Economics of Education Review 44 (2015) 114–125

Contents lists available at ScienceDirect

Economics of Education Review

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

Performance pay, test scores, and student learning objectives

Ryan Balch[*], Matthew G. Springer

Vanderbilt University, United States

ARTICLE INFO

Article history:
Received 1 December 2012
Revised 5 November 2014
Accepted 8 November 2014
Available online 9 December 2014

JEL:
I22

Keywords:
Pay for performance
Teacher salaries
Educational economics

ABSTRACT

Austin Independent School District's (AISD) REACH pay for performance program has become a national model for compensation reform. This study analyzes the test scores of students enrolled in schools participating in the REACH program to students enrolled in schools within AISD not participating in the program. We also investigate the relationship between student learning objectives (SLOs), the program's primary measure of individual teacher performance, and teacher performance as measured by value-added student test scores. The AISD REACH program is associated with positive student test score gains in both math and reading during the initial year of implementation. Student test score gains are maintained in the second year, but we do not find any additional growth. We also find that SLOs are not significantly correlated with a teacher's value-added student test scores.

1. Introduction

Many districts and states have initiated performance pay policies to identify and reward teachers that lead students to significant gains in achievement (Goldhaber, 2009; Johnson & Papay, 2009; Podgursky & Springer, 2007, 2011; Springer, 2007;). A driving factor comes in part from research indicating that the most important determinant of a teacher's pay, years of experience and advanced degrees, are not closely related to student performance or school outcomes (Goldhaber, 2002; Gordan, Kane, & Staiger, 2006; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Many stakeholders have focused on pay for performance programs as a means to correct inefficiencies found in traditional salary schedules. However, how they influence student test scores is still not fully understood in practice. Nor do we know much empirically on the relationship between individual performance pay program components and outcomes of interest (Springer and Balch, 2010).

The theory of action behind performance pay holds that making pay in part contingent on performance provides strong incentives to improve outcomes of interest. It can help motivate teachers to higher levels of performance and align their behaviors and interests with institutional goals. Incentives may help attract teachers into the workforce who are relatively more effective at meeting the performance targets while inducing high-performers to remain in the teaching profession. Select observational studies have documented the compositional effect of pay for performance programs (Clotfelter et al., 2008; Fulbeck, 2014; Steele, Murnane, & Willett, 2010; Taylor & Springer, 2010; Wiley, Spindler, & Subert, 2010).[1]

Opponents of performance pay systems contend that compensating teachers for the performance of their students or for the subject and/or location in which they teach can compromise collegiality and cooperation among teachers (Murnane & Cohen, 1986). An analysis of data from the schools and staffing survey further suggests that job satisfaction is lower for those teachers who receive merit pay awards

[*] Corresponding author at: 1111 Argonne Ave, Baltimore, MD 21218, United States . Tel.: +1 404 759 3085.
E-mail address: ryan.balch@vanderbilt.edu (R. Balch).

[1] Using a unique dataset on test scores for students admitted into an unidentified Australian university, Leigh (2012) finds that a 1% rise in the salary of a starting teacher boosts the average aptitude of students entering teacher education courses by 0.6 percentile ranks, with the effect being strongest for those at the median.

http://dx.doi.org/10.1016/j.econedurev.2014.11.002
0272-7757/© 2014 Published by Elsevier Ltd.

(Belfield & Haywood, 2008).[2] If a breakdown occurs in collegiality, cooperation, and job satisfaction opponents argue that compensation reform efforts may render schools less effective.

In this paper, we analyze the test scores of students enrolled in schools participating in the REACH program to students enrolled in schools within AISD not participating in the program. We also investigate the relationship between student learning objectives (SLOs), the program's primary measure of individual teacher performance, and teacher performance as measured by value-added student test scores. The REACH program has become a national model for compensation reform. Initially designed as a four-year pilot program that would involve 15 schools, the REACH program was awarded $62.3 million from the United States Department of Education's Teacher Incentive Fund, which helped finance expansion to 43 high-needs schools located primarily in East Austin. AISD REACH offers a unique opportunity to evaluate a comprehensive compensation reform program that incorporates a variety of elements including school-level performance based bonuses, teacher-level performance bonuses, hard to staff bonuses, and professional development incentives.

While the school-level performance bonuses in REACH were based on relatively conventional test score gains measures (see for example, Marsh et al., 2012), the teacher-level incentives were tied to the accomplishment of a set of individualized student learning objectives (SLOs) that each teacher established with their principal. SLOs are targets of student growth that teachers set at the start of the school year and work toward by the end of the school year. SLOs are intended to be data-based instructional goals that are collaboratively constructed with principals based on the needs of students in their classroom. However, it is unclear whether these serve as a valid measure for effective teaching as defined by student achievement data as no research study has yet analyzed their relationship to student achievement. The current investigation will assess the relationship between SLOs and value-added student achievement in an effort to inform choices for measuring teacher effectiveness. This should prove informative given the growing number of districts and states interested in SLOs as a measure of teacher effectiveness.

The simultaneous implementation of the broad set of program components makes it difficult to isolate the effects of any of the particular policies that compose REACH. However, we are able to examine the overall effect of the comprehensive compensation reform and offer exploratory evidence with respect to its most unique component – student learning objectives (SLOs). More specifically, we describe research on two primary questions related to AISD REACH:

- How does student achievement change in the first two years of REACH implementation?
- How well do student learning objectives (SLOs) relate to teacher performance?

We find significant student test score gains in both math and reading associated with the initial year of AISD REACH implementation. Student test score gains are maintained in the second year, but we do not find any additional growth. We also find that SLOs are not significantly correlated with teacher value added. These estimates are consistent across a number of different analytic specifications, sample choices, and sensitivity tests.

In the next two sections, we outline teacher compensation reform efforts in AISD and provide a brief review of relevant literature on teacher compensation reform programs. Section 4 describes our data sources, measures, and analytic strategy, and Section 5 reports our main findings along with a series of sensitivity analyses to verify the robustness of the reported results. The paper concludes with a discussion of our findings, their limitations, and implications for policymakers and research moving forward.

2. REACH background

During the 2004–2005 school year, the AISD Board of Trustees developed a strategic plan that directed then-Superintendent Pascal Forgione to design a new compensation system to recruit, develop, retain, and reward highly effective teachers and administrators. Following an extended planning period, AISD implemented the Strategic Compensation Initiative (now referred to as AISD REACH) in the 2007–2008 school year using $4.3 million in local revenue to fund the program. An additional $5.4 million from the state-run grant program, District Awards for Teacher Excellence, and $1 million from the Texas Education Agency's Beginning Teacher Induction grant, financed REACH the following school year, during which nine schools participated.[3]

As displayed in Table 1, the AISD REACH program includes six components organized within three program elements: student growth, professional development, and recruitment and retention of teachers at the highest needs schools. Student growth includes teacher developed SLOs and school-wide growth on the Texas Assessment of Knowledge and Skills (TAKS). Professional development is addressed through Take One!®, a single element of the National Board for Professional Teaching Standards (NBPTS) certification process, and novice teacher mentoring, a mentoring program for teachers serving in highest needs schools who are in their first three years of teaching. Recruitment and retention of teachers is addressed by two stipends – one for teachers new to a highest needs school and one for teachers who remain at that school.

The student growth incentives were designed to recognize teachers and principals for student growth, both at the classroom level and at the school level. The student growth component included rewards for all teachers and principals for meeting growth targets in school-wide performance on TAKS testing, and rewarding of individual teachers for meeting their teacher-developed SLOs, which measured performance at the classroom level. The school-wide TAKS growth stipend rewards whole campuses for the performance of students on

[2] Ballou (2001) suggests that the presence of teacher associations is more directly related to the failure of early experiences with merit pay as opposed to the nature of teaching itself.

[3] Springer, Ballou et al. (2012) and Springer, Pane et al. (2012) provide a thorough discussion of performance pay programs in Texas.

Table 1
Overview of AISD REACH.

Program element	Student growth		Professional growth		Retirement and retention	
	School-wide TAKS growth	Student learning objectives	Take One!®	Novice teacher mentoring	New school stipend	Retention stipend
Description	School level growth on state achievement assessment	Teacher designed data-based instructional goals	Completion of one component of NBPTS certification	Full-time mentoring to novice teachers (1–3 years exp.)	Teachers in years 1–3 of service at highest needs schools	Teachers and principals retention at highest-need schools
Stipend amount	Principals = $3000–4000 Teachers = $1000–3000	Principals = $3000–4500 Teachers = $1000–3000	Teachers = $395–795 NBPTS facilitators = $1000	Mentors = $3000–7000	Teachers = $1000	Principals = $3000 Teachers = $3000–6000

the TAKS. Stipends are determined based on the Comparable Improvement Index, a state-developed index that compares groups of 40 match-comparison schools. Teachers earn $1000 per subject ($2000 at highest needs schools), with half the stipend distributed only if the teacher returns to service the following year. Principals receive $2000 for each subject in the year achieved, and another $2000 per subject if they return the following school year.

SLOs are teacher designed data-based instructional goals. Through examination of student achievement data, teachers work with their principal to develop two SLOs based on student need. At least one SLO must target the teacher's class as a whole, while the second SLO can focus on a particular sub-group of students. SLOs must be based on the state standards, address classroom needs, align with the goals of the Campus Improvement Plan, and satisfy standards of rigor for both performance and assessment. Each SLO must also be approved by both the teacher's principal and AISD REACH staff to ensure objectives are appropriate and rigorous. At the end of the school year, students are evaluated to determine whether the performance objective has been met. Teachers can earn $1000 for each SLO met ($1500 if at highest-needs school). Principals receive a $3000 stipend for facilitating the SLO process on their campus ($4500 at highest-needs schools).

The second major element of REACH promotes the professional growth of AISD faculty members in two ways: development opportunities through Take One!®, and novice teacher mentoring programs for highest needs schools. Take One!® is a unique offering from the NBPTS that allows teachers and administrators to complete one element of the National Board certification process. Take One!® focuses on examination of classroom practices through the use of guided self-reflection and videotaping. Take One!® participants develop and submit to the National Board a portfolio detailing their instructional practices, including videotaped lessons, providing evidence of how their instruction improves student learning. In addition to the $395 portfolio submission fee being waived, participants receive a stipend of $200 for submitting their portfolio and an additional $200 if they receive a passing score from the National Board. Teacher facilitators, who work with participants on portfolio development, receive a $1000 stipend for service.

The novice teacher mentoring component provides a dedicated mentor for new teachers within their first three years in the profession at the five highest-needs pilot schools. Mentors are teachers with seven or more years of experience that are released from classroom assignments for a period of up to two years. The senior teachers assist new teachers with instructional planning, setting SLOs, classroom management, and other support. Mentors receive a $3000 stipend for service and may receive an additional $2000 if they receive a satisfactory evaluation.

The final two components of the REACH program are designed specifically for the five highest needs schools participating in the pilot. Teachers and principals in these schools are eligible for a new to school stipend (for teachers only) or a retention stipend (teachers and principals) starting in the 2008–2009 school year for each year of service. Teachers who remain at highest-needs schools for one to three years receive an annual $1000 new to school stipend; those who stay for four to six years receive a $3000 retention stipend annually. Principals receive $3000 for each year of service. Stipends are paid in two parts: half at the beginning of the school year, and half for completing the school year.

In total, principals in AISD REACH schools can earn annual awards between $3000 and $15,500. Teacher awards can range from $200 to $14,795. In practice, principals earned between $3000 and $14,000 with an average award of $8244 during the 2007–2008 school year. Teachers received about half as much on average ($4620), with individual award amounts ranging between $200 and $8700.

While we do not directly assess the effects of REACH retention and hiring incentives on teacher turnover or the characteristics of the composition of teacher labor force in participating schools, this is an important component of the program worthy of future study. The influence of the incentives on the composition of the workforce is critical to understand as the theory suggests incentives may help attract and retain teachers who are relatively more effective at meeting the performance targets while inducing high-performers to remain in the teaching profession.

3. Review of relevant literature

A long body of literature has examined the rigid structure of teacher compensation policies, their foundations (particularly with respect to collective bargaining agreements), the incentive structure they imply for teachers, and their role in promoting or hindering student achievement (e.g., Hanushek, Kain, & Rivkin, 2004; Jackson, 2012; Lankford, Loeb, & Wyckoff, 2002; Podgursky, 2006; Podgursky, 2010; Hendricks, 2011; Gilpin, 2012). Recent rigorous studies have examined policies similar to all of the components of REACH. For example, Marsh et al. (2012) study of school-level performance incentives in New York City employed a randomized controlled design and found no significant effects on student achievement for the opportunity to earn bonuses far larger than those offered in REACH. Alternatively, a number of evaluations of recruitment and retention bonuses for hard to staff schools have found significant effects on teacher mobility patterns (e.g. Glazerman et al., 2013; Springer, Rodriguez, & Swain, 2014), even for relatively small bonuses (Clotfelter et al., 2008). Other studies have found significant positive associations between student achievement, formative teacher evaluation, mentorship and professional development (e.g., Jackson & Bruegmann, 2009; Taylor & Tyler, 2012).

However, for the purpose of this study we focus primarily on the teacher-level incentives and measures of teacher effectiveness. We first provide a brief overview of the literature on the effect of financial incentives on student outcomes. We then describe recent developments and research on approaches to measuring teacher effectiveness in teacher evaluation systems, which we use to inform our analysis of the relationship between student learning objectives (SLOs) and teacher performance.

