NATIONAL CENTER ON
Performance Incentives

New York City's

School-Wide Bonus Pay Program:

Early Evidence from a Randomized Trial

Matthew G. Springer
Marcus A. Winters

Working Paper 2009-02
April 2009

LED BY
VANDERBILT PEABODY COLLEGE

IN COOPERATION WITH:
RAND CORPORATION
Mizzou University of Missouri - Columbia

The National Center on Performance Incentives (NCPI) is charged by the federal government with exercising leadership on performance incentives in education. Established in 2006 through a major research and development grant from the United States Department of Education's Institute of Education Sciences (IES), NCPI conducts scientific, comprehensive, and independent studies on the individual and institutional effects of performance incentives in education. A signature activity of the center is the conduct of two randomized field trials offering student achievement-related bonuses to teachers. The Center is committed to air and rigorous research in an effort to provide the field of education with reliable knowledge to guide policy and practice.

The Center is housed in the Learning Sciences Institute on the campus of Vanderbilt University's Peabody College. The Center's management under the Learning Sciences Institute, along with the National Center on School Choice, makes Vanderbilt the only higher education institution to house two federal research and development centers supported by the Institute of Education Services.

This working paper was supported, in part, by the National Center on Performance Incentives at Vanderbilt University, which is funded by the United States Department of Education's Institute of Education Sciences (R30SA06034). We appreciate helpful comments and suggestions from Dale Ballou, Julie Marsh, Dan McCaffrey, and Patrick Wolf. We would also like to acknowledge Terry Bowman and the many individuals atthe New York City Public Schools for providing data and technical support to conduct our analyses. Any errors remain the sole responsibility of the authors. The views expressed in this paper do not necessarily reflect those of sponsoring agencies or individuals acknowledged.

Please visit **www.performanceincentives.org** to learn more about our program of research and recent publications.

New York City's School-Wide Bonus Pay Program: Early Evidence from a Randomized Trial

MATTHEW G. SPRINGER
Vanderbilt University
National Center on Performance Incentives

MARCUS A. WINTERS
Manhattan Institute

Abstract

In this paper, we examine the impact of New York City's School-Wide Performance Bonus Program (SPBP) on student outcomes and the school learning environment. The SPBP is a pay-for-performance program that was implemented in approximately 200 K-12 public schools midway into the 2007-08 school year. Participating schools can earn bonus awards of up to $3,000 per full-time union member working at the school if the school meets performance targets defined by the city's accountability program. Our sample includes 186 SPBP-eligible elementary, K-8, and middle schools and 137 control-condition schools in New York City over a two-year period. Overall, we find that the SPBP had little impact on student proficiency or school environment it its first year. However, it is important to remember the short-run results reported in this study provide only very limited evidence of the SPBP's effectiveness. An evaluation of the program's impact after two years should provide more meaningful information about the impact of the SPBP

1. Introduction

Teacher pay for performance has resurfaced as a popular reform strategy in the United States and abroad.[1] The basis for these proposals is grounded in the argument that current compensation policies provide weak incentives to teachers to act in the best interest of their students and that inefficiencies arise from rigidities in current compensation policies. Proponents claim that linking teacher pay to student performance is a powerful way to affect teacher motivation and labor-market selection. Critics, on the other hand, contend that extrinsic incentives may compromise the intrinsic motivation of teachers and possibly lead to dysfunctional behaviors or negative spillover effects.[2] Another frequent criticism of this reform strategy is that output in the education sector is difficult to define because it is not readily measured in a reliable, valid, and fair manner.

Recent experimental and quasi-experimental evidence paints a mixed picture of the impact of teacher pay-for-performance programs. Muralidharan and Sundararaman (2008) and Lavy (2002, 2007) found that teacher incentive programs in India and Israel, respectively, improved student outcomes and promoted positive changes in teacher behavior and/or classroom pedagogy. Glewwe, Ilias, and Kremer (2008) similarly reported that students instructed by teachers eligible to receive an award in a teacher incentive program in Kenya demonstrated better scores on high-stakes tests; however, no discernible impact was found on low-stakes tests taken by a sample of students or on

---

[1] A number of school districts and states in the United States have recently adopted performance-related compensation reforms. Performance is part of compensation packages in the Denver, New York City, Dallas, and Houston public school systems. Florida, Minnesota, and Texas allocate over $550 million to incentive programs that reward teacher performance. The U.S. Congress advanced policy dialogues around teacher compensation reform: first, in 2006, with the appropriation of $495 million over a five-year period to provide Teacher Incentive Fund grants to select districts and states across the country; and in 2009, with part of a massive economic stimulus package earmarking around $200 million for the development and implementation of teacher pay-for-performance programs.

High-profile teacher pay-for-performance plans have also been implemente1d abroad: for example, Chile's Sistema Nacional de Evaluación del Desempeño de los Establecimientos Educacionales (SNED) (Mizala and Romaguera, 2003); Mexico's Carrera Magisterial (McEwan and Santibanez, 2005; Santibanez et al., 2007); programs developed by Israel's Ministry of Education (Lavy, 2002, 2007); and experiments in Andhra Pradesh, India (Muralidharan and Sundararaman, 2008), and in the Busia and Teso districts of western Kenya (Glewwe, Ilias, and Kremer, 2008).

[2] See, for example, Holmstrom and Milgrom (1991), Baker (1992), and Prendergast (1999).

the same students when they took high-stakes tests during the post-intervention school year. Furthermore, a comprehensive evaluation of a long-standing incentive program in Mexico detected a negligible impact on elementary students' test scores and small, positive effects at the secondary level (Santibanez et al., 2007).

This paper contributes to the evaluation literature on teacher incentive systems by assessing the short-run impact of a group incentive program designed by the New York City Department of Education (NYCDOE). The School-Wide Performance Bonus Program (SPBP) was implemented midway into the 2007–08 school year and was designed to provide financial rewards to educators in schools serving disadvantaged students. The SPBP sets expected incentive payments as a fixed performance standard, meaning that schools participating in the program are not competing against one another for a fixed sum of money. All participating schools can earn bonus awards of up to $3,000 per full-time union member working at the school if the school meets predetermined performance targets defined by the NYCDOE's accountability program, with the idea that this sum will be used to award bonuses to teachers and staff found to be deserving. The SPBP rules further mandate that schools participating in the program establish a four-person site-based compensation committee to determine how bonus awards will be distributed to school personnel.

The SPBP is interesting to study for a number of reasons. First, the NYCDOE randomly assigned schools qualifying for the program to either treatment or control status. Since true random assignment will remove unobserved factors that can lead to systematic differences between schools receiving the SPBP treatment and those not eligible to do so, any significant differences in future outcomes can be attributed to the SPBP intervention rather than to other confounding factors associated with outcomes of interest. Moreover, even though the United States has a long history of

testing various teacher compensation reforms, this study is the first to report the causal effect of a domestic teacher pay-for-performance program.[3]

Second, design and implementation of the SPBP addressed potential obstacles that can diminish teachers' receptiveness to compensation reform. The SPBP was developed collaboratively by the NYCDOE and the United Federation of Teachers (UFT), which is the sole bargaining agent for school district personnel.[4] Program guidelines that they developed required that at least 55 percent of school personnel in SPBP-eligible schools vote in favor of participation and that all school personnel within a school be eligible to receive an award.[5] On the other hand, some observers contend that using a school as the unit of accountability makes for a weak incentive policy, since school personnel may feel unable to influence the chances that their school qualifies for an award.[6] And despite the SPBP's assignment of responsibility to site-based "compensation committees" for determining how bonus awards are distributed, similar reforms in Texas suggest that schools tend to adopt very egalitarