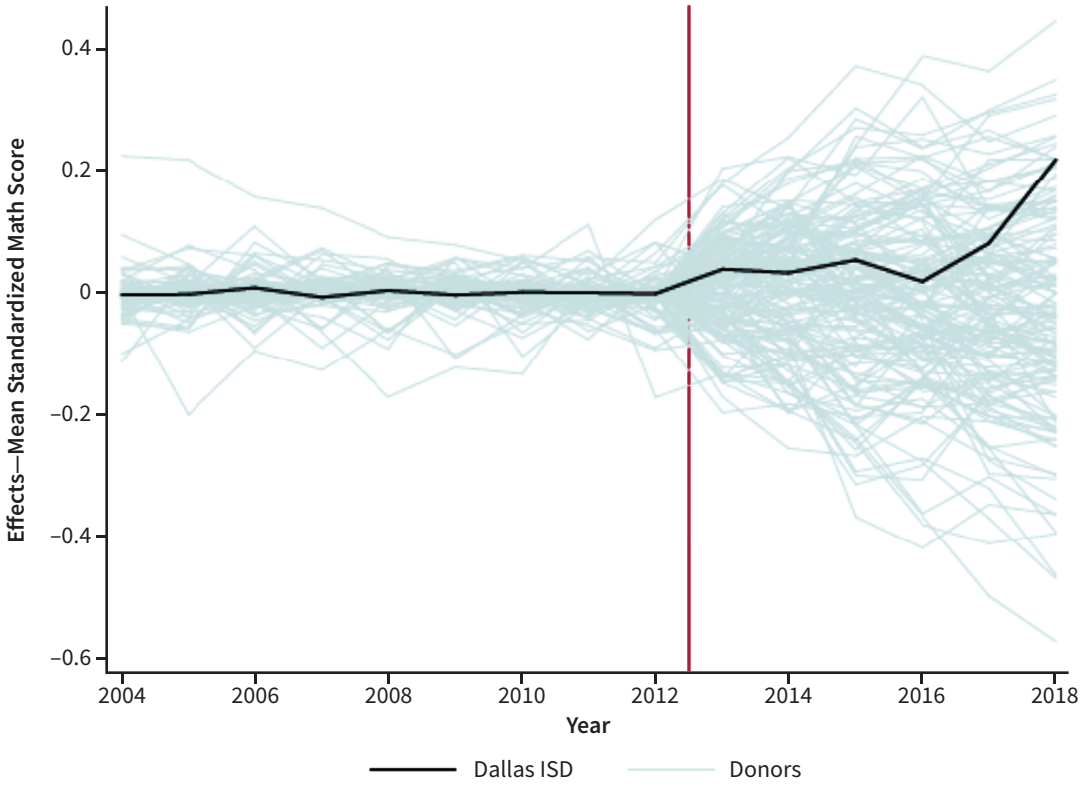
Finally, there are differences in the determination of salaries. Specifically, salaries differ by level: conditional on effectiveness rating, high school principals earn more than middle school principals, who earn more than elementary school principals.

Preliminary Evidence

Estimation of the effects of PEI and TEI on the quality of instruction requires a valid control group and is complicated in the absence of a randomized or quasi-randomized group of control districts. Other school districts adopt myriad programs and approaches, and we want to understand whether the Dallas ISD reforms led to larger improvements in education outcomes than policies and practices embraced by other districts. The special case of a single treated unit (in our case a school district) is quite common, and Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010) make a compelling case that synthetic control estimation is appropriate with a single treated unit. Rather than selecting a comparison group based on observable characteristics such as poverty level or racial composition, the synthetic control method selects weights for the potential control districts to match as closely as possible the pre-period achievement trend. In this case we restrict the potential control districts to those with at least five thousand students. This maintains a large donor pool but removes very small, mostly rural districts from the analysis.

The solid dark line in figure 1, taken from Hanushek et al. (2019), illustrates the synthetic control estimates for Dallas ISD—that is, how math achievement evolved in Dallas ISD relative to the weighted average for the districts that make up the synthetic control. The analysis is an

Figure 1. Synthetic control estimated treatment effects for Dallas ISD and placebo synthetic-control estimates for the other districts in the synthetic control donor pool



Source: Hanushek et al. (2019).

assessment of how Dallas performed following the reforms compared to a synthetic control that was trending similarly to Dallas prior to the reform. The line shows little change following the adoption of PEI, but mathematics achievement in Dallas ISD relative to the synthetic control increases dramatically following 2016, the second year of TEI. By 2018 achievement had increased more than 0.2 standard deviations in Dallas ISD relative to the controls.

Improvement that exceeds 0.2 standard deviations relative to the control is certainly large, but it is important to examine the probability that such outperformance relative to the control could have occurred by chance. To assess this possibility, we follow Abadie, Diamond, and Hainmueller (2010) and use a permutation test for inference. Essentially each district in the donor pool is considered to be a placebo treatment, and a synthetic control estimate is produced for each. The lighter lines represent the placebo effects for each of the other districts, and some exhibit even larger achievement increases than Dallas ISD. Only 10 percent of the districts outperform their synthetic control by more than the 0.2 standard deviations produced by Dallas ISD. This suggests that the Dallas ISD reforms were effective.

We look forward to examining an additional year of data to see if the trend continues. The reforms in Dallas are expected to take time to affect student outcomes both because of the initial disruption and because several key mechanisms are expected to act on a delay. For example, if higher-quality teachers are attracted to Dallas because of the possibility of merit pay increases, this effect will compound over time as these teachers become a larger proportion of the district. Though not definitive, figure 1 suggests that Dallas began improving substantially in 2016 and the rate of increase has been growing since then. The central question for researchers and policy makers interested in the Dallas reforms is whether the rapid improvements of 2016–18 will continue, level off, or reverse in the coming years.

Accelerating Campus Excellence

Particular concern about the lowest-performing schools and early evidence that outcome-based evaluation could exacerbate their difficulties in attracting and retaining effective educators led Dallas ISD to develop and implement the Accelerating Campus Excellence (ACE) initiative, which paid effective educators as measured by PEI and TEI substantial stipends to work in the lowest-performing schools. In academic year 2015–16, one year after TEI adoption, Dallas ISD implemented the ACE program to raise the quality of instruction and achievement in Dallas ISD's chronically low-performing schools. This intervention incorporates several components, but the cornerstone of ACE is the dedication of substantial resources to attract and retain highly effective teachers and leadership teams.[7] Educators who apply and are selected to work at ACE campuses receive signing bonuses of $2,000 and stipends that depend upon position and, in the case of teachers, on TEI effectiveness ratings for the prior year. Stipend amounts equal $13,000 for a principal, $11,500 for an assistant principal, $8,000 for a counselor, $6,000 for an instructional coach, and between $6,000 and $10,000 for teachers. Note that classroom teachers and specialists were eligible for the ACE payments. The ACE program had a total budget of $4,720,200 for the 2015–16 academic year, which came out of general operating funds from Dallas ISD. The signing bonuses and stipends constituted roughly 85 percent of the budget, with the remainder divided among professional development ($350,000), transportation ($246,000), and uniforms ($125,000) for schools that decided to require them.[8]

Based on the target distribution of ratings, approximately 20 percent of Dallas ISD teachers qualify for the $10,000 pay premium by having passed distinguished teacher review, 40 percent of teachers qualify for an $8,000 pay premium by obtaining a proficient rating, and 37 percent qualify for a $6,000 premium by receiving a progressive rating due to either being inexperienced or failing to reach proficiency. In the first year of the ACE program, 40 percent of ACE teachers qualified for a $10,000 stipend, 28 percent for an $8,000 stipend, and 32 percent for a $6,000 stipend. In addition to raising the level of compensation, the structure of these stipends amplifies the TEI pay-for-performance incentive for teachers in an ACE school by increasing the differential between rating categories. For example, at a non-ACE school, moving from the level just below distinguished up to the first rung in the distinguished category raises salary by $5,000. At an ACE school, the same rating change raises salary by $7,000.

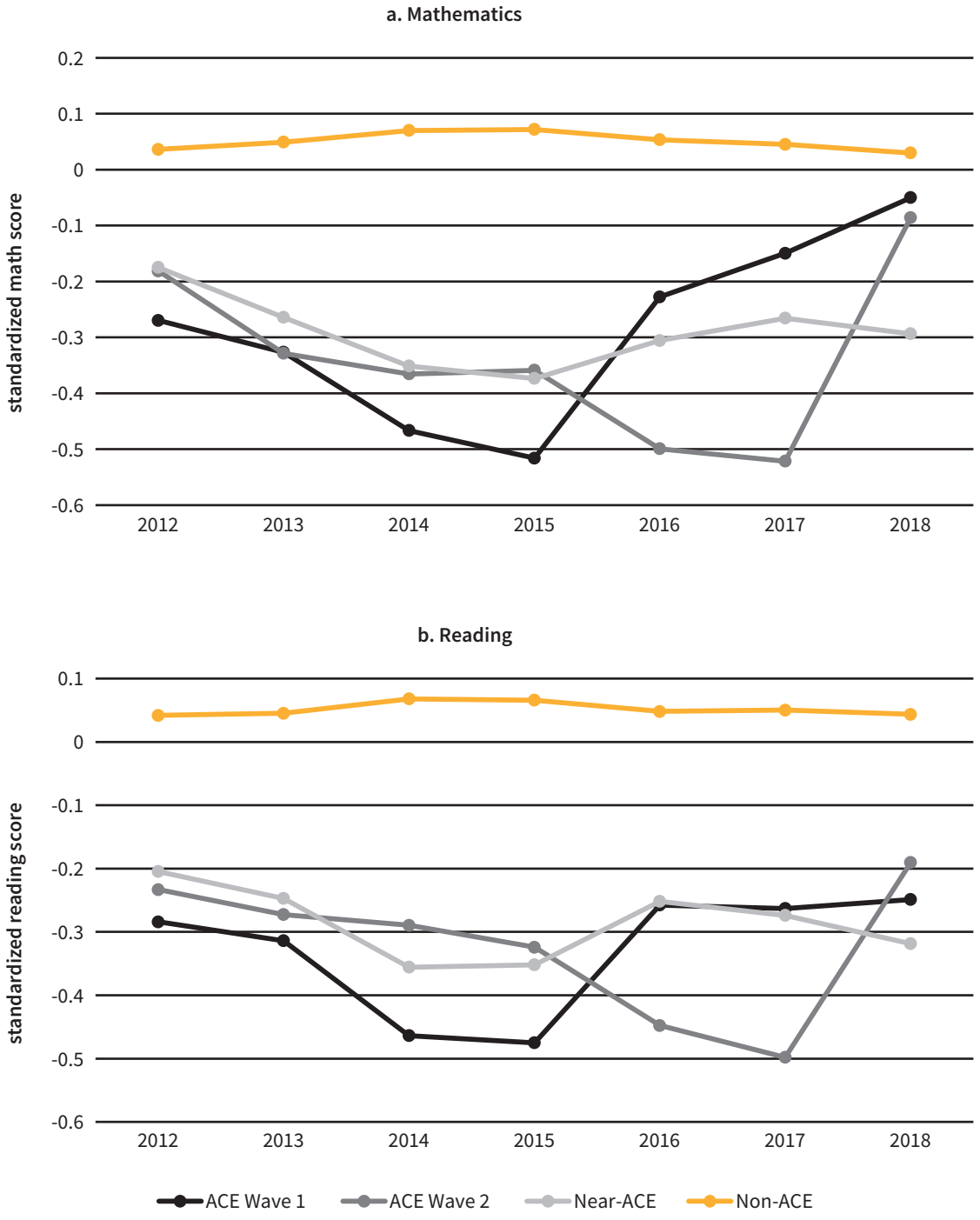
Dallas identified a total of twenty-five low-performing elementary and middle schools that they considered for the ACE intervention but ultimately designated seven as ACE schools in 2015–16 based on persistently low achievement; the remaining eighteen schools were designated as ISN ("near ACE"). Another six schools were selected for the second wave of ACE in 2017–18. All potential ACE teachers (including those who were effective) were required to interview or were evaluated to stay at an ACE campus. Some teachers decided to leave even though offered the opportunity to stay, perhaps in response to the requirement to contribute three hours per week to the after-school program. To the extent possible, campuses were reconstituted with teachers who had earned high evaluation ratings. Over 60 percent of teachers and all principals in schools newly designated as ACE were different from the teachers and principals who had been in the school the previous year.