3.1. Evaluations of performance pay programs

Some of the most rigorous scientific evidence on teacher compensation reform comes from abroad. Although most report generally positive effects on student achievement (e.g., Glewwe, Ilias, & Kremer, 2010; Lavy, 2002, 2009; Muralidharan & Sundararaman, 2011a; Santibañez et al., 2007), it is less clear whether these programs actually promoted long-run learning.[4] Some studies find the incentive pay effects do not persist over time or document opportunistic behaviors on the part of treatment teachers that account for increased student achievement (Glewwe, Ilias, & Kremer, 2010). Furthermore, these findings are not necessarily generalizable to the U.S. context. The incentive structure facing teachers and schools (e.g., Andhra Pradesh, India or rural Kenya) are very different from the operational context found within the U.S. public school system.[5]

There are a growing number of rigorous evaluations of performance pay programs in U.S., with evidence generally showing negligible effects on student performance. Studies by Springer, Ballou et al. (2012) in Nashville, Springer, Pane et al. (2012) in Round Rock, Texas, Goldhaber and Walch (2012) in Denver, Colorado, and Glazerman and Seifullah (2010) in Chicago found that performance pay programs were not associated with improved student outcomes or positive changes in teacher practice.[6] Several evaluations of New York City's School-Wide Performance Bonus Program also report a null (Goodman & Turner, 2011; Marsh et al., 2011) or negative (Fryer, 2011) finding. A more recent experiment conducted by Fryer et al. (2012) in Chicago found positive effects for both individual and team-based incentive programs, which a number of correlational analyses have also concluded (Figlio and Kenny, 2007; Springer, Ballou, & Peng, 2014; Winters et al., 2006).

3.2. Evaluating educators in performance pay programs

Measuring teacher performance is a formidable challenge due to the lack of specificity regarding ideal teaching behaviors and desired outcomes of education. There are a number of different strategies employed in performance pay programs to evaluate educator performance, including classroom observations of teaching, student perceptions of teaching practices, value-added models of teacher effectiveness, and teacher-defined student learning objectives.

Classroom observations represent one of the most commonly used evaluation systems for teachers (Goe, 2008). Examples of instruments that have been validated for their relationship to student achievement include Charlotte Danielson's (1996) Framework for Teaching and the Classroom Assessment Scoring System (CLASS) for grades K-5 (Pianta, La Paro, & Hamre, 2006).[7] Kane et al. (2010) report that a student with a teacher in the top quartile according to Danielson's Framework for Teaching would score 0.10 standard deviations higher in math and 0.125 standard deviations higher in reading than a student assigned to a teacher in the bottom quartile. Similarly, in a three-year study designed to determine how to best identify and promote teaching excellence, called the measure of effective teaching (MET) and funded by the Gates Foundation, teacher ratings from the Danielson Framework had a 0.19 correlation with student achievement in math and a 0.11 correlation with student achievement in ELA. For the CLASS instrument, the correlations were 0.24 and 0.10 respectively.

A further measure of teacher practice comes from student perceptions. When considering possible measures of teacher effectiveness in K-12 education, it can be argued that student perceptions of a teacher are an important consideration in any teacher evaluation system as students have the most contact with teachers and are the direct consumers of a teacher's service (Goe, Bell, & Little, 2008). Findings from the MET project report a significant correlation between a teacher's total score on the student survey and value-added

[4] Using PISA-2003 international achievement micro data, Woessmann (2011) reports that teacher salary adjustments for outstanding performance are significantly associated with math, science, and reading achievement across countries.

[5] A study by Muralidharan and Sundararaman (2011b) report teacher opinions about performance pay programs.

[6] In reporting findings from three randomized studies on incentive pay programs (New York City, Nashville, TN, and Round Rock, TX), Yuan et al. (2013) conclude that these programs do not affect teacher motivation or reported instructional practices as measured by unique survey instruments developed for evaluations of these programs.

[7] There are countless variations in frequency, instrument, rating scales, and protocol. Some of the main issues to consider with observations are the validity of the instrument and the reliability of rating, particularly if ratings are attached to financial rewards or job security.

achievement on state tests in both math and ELA. These are similar to correlations between value-added and observation rubrics that look at general teaching practices such as Danielson's Framework for Teachers and CLASS with student surveys showing a 0.218 correlations with value-added in math and a 0.095 correlation with value-added in ELA. While correlations for value-added are significant but small, research findings suggest that student surveys had higher levels of reliability.

On the quantitative side, districts and states are increasingly using output measures to judge the performance of individual teachers, teams of teachers, and/or schools. Advances in technology and statistical modeling now allow for student achievement results to be attributed to individual teachers, potentially indicating a teacher's contribution to student achievement. Generally referred to as value-added models, these growth measures have the potential to identify unique contributions of the school or teacher to students' progress over the course of a year rather than cumulative education effects or student background factors. Researchers and practitioners have employed an increasing number of methods to calculate teacher effect estimates (e.g., Ballou, Sanders, & Wright, 2004; McCaffrey, Koretz, Lockwood, & Hamilton, 2003; Rubin, Stuart, & Zanutto, 2004).

SLOs, also known as student growth objectives, are measurable, long-term growth targets that a teacher sets at the beginning of the school year for all students or for subgroups of students. SLOs are designed to demonstrate a teacher's impact on student learning. A number of states currently use, or are considering using, SLOs in teacher evaluation systems, including Colorado, Georgia, Maryland, New York, and Ohio. The only evidence on the relationship between SLOs and teacher effectiveness that we are aware comes from an evaluation of Denver ProComp, where Goldhaber & Walch (2012) observe a slightly higher proportion of teachers that earned SLO bonuses at the top of the teacher effectiveness distribution. By investigating the relationship between SLOs and value-added student test scores we are better able to inform choices policymakers and practitioners may make when designing performance pay programs, particularly for school systems interested in evaluating teachers in non-tested grades and subjects.

4. Data, sample, and analytical methods

This section provides a description of the AISD administrative data utilized in this study, the process by which restrictions differentiate our analytic sample from the full population, and the methods applied to assess the overall impact of REACH and associations between SLOs and value-added measures of teacher effectiveness. We also describe the matching procedure we use to construct a comparison group of schools that is comparable to REACH participating schools on observed characteristics.

4.1. Data

Our analysis of the REACH program requires school, classroom, teacher, and student-level data. Several school- and classroom-level data elements were calculated by aggregating student-level information to create classroom or school averages for characteristics such as percent minority, percent at-risk, and prior achievement scores at the classroom and building level.

AISD provided two types of student-level data files: student biographical information and student achievement test scores. The student biographical files contained information on the universe of students enrolled in AISD by school year. The biographical files also contained student background information, including gender, race/ethnicity, and participation in special education services and English language learner programs, and eligibility for free or reduced priced school lunches. Both the biographical and achievement data files include a unique, longitudinally consistent identifier for each student.

AISD also provided a series of data files containing test scores for the universe of students enrolled in elementary, middle, and high schools. Test data comes from the TAKS that is based on the state-mandated curriculum, the Texas Essential Knowledge and Skills (TEKS). All students in grades 3 through 9 take the standardized reading assessment, while students in grade 10 take an English language arts examination. All students enrolled in grades 3 through 10 are administered a standardized mathematics assessment.

