Educational Evaluation and Policy Analysis
http://eepa.aera.net

Incentive Pay Programs Do Not Affect Teacher Motivation or Reported Practices : Results From Three Randomized Studies

Kun Yuan, Vi-Nhuan Le, Daniel F. McCaffrey, Julie A. Marsh, Laura S. Hamilton, Brian M. Stecher and Matthew G. Springer

EDUCATIONAL EVALUATION AND POLICY ANALYSIS 2013 35: 3 originally published online 12 November 2012
DOI: 10.3102/0162373712462625

The online version of this article can be found at:
http://epa.sagepub.com/content/35/1/3

Published on behalf of

American Educational Research Association

and

http://www.sagepublications.com

Educational Evaluation and Policy Analysis
March 2013, Vol. 35, No. 1, pp. 3–22
DOI: 10.3102/0162373712462625

Incentive Pay Programs Do Not Affect Teacher Motivation or Reported Practices: Results From Three Randomized Studies

Kun Yuan
Vi-Nhuan Le
Daniel F. McCaffrey
RAND Corporation

Julie A. Marsh
University of Southern California, Los Angeles

Laura S. Hamilton
Brian M. Stecher
RAND Corporation

Matthew G. Springer
Vanderbilt University

This study drew on teacher survey responses from randomized experiments exploring three different pay-for-performance programs to examine the extent to which these programs motivated teachers to improve student achievement and the impact of such programs on teachers' instruction, number of hours worked, job stress, and collegiality. Results showed that most teachers did not report their program as motivating. Moreover, the survey responses suggest that none of the three programs changed teachers' instruction, increased their number of hours worked or job stress, or damaged their collegiality. Future research needs to further examine the logic model of pay-for-performance programs and test alternative incentive models such as rewarding teachers based on their practices and job responsibilities rather than on student outcomes.

Keywords: *pay-for-performance program, randomized experimental studies, teacher*

Introduction

Traditionally, teachers have been paid on a single salary schedule, in which teacher salary is based on their years of experience and education level (Podgursky & Springer, 2006). Recently, there have been a number of reform initiatives that have attempted to change the manner in which teachers are compensated. In particular, pay-for-performance programs, also referred to as incentive pay programs, in which teachers' pay is linked to some aspects of their performance, have gained substantial popularity in the last decade (Podgursky & Springer, 2006). Various efforts at

the national, state, and district levels have been made to design, implement, and study the effectiveness of performance pay programs on educators' attitudes and students' learning (Center for Educator Compensation Reform, n.d.; Koppich & Rigby, 2009; Schuermann, 2009).

For incentive pay programs to lead to improved student outcomes in the short run, the programs must motivate teachers to make changes that effectively lead to the accomplishment of program goals (Kelley, Odden, Milanowski, & Heneman, 2000). However, the literature paints a mixed picture of the extent to which incentive pay programs motivate teachers. Some studies suggest that most teachers are supportive of tying pay to performance and feel motivated by incentive pay programs (Langdon & Vesper, 2000), whereas other studies find that the majority of teachers oppose linking pay increases to performance (Teaching Commission, 2004).

Purpose and Structure of This Article

Despite the increasingly popular practice of using pay-for-performance programs to reward educators in the United States (Fryer, 2012) and the fact that the impact of incentive pay programs on student achievement depends in part on how teachers respond to such programs, only a few studies have examined teachers' responses to incentive pay programs (notable exceptions include Glewwe, Ilias, & Kremer, 2010; Marsh et al., 2011; Muralidharan & Sundararaman, 2008; Springer et al., 2010; Springer et al., 2012; Wiley, Spindler, & Subert, 2010). The goal of this study is to examine whether three recently implemented pay-for-performance programs had similar effects on teachers' motivation and reported practices. These programs were (a) the Project on Incentives in Teaching (POINT) in Tennessee's Metropolitan Nashville Public Schools (MNPS) (Springer et al., 2010), (b) the Pilot Project on Team Incentives (PPTI) in Texas's Round Rock Independent School District (RRISD) (Springer et al., 2012), and (c) the School-Wide Performance Bonus Program (SPBP) in New York City Public Schools (Marsh et al., 2011).

All three programs rewarded teachers primarily on the basis of student achievement gains but differed in program design, such as unit of accountability and incentive structure. Across the three sites, teachers were surveyed about their attitudes toward the incentive pay programs, their instructional practices, and about ways in which the implementation of the incentive pay programs changed their teaching behaviors or working conditions. Drawing on teacher survey data to explore commonalities and differences across the three different incentive programs, we addressed the following research questions:

1. Did teachers find these three incentive pay programs to be motivating?
2. In response to the implementation of these programs, did teachers report changes in their practices or their working conditions?

In the following sections, we review the rationale for pay-for-performance programs and prior research findings about their impact on teachers' motivation, practices, and working conditions.[1] Next, we describe the three programs as well as the surveys and analytical approach used in this study. Finally, we present findings and discuss the implications for future policy and research on incentive pay programs.

Rationale for Pay-for-Performance Programs

There are three primary methods by which pay-for-performance programs are thought to improve student achievement. The first mechanism by which pay-for-performance programs can improve student achievement is by motivating existing teachers to improve or innovate their teaching practices. The theory posits that teachers will respond to financial incentives by altering their teaching practices, which may include changing the way they teach, modifying the content of their curriculum, working longer hours, undergoing more professional development, or engaging in different types of professional development activities. These efforts are believed to result in better student achievement (Milanowski, 2003; Wiley et al., 2010).

The second mechanism is to improve student learning by changing teachers' working environment. Although much of the rationale for pay-for-performance programs focuses on improvement of student achievement, there has been an increasing recognition that such programs may

have consequences on teachers' work environment. Namely, the extent to which teachers are motivated to improve their teaching will depend on the outcomes that arise from trying to attain the program goals (Goodman & Turner, 2011). In theory, pay-for-performance programs that award bonuses based on teams of teachers may encourage cooperation and teamwork, which in turn could support improvements in instruction and ultimately student learning.[2]

The third method is by changing the supply of teaching candidates through incentives that attract or retain more highly qualified candidates. The prospect of a salary structure that awards bonuses for higher student achievement may attract a more highly qualified pool of candidates who feel they could thrive under this type of salary structure (Barnett & Ritter, 2008). Alternatively, given that teachers who tend to leave the field are the most highly skilled (Margolis, 2008) and that teachers who had higher compensation stayed in teaching longer (Clotfleter, Glennie, Ladd, & Vigdor, 2008; Guarino, Santibanez, & Daley, 2006), pay-for-performance programs may retain more talented teachers who would otherwise leave the field for other higher paying opportunities.

The first and second mechanisms were the primary basis for the adoption of the pay-for-performance programs included in our study. Although an important rationale for pay-for-performance programs, the third mechanism was not tested in the evaluation of any of the three programs given the short duration of the programs and limited data, and we will not consider this mechanism for change in this article.

It is important to recognize that the mechanisms examined by our study and the tension between rewarding individuals versus groups reflect a long-running debate in social theory. The "utilitarian tradition"—associated with John Locke, Adam Smith, and in later years economists such as Milton Friedman—argues that the individual is the starting point of society and that rational individuals attempt to maximize their utility, calculating their self-interest and how they make social exchanges (Collins, 1994). In contrast, the "Durkheimian tradition" rejects the idea that individual action can explain societal features and instead asserts that society determines the individual. In other words, structural relations among people—social and moral ties, rituals, norms—hold society together and greatly affect individual behavior and ideas (Collins, 1994).

The study of individual versus group incentives is one particular case of this deeper, theoretical discussion and set of tensions. Echoing the utilitarian tradition, the first mechanism asserts that individual rewards will have the greatest motivational effect: Individuals will alter their behavior to maximize their probability of earning a reward. Falling clearly within the Durkheimian tradition, the second mechanism argues that collective rewards will show more of an effect. Group-based rewards rest on a belief that the quality of the school as a whole takes priority over the success of the individual teachers (Ford, 2012) and that individual performance is dependent on the collective efforts of a school community and the quality of the school environment. Those advocating for collective rewards further argue that although the first mechanism may motivate individual teachers, it may not improve the educational experiences for all students in the schools. Depending on program design, the pursuit of individual rewards may also create a competitive environment that negatively affects collective teacher motivation and performance. Although the second mechanism has the possibility of improving the educational experiences for groups of students, it also suffers from a collective action problem, in which individuals may still benefit, despite not contributing to the group (Oliver, 1980). This may lead to teacher inaction due to the belief that individual contributions will have little influence on obtaining a bonus or to teacher resentment due to the perceived free-rider problem (Willer, 2009). In this article, we examine these two competing strategies for incentivizing teachers, albeit with limited data from three programs.