The analysis of ACE effects and evaluation inflation uses a difference-in-differences approach, and we have begun the analysis of ACE with the early years of data. Figure 2, taken from Morgan et al. (2019), plots mathematics and reading achievement for two cohorts of ACE schools, the near-ACE control group, and the remaining elementary and secondary schools in Dallas ISD. The parallel mathematics and reading achievement trends for the wave 1 ACE and near-ACE schools constitute evidence in support of the common trends assumption, and the large and growing effects provide preliminary evidence in support of program effectiveness; the absence of parallel trends for wave 2 raises concerns about confounding factors, though the achievement jump following program implementation nevertheless supports the belief that ACE substantially increased the quality of instruction.

Though figure 2 already provides compelling evidence that ACE improved student outcomes in the short term, there are several critical questions moving forward. The ACE program represents a major financial investment by the district, and a key question is whether this investment needs to be ongoing to maintain the improvements that ACE schools observed. In other words, ACE schools improved dramatically, likely because of an influx of high-quality teachers attracted by the pay premium associated with working in ACE schools. In 2018, Dallas removed this pay premium for wave 1 ACE schools, and we intend to examine achievement trends post-2018 to gain a better understanding of the consequences of removing the stipends, given the presence of TEI and PEI. It is possible that ACE schools will revert back to their initial low performance, or they may continue their improvement. The answer to this question is crucial for understanding the likely long-term effects of a program that provides stipends to poorly performing schools and stops providing stipends once the schools improve. A second key goal is to assess a third wave of ACE schools designated in 2018. Substantial changes were introduced in the ACE program for this group in order to investigate the efficacy of a more limited and less costly program. Comparisons with effects for the earlier waves will provide valuable information on the importance of different program components.

Although TEI and PEI have undergone some changes following initial adoption, the decision to make major changes in the most recent ACE cohort based on the promise shown by the first cohort will illuminate the contributions of various program components.

Figure 2. Trends in mathematics and reading achievement, by school category, 2012–18



Source: Morgan et al. (2019). Note that non-ACE refers to all schools other than ACE and near-ACE. ACE Wave 1 adopts ACE in 2016, and ACE wave 2 adopts ACE in 2018.

Moreover, the removal of stipends for educators at the initial seven ACE schools following the 2017–18 academic year will enable us to learn much more about the dynamics of educator transitions and effectiveness following a temporary, targeted intervention. A major hurdle to the development of a program that provides inducements for educators to work in more challenging schools is the identification of the factors related to the difficulty of working conditions. Rather than using student demographic characteristics, Dallas ISD opted to use chronically low achievement, which is almost certainly more closely related to challenging working conditions. Following the apparent success of a program that provided stipends for all teachers and administrators, the next ACE generation will pay stipends only to the principal and lead teachers. An understanding of the effects of such targeted stipends will provide important evidence on program structure and the factors that influence hiring and retention of effective teachers.

Evidence and Policy

It would appear to be a straightforward proposition to reform educator evaluation and pay in ways that elevate the quality of instruction and schooling, but such has not proved to be the case. Nonetheless, the evidence on MA degrees, the benefits of supervisor feedback, the conceptual appeal of rewarding effective performance, and the promise shown by recently adopted evaluation and compensation reforms support continued efforts to develop and refine systems of evaluation and pay as a means to raise the quality of instruction and reduce inefficiencies in the provision of schooling. The evidence on the return to targeted pay for educators in disadvantaged schools seems even more promising, notwithstanding the absence of interventions limited to comprehensive changes in personnel practices. A randomized controlled trial found that teachers identified as highly effective in other schools remain effective when induced to work in disadvantaged schools, and the evidence on teacher retention, peer effects, school leadership, and working conditions is consistent with the belief that attracting and retaining effective educators in chronically low-performing schools can dramatically increase the quality of instruction.

The following points highlight key recommendations for policies and practices that have the potential to raise the quality of instruction and improve the valued educational, social, and labor-market outcomes that are of primary interest.

Overall Reforms

1. There is little support for higher pay for an MA degree. It raises the cost of becoming an educator without increasing effectiveness.

2. Rigorous observation-based evaluation informed by thoughtful information on educator contributions to achievement and other student outcomes should be the primary professional development activity for teachers and administrators. Evaluations should collect information from multiple sources. These may include student and parent feedback in addition to supervisor evaluations and evidence on student achievement.

The desirability of employing outside evaluators, as was the practice in the early IMPACT years, rather than relying only on school administrators remains uncertain.

3. Administrators should have a primary role to play and bear substantial risk with regard to the performance of the school. Teacher pay-for-performance programs in isolation are unlikely to be successful.

4. The selection of assessment instruments and formulae for the incorporation of student outcomes into evaluation and compensation systems is of crucial importance and carries many challenges. Easily manipulated outcomes, including high-school graduation, should be used with great care.

5. The desirability of including categories that place teachers on probation or terminate employment, such as those adopted as part of IMPACT, merit consideration.

6. Information on longer-term outcomes would be valuable for districts, but most do not have the capacity to produce information on college attendance, persistence, involvement with the criminal justice system, earnings, and employment. The provision of information to districts on their trends over time in these longer-term outcomes is a potentially important role for states, particularly large states like Texas, where most students remain in state after high school. National Assessment of Education Progress (NAEP) districts should also monitor NAEP score trends.

7. Despite evidence on the value of behavioral skills, incorporation of these measures into the accountability system should be done with great care so as not to incentivize behaviors that lack content in terms of raising important, longer-term outcomes.

8. The challenges of identifying good assessments and incentive structures highlight the importance of learning from experience and ongoing improvement. Fuzzy, subjective outcomes are not likely to raise the quality of instruction.

Supporting Chronically Low-Achieving and Disadvantaged Schools

9. Leveraging evaluation systems to induce effective educators to relocate to low-performing schools appears to have great promise based on the evidence to date.

10. Systems that identify educationally disadvantaged schools on the basis of student demographics alone are unlikely to target resources well and reverse the root causes of low quality of instruction and difficulties attracting and retaining effective teachers. Chronic low performance and high turnover among effective educators provide better measures but also complicate program design in terms of how to handle targeted schools that recover from low performance. We know very little about such dynamics, and the experiences of Dallas ISD will be informative.

11. The hiring and retention of ineffective principals and teachers who contribute negatively to the school environment not only adversely affects the current quality of instruction and achievement in their classrooms but also negatively affects other educators through peer effects and diminishes future school quality through educator transitions and development. This should be taken into consideration in hiring and retention decisions.

Recommendation 6 highlights a potential role of states as the provider of longer-term outcome information for districts, and this raises the general issue of the division of responsibilities for states vis-à-vis districts. Much of such discussion has focused on state redistribution of revenue, and the possibility that states may seek to use resources to mitigate district disadvantages in the educator labor market is certainly related to the issues raised in this paper. There is also debate over state-required testing, mandates to use measures of value added in personnel decisions, including tenure, and school district takeovers. The provision of information on teacher and school value added or student growth percentiles to schools or districts is also informative, though a determination of which information to make public and which to keep private raises complex issues. States have greater technical capacity than most districts, and there is great value in enabling stakeholders to compare their schools and districts with others in the state. States may also want to support research on best practices given that the results could serve children across the state.

NOTES

1 Hanushek (1986) and Hanushek, Rivkin, and Taylor (1996) find little or no evidence that teachers with an MA are more effective than those without an MA. There is a broad consensus that effective teachers have powerful impacts not only on test scores but also on educational attainment and future earnings (Chetty, Friedman, and Rockoff 2014a, 2014b; Rivkin, Hanushek, and Kain 2005; Koedel, Mihaly, and Rockoff 2015; Hanushek and Rivkin 2012).

2 Murnane and Cohen (1986) highlight the deficiencies of merit pay.

3 Three US Department of Education, Institute for Education Sciences studies of content-focused professional development using random assignment methods find little or no evidence of positive effects on achievement in second-grade reading (Garet et al. 2008), seventh-grade mathematics (Garet et al. 2011), and fourth-grade mathematics (Garet et al. 2016).