Raw scores from TAKS are not expressed on the same developmental scale from one year to the next or from one grade to the next. Since the structure of the TAKS tests may lead to smaller or larger gains at various points on the test score distribution, we created a normalized test scale scores of students enrolled in elementary and secondary schools by subject, grade, and school year to ease threats associated with potential variations in content standards, grade level expectations, test constructs, and performance standards. Using data for the universe of students enrolled in AISD, we estimated the means and standard deviations of scores by grade level for each school year and subject. The standardized score equals the quantity of the scale score minus the mean for the corresponding grade-level, subject, and year divided by the standard deviation for the grade-level, subject and year. This transformation makes the year-to-year student test score gains reflect a change in their relative position within the distribution of test takers for the grade-level, subject and year.

To measure an individual teacher's contribution to student learning, as defined by a value-added measure of teacher effectiveness, we use an adaptation of Webster and Meandro (1997) as suggested by Thum and Bryk (1997). Generally speaking, we use a two-level hierarchical linear model where students are nested in classrooms while controlling for a series of student- and classroom-level compositional variables, including prior test scores, race/ethnicity, language proficiency, socioeconomic status, and gender. This class of mixed method modeling performs well on various noise and error metrics (McCaffrey, Han, & Lockwood, 2009).

AISD also provided a series of teacher-level data files regarding the number of SLOs a teacher met. As indicated earlier, SLOs are targets of student growth that teachers set at the start of the school year and work toward by the end of the

Table 2
Percentage of math and reading teachers meeting student learning objectives (2008–2009 school year).

	Mathematics	Reading
Met 0% of SLOs set	46.70%	37.30%
Met 50% of SLOs set	14.70%	30.50%
Met 100% of SLOs set	38.70%	32.20%
# of observations	75	59

school year.[8] Our investigation into SLOs is limited to math and reading teachers in REACH schools that had more than 15 students with valid TAKS test scores.[9] Table 2 displays the percentage of teachers meeting the SLOs that they set by content area in which they teach. SLOs appear to be fairly difficult to meet, with less than 40% of teachers in either subject meeting all of their objectives. While teachers are fairly evenly distributed in reading, math teachers tended to either meet all of their objectives or none.

4.2. Analytic sample and matching procedure

Table 3 displays select summary statistics on student background characteristics for schools participating in the REACH program, a matched comparison sample of schools, and all AISD schools not participating in the REACH program at baseline. To construct the matched comparison sample of schools, we adopted a nearest neighbor matching approach using Mahalanobis distances, which is widely used in the empirical program evaluation literature (Cameron & Trivedi, 2005). We match each REACH school to the best look-alike school on a vector of observable background characteristics prior to REACH implementation, including prior test scores in math, reading, science, and social studies and the percentages of at-risk, minority, low socio-economic status, ESL, and special education students. Specifically, the algorithm first calculated the Mahalanobis distance between each REACH school to the other schools in AISD. This optimal matching algorithm then pairs schools in a way that minimizes the sum of the Mahalanobis distances among REACH and matched sample schools.

Table 3
Select summary statistics on student characteristics by year.

	REACH	Matched sample	All non-REACH
Enrollment	3030	9510	35560
% At-risk	60.60	60.40	49.00
% Minority	75.60	81.50	64.90
% Low SES	68.50	71.50	53.20
% Female	49.80	44.80	49.30
% ESL	18.00	15.10	9.40
% Special Ed	12.10	12.80	10.20
Math scale score	2194.20	2181.10	2239.00
	2213.30	2180.80	2246.40
Reading scale score	2238.40	2225.20	2276.10
	2255.80	2237.70	2280.70

Note: REACH schools: $N = 9$ in 2008 and 11 in 2009. We include all 11 schools in baseline estimates. Non-REACH schools: $N = 104$. Matched comparison schools: $N = 22$. At-risk is a composite variable that includes factors that contribute to being at risk of dropping out (achievement, failing a class, low attendance, being suspended, and being old for the grade), minority includes students that are African-American and Hispanic, low SES includes students eligible for the school lunch program. All math and reading scores are from normed results of TAKS testing.

Table 4
Select summary statistics on teachers and teacher's classrooms in SLOs analysis (2008–2009).

	Mathematics	Reading
Teacher characteristics		
Years of experience	8.33	11.86
% Master's degree	28	35.1
% Novice	42.7	30.5
Classroom characteristics		
Test score (2008–2009)	2238	2280.9
Test score (2007–2008)	2236.2	2264.8
% Minority	75.6	75.4
% Special education	1.6	2
% Econ. disadvantaged	68.4	68.6
# of observations	75	59

Schools in the matched comparison sample look similar to REACH schools on several observable characteristics, including the percentage of students designated at risk and the percentage receiving special education services, while a greater percentage of students are classified as non-white and economically disadvantaged in the matched comparison sample. Student test scores in the matched comparison sample are lower than student test scores in REACH schools, but more similar to REACH schools than the all non-REACH sample. None of these differences at the school-level are significant at traditional levels.

To investigate the relationship between SLOs and teacher effectiveness, our sample is restricted to REACH schools and the 75 mathematics and 59 reading teachers with valid SLO data during the 2008–2009 school year. As displayed in Table 4, the average years of teaching experience for math and reading teachers is 8.33 and 11.86, respectively. Slightly more reading teachers have obtained a master's degree (35.1% vs. 28%), while a larger percentage of math teachers are identified as novice teachers (42.7 vs. 30.5). The composition of students enrolled in math and reading teachers' classrooms are very similar.

4.3. Analytic method

To examine how test scores of students enrolled in schools participating in the REACH program compare to test scores

[8] According to AISD's SLO guide, SLOs should support the goals of the school's Campus Improvement Plan, be rigorous and measurable, provide clear focus for instruction and assessment, and be a good example of ongoing, reflexive practice. Each SLO must be approved by both the teacher's principal and district staff to ensure objectives meet these criteria. Objectives are scored on a scale of 1–4, with approval granted if objectives earn a 3 or 4 in each of the designated areas. Two criteria that are particularly relevant to the current investigation are the learning objective and assessment. The AISD guide states that SLOs should be aligned with state teaching standards, known as the Texas Essential Knowledge and Skills (TEKS). Further, teachers have the option of either using a pre-made test or creating their own assessment. One might expect there to be a relationship between meeting SLOs and the TAKS student achievement test because the test is based on the TEKS standards. On the other hand, the lack of standardized assessment that is connected to the TAKS test means that it is possible for there to be no relationship.

[9] McCaffrey, Han, and Lockwood (2009) and Schochet and Hanley (2010) report on error rates in teacher value-added models applied to student test score gains where class sizes vary. States have also adopted business rules for setting minimum class size requirements, e.g., Ohio and Tennessee.

of students enrolled in schools within AISD not participating in the program, we estimate variants of the following OLS regression model:

$$testscore_{it} = \beta_0 + \beta_1 REACH_t + student_{it}\beta_2 + \beta_1 testscore_{it-1} + \varphi_{it} + \varepsilon_{it} \qquad (1)$$

where, $testscore_{it}$ represents the standardized test score for student i at time t in math or reading as measured by TAKS; *REACH* is an indicator variable that equals one if a school participated in the REACH program and zero if the school was part of the matched comparison sample; *student* is a vector of baseline observable student-level characteristics, including binary variables for gender, free lunch status, ELL status, SPED status, and a series of race/ethnicity classifications; *testscore* represents the standardized test score for student i at time $t - 1i$ in math or reading as measured by TAKS; and φ is a series of grade level dummy variables, eliminating across grade variation from the estimates. Standard errors are clustered at the school level to account for within school correlation.

Here, we are most interested in the estimates on β_1, which identifies the average difference in student test score gains associated with implementation of REACH as compared to test score gains in the matched comparison schools. It is important to note that the REACH estimate should be thought of as being broader than simply compensation reform. AISD REACH is a comprehensive program that incorporates a variety of reform elements including school based bonuses, teacher level bonuses, hard to staff bonuses, and professional development. Additionally, at the time of this study, there were other reforms being implemented in the district, namely, the Texas Educator Excellence Grant (TEEG) program. Thus we estimate a variant of (1) that controls for the two REACH schools that concurrently were participating in the TEEG program. By including this control we further isolate the specific type of performance pay program of interest (in this case, REACH).

We also estimate a series of alternative specifications to test the robustness of findings. Select specifications include prior achievement expressed as quadratic and cubic function to account for a non-linear relationship between prior and current test scores. We also test models controlling for two prior test scores. Using a single prior score is subject to regression to the mean and other temporary shocks that can be minimized when multiple prior scores are used (McCaffrey et al., 2003). Although the sample of valid observations drops 23.6% in math and 23.7% in reading when controlling for two prior test scores, the estimates are qualitatively similar across all specifications. Finally, we re-estimate all model specifications using data from all students and schools in AISD and, once again, find that the estimates are similar to those produced using the matched comparison sample of schools.

It is also important to note a number of internal and external validity-related issues. First, we are comparing schools that were and were not subject to the REACH program, which was targeted to schools based on level of student need, representation of the AISD population, and the presence of an experienced principal who was willing to facilitate the program. Thus, if one assumes there is a clear demand for the policy in the schools where REACH is implemented then this demand may bias toward finding a treatment effect. Second, it may be possible that there is a heterogeneous treatment effect finding for areas that have a particular need for the REACH policy and have a strong leader who can carry out the policy. Finally, the district under study might be considered a highly “treated” school district (or, a reform-oriented district) with many different program under-development and underway. While we attempt to account for some competing reforms in our evaluation, it is possible that the reform orientation of the district may attract certain types of administrators and teachers to the district, as well as a certain type of students. As such, the generalizability of our findings needs to be interpreted with this context in mind.

4.4. Student learning objectives (SLOs) and teacher effectiveness

To examine the relationship between SLOs and a teacher's value-added effect estimates, we estimate variants of the following OLS regression model:

$$valueadded_{jk} = \beta_0 + \beta_1 SLO_{jk} + \varphi_j + \varepsilon_{jk} \qquad (2)$$

where, $valueadded_{jk}$ represents the value-added effect estimate for teacher j in school k in math or reading; *SLO* is a binary indicator that equals one if a teacher met one or more of their SLO goals and zero if they did not meet any; and φ is a series of grade level dummy variables, eliminating across grade variation from the estimates. Standard errors are clustered at the school level to account for within school correlation.

Here, we are most interested in the estimates on β_1, which identifies the average difference in teacher effectiveness among teachers that met at least one SLO and those that did not meet any of their SLOs. We also estimate two other models where we replace the *SLO* indicator with (1) a dummy variable for met one SLO and a second dummy variable for met both SLOs; or (2) the percentage of SLOs met by a teacher. We also estimate a series of equations controlling for school-level characteristics.

5. Results

We find that controlling for a set of observable characteristics, students attending schools that participated in the REACH program experienced significantly larger test score gains than students in matched comparison schools that did not participate. In the second year of program implementation, we find no additional gains associated with attending REACH schools, but no indications that REACH school students fall back to their previous achievement levels. With respect to the second primary research question, we find no significant relationship between the percentage of SLOs a teacher meets and the teacher's effectiveness estimated by test score value-added. There is some indication, however, that teachers in the bottom quartile of value-added effectiveness are more likely to meet a larger percentage of their SLO targets.

5.1. Student test scores and REACH implementation

The results of our first set of regression estimates are displayed in Table 5. We find that, in the 2007–2008 school

Table 5
Estimates of the relationship between REACH and math test scores by year.

	2007–2008 school year		2008–2009 school year	
	Model (1)	Model (2)	Model (3)	Model (4)
REACH	0.1344 (0.0399)**	0.1761 (0.03678)**	0.0347 (0.0275)	0.0122 (0.0305)
Student controls	√	√	√	√
Grade effects	√	√	√	√
Prior test scores	√	√	√	√
Other reforms		√		√
Number of students	8535	8535	9148	9148
Number of schools	31	31	33	33

Notes: Clustered standard error in parentheses. * significant at the 10% level; ** 5% level; *** 1% level.

Table 6
Estimates of the relationship between REACH and reading test scores by school year.

	2007–2008 school year		2008–2009 school year	
	Model (1)	Model (2)	Model (3)	Model (4)
REACH	0.1001 (0.0605)*	0.0554 (0.0610)*	0.0238 (0.0399)	−0.0342 (0.0246)
Student controls	√	√	√	√
Grade effects	√	√	√	√
Prior test scores	√	√	√	√
Other reforms		√		√
Number of students	8321	8321	8835	8835
Number of schools	31	31	33	33

Notes: Clustered standard error in parentheses. * significant at the 10% level; ** 5% level; *** 1% level.

year, the sign on the REACH coefficient for math is positive and statistically different from zero. Students enrolled in AISD REACH schools demonstrate between 0.1344 and 0.1761 standard deviations greater gains in math. It is also apparent that the magnitude of the coefficient on the REACH variable increases when we control for the schools that participated in both REACH and the state-level TEEG program, suggesting these schools attenuated the relationship between REACH program implementation and student test scores.

However, this relationship does not persist in the 2008–2009 school year. The relationship between REACH program implementation and student test score gains is no longer significant, although the sign on the REACH coefficient is still in the expected direction. These results suggest that gains made during the first year of the program were maintained during the second year of the program but did not increase despite the addition of two schools to the program.

As an added check to ensure schools that participated in REACH for a second year were not more likely to increase test scores, a dummy variable was added to Eq. (1) that indicated whether a school would ever participate in the REACH program. The coefficient on the interaction variable was larger in magnitude but not statistically different from zero, indicating that the null finding in 2008–2009 was a result of the program rather than any pre-existing factors that led the school's participation in the program.

The results for reading, as displayed in Table 6, follow a similar pattern. For the 2007–2008 school year, the coefficient on the REACH variable is positive and significant. However, this relationship is only marginally significant when we control for the schools that participated in both REACH and the state-level TEEG program. Further, as in math, this relationship does not persist in the 2008–2009 school year, although the sign on the REACH coefficient is still in the expected direction, except when we control for other reforms. Nonetheless, the size of the value on the coefficients are very small.