Motivational Theories Underlying Pay-for-Performance Programs

Although none of the programs in our study included an explicit theoretical framework underlying their designs, the literature generally puts forth two theoretical frameworks that help us understand the possible effects of performance incentives on teachers' motivation: expectancy

theory (Vroom, 1964) and goal-setting theory (Locke, 1968).

We selected these psychological theories to guide our analysis for several reasons. First, we believed these theories could apply to the heterogeneity of the programs studied and of the underlying rationales. Although subscribing to slightly different ideas of how to leverage change, all three programs emphasized the motivational aspects of incentives. One could argue expectancy theory applies not only to individuals—rewards boost individual expectancy—but also to groups—rewards could also boost collective expectancy. Our surveys were also developed to test individual psychological differences, making the choice of these theories appropriate for framing. Finally, our selection of these theories was further influenced by past studies of similar pay-for-performance programs that relied on these same theoretical frameworks (Adkins, 2004; Kelley, Heneman, & Milanowski, 2002).

Expectancy theory identifies three factors that contribute to motivating teachers to engage in certain behaviors: expectancy, instrumentality, and valence. According to this theory, the amount of effort that teachers expend to meet the program goals is a function of teachers' perceptions that their personal efforts will lead to their students reaching the expected achievement goals (expectancy), the likelihood that meeting the achievement goals will result in a financial reward (instrumentality), and the desirability of attaining the bonus (valence).

Goal-setting theory complements expectancy theory by further describing the conditions under which teachers' actions are likely to be changed by the incentive. Under goal-setting theory, teachers are predicted to respond to the incentives if the goals under the performance pay systems are clearly defined, moderately challenging, and accepted by teachers. In addition, goal-setting theory suggests that teachers' effort will be influenced by a number of factors, such as their understanding of the incentive pay program and its goals (Kelley, 1999) and perceptions of the fairness of the program (Heneman & Milanowski, 1999).

Evidence related to expectancy theory. Recent evaluations of pay-for-performance programs suggest that incentive pay programs may have mixed effects on teachers' motivation. For instance, Kelley et al. (2002) reported that 56% of teachers in the Charlotte-Mecklenburg Schools (CMS) and 39% in Kentucky estimated that their school would meet its program goals if they put in a high level of personal effort. The fact that the rewards were given at the school level may have contributed to the modest expectancy estimates from individual teachers. Examining a program in a Florida district that rewarded individual teachers based on performance measures such as student achievement gains and participation in staff development, Adkins (2004) noted that 65% of teachers believed that their individual performance as an educator had little influence on whether or not they would earn a bonus. In a survey with Maryland teachers in a school district where the compensation system included incentives for various aspects of teaching, such as school-wide achievement growth and teaching evaluations, approximately one-third of teachers did not think their payouts were linked to their performance in the classroom (Rice, Jackson, Hoyer, Malen, & Hyde, 2011).

The relatively low expectancy may stem from teachers' concerns that student achievement is influenced by factors that are outside of their control (Heneman, 1998). For example, Adkins (2004) found that although nearly all teachers believed that their individual performance as an educator could significantly influence their students' achievement, the majority also believed that the classroom composition would have a greater impact on whether they met their achievement goals than did their individual effort as educators.

Studies also suggest that teachers do not always have high estimates of instrumentality with respect to performance pay systems. Kelley et al. (2002) asked teachers to estimate the probability of receiving a bonus if the school met its achievement goals. Nearly 30% of CMS teachers and 45% of Kentucky teachers did not believe a bonus would be awarded, even if the school attained its achievement goals.

With respect to valence, the literature generally suggests that teachers value the opportunity to win a bonus. The majority of teachers in the Kelley et al. (2002), Adkins (2004), and Rice et al. (2011) studies indicated they wanted to receive a bonus, although teachers also indicated

that their efforts to improve student achievement were driven by other nonfinancial reasons (e.g., helping students learn). In some instances, although teachers valued the award, they also believed the award was too small to be worth any additional effort (Kelley et al., 2002; Rice et al., 2011).

Evidence related to goal-setting theory. Similar to the results from an expectancy theory perspective, studies using a goal-setting theory framework paint a mixed picture as to whether incentive pay programs motivate teachers. In general, teachers understand the overarching goals of the pay-for-performance programs. For example, the majority of teachers in the Adkins (2004) study agreed with the premise that educators should receive additional compensation for meeting achievement goals and indicated that they understood the process for awarding incentive pay to teachers. Similarly, most teachers in CMS and Kentucky believed the program goals were clear (Kelley et al., 2002). Examining Denver Public Schools' Professional Compensation System for Teachers (ProComp), Wiley et al. (2010) reported that most teachers felt the program's goals of improving student achievement and retaining and attracting qualified teachers were aligned with the district's and their own goals.

However, teachers were more cautious with respect to whether the programs were fair. For example, most Kentucky teachers disagreed that it was fair to pay bonuses to teachers for achievement gains and to hold teachers accountable for student achievement (Kelley et al., 2002). Similarly, three-quarters of responding teachers in the Adkins (2004) study believed that their program was not fair in how it distributed the bonuses. In the Rice et al. (2011) study, about one-third of the teachers questioned the fairness in how teaching was evaluated and the bonus was distributed.

Effects of Pay-for-Performance Programs on Teacher Practices and Working Conditions

Teacher practices. The literature on whether teachers change their practices in response to incentive pay programs is mixed. Two experimental studies in India and Kenya that examined the effect of incentive pay programs on teachers' practices found no significant differences in teachers' practices between treatment and control group teachers based on observations of instruction (Glewwe et al., 2010; Muralidharan & Sundararaman, 2008). In contrast, using a regression discontinuity and matching methods design to study a pay-for-performance program in Israel, Lavy (2009) found that treatment group teachers reported greater use of individualized instruction, more tracking in the classroom by ability, and longer instructional time than control group teachers.

Bonus-eligible teachers in observational studies of incentive pay programs conducted in the United States also report mixed results. For example, only 27% of teachers who were required to participate in the Denver ProComp system reported using different teaching methods and even fewer (14%) reported changing the content of instruction (Wiley et al., 2010). Similarly, about 80% of teachers in the Adkins (2004) study reported that they did not change their instruction, assessment methods, or professional development activities as a result of teacher performance pay.

In contrast, more than two-thirds of teachers in Kentucky and CMS reported making changes in instruction, including changing instructional content, increasing instructional time on teaching tasks, and collaborating more with colleagues (Kelley, 1999). However, the teachers in these studies were concurrently undergoing standards and assessment-based reforms. In the absence of responses from a control group, it is unclear the extent to which the changes were due to the pay-for-performance programs, as opposed to the other reform efforts.

Number of hours worked. Many studies have found that the implementation of pay-for-performance programs is associated with increases in the number of hours teachers report working. For example, 37% of responding teachers in the Adkins (2004) study reported working longer hours since the implementation of the program. Upwards of two-thirds of CMS teachers and three-quarters of Kentucky teachers indicated they worked longer hours since their program started in the Kelley et al. (2002) study. However, again, because these studies only examined bonus-eligible teachers, changes in the number of

hours worked may have partly resulted from other reform efforts that happened concurrently with the incentive pay programs.

Job stress. Many studies have also found that teachers report an increase in their job stress after the implementation of incentive pay programs. For instance, 72% of CMS teachers and 87% of Kentucky teachers reported the award program put more job pressure and stress on them (Kelley, 1999). Nearly half of the responding teachers in the Adkins (2004) study reported they experienced increased stress as a result of the pay-for-performance program. Half of the teachers who were required to participate in ProComp reported increased stress (Wiley et al., 2010). However, given the lack of control group teachers' responses, it is unknown whether the increased job stress was due solely to the incentive pay programs.

Collegiality. Although there are few studies that have examined the issue of collegiality, the general consensus of these studies is that the implementation of pay-for-performance program is associated with slightly higher levels of collegiality. Kelly (1999) reported that the school-wide performance pay programs might have helped teacher collaboration and increased teacher collegiality in the CMS and Kentucky programs. Hall and Caffarella (1997) reported similar findings about the positive association between the implementation of a group-based incentive pay program and collegiality in Douglas County, Colorado. About half of the incentive eligible teachers in the individual-based ProComp program reported that the program helped create a collaborative working environment (Wiley et al., 2010).

In summary, existing findings about the impact of incentive pay programs on teachers' practices are inconclusive. The literature is more consistent with respect to the effects on number of hours worked, job stress, and collegiality, with studies generally reporting increases in these areas after the implementation of pay-for-performance programs. However, it is difficult to disentangle the effects of the pay-for-performance programs from the effects of other concurrent reform efforts due to the lack of responses from control group teachers in the latter findings.

Method

In this study, we examined three randomized controlled trials of pay-for-performance programs, in which the units of accountability were randomly assigned to either the treatment group (who were eligible to receive the bonus) or the control group (who were not eligible to receive the bonus). Below we provide an overview of each program.

Sample

POINT. The POINT experiment was conducted in the MNPS from 2006–2007 through 2008–2009, during which the district served about 73,000 students annually. Slightly more than one-third of the students were White. Half were Black. Hispanic students were 15% of the student population. About 65% of the students in the district were eligible for free and reduced-price lunch (FRPL). Ten percent were English language learners (ELLs). The district's performance on the statewide mathematics and reading tests was below the state average.

POINT is an individual-based bonus program that rewarded teachers based on their value-added scores, which were calculated as the average of their students' year-to-year growth on the statewide mathematics test, adjusted by the statewide average gain. All teachers who met fixed performance thresholds would earn bonuses. Teachers in the treatment group were eligible for a financial reward of $5,000, $10,000, or $15,000, if their value-added scores, respectively, reached the 80th, 85th, and 95th percentile of the historic, school year, and distribution of teachers' value-added scores from the 2004–2005 and 2005–2006 school years. Teachers in both the treatment and the control groups received a stipend of $750 for each year of participation.

In the initial year of the study, 296 mathematics teachers in fifth through eighth grade who taught at least 10 students volunteered to participate in POINT. In total, 143 and 140 teachers[3] were randomly assigned to the treatment and control group, respectively, and were expected to remain in the same experimental condition for all three program years. By the final program year, only 84 treatment group teachers and 64 control group teachers remained in the study. The rate of

attrition observed among the sample participants was consistent with historical rates of turnover among middle school mathematics teachers in the district and reflected attrition due to teachers leaving the district, teachers moving to elementary or high school, teachers no longer teaching mathematics, or teachers teaching fewer than 10 mathematics students.[4]

The number of bonus winners in each year ranged from 41 to 44. In total, 51 treatment group teachers received a bonus during the three years of the POINT study. Among them, 16 teachers won a bonus once, 17 won twice, and 18 won a bonus in all three years. The average bonus award distributed to teachers each year ranged from $9,623 to $11,370.

PPTI. The RRISD is a suburban district in Texas with above average achievement on the state accountability tests. During the implementation of PPTI, it served about 43,000 students annually, which comprised of 46% White, 30% Hispanic, 10% Asian, and 8% Black students. About one-quarter of the students were eligible for FRPL. Seven percent of the district's student population was ELL.

RRISD organized middle school teachers into multiple grade-level interdisciplinary teams in each school consisting of at least one teacher for each core subject of mathematics, reading/English language arts, science, and social studies. Some teams also included special education teachers and specialists for ELL. Teams could change across years but always included at least one teacher from each of the core subject areas. Teams were randomized to either the bonus intervention or control condition using a block-randomized design. Blocks were defined by grades within school. Within each block, there were multiple teams. When there was an even number of teams, half the teams in each block were randomized to treatment and half to control. In blocks with three teams (no blocks had more than four teams), two teams were randomly assigned to treatment or control, and the remaining team was assigned to the other condition. The randomizations were constrained so that the number of treatment and control teams was balanced at each grade level.

PPTI rewarded teaching teams based on the team's average value-added scores. Value-added scores for teachers in each subject area were calculated, and the overall performance measure for the team was the average of its contributions to each of the four subject areas. Bonuses were awarded to a team whose score ranked in the top third of all teams at the same grade level in the treatment group. If a team fell just below the bonus threshold and would have earned a bonus had another team in the same school not outperformed it, the nonqualifying team was also designated a bonus winner. This stipulation ensured that no team close to earning a bonus would be denied a bonus because another team in the same school had outperformed it.

In 2008–2009, 78 teaching teams at the sixth-, seventh-, and eighth-grade levels participated in PPTI. They were randomly assigned to either the treatment or control group. Thirty-nine out of 78 participating teams were assigned to the treatment group. In the first program year, 67 teachers on 14 teams won bonuses, with an average bonus of $5,373 per teacher. Teams of teachers were rerandomized to the treatment or the control group in the second program year, in which 40 out of 81 teams were assigned to the treatment group. In the second program year, 52 teachers on 12 teams received a bonus, with an average payout of $5,862 per teacher.

SPBP. From 2007–2008 to 2009–2010, the New York City Department of Education (NYCDOE) and United Federation of Teachers (UFT) implemented SPBP, a school-based performance pay program for high-needs K-12 schools, as defined by their poverty rates, student demographic characteristics, and fourth- and eighth-grade scores on the statewide mathematics and English language arts tests (Marsh et al., 2011). The district identified 427 eligible schools at the beginning of the program. The percentage of Hispanic and Black students was 55% and 40%, respectively. Over 80% of the students were eligible for FRPL, and 45% were ELL.

SPBP rewarded schools according to their Progress Report scores issued by the NYCDOE. Schools' Progress Report scores were calculated based on a number of criteria, including student test scores, graduation rates (for high schools), student attendance, and school environment, as measured by student, teacher, and parent responses to a district-administered survey.

However, the most important factor for awarding bonuses was student achievement. Schools that met their annual performance targets, as defined by the Progress Reports, were eligible to receive full bonuses equal to $3,000 per full-time UFT-represented staff member working at each school. Schools that met 75% of their targets received a school-level partial bonus of $1,500 per full-time UFT-represented staff member. Within each school, a four-person compensation committee was established to determine how to distribute the bonus among staff members. Most compensation committees decided to distribute the bonus equally among eligible staff members or give smaller awards to teachers who did not work in the school all year.

The initial randomization of schools resulted in 234 treatment schools and 168 control schools,[5] and they remained in their assigned group for the entire duration of the study. Each year, eligible treatment schools needed to secure approval of at least 55% of its UFT-represented staff to participate. After staff voting and some withdrawals and school closings in later years, the number of treatment group schools was 199, 191, and 189 in each of the three program years, respectively. The number of schools in the control group was 168, 167, and 167, respectively, in each of the three program years. In each of the three years of SPBP, 62%, 84%, and 13% of schools won full or partial bonuses.[6] The within-school average payout was $2,857, $2,841, and $2,812 per staff member in each of the three program years, respectively.

Surveys

All three studies conducted online surveys of participating teachers from both groups. For POINT, teachers were surveyed once during each of the three program years. For PPTI, teachers were surveyed twice during each of the two program years, once in the fall or winter and once in the spring. For SPBP, seven classroom teachers were randomly selected from each participating school and surveyed during the last program year.[7] Four of them taught subjects or grades that were included in the statewide achievement assessment program, including mathematics or English language arts teachers at the elementary and middle school levels and high school teachers who taught subjects tested by the Regents exam. The remaining three teachers taught nontested subjects or grades.

Teachers took the surveys online. An initial email contained information about survey content, details about the stipend offered for participating in the survey, and an explanation of how confidentiality would be protected. Response rates for the final program year were 98% for POINT, 91% for PPTI, and 58% for SPBP.

Surveys used in all three programs asked teachers about their understanding of the program, motivation, and practices. Most of the survey questions were either on a 4-point Likert-type scale that asked about the extent to which teachers agreed with the item (e.g., where 1 indicated *strongly disagree* and 4 indicated *strongly agree*) or a 6-point Likert-type scale that asked about the extent that teachers engaged in a particular practice (e.g., where 1 indicated *never* and 6 indicated *almost daily*). Surveys used in PPTI and SPBP also collected teachers' demographic information such as gender, ethnicity, highest education level attained, and years of teaching experience. We collected the same type of information from district records for POINT teachers.

Across the three studies, there were similarly worded items that assessed teachers' perceptions and understanding of the incentive pay program, as well as various practices that teachers may be likely to change as a result of being eligible to receive a bonus. Based upon these common survey items, we created 10 scales measuring key constructs related to teachers' motivation and practices. To create these composite measures, we reviewed each of the survey questions, computed descriptive statistics for all item-level responses, and conducted exploratory factor analyses where appropriate. Scale scores were calculated by averaging item scores over all responded items in a scale. In other instances, we created item-level scales by dichotomizing the 4-point Likert-type responses (*strongly agree/agree* vs. *strongly disagree/disagree*). Thus, these scales represented the percentage of teachers who endorsed the items. The complete list of items and scales, along with the internal consistency reliability of each scale, is presented in Table 1.[8]

Analytic Approach

We analyzed teachers' motivation and attitudes toward incentive pay programs based on responses from incentive eligible teachers in

TABLE 1

Common Survey Scales and Reliability Coefficients for Composite Scales

Scale	Description of the Scale/Item With Score Scale in Parenthesis	Number of Items	Alpha
Goal acceptance			
Skill-based pay	Teachers should be compensated for demonstrating outstanding teaching skills (A)[a]	1	—
Performance measure	Rewarding teachers based on test score gains is problematic (A)	1	—
Program understanding	I have a clear understanding of the performance criteria for earning a bonus award (A)	1	—
Program fairness	The method used to award bonuses is fair to teachers (A)	1	—
Expectancy			
Teachers' self-efficacy	Teachers' confidence in teaching (A)	4	0.65
Teachers' impact on student achievement	Teachers' impact on student achievement is limited due to the effect of home environment (A)	1	—
Chance to win a bonus	What is the chance you/your team will receive a bonus based on this year's performance (N)	1	—
Valence			
Desire to earn a bonus	I have a strong desire to earn a bonus (A)	1	—
Motivating effect of bonuses	The chance to earn a bonus award has energized me to improve my teaching (A)	1	—
Desirability of bonuses	The bonus amount is not large enough to motivate extra effort (for POINT and SPBP)[b] (A)	1	—
Instruction			
Focus on standards	Frequency of activities to align instruction with state standards (F) (G)[c]	2	0.51
Data-driven decision-making	Use of student test scores to guide instruction (E)	9	0.89
Test preparation	Importance teachers placed on test-preparation activities (I)	4	0.81
Changes in instruction	Changes made in classroom emphasis on state standards and tests (C)	6	0.84
Changes in student learning	Changes made in the emphasis on hands-on activities and having students work in groups (C)	2	0.77
No program impact on practices	The program did not affect my teaching practices or professional behaviors (A)	1	—
Number of hours worked			
Total extra working hours per week	Number of hours worked outside of formal school hours on a weekly basis (N)	1	—
Portion of extra working hour on routine tasks	Portion of extra working hours spent on preparing lessons, grading, and administrative tasks (P)	3	0.53
Portion of extra working hour on other tasks	Portion of extra working hours spent on professional development, meeting with students, colleagues, and parents, and maintaining class website (P)	6	0.67
Job stress			
Program impact on job stress	I have experienced increased job stress as a result of the program (for POINT)[d] (A)	1	—
Collegiality			
Positive relationship with colleagues	Positively worded items about relationship with colleagues such as cooperation and mutual help (A)	2	0.76
Negative relationship with colleagues	Negatively worded items about relationship with colleagues such as competitiveness and lack of trust (A)	2	0.72
Negative program impact on relationships among colleagues	I have noticed increased resentment among teachers since the start of the program (A)	1	—

Note. POINT = Project on Incentives in Teaching; SPBP = School-Wide Performance Bonus Program; PPTI = Pilot Project on Team Incentives.

[a]Please refer to the following notation for the score scale used for each scale/item. A: 4-point Likert-type scale, with 1 = *strongly disagree,* 2 = *disagree*, 3 = *agree*, and 4 = *strongly agree*. C: 5-point Likert-type scale, with 1 = *much less than last year*, 2 = *a little less than last year*, 3 = *the same as last year*, 4 = *a little more than last year*, and 5 = *much more than last year*. E: 4-point Likert-type scale, with 1 = *not used in this way*, 2 = *used minimally*, 3 = *used moderately*, and 4 = *used extensively*. F: 6-point Likert-type scale, with 1 = *never*, 2 = *once or twice a year*, 3 = *once or twice a semester*, 4 = *once or twice a month*, 5 = *once or twice a week*, and 6 = *almost daily*. G: 5-point Likert-type scale, with 1 = *never*, 2 = *every few months*, 3 = *once or twice a month*, 4 = *once or twice a week*, 5 = *almost daily*. N: numerical scale. P: 4-point Likert-type scale, with 1 = *none*, 2 = *a small portion*, 3 = *a moderate portion*, and 4 = *a major portion*.

[b]The corresponding PPTI survey item was "The size of the bonus award is too small to motivate me to work harder" (A).

[c]POINT and PPTI survey items used a 6-point scale, whereas SPBP survey items used a 5-point scale.

[d]The corresponding SPBP survey items asked teachers to report changes in their relationships with administrators, other teachers, and other non-classroom teaching staff members. The score scale was a 5-point Likert-type scale, with 1 = *changed significantly for the worse*, 2 = *changed slightly for the worse*, 3 = *did not change*, 4 = *changed slightly for the better*, and 5 = *changed significantly for the better*.

three programs. To allow teachers maximum time to fully understand the program and its effects, we examined the survey responses obtained in the final year for each program. In addition, we examined the effects of incentive pay programs on teachers' instruction, number of hours worked, job stress, and collegial relationships based on three types of results, including differences in teachers' practices between the treatment and control group in each program,[9] teachers' reported changes in practices, and changes in the responses of POINT teachers who remained in the program for three years (referred to as POINT persistent teachers below) between the first and third program year surveys.

We used a two-level mixed-effects hierarchical linear model to analyze differences between treatment and control group teachers within each program. Level 1 models individual teacher survey responses (Y_{ij}) as a function of a team or school mean (θ_j), teacher-level covariates (X_i^T) such as teachers' demographic characteristics (i.e., gender, ethnicity, years of teaching experience, whether the teachers had a master's degree or higher), and a teacher-specific error term (ε_{ij}). The Level 1 model used for PPTI included indicators for whether the teacher was an English language arts, mathematics, science, and social studies teacher. For SPBP, whether the teacher taught tested subject or grade was also included as a Level 1 covariate.

Level 2 models the team or school mean (θ_j) as a function of the team's or school's treatment status (T_j) and random effects for teams or schools (ξ_j) to account for the clustering of teachers within course-groups, teams, or schools. For POINT, the Level 2 models included fixed effects (u_{jg}) for the course-group randomized block to which teachers were assigned and an indicator for treatment condition. For PPTI, the Level 2 modeled the team components as a function of the team's intervention status, fixed effects (u_{jg}) for the blocks with which teams were randomly assigned to interventions, and random effects for team, which accommodates the fact that responses from teachers on the same team may be correlated. For SPBP, we also controlled for grade- and school-level demographic variables (G_j^T & S_j^T), including enrollment size, percentage of ELL, percentage of students eligible for individual education program, percentage of Hispanic and Black students, and percentage of students eligible for FRPL.

$$Y_{ij} = \theta_j + X_i^T\beta + \varepsilon_{ij}\ , \tag{1}$$

$$Level\ 2\ (POINT\ and\ PPTI):\ \theta_j = \mu + T_j\delta + \sum \gamma_g u_{jg} + \zeta_j\ , \text{ and} \tag{2}$$

$$Level\ 2\ (SPBP): \theta_j = \mu + T_j\delta + G_j^T\gamma + S_j^T\lambda + \zeta_j\ . \tag{3}$$

Equations 1 to 3 summarize the model used for composite outcome measures. *i* index teachers and *j* index course-groups (POINT) or teaching teams (PPTI) or schools (SPBP). β is the vector of parameters for teacher-level covariates. ε_{ij} are independently normally distributed residual errors. μ is the population mean. δ is the parameter for the treatment effect. u_{jg} equals 1 if the team is in randomization course-group (POINT) or block (PPTI) *g* and 0 otherwise. γ and λ are the coefficient vectors for grade- and school-level covariates, respectively. ξ_j are the random effects for the course-groups, teaching teams, or schools.

We applied an ologit (ordinal logit) regression model for 4-point Likert-type scale items and a logit regression model for the dichotomized scores, with the same set of teacher- and school-level covariates and adjustment for clustered data (Long & Freese, 2006).[10] When analyzing changes in POINT persistent teachers' responses between the first and final program year, we used the same model to compare treatment and control group teachers in POINT with teachers' responses in the final program year as the outcome variable and their first year responses as one additional Level 1 covariate. We used the Benjamini and Hochberg (1995) method to control for false significant rate at 5% across all tests to adjust for multiple comparisons.