4 Neal (2011) summarizes research that highlights potential problems introduced by test-score-based incentives, and Cunha and Heckman (2007, 2008) provide evidence on the importance of noncognitive skills and discuss the possibility that test-based accountability leads to underproduction of these skills.

5 Research that investigates the effects of programs designed to attract educators to hard-to-staff schools includes Clotfelter et al. (2007); Clotfelter, Ladd, and Vigdor (2011); Steele, Murnane, and Willett (2010); Cowan and Goldhaber (2018); Springer et al. (2010); Springer, Swain, and Rodriguez (2016); and Glazerman et al. (2013).

6 Sources for the discussion of TEI include Miles (2015), Dallas ISD (2014b, 2014c, 2015a, 2015b), and Weerasinghe (2008). Sources for the discussion of PEI include Dallas ISD (2012a, 2012b, 2013b, 2014d, 2014e).

7 Other components of the program include enhanced professional development, tools and commitment to data-driven instruction and ongoing assessment, an extra hour in the school day, and after-school enrichment programs.

8 Information on costs and programs comes from Palladino (2017).

REFERENCES

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105, no. 490: 493–505.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93, no. 1: 113–32.

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

Brown, Emma, Valerie Strauss, and Perry Stein. 2018. "It Was Hailed as the National Model for School Reform. Then the Scandals Hit." *Washington Post,* March 10, 2018.

Chetty, Raj, John N. Friedman, and Jonah Rockoff. 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104, no. 9 (September): 2593–632.

———. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104, no. 9 (September): 2633–79.

———. 2017. "Measuring the Impacts of Teachers: Reply." *American Economic Review* 107, no. 6: 1685–717.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2011. "Teacher Mobility, School Segregation, and Pay-Based Policies to Level the Playing Field." *Education Finance and Policy* 6, no. 3: 399–438.

Clotfelter, Charles T., Helen F. Ladd, Jacob L. Vigdor, and Justin Wheeler. 2007. "High Poverty Schools and the Distribution of Teachers and Principals." *North Carolina Law Review* 85 (June): 1345–79.

Cowan, James, and 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: 138–52.

Cunha, Flavio, and James J. Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97, no. 2: 31–47.

———. 2008. "Formulating, Identifying, and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources* 43, no. 4 (Fall): 738–82.

Dallas ISD. 2012a. "Professional Development Meeting." Internal Document.

Dallas ISD. 2012b. "Principal Performance Rubric." Internal Document.

Dallas ISD. 2013a. "Measuring Principal Effectiveness." Internal Document.

Dallas ISD. 2013b. "Principal Professional Development." Internal Document.

Dallas ISD. 2014a. "Teacher Guidebook 2014–2015." Internal document.

Dallas ISD. 2014b. "Student Learning Objective (SLO) Rubric." Internal Document.

Dallas ISD. 2014c. "Teacher Performance Rubric." Internal Document.

Dallas ISD. 2014d. "2014–2015 Principal Handbook." Internal Document.

Dallas ISD. 2014e. “Compensation Resource Book.” Internal Document.

Dallas ISD. 2015a. “Rule and Procedures for Calculating TEI Evaluation Scores and Effectiveness Levels.” Internal Document.

Dallas ISD. 2015b. “Student Achievement Templates.” Internal Document.

Dee, Thomas S., and James Wyckoff. 2015. “Incentives, Selection, and Teacher Performance: Evidence from IMPACT.” *Journal of Policy Analysis and Management* 34, no. 2: 267–97.

———. 2017. “A Lasting Impact: High-Stakes Teacher Evaluations Drive Student Success in Washington, D.C.” *Education Next* 17, no. 4 (Fall): 58–66.

Fryer, Roland G. 2013. “Teacher Incentives and Student Achievement: Evidence from New York City Public Schools.” *Journal of Labor Economics* 31, no. 2: 373–427.

Garet, M. S., S. Cronen, M. Eaton, A. Kurki, M. Ludwig, W. Jones, et al. 2008. *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement,* NCEE 2008-4030. Washington, DC: US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://ies.ed.gov/ncee/pdf/20084030.pdf.