Although clustering at the school level addresses the possibility of intra-school correlation, we also employ a hierarchical linear modeling approach to further assure that error is properly distributed between students and schools. This approach reveals what changes should be attributed to differences in students or differences in schools. Overall, as displayed in Table 7, results from the HLM are qualitatively similar to the OLS estimates. The main difference is that the magnitude of the estimate on the REACH variable when math is the dependent variable is smaller but still statistically different from zero. Point estimates are very similar when reading is the dependent variable as reported in Panel B of Table 7. Additionally, in the 2008–2009 school year, the relationship between student test score gains and REACH implementation is no longer statistically different from zero.

Regression to the mean may be an explanation for why REACH schools obtained better results. It is possible that schools participating in the REACH program would have improved even without program participation due to unobserved factors that were already influencing test scores. To test for this, a REACH dummy was created that identified REACH schools one year prior to the start of the program in 2007. When tested in the similar final model shown in Tables 3 and 4, the false REACH dummy variable was not significant in either math or reading when compared to all

Table 7
HLM estimates of the relationship between REACH and test scores by subject and school year.

	Panel A: DV – math			
	2007–2008 school year		2008–2009 school year	
	Model (1)	Model (2)	Model (3)	Model (4)
REACH	0.1073 (0.057)*	0.1133 (0.061)*	0.0217 (0.062)	0.0133 (0.072)
Student controls	√	√	√	√
School controls	√	√	√	√
Prior student test scores	√	√	√	√
Prior school performance	√	√	√	√
Other reforms		√		√
Number of students	8535	8535	9148	9148
Number of schools	31	31	33	33
	Panel B: DV – reading			
	2007–2008 school year		2008–2009 school year	
	Model (5)	Model (6)	Model (7)	Model (8)
REACH	0.1160 (0.035)**	0.1100 (0.036)**	0.025 (0.039)	−0.0159 (0.045)
Student controls	√	√	√	√
School controls	√	√	√	√
Prior student test scores	√	√	√	√
Prior school performance	√	√	√	√
Other reforms		√		√
Number of students	8321	8321	8835	8835
Number of schools	31	31	33	33

Notes: Clustered standard error in parentheses. * significant at the 10% level; ** 5% level; *** 1% level.

schools in AISD. When using the limited comparison sample, the coefficients were positive but not significant at traditional levels.[10]

5.2. Student learning objectives (SLOs) and teacher effectiveness

We now turn to an investigation of the relationship between SLOs, the individual measure of teacher effectiveness in AISD REACH, and a teacher's value-added student test scores. It is worth noting that teachers in REACH schools identify student areas for improvement with the assistance of their principal. Similar to Denver's student growth objectives, teachers must specify a rationale, target population, learning content, outcome assessment, and student growth target and these decisions must be informed by baseline student achievement data.

At least one SLO must target the teacher's class as a whole, while the second SLO can focus on a particular sub-group of students. In both cases, 75% of students in the class must meet the student learning objective in order for the teacher to receive credit. Teachers are evaluated to determine whether the performance objectives have been met by submitting post-assessment data and a copy of both pre- and post-assessments if they are teacher-developed. Principals then verify that 75% of students have met the growth target, with district staff conducting an audit of a random sample of SLOs in an effort to prevent any system-gaming or errors.

Table 8 reports estimates on the relationship between SLOs and teacher effectiveness by subject. Although there appears to be a relationship in math between the percent of SLOs that a teacher meets and student achievement results in math, this relationship disappears when student level controls are included in the model. In reading there also appears to be no relationship. These results suggest that a teacher is no more likely to meet his or her SLO targets if their students have higher levels of achievement.

Although there does not appear to be an overall relationship between a teacher's value-added effect estimate and SLOs, Table 9 further delineates teachers into teacher effectiveness quartiles to investigate what types of teachers are meeting their SLOs.[11] Quartile 1 represents the teachers with the lowest residuals (the lowest mean value-added student achievement) and quartile 4 represents the teachers with the highest residuals. We see that many of the most effective

[10] Another way to test this is with both once and twice lagged test scores and another model with just twice lagged test scores. If REACH schools were chosen because of low achievement in the prior year, then it would be expected that these schools would show an improvement in test scores the following year even without the REACH program. However, we find that the coefficient increases slightly (along with a slight increase in the standard error), which suggests that increases in student test scores were not a normal correction from the prior year scores. We also see that coefficients are slightly larger when only using twice lagged prior achievement casting doubt on the idea that effects are simply regression to the mean. Another possibility is that including a twice lagged test score alters the available sample in a systematic way. To test for this, the model using only the previous test score was run with the sample of student with a non-missing value for twice lagged test scores. The coefficients maintain the same direction and significance in both math and reading.

[11] We use teacher-level residuals from our value-added model to separate teachers into quartiles.

Table 8
OLS estimates of the relationship between SLOs and teacher effectiveness by subject.

	Panel A: DV – math				
	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
Percentage of SLOs met	0.879 (0.337)**	0.209 (0.195)	0.154 (0.173)	0.125 (0.162)	0.302 (0.203)
Student prior achievement		√	√	√	√
Student demographics			√	√	√
Teacher controls				√	√
School controls					√
Number of teachers	75	75	75	75	75
	Panel A: DV – reading				
	Model (7)	Model (8)	Model (9)	Model (10)	Model (11)
Percentage of SLOs met	−0.127 (0.431)	0.014 (0.382)	0.034 (0.305)	0.122 (0.194)	0.073 (0.273)
Student prior achievement		√	√	√	√
Student demographics			√	√	√
Teacher controls				√	√
School controls					√
Number of teachers	57	57	57	57	57

Notes: Clustered standard error in parentheses. * significant at the 10% level; ** 5% level; *** 1% level.

Table 9
Percent of teachers meeting SLOs by teachers' value-added quartile in 2008–2009.

	Teacher effectiveness quartile			
	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Math				
Meet 100%	44%	37%	37%	37%
Met 50%	0%	21%	21%	16%
Met 0%	56%	42%	42%	47%
Number of teachers	18	19	19	19
Reading				
Met 100%	43%	21%	21%	33%
Met 50%	14%	36%	43%	33%
Met 0%	43%	43%	36%	33%
Number of teachers	14	14	14	15

teachers, as measured by having the highest mean value-added, did not meet any of their SLOs. In math, for example, 37% of quartile 4 teachers did not meet any of their SLOs and in reading, 33% of quartile 4 teachers did not meet any SLOs. On the other end, many teachers with the lowest mean-value added student achievement met both of their SLOs, which may suggest that higher-performing teachers are more likely to set more difficult performance standards.[12]

This lack of relationship may be due to a variety of factors. First, the baseline and final data used for SLOs did not include district administered standardized tests. This was a policy decision made by the district because teachers were already evaluated using school-wide test scores. Next, teachers were able to choose specific aspects of the curriculum as the focus of an SLO. For instance, a teacher could write an SLO about fractions instead of basing the objective on aspects of the curriculum that span the entire school year. While conceptually related, the performance of students on this narrow portion of the curriculum and overall test results may not show a statistical relationship in practice. SLOs that are aligned to district administered tests and cover more comprehensive topics may show a stronger relationship to estimates of teacher value-added. Educational policy and practice could benefit from future research that examines if teachers change their SLOs from one year to the next. For instance, do teachers set SLO performance targets that are easier (or harder) depending upon how they performed during the prior school year?