Results showed some differences among programs on certain measures we examined. Multiple reasons might have contributed to these differences, such as differences in the program features, site effects, program implementation, and other unobserved factors that might have affected teachers' motivation and behavior. Unfortunately, it is impossible to specifically link any differences in teachers' motivation

TABLE 2
Incentive Eligible Teachers' Responses About Goal Acceptance, Expectancy, and Valence

Scale	Program	*N*	% Agree / *M* (*SD*)
Goal acceptance			
Teachers should be rewarded for demonstrating outstanding teaching skills	POINT	80	80%
	PPTI	163	75%
Rewarding teachers based on test score gains is problematic	POINT	82	90%
	SPBP	747	81%
I have a clear understanding of the performance criteria for earning a bonus award	PPTI	172	48%
	SPBP	773	85%
The method used to award bonuses is fair to all program participants	POINT	81	55%
	PPTI	172	34%
	SPBP	648	55%
Expectancy			
Teachers' confidence in teaching	POINT	81	3.17 (0.39)
	PPTI	164	3.05 (0.36)
Teacher's impact on student achievement is limited due to the effect of home environment	POINT	82	50%
	PPTI	165	40%
What is the chance you/your team will receive a bonus based on this year's performance	POINT	81	46.05 (28.72)
	PPTI	164	51.68 (24.87)
Valence			
I have a strong desire to earn a bonus	POINT	81	77%
	SPBP	753	64%
The chance to earn a bonus award has energized me to improve my teaching	POINT	82	42%
	PPTI	174	19%
The bonus amount is not large enough to motivate extra effort (PPTI survey item was "The size of the bonus award is too small to motivate me to work harder")	POINT	81	49%
	PPTI	169	16%
	SPBP	684	42%

Note. POINT = Project on Incentives in Teaching; SPBP = School-Wide Performance Bonus Program; PPTI = Pilot Project on Team Incentives.

and behavioral changes to differences in program features based on the data we had. Thus, when reporting the results, we focused on the commonalities across programs by studying the general patterns in the results among programs.

We present results on the common items and scales generated for this comparative analysis. Where applicable, we also refer to results from the evaluation of each program, in which the same teaching practice variables as those we examined (e.g., instruction) were measured using slightly different scales. The original scales usually included a few more items than the scales we generated based on common items across surveys and had higher internal consistency reliability than the corresponding measures used in this study.

Results

The number of teachers included in the analysis was 145 for POINT (N_{treat} = 82, $N_{control}$ = 63), 355 for PPTI (N_{treat} = 175, $N_{control}$ = 180), and 1,407 for SPBP ($N_{incentive\ eligible}$ = 798, $N_{control}$ = 609). Seventy-seven percent of the respondents were female. The percentage of teachers with a master's degree or higher was 66%, 29%, and 55% in POINT, PPTI, and SPBP, respectively. More than two-thirds of teachers in POINT and PPTI were White. In SPBP, 36% of teachers were White and 27% were Black. The average years of teaching experience were 14 (*SD* = 8), 10 (*SD* = 8), and 13 (*SD* = 8) for POINT, PPTI, and SPBP, respectively.

Teachers' Motivation

Goal acceptance. As shown by Table 2, teachers accepted the general idea of rewarding teachers based on teaching performance. Namely, 80% of POINT teachers and 75% of PPTI teachers believed teachers should be rewarded for outstanding teaching skills.[11] However, teachers

do not appear to equate teaching skills with student achievement on standardized tests. The majority of POINT (90%) and SPBP (81%) teachers agreed that rewarding teachers based on student test score gains was problematic because student test scores did not capture important aspects of teaching performance. In addition, teachers' understanding of the program and perceptions of fairness differed by program. The majority of SPBP teachers (85%) reported that they understood the Progress Report and contributing factors. However, about one-third of SPBP teachers reported not understanding several aspects of the program very well, such as the amount of funding a school would receive if it met 100% of the target, the criteria for receiving a partial bonus, and the source of funding for this program (Marsh et al., 2011). Only half of PPTI bonus eligible teachers (48%) reported having a good understanding of the program's criteria. With respect to fairness, many teachers in all three programs had concerns. Only slightly over half (55%) of the bonus eligible teachers in POINT and SPBP and one-third of the PPTI treatment group teachers agreed that the method for awarding bonuses in their programs was fair to all program participants.[12]

Expectancy. Although teachers had confidence in their teaching abilities, they doubted that their personal efforts would lead to the expected student achievement goals. Moreover, teachers had concerns about the influence of family on student achievement. Overall, they did not have high expectancy of achieving the program goals through personal efforts.

Specifically, POINT and PPTI treatment group teachers were confident about their abilities to reach the most difficult students, help students retain information learned, manage disruptive behaviors during class, and evaluate the appropriate level of assignments for students. The average score of the teachers' confidence in teaching scale was above 3 on a 4-point scale (see Table 2). However, about half of POINT (50%) and PPTI (40%) teachers agreed that teachers were limited in what they could achieve because family environment had a large influence on student achievement. When asked to estimate the probability that they would win a bonus, both POINT and PPTI teachers estimated their chances to be around 50%. This indicates that teachers' expectancy was at breakeven. Given that people often report 50% for an unknown or uncertain probability, this may indicate that many teachers did not have a strong sense of the likelihood they would win a bonus.

Valence.[13] Results showed that although about two-thirds of teachers in POINT (77%) and SPBP (64%) wanted to earn a bonus (see Table 2), teachers' interest in earning a bonus was not associated with taking actions to win it. Only slightly over 40% of POINT teachers and 20% of PPTI teachers reported that the chance to earn a bonus energized them to improve teaching. In addition, 42% of SPBP bonus eligible teachers and about half of POINT treatment group teachers reported that the bonus was not large enough to motivate extra effort. In contrast, only 16% of PPTI teachers reported that the bonus was too small to motivate them to work harder.[14]

Teachers' Practices and Working Conditions

Instruction. Comparisons between treatment and control group teachers within each program on three measures of teaching practices (i.e., focus on standards, data-driven decision-making, and test preparation) found only one significant difference between two groups in POINT (see Table 3). Treatment group teachers reported greater emphasis on test preparation than control group teachers in the last program year. However, analysis did not find a positive, significant association between teachers' reported classroom time on test preparation and student achievement (Springer et al., 2010). Both treatment and control group teachers in POINT and PPTI reported making little change to their instruction and student learning due to the implementation of the incentive pay program. The average scores of the scales on changes made in classroom emphasis on state standards and tests and on student engagement in hands-on activities and group learning represent a level of change that was between no change and a little more than last year. In addition, the majority of incentive eligible teachers in all three programs reported that their programs had no effect on teaching, with 85% in POINT, 78% in PPTI, and 90% in SPBP. Moreover, analyses of differences

TABLE 3

Teachers' Responses About Instruction, Number of Hours Worked, Job Stress, and Collegiality

	POINT[a]						PPTI			SPBP		
Scale	T_3	C_3	Dif	T_31	C_31	Dif	T	C	Dif	T	C	Dif
Instruction												
Focus on standards												
M	5.1	5.02	0.08	–0.03	0.16	–0.19	5.04	5.12	–0.08	4.43[b]	4.52	–0.09
SD	0.81	0.97		0.89	0.95		0.75	0.72		0.68	0.6	
N	82	63		81	61		172	171		793	608	
Data-driven decision-making												
M	3.04	2.98	0.06	0.17	0.15	0.02	3.21	3.25	–0.04	3.55	3.57	–0.02
SD	0.68	0.55		0.64	0.58		0.53	0.52		0.7	0.69	
N	80	63		78	62		172	172		723	542	
Test preparation												
M	3.55	3.28	0.27*	0.06	0.03	0.03	3.34	3.29	0.05	—	—	—
SD	0.5	0.56		0.42	0.58		0.57	0.59		—	—	
N	82	61		81	60		173	172		—	—	
Changes made in instruction												
M	3.66	3.59	0.07	—	—	—	3.43	3.41	0.02	—	—	—
SD	0.57	0.54		—	—		0.52	0.52		—	—	
N	66	49		—	—		169	167		—	—	
Changes made in student learning												
M	3.48	3.34	0.14	—	—	—	3.51	3.57	–0.06	—	—	—
SD	0.78	0.6		—	—		0.66	0.69		—	—	
N	82	63		—	—		170	167		—	—	
No program impact on practices												
Pct	85%	89%	–4%	—	—	—	78%	—	—	90%	—	—
N	82	63		—	—		174	—		759	—	
Number of hours worked												
Total extra working hours per week												
M	14.26	12.83	1.43	2	–0.78	2.78	12.79	11.97	0.82	12.77	13.18	–0.41
SD	10.11	8.9		11	7.46		9.56	7.15		7.96	8.56	
N	54	46		52	45		172	171		788	603	
Portion of extra working hours on routine tasks												
M	3.15	3.1	0.05				3.15	3.08	0.07	—	—	—
SD	0.66	0.66					0.52	0.55		—	—	
N	81	63					172	172		—	—	
Portion of extra working hours on other tasks												
M	2.26	2.14	0.12				2.28	2.3	–0.02	—	—	—
SD	0.53	0.5					0.44	0.53		—	—	
N	81	63					172	172		—	—	
Job stress												
Program impact on job stress[c]												
Pct	17%	—	—	12%	—	—	—	—	—	—	—	—
N	82	—		81	—		—	—		—	—	
Collegiality												
Positive relationships with colleagues												
M	3.01	2.69	0.32*	–0.01	0.01	–0.02	3.22	3.32	–0.1	2.94	2.96	–0.02
SD	0.61	0.62		0.82	0.58		0.66	0.62		0.69	0.69	
N	82	63		81	61		166	176		782	600	
Negative relationship with colleagues												
M	1.92	2.08	–0.16	0.14	0.01	0.13	1.72	1.72	0	2.19	2.19	0
SD	0.67	0.58		0.83	0.78		0.65	0.69		0.66	0.67	
N	82	63		81	61		166	176		781	598	
Negative impact on relationships among colleagues[d]												
M	13%	14%	–1%				36%	—	—	—	—	—
N	81	63					171	—		—	—	

[a]T = Treatment; C = Control; 3 = the third program year; 31 = difference in POINT persistent teachers' responses between the first and third program year; POINT = Project on Incentives in Teaching; SPBP = School-Wide Performance Bonus Program; PPTI = Pilot Project on Team Incentives.