Garet, M. S., J. B. Heppen, K. Walters, J. Parkinson, T. M. Smith, M. Song, et al. 2016. *Focusing on Mathematical Knowledge: The Impact of Content-Intensive Teacher Professional Development,* NCEE 2016-4010. Washington, DC: US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://ies.ed.gov/ncee/pubs/20164010/pdf/20164010.pdf.

Garet, M. S., A. Wayne, F. Stancavage, J. Taylor, M. Eaton, K. Walters, et al. 2011. *Middle School Mathematics Professional Development Impact Study: Findings after the Second Year of Implementation,* NCEE 2011-4024. Washington, DC: US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://ies.ed.gov/ncee/pubs/20114024/pdf/20114024.pdf.

Glazerman, Steven, Ali Protik, Bing-ru Teh, Julie Bruch, and Jeffrey Max. 2013. *Transfer Incentives for High Performing Teachers: Final Results from a Multisite Randomized Experiment,* NCEE 2014-4003. Washington, DC: US Department of Education.

Goldhaber, Dan, and Joe Walch. 2012. “Strategic Pay Reform: A Student Outcomes-Based Evaluation of Denver’s Procomp Teacher Pay Initiative.” *Economics of Education Review* 31, no. 6: 1067–83.

Hanushek, Eric A. 1986. “The Economics of Schooling: Production and Efficiency in Public Schools.” *Journal of Economic Literature* 24, no. 3 (September): 1141–77.

Hanushek, Eric, Jin Luo, Andrew Morgan, Minh Nguyen, Ben Ost, and Steven Rivkin. 2019. “The Effects of Comprehensive Educator Evaluation and Pay Reforms on the Quality of Instruction and Leadership.” Unpublished manuscript.

Hanushek, Eric, Steven Rivkin, and Lori Taylor. 1996. “Aggregation and the Estimated Effects of School Resources,” *Review of Economics and Statistics* (November).

Hanushek, Eric A., and Steven G. Rivkin. 2012. “The Distribution of Teacher Quality and Implications for Policy.” *Annual Review of Economics* 4: 131–57.

Hoxby, Caroline M., and Andrew Leigh. 2004. “Pulled Away or Pushed Out? Explaining the Decline of Teacher Aptitude in the United States.” *American Economic Review* 94, no. 2 (May): 236–40.

Jackson, C. Kirabo, and Elias Bruegmann. 2009. “Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers.” *American Economic Journal: Applied Economics* 1, no. 4 (October): 85–108.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff. 2015. “Value-Added Modeling: A Review.” *Economics of Education Review* 47: 180–95.

Lavy, Victor. 2009. “Performance Pay and Teachers’ Effort, Productivity, and Grading Ethics.” *American Economic Review* 99, no. 5 (December): 1979–2011.

Leigh, Andrew. 2013. “The Economics and Politics of Teacher Merit Pay.” *CESifo Economic Studies* 59, no. 1 (March): 1–33.

Lewin, Tamar. 2010. “School Chief Dismisses 241 Teachers in Washington.” *New York Times,* July 23, 2010.

Marsh, Julie A., Matthew G. Springer, Daniel F. McCaffrey, Kun Yuan, Scott Epstein, Julia Koppich, Nidhi Kalra, Catherine DiMartino, and Art (Xiao) Peng. 2011. *A Big Apple for Educators: New York City’s Experiment with Schoolwide Performance Bonuses, Final Evaluation Report.* Santa Monica, CA: RAND Corporation.

Miles, M. 2015. “Teacher Excellence Initiative—Refresher.” Presentation.

Morgan, Andrew, Minh Nguyen, Eric Hanushek, Ben Ost, and Steven Rivkin. 2019. “Getting Effective Educators in Hard-to-Staff Schools.” Unpublished manuscript.

Murnane, Richard J., and David K. Cohen. 1986. “Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive.” *Harvard Educational Review* 56, no. 1 (February): 1–17.

Neal, Derek. 2011. “The Design of Performance Pay in Education.” In *Handbook of the Economics of Education,* edited by Stephen Machin, Eric A. Hanushek, and Ludger Woessmann, 495–550. Amsterdam: Elsevier.

Palladino, D. 2017. “Updated Evaluation of Accelerating Campus Excellence (ACE) 2015-16.” https://www.dallasisd.org/site/handlers/filedownload.ashx?moduleinstanceid=66303&dataid=88235&FileName=Evaluation%20of%20Accelerating%20Campus%20Excellence%20ACE%202015-16.pdf.

Podgursky, Michael J., and Matthew G. Springer. 2007. “Teacher Performance Pay: A Review.” *Journal of Policy Analysis and Management* 26, no. 4: 909–49.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. “Teachers, Schools, and Academic Achievement.” *Econometrica* 73, no. 2: 417–58.

Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and 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., John F. Pane, Vi-Nhuan Le, Daniel F. McCaffrey, Susan Freeman Burns, Laura S. Hamilton, and Brian Stecher. 2012. “Team Pay for Performance: Experimental Evidence from the Round Rock Pilot Project on Team Incentives.” *Educational Evaluation and Policy Analysis* 34, no. 4 (December): 367–90.

Springer, Matthew G., Walker A. Swain, and Luis A. Rodriguez. 2016. “Effective Teacher Retention Bonuses: Evidence from Tennessee.” *Educational Evaluation and Policy Analysis* 38, no. 2 (June): 199–221.

Steele, Jennifer L., Richard J. Murnane, and 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, no. 3: 451–78.

Steinberg, Matthew P., and Lauren Sartain. 2015. “Does Better Observation Make Better Teachers?” *Education Next* 15, no. 1 (Winter): 70–76.

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

Weerasinghe, D. 2008. “How to Compute School and Classroom Effective Indices: The Value-Added Model Implemented in Dallas Independent School District.” Internal Document.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness.* New York: New Teachers Project.

The views expressed in this publication are entirely those of the authors and do not necessarily reflect the views of the staff, officers, or Board of Overseers of the Hoover Institution.

Copyright © 2020 by Steven Rivkin, Ben Ost, Andrew Morgan, and Minh Nguyen. All rights reserved.

26 25 24 23 22 21 20 7 6 5 4 3 2 1

About the Authors

STEVEN RIVKIN

Steven Rivkin is professor and head of the Department of Economics at the University of Illinois at Chicago, director of the John F. Kain Center for Education Research at the University of Texas at Dallas, research associate at the National Bureau of Economic Research, and fellow at the CESifo Research Network in Munich. Rivkin has served on government advisory committees and task forces and served as a member of the Amherst and Amherst Regional School Boards.

BEN OST

Ben Ost is associate professor of economics at the University of Illinois at Chicago. His research explores a variety of issues related to education policy including teacher quality measurement, within-school resource allocation, and the returns to higher education. He received his PhD in economics from Cornell University.

ANDREW MORGAN

Andrew Morgan is a PhD candidate in economics at the University of Illinois at Chicago. His research focuses on the economics of education with an emphasis on educator evaluation and quality. Andrew received his bachelor's degree from the University of Oklahoma and his master's degree in economics from the University of New Hampshire.

MINH NGUYEN

Minh Nguyen is an assistant professor at Ball State University. Her research interests lie in health, education, and development economics. Her current research examines how public policies have affected the effectiveness of school administrators and teachers and the accumulation of human capital. She received her PhD in economics from the University of Illinois at Chicago.

Hoover Education Success Initiative

With passage in 2015 of the Every Student Succeeds Act (ESSA), states are again in charge of American education policy. To support them in this undertaking, the Hoover Education Success Initiative (HESI), launched in 2019, seeks to provide state education leaders with policy recommendations that are based upon sound research and analysis.

HESI hosts workshops and policy symposia on high-impact areas related to the improvement and reinvention of the US education system. The findings and recommendations in each area are outlined in concise topical papers.

The leadership team at HESI engages with its Practitioner Council, composed of national policy leaders, and with interested state government leaders. HESI's ultimate goal is to contribute to the ongoing transformation of the nation's education landscape and to improve outcomes for our nation's children.

For more information about the Hoover Education Success Initiative, visit us online at hoover.org/hesi.

Hoover Institution, Stanford University
434 Galvez Mall
Stanford, CA 94305-6003
650-723-1754

Hoover Institution in Washington
The Johnson Center
1399 New York Avenue NW, Suite 500
Washington, DC 20005
202-760-3200

HOOVER INSTITUTION