6. Discussion

The primary findings of this study are that AISD's REACH reforms were associated with improved student test score gains, and that the percentage of SLOs teachers meet has no significant association with their test score based

[12] Another way to think about the relationship between SLOs and teacher effectiveness is to examine how effectively SLOs are targeted to highly effective teachers. We estimate a series of multinomial logit models that take the following form: $SLOsmet_{jk} = \alpha + \pi_1 valueadded_{jk} + \varphi_i + \varepsilon_{jk}$, where, $SLOsmet_{jk}$ is a categorical variable taking on a value of zero if a teacher did not meet any SLOs, a value of one if a teacher met one of their SLOs, and value of two if a teacher met both of their SLOs; $valueadded_{jk}$ is an effectiveness estimate for teacher j in school k; and φ is a series of grade level dummy variables, eliminating across grade variation from the estimates. Standard errors are clustered at the school level to account for within school correlation. Here, we are most interested in the estimates on π_1, which indicates whether more effective teachers are more likely to meet one or both of their SLOs. We find no significant relationships, irrespective of whether we control for school-level characteristics.

value-added effectiveness. A naïve reading of these two findings could lead to the improper conclusion that individually determined teacher-level incentives improve overall achievement even if they have no relationship to individual teacher value-added. However, that interpretation ignores two important limitations of this study.

First, as noted above, AISD REACH has a number of important components that might have plausibly contributed to the test score gains of students attending participating schools. While there is no indication in extant literature that school-level value-added performance pay systems improve student outcomes (Fryer, 2011; Goodman & Turner 2011; Marsh et al., 2012), prior research indicates that retention bonuses for hard-to-staff schools can significantly mitigate the negative effects of teacher turnover and improve the average quality of the teacher workforce (Clotfelter et al., 2008; Glazerman et al., 2013; Springer et al., 2014). Other studies have found positive associations between formative teacher evaluation and student achievement (Taylor & Tyler 2012). The same is true for National Board Certified teachers, though evidence for an effect of the National Board Certification process is sparse (Goldhaber & Anthony, 2007; Harris & Sass, 2008). Thus, while REACH participant schools' SLO implementation was associated with improved overall student gains, it is impossible to separate the effects of the teacher-level performance incentives from the other simultaneously implemented components of the comprehensive reform.

Second, there are still important concerns about the endogeneity of assignment to program participation. These are legitimate worries even though estimates of REACH effects are statistically significant and robust to a variety of model specifications and sample restrictions (and they do not appear to be the result of regression to the mean). If the selection criteria for participation in the REACH program are based on some unobserved characteristics of the students, teachers, or administration that made them more likely to improve achievement even in the absence of the reforms (e.g., school leadership), then the matching strategy employed to construct a comparison group would fail to account for this potential omitted variable bias.

Additionally, the finding that the percentage of SLOs completed is uncorrelated with individual teachers value-added estimates does not necessarily mean that the use of SLOs in AISD REACH did not pose an effective incentive, or that SLOs should not be considered by other locations interested in performance pay policies. It is plausible that the lack of correlation between the percent of SLOs completed and value-added estimates reflects the differential difficulty of the goals set by individual teachers and their administrators. In fact, we see some evidence that the least effective teachers, based on their value-added estimates, met higher percentages of their SLO targets. If this differential goal setting resulted in more teachers feeling that their targets were attainable, the use of SLOs could have presented a stronger incentive for teachers than a conventional universal value-added target that teachers could perceive as too difficult to meet.

The inability of SLOs to identify teacher effectiveness was previously suggested by district teachers, as 61% of teachers in the AISD REACH program disagreed that the program was able to distinguish between effective and ineffective teachers (Burns, Gardner, & Meeuwsen, 2009). While the process of setting objectives may be a helpful endeavor that establishes a culture of using data and setting goals, the lack of relationship to other measures of teacher effectiveness, in this case value-added, raises some concerns. Further research is necessary to assess whether the process of individual goal setting itself results in improved performance. This study highlights the need for more rigorous evaluations of the use of SLOs. Researchers could examine the effects of the SLO practice both in the presence and absence of financial incentives and in direct contrast to conventional test-score based performance incentives.

Often districts, like other policymaking bodies, do not institute reforms one-by-one to facilitate evaluation of the effectiveness of each component, but rather simultaneously institute a comprehensive set of related reforms. This practice leads many policy researchers to either ignore reforms that are too broad in scope to be evaluated as a clearly defined construct, or downplay the potential for simultaneous treatments to bias their estimations of the effects of a specific policy. Alternatively, this study attempts to estimate the overall effect of a comprehensive set of reforms (AISD REACH) and glean some insight about a novel component (SLOs), while fully acknowledging the inherent limitations of such an exercise.

References

Ballou, D. (2001). Pay for performance in public and private schools. *Economics of Education Review*, *20*, 51–61.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65.

Belfield, C. R., & Heywood, J. S. (2008). Performance pay for teachers: Determinants and consequences. *Economics of Education Review*, *27*, 243–252.

Burns, S. F., Gardner, C., & Meeuwsen, J. (2009). *An interim evaluation of teacher and principal experiences during the pilot phase of AISD REACH*. Nashville, TN: National Center on Performance Incentives.

Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics using STATA. College Station, TX: STATA press.

Clotfelter, C., Glennie, E., Ladd, H., & Vigdor, J. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics*, *92*, 1352–1370.

Figlio, D. N., & Kenny, L. (2007). Individual teacher incentives and student performance. *Journal of Public Economics*, *91*, 901–914.

Fryer, R. G. (2011). *Teacher incentives and student achievement: Evidence from New York City public schools*. Cambridge, MA: Harvard University.

Fryer, R. G., Levitt, S. D., List, J., & Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. NBER Working Paper No. 18237. Cambridge, MA.

Fulbeck, E. S. (2014). Teacher mobility and financial incentives: A descriptive analysis of Denver's ProComp. *Educational Evaluation and Policy Analysis*, *36*(1), 67–82.

Glazerman, S., Protik, A., The, B., Bruch, J, & Max, J (2013). *Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment (NCEE 2014-4003)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Glazerman, S., & Seifullah, A. (2010). *An evaluation of the Teacher Advancement Program (TAP) in Chicago: Year two impact report*. Washington, DC: Mathematica Policy Research, Inc.

Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Review*, 205–227.

Gilpin, G. A. (2012). Teacher salaries and teacher aptitude: An analysis using quantile regressions. *Economics of Education Review*, *31*(1), 15–29.

Goe, L. (2008). *The link between teacher quality and student outcomes: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D. (2002). The mystery of good teaching. *Education Next*, *2*(1), 50–55.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National Board certification as a signal of effective teaching. *The Review of Economics and Statistics*, *89*(1), 134–150.

Goldhaber, D. (2009). *Teacher pay reforms: The political implications of recent research*. Washington, DC: Center for American Progress.

Goldhaber, D., & Walch, J. (2012). Strategic pay reform: A student outcomes-based evaluation of Denver's ProComp teacher pay initiative. *Economics of Education Review*, *31*, 1067–1083.

Goodman, S., & Turner, L. (2011). Does wholeschool performance pay improve student learning? Evidence from the New York City schools. *Education Next*, *11*, 66–71.

Gordon, R., Kane, T., & Staiger, D. (2006). *Identifying effective teachers using performance on the job*. Washington, DC: Brookings Institution.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *Journal of Human Resources*, *39*(2), 326–354.

Harris, D. N. & Sass, T. R. (2008). *The effects of NBPTS-certified teachers on student achievement*. CALDER Working Paper #4. Washington, DC.

Hendricks, K. (2011). Examining the impact of school climate on student achievement: A retrospective study (Doctoral disseration). Retrieved from Proquest Dissertations and Theses. (Accession Order No. AAT 1015363752).

Jackson, C. K. (2012). Recruiting, retaining, and creating quality teachers. *Nordic Economic Policy Review*, *3*(1), 1–52.

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

Johnson, S. M., & Papay, J. P. (2009). *Redesigning teacher pay: A system for the next generation of educators*. Washington, DC: Economic Policy Institute.

Kane, T., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010). *Identifying effective classroom practices using student achievement data*. NBER Working Paper No. 15803. Cambridge, MA.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, *24*(1), 37–62.

Lavy, V. (2002). Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of Political Economy*, *110*, 1286–1317.

Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review*, *99*, 1979–2011.

Leigh, A. (2012). Teacher pay and teacher aptitude. *Economics of Education Review*, *31*(3), 41–53.

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., Epstein, S., Koppich, J. E., et al. (2011). *A big apple for educators: New York City's experiment with schoolwide performance bonuses*. Santa Monica, CA: RAND Corporation.

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., Epstein, S., Koppich, J., Kalra, N., DiMartino, C., & Peng, A. (2012). *A big apple for educators: New York City's experiment with schoolwide performance bonuses*. Santa Monica, CA: RAND.

McCaffrey, D., Han, B., & Lockwood, J. R. (2009). Turning student test scores into teacher compensation systems. In M. G. Springer (Ed.), *Performance incentives: Their growing impact on American K-12 education*. Washington, DC: Brookings.

McCaffrey, D., Koretz, D., Lockwood, J. R., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.

Muralidharan, K., & Sundararaman, V. (2011a). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, *119*(1), 39–77.

Muralidharan, K., & Sundararaman, V. (2011b). Teacher opinions on performance pay: Evidence from India. *Economics of Education Review*, 394–403.

Murnane, R., & Cohen, D. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review*, *56*(1), 1–17.

Pianta, Paro L., & Hamre (2006). *Classroom Assessment Scoring System: Preschool version*. Charlottesville, VA: Center for Advanced Study of Teaching and Learning.

Podgursky, M. (2006). Teams versus bureaucracies: Personnel policy, wage setting, and teacher quality in traditional public, charter, and private schools. In M. Berends, M. Springer, & H. Walberg (Eds.), *Charter school outcomes* (pp. 61–84). New York, NY: Lawrence Erlbaum Associates.

Podgursky, M. (2010). Teacher compensation and collective bargaining. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.). *Handbook of the economics of education: Vol. 3* (pp. 279–313). Amsterdam: North-Holland.

Podgursky, M., & Springer, M. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management*, *26*, 909–950.

Podgursky, M., & Springer, M. (2011). Teacher compensation systems in the United States K-12 Public School System. *National Tax Journal*, *64*(1), 165–192.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*, 417–458.

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*, 247–252.

Rubin, D., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103–116.

Santibañez, L., Martinez, J. F., Datar, A., McEwan, P., Setodji, C., & Basurto-Davila, R. (2007). *Breaking ground: Analysis of the assessment system and impact of Mexico's teacher incentive program 'Carrera Magisterial'*. Santa Monica, CA: RAND Corporation.

Schochet, P. Z., & Hanley, C. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Springer, M. G. (2007). *Performance incentives: Their growing impact on American K-12 education*. Washington, DC: Brooking Institution Press.

Springer, M. G., & Balch, R. (2010). Design components of incentive pay programs in the education sector. In S. Sclafani (Ed.), *Teacher Incentives and Stimuli*. Paris: Organisation for Economic Co-Operation and Development.

Springer, M. G., Ballou, D., Hamilton, L., Le, V., Lockwood, J. R., McCaffrey, D. F., et al. (2012). *Teacher pay for performance: Experimental evidence from the project on incentives in teaching*. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Springer, M. G., Ballou, D., & Peng, A. (2014). Estimated effect of the teacher advancement program on student test score gains. *Education Finance and Policy*, *9*(2), 193–230.

Springer, M. G., Pane, J. F., Le, V., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., et al. (2012). Team pay for performance: Experimental evidence from the Round Rock Pilot Project on team incentives. *Educational Evaluation and Policy Analysis*, *34*(4), 367–390.

Springer, M. G., Rodriguez, L., & Swain, W. (2014). Effective teacher retention bonuses: evidence from tennessee. Tennessee Consortium on Research, Evaluation, and Development Working Paper. Nashville, Tennessee.

Steele, J. L., Murnane, R. J., & Willett, J. B. (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.

Taylor, L. L., & Springer, M. G. (2010). *Optimal incentives for public sector works: The case of teacher-designed incentive pay in Texas*. Nashville, TN: National Center of Performance Incentives.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(7), 3628–3651.

Thum, Y. M., & Bryk, A. S. (1997). Value-added productivity indicators: The Dallas system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin Press.

Webster, W. J., & Meandro, R. L. (1997). The Dallas value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin Press.

Wiley, E.W., Spindler, E.R., & Subert, A.N. (2010). *Denver ProComp: An outcome evaluation of Denver's alternative teacher compensation system*. Boulder Working Paper. University of Colorado.

Winters, M., Ritter, G., Barnett, J., & Greene, J. (2006). *An evaluation of teacher performance pay in Arkansas*. University of Arkansas, Department of Education Reform.

Woessmann, L. (2011). Cross-country evidence on teacher performance pay. *Economics of Education Review*, *30*(3), 404–418.

Yuan, K., Le, V., McCaffrey, D., Marsh, J. A., Hamilton, L. S., Stecher, B., et al. (2013). Incentive pay programs do not affect teacher motivation or reported practices: Results from three randomized studies. *Educational Evaluation and Policy Analysis*, *35*(1), 3–22.

Further reading

Hanushek, E. A. (2007). The single salary schedule and other issues of teacher pay. *Peabody Journal of Education*, 574–586.