[b]Items used in SPBP were on a 5-point Likert scale.

[c]The percentage of SPBP teachers who reported that their level of job stress had changed for worse, did not change, or changed for better as a result of their school's participation in the program was 24%, 50%, and 21%, respectively.

[d]The percentage of SPBP teachers who reported that their relationships with administrators, other teachers, and other nonclassroom staff members changed for worse during the program was 9%, 3%, and 4%, respectively.

*indicates significant differences at 0.05 level after adjustment for multiple comparisons, ** indicates significant differences at 0.01 level after adjustment for multiple comparisons. The *p* values are available from the authors upon request.

in POINT persistent teachers' responses between the first and the final program year did not find any significant changes in teachers' reported instructional practices during the program. Overall, there is limited evidence that any of the three programs changed teachers' instructional practices, especially practices significantly associated with student achievement.

Number of hours worked. Analyses did not find any evidence that incentive pay programs increased teachers' number of hours worked (see Table 3). Teachers in the treatment and control groups of all three programs reported that they worked for 2.6 to 2.8 extra hours beyond their contracted hours per work day, on average. No difference was found in the reported numbers of extra working hours between two groups in any program. Treatment and control group teachers in POINT and PPTI were also asked about how they used the extra working hour and reported similar ways of allocating this time. They spent a moderate to major portion of their extra working hours on routine tasks such as preparing lessons, grading, and completing administrative tasks and a small to moderate portion of extra working hours on other tasks such as professional development training and meeting with students, colleagues, and parents. In addition, no significant difference was found in the reported number of extra working hours per week between the first and final program year among POINT persistent teachers. Overall, results show that these incentive pay programs did not affect teachers' number of hours worked.

Job stress. Results about job stress are available only from incentive eligible teachers in SPBP and POINT (see Table 3). Only a minority of teachers in both programs reported their job stress increased as the result of the implementation of the program. For SPBP, 24% of the incentive eligible teachers indicated their level of job stress changed for worse due to the implementation of the program. Half of the incentive eligible teachers reported that participation in the program did not increase their level of job stress. In POINT, 17% of treatment group teachers reported that job stress level increased as a result of this program in the final program year, which is up 12 percentage points from the 5% who felt the program increased stress in the first program year.

Collegiality. Analyses of the impact of pay-for-performance programs on collegiality did not find evidence that such programs damaged teachers' relationships with their colleagues (see Table 3). Comparisons between treatment and control group teachers' reported relationships with colleagues did not find any significant difference between two groups in any programs except that POINT treatment group teachers reported greater collaboration than control group teachers. Additionally, analyses of changes in POINT persistent teachers' report of collegial relationships between the first and last program year did not find any evidence that POINT damaged the relationships among teachers.

Discussion

This study examines the extent to which three incentive pay programs motivated teachers to achieve program goals and changed teachers' instruction, number of hours worked, job stress, and collegiality. Results showed that teachers did not consider their programs as motivating. First, teachers' level of goal acceptance was not high due to a lack of understanding of the program among some teachers and teachers' concerns about using student test scores to measure teaching performance and the fairness of the program. Second, teachers did not have high expectancy that their personal efforts would lead to student achievement gains due to concerns about the influence of family environment on student achievement. Third, although teachers would have liked to earn a bonus, they did not see the opportunity as worthy of changing behavior.

With respect to the impact of three incentive pay programs on teachers' practices and working conditions, our analyses did not find that any of the three programs had affected teachers' reported instructional practices, number of hours worked, or collegiality, except that POINT treatment group teachers reported greater emphasis on test preparation and collaboration among colleagues than their counterparts in the control group. However, classroom time on test preparation was not associated with student achievement. Although a minority of bonus eligible teachers in three programs reported increased job stress, no

comparison data were available for the control group. Therefore, we could not conclude that the implementation of these pay-for-performance programs increased teachers' job stress.

Notably, our findings did not differ across the different types of programs. We found no effects on teachers regardless of the choice of individual versus collective motivational mechanisms. Of course, a study of three programs cannot definitively answer the empirical question and more research is needed to shed light on this enduring social debate.

Connecting our findings with results from prior studies, we found both similarities and differences. For instance, we found that teachers had doubts about their control over student achievement, given the influence of family environment on student learning, and challenged the fairness of the programs' methods used to determine the bonuses. Although studies using value-added modeling to examine student performance on standardized tests showed that teachers have a large impact on student test score growth (Rivkin, Hanushek, & Kain, 2005; Rowan, Correnti, & Miller, 2002), prior studies also found a variety of family, student, classroom, and school factors associated with student performance on standardized tests, such as student socioeconomic status (Caro, 2009), parents' expectation for student academic performance and involvement in student education (Davis-Kean & Sexton, 2009), and composition of peers (Rubin, Bukowski, & Parker, 2006). Because teachers lack control over many factors that might affect student achievement, it is not surprising that teachers question the strength of the link between their personal effort and the increase in student achievement.

Related to teachers' doubts about their control over student achievement, teachers questioned using student test score gains to measure teaching performance. Most prior studies on teachers' attitudes about pay-for-performance programs reported that teachers objected to the idea of measuring teaching performance based on student test scores or test score gains (Ballou & Podgursky, 1993; Farkas, Johnson, Duffet, Moye, & Vine, 2003; Goldhaber, DeArmond, & DeBurgomaster, 2011; Langdon & Vesper, 2000), although a study on teachers in the Texas Governor's Educator Excellence Grant found that more than 90% of respondents considered student test score gains as moderately or highly important for evaluating teaching performance in an incentive pay program (Springer et al., 2007). It is difficult to obtain teachers' support of incentive pay programs if they think the performance measure is problematic. However, given the current emphasis on educational accountability, it is also difficult for incentive pay programs to totally ignore student test scores or test score gains when measuring teaching performance. Combinations of multiple teaching performance measures might help to gain teachers' support of the program and boost their motivation to make behavioral changes. However, which teaching performance measures should be included and what formula should be used to combine them remain as questions to be answered.

We found that although teachers had a strong desire to earn a bonus, bonuses had a limited reported motivating effect, which is consistent with findings from prior studies (Heneman, 1998; Kelley, 1999). Bonuses may have a limited motivating effect on teachers because teachers view the receipt of a bonus as an acknowledgement of their hard work rather than an incentive to work harder (Marsh et al., 2011). Teachers may also consider meeting performance goals and the associated pride more important than the bonus. We found that the large size of an award, such as in POINT, is not enough to offset other problems with the program such as teachers' concerns about the teaching performance measure. Although the average bonus size of POINT tripled that of SPBP, teachers in the two programs showed similar attitudes about the motivating effect of bonuses. Program features and other contextual factors in the implementation of the program might affect the relationship between bonus size and the motivational effect of the bonus (Kelley, 1999).

Teachers included in this study reported little impact of the program on their instructional practices, which is a departure from findings in several previous observational studies that did not include control groups (Adkins, 2004; Kelley, 1999) but consistent with results based on observation of instruction in international experimental studies on pay-for-performance programs that also compared treatment and control group teachers (Glewwe et al., 2010; Muralidharan & Sundararaman, 2008). A few factors might have contributed to the

lack of changes in instruction. First, the lack of motivation among teachers to achieve program goals might have led to little program effect on teachers' practices. Second, existing accountability pressure or intrinsic motivators may have had a stronger motivational effect on teachers than financial incentives (Marsh et al., 2011). For instance, test scores increased in Nashville, Tennessee, for students taught by teachers in both the treatment and control groups, as well as for students taught by other middle school teachers who did not participate in the program and for fourth-grade students during the study period, a period when the district was threatened with state takeover due to the No Child Left Behind Act of 2001 (20 U.S.C. 6311 *et seq*.). Third, although these programs lasted for two or three years, it still might not be enough time for teachers to make substantial changes in instruction. Fourth, it is questionable whether teachers have the time and energy to make any additional changes in instruction given that they report working an additional two or three hours per day to fulfill their routine responsibilities, such as grading, preparing for lessons, and completing administrative tasks. Fifth, teachers might not be well equipped with the knowledge and skills required to improve student achievement. If teachers had known how to effectively improve student achievement, they might have done their best to achieve this goal, given that most teachers reported that seeing students increase their achievement gives them high motivation to work hard (Marsh et al., 2011).

We did not find evidence that the implementation of incentive pay programs increased teachers' number of hours worked or job stress, which is different from findings of prior observational studies (Adkins, 2004; Kelley et al., 2000; Wiley et al., 2010). The types of analyses conducted might have contributed to differences in results among studies (Muralidharan & Sundararaman, 2008). Previous studies mainly analyzed incentive eligible teachers' reported data, whereas our study examined not only incentive eligible teachers' reported data but also compared differences between treatment and control groups and, in one case, analyzed changes in POINT persistent teachers' responses between the first and final program year.

We also did not find evidence that the implementation of incentive pay programs damaged teachers' collegial relationships. This is consistent with findings from prior studies of both school-based and individual-based pay-for-performance programs (Kelley, 1999; Wiley et al., 2010). However, it is important to understand these findings in the context of the teachers' responses to these programs. Teachers' collegial relationships might have changed had they actively competed for awards, which is not suggested by the lack of changes in instructional practices and the numbers of hours worked.

Overall, our findings showed that these three incentive pay programs did not motivate teachers to make behavioral changes that lead to student achievement gains, nor did these programs increase teachers' number of hours worked or damage teachers' collegial relationships. The lack of program impact on teachers' practices suggests that more careful thinking about the logic model of incentive pay programs is necessary. For instance, is the marginal motivational effect of bonus strong enough to lead to teachers' behavioral changes in instruction, especially in the context of high-stakes accountability? Should teachers' knowledge and skills be incorporated in the logic model of pay-for-performance program—for example, by adding professional development and other capacity-building mechanisms to the program? If so, how do these new mechanisms interact with other components of the logic model? How should teaching performance be measured to maximize teachers' acceptance of the program and its goals?

Given our findings and the previous literature that finds weak effect of performance pay for teachers, policymakers might favor other reforms. For instance, compensation tied to career lattice and other professional growth and goals or compensation for work in challenging schools might be alternatives to bonus-based compensation reform. If bonus-based policy is pursued, policymakers need to recognize this lack of evidence and take steps to monitor program implementation and evaluate program impact on targeted outcomes.

Alternatively, policymakers can try out some untested new designs that reward teachers based on specific teaching behavior. Recent studies suggest that when people do not know how to achieve a complex goal, incentives on specific input behaviors that are directly linked to the

achievement of the final goal are more effective than incentives on the output in reaching the final program goal (Mehrotra, Sorbero, & Damberg, 2010). For instance, Fryer (2012) conducted school-based randomized trials in four large urban school districts to examine whether incentives on students' specific learning input behaviors such as reading a book or turning in homework on time are more effective than incentives on the output of student behaviors (e.g., increased test scores) in increasing student achievement. Results showed that incentives based on students' inputs were more effective than incentives on the output in increasing student achievement. Based on interviews with students, Fryer concluded that the major reason for differences in the effectiveness of incentives on input versus output might be because students did not know the strategies they could use to increase their achievement. Thus, students in the incentives on input group acted upon behaviors that are useful for increasing their engagement and learning outcomes, whereas students in the incentives on output group could not transform their excitement about the incentives into effective actions that lead to increased test scores, although they were excited about the incentives on increased achievement. Based on these findings and the assumption that increasing student achievement is a complex goal that teachers might not know how to achieve, Fryer inferred that teacher incentive pay based on input (i.e., certain instructional behaviors) might be more effective than rewards on output (e.g., increase in their students' test scores) in increasing students' achievement. This new model of teacher incentive pay programs might be beneficial, but it would require identifying specific practices that teachers should follow to yield better student outcomes. If such practices can be found, Fryer's and others' work suggests financial rewards might help in the uptake of the practices, but other means of motivation might also be explored.

The current study had several limitations. First, it is based on teacher self-reported data, which suffer from a variety of response biases such as biases caused by social desirability (Heneman & Milanowski, 1999). However, it is usually necessary to rely on self-reported data when it comes to measuring teachers' opinions. Future studies could consider multiple data collection methods, such as teaching logs and classroom observation, which may provide more detailed information about teachers' behavioral changes and opportunities for triangulation. Second, teacher attrition in POINT and the relatively low response rate in SPBP might have affected the results. Our analyses on differences in the background characteristics between participating teachers in the first and last year of POINT and differences between the weighted and unweighted results for SPBP show that any such differences are likely to be small. Third, it is possible that teachers in different programs interpreted the same item differently due to other contextual factors such as district and program policies and professional development. It is also possible that the experimental nature of the intervention in POINT and PPTI influenced teachers' responses, since they knew that the program was short-term and was not intended to become district policy. Finally, all three studies focus only on existing teachers and cannot comment on the potential effects that pay-for-performance programs might have on the composition of the teacher labor force.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Center on Performance Incentives, which is funded by the United States Department of Education's Institute of Education Sciences (R305A06034) and the New York City Fund for Public Schools (071609). We are grateful for the officials and teachers in three participating districts. The content or opinions expressed do not necessarily reflect the views of the funders of this project. A version of this research was presented at the Annual Meeting of the American Educational Association (April, 2012).

Notes

1. None of the three programs found any significant impact of teacher incentive pay programs on student achievement. Because the focus of this study is on teachers, we did not review results regarding the

effect of teacher incentive pay programs on student achievement across three projects. Interested readers are referred to Springer et al. (2010), Marsh et al. (2011), and Springer et al. (2012) for the details of analyses on student achievement.

2. On the other hand, as teachers put in additional effort to win a bonus, teachers may experience increased job stress. These types of externalities bear on the extent to which teachers will feel motivated by incentive pay programs.

3. These numbers represent the number of teachers in each group at the end of each academic year.

4. We compared POINT teachers who persisted through Year 3 with those who dropped out during the program on their Year 1 and Year 3 survey responses. We did not find any significant differences between these two groups of teachers.

5. The district removed 25 eligible schools before the randomization. We have very limited information about school removal decisions.

6. The percentage of schools winning bonuses greatly decreased in the final year due to the state's decision to change proficiency cutoff scores.

7. All classroom teachers were surveyed in schools where the total number of classroom teachers was less than seven.

8. Readers interested in details of each survey are referred to Springer et al. (2010), Marsh et al. (2011), and Springer et al. (2012).

9. We examined differences between the treatment and control groups at baseline based on teacher administrative data and student achievement data for each program. In POINT and PPTI, treatment group teachers' classes contained slightly greater portions of ELLs than those of control teachers. In PPTI, the percentage of talented and gifted students was greater in the control group than that in the treatment group. No significant difference was found between the treatment and control groups for SPBP. We also tested whether the treatment and control groups had differential nonresponse patterns based on teacher administrative data and student achievement data. Overall, we did not find any evidence of differential nonresponse patterns except on one variable for POINT. POINT treatment teachers who stayed in the program for 3 years had about half a day more of absences than their counterparts in the control group. These differences are substantively small. Moreover, we controlled for available covariates at the teacher and school levels in the analysis to mitigate the influence of potential imbalance and differential nonresponse patterns between two groups of teachers. Interested readers are referred to Springer et al. (2010), Marsh et al. (2011), and Springer et al. (2012) for the details of these analyses and results. Results are also available from the authors.

10. Details of the equations used for items scales are available from the authors upon request.

11. Because the response rates of POINT and PPTI were high, and there were few differences between weighted and unweighted results in SPBP (Marsh et al. 2011), we reported the unweighted results in this study.

12. The response rate for POINT was calculated based on teachers who remained in the program. There were limited differences between teachers who remained in or dropped out of the study.

13. None of the surveys used in three programs asked teachers questions about instrumentality (i.e., the likelihood that meeting the achievement goals will result in a financial reward).

14. It is important to note that the wording of the PPTI item was different from the wording of the SPBP and POINT items, which renders it difficult to directly compare the motivating effects of the bonus across programs.

References

Adkins, G. K. (2004). *Teacher performance pay: The perceptions of certified school-based personnel* (doctoral dissertation, University of Central Florida). Retrieved from http://accountability.lee-schools.net/research_projects/pdf/Adkins_Gregory_K_200412_EdD.pdf

Ballou, D., & Podgursky, M. (1993). Teacher attitudes toward merit pay: Examining conventional wisdom. *Industrial and Labor Relations Review*, *47*, 50–61.

Barnett, J. H., & Ritter, G. W. (2008, October). When merit pay is worth pursuing. *Educational Leadership*, *66*. Retrieved from http://www.ascd.org/publications/educational_leadership

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289–300.

Caro, D. H. (2009). Socio-economic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education*, *32*, 558–590.

Center for Educator Compensation Reform. (n.d.). *List of TIF grants*. Retrieved from http://cecr.ed.gov/TIFgrantees/list.cfm

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.

Collins, R. (1994). *Four sociological traditions*. New York: Oxford University Press.

Davis-Kean, P. E., & Sexton, H. R. (2009). Race differences in parental influences on child achievement: Multiple pathways to success. *Merrill-Palmer Quarterly: Journal of Developmental Psychology*, *55*, 285–318.

Farkas, S., Johnson, J., Duffet, A., Moye, L., & Vine, J. (2003). *Stand by me: What teachers really think about unions, merit pay, and other professional matters*. Washington, DC: Public Agenda.

Ford, M. (2012). *A modern teacher compensation for Wisconsin* (Vol. 25, No. 1). Hartland: The Wisconsin Policy Research Institute. Retrieved from http://www.wpri.org/Reports/Volume25/Vol25No1/Vol25No1.html

Fryer, R. G. (2012). *Aligning student, parent, and teacher incentives: Evidence from Houston Public Schools* (Working Paper 17752). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w17752

Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, *2*, 1–25.

Goldhaber, D., DeArmond, M., & DeBurgomaster, S. (2011). Teacher attitudes about compensation reform: Implications for reform implementation. *Industrial and Labor Relations Review*, *64*, 441–463.

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

Guarino, C. M., Santibanez, L., & Daley, G. (2006). Teacher recruitment and retention: A review of the recent empirical literature. *Review of Educational Research*, *76*, 173–208.

Hall, G. E., & Caffarella, E. P. (1997). *Third-year implementation assessment of the Douglas County, Colorado School District Performance Pay Plan for Teachers*. Castle Rock, CO: Douglas County School District.

Heneman, H. G. (1998). Assessment of the motivational reactions of teachers to a school-based performance award program. *Journal of Personnel Evaluation in Education*, *12*, 43–59.

Heneman, H. G., & Milanowski, A. T. (1999). Teachers attitudes about teacher bonuses under school-based performance award programs. *Journal of Personnel Evaluation in Education*, *12*, 327–341.

Kelley, C. (1999). The motivational impact of school-based performance awards. *Journal of Personnel Evaluation in Education*, *12*, 309–326.

Kelley, C., Heneman III, H. G., & Milanowski, A. (2002). Teacher motivation and school-based performance awards. *Educational Administration Quarterly*, *38*, 372–401.

Kelley, C., Odden, A., Milanowski, A. T., & Heneman, H. G. (2000). *The motivational effects of school-based performance awards* (RB-29). Philadelphia: Consortium for Policy Research in Education.

Koppich, J. E., & Rigby, J. (2009). *Alternative teacher compensation: A primer*. Berkeley, CA: Policy Analysis for California Education.

Langdon, C., & Vesper, N. (2000). The sixth Phi Delta Kappa poll of teachers' attitudes toward the public schools. *Phi Delta Kappan*, *81*, 607–611.

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

Locke, E. A. (1968). Toward a theory of task motivation and incentives. *Organizational Behavior and Human Performance*, *3*, 157–189.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (2nd ed.). College Station, TX: StataCorp.

Margolis, J. (2008). What will keep today's teachers teaching? Looking for a hook as a new career cycle emerges. *Teachers College Record*, *110*, 160–194.

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

Mehrotra, A., Sorbero, M. E., & Damberg, C. L. (2010). Using the lessons of behavioral economics to design more effective pay-for-performance programs. *American Journal of Managed Care*, *16*, 497–503.

Milanowski, A. (2003). The varieties of knowledge and skill-based pay design: A comparison of seven new pay systems for K-12 teachers. *Education Policy Analysis Archives*, *11*. Retrieved from http://epaa.asu.edu/epaa/v2011n2014/

Muralidharan, K., & Sundararaman, V. (2008). *Teacher incentives in developing countries: Experimental evidence from India* (Working Paper No. 2008-13). Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

No Child Left Behind Act of 2001. 20 U.S.C. 6311 et seq.

Podgursky, M. J., & Springer, M. G. (2006). *Teacher performance pay: A review*. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Oliver, P. E. (1980). Rewards and punishments as selective incentives for collective action: Theoretical investigations. *American Sociological Review*, *85*, 1356–1375.

Rice, J. K., Jackson, C., Hoyer, K. M., Malen, B., & Hyde, L. (2011, March). *Time to pay up: Distribution patterns and perceived effects of financial awards in a Teacher Incentive Fund (TIF) program*. Paper presented at the Annual Meeting of the Association for Education Finance and Policy, Seattle, Washington.

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

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher

effects on student achievement: Insights from the Prospects study of elementary schools. *Teacher College Record*, *104*, 1525–1567.

Rubin, K. H., Bukowski, W., & Parker, J. G. (2006). Peer interactions, relationships, and groups. In N. Eisenberg (Ed.), *Handbook of child psychology: Social, emotional, and personality development* (Vol. 3, pp. 571–645). New York: Wiley.

Schuermann, P. (2009). *Current status of district and state performance pay initiatives in the U.S.* Retrieved from http://www.learningpt.org/rel/archive/Schuermann_Oct15_162009.pdf

Springer, M. G., Podgursky, M., Lewis, J. L., Ehlert, M. W., Gardner, C. D., Gosh-Dastidar, B., Lopez, O., Patterson, C. H., & Taylor, L. L. (2007). *Governor's Educator Excellence Grant (GEEG) Program: Year one evaluation report.* Austin, TX: Texas Education Agency.

Springer, M. G., Lewis, J. L., Podgursky, M., Ehlert, M. W., Gronberg, T. J., Hamilton, L., Jansen, D. W., Stecher, B. M., Taylor, L. L., Lopez, O., & Peng, X. (2009). *Texas Educator Excellence Grant (TEEG) Program: Year three evaluation report*. Austin, TX: Texas Education Agency.

Springer, M. G., Ballou, D., Hamilton, L., Le, V., Lockwood, J. R., McCaffrey, D. F., Pepper, M., & Stecher, B. (2010). *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., Pane, J. F., Le, V., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., & Stecher, B. (2012). Team pay for performance: Experimental evidence from the Round Rock Pilot Project on Team Incentives. *Educational Evaluation and Policy Analysis*, *34*, 367–390.

Teaching Commission. (2004). *Teaching at risk: A call to action*. Washington, DC: Author.

Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.

Wiley, E. W., Spindler, E. R., & Subert, A. N. (2010). *Denver ProComp: An outcomes evaluation of Denver's Alternative Teacher Compensation System*. Boulder: University of Colorado at Boulder.

Willer, R. (2009). A status theory of collective action. In S. R. Thye & E. J. Lawler (Eds.), *Advances in group processes* (Vol. 26, pp. 133–163). London: Emerald.

Authors

KUN YUAN is a Behavioral Scientist at RAND Corporation, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213; kyuan@rand.org. Her areas of specialization are educational measurement and program evaluation.

VI-NHUAN LE is a Behavioral Scientist at RAND Corporation, 1776 Main Street, Santa Monica, CA 90407; vinhuan@rand.org. Her areas of specialization are educational measurement and early childhood education.

DANIEL F. MCCAFFREY is a Senior Statistician at RAND Corporation, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213; danielm@rand.org. His areas of specialization are value-added modeling and program evaluation.

JULIE A. MARSH is an Associate Professor at the University of Southern California Rossier School of Education, 3470 Trousdale Parkway, Waite Phillips Hall, 904B, Los Angeles, CA 90089; julie.marsh@usc.edu. Her areas of specialization are accountability policy and policy implementation.

LAURA S. HAMILTON is a Senior Behavioral Scientist at RAND Corporation, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213; laurah@rand.org. Her areas of specialization are assessment, accountability, and school reform evaluation.

BRAIN M. STECHER is a Senior Behavioral Scientist at RAND Corporation, 1776 Main Street, Santa Monica, CA 90407; stecher@rand.org. His areas of specialization are assessment, accountability, classroom practices, and program evaluation.

MATTHEW G. SPRINGER is an Assistant Professor of Public Policy and Education at the Peabody College of Vanderbilt University, Peabody #43, 230 Appleton Place, Nashville, TN 37203; matthew.g.springer@vanderbilt.edu. His areas of specialization are teacher performance incentives, school choice, and education policy.

Manuscript received April 3, 2012
Revision received August 29, 2012
Accepted September 4, 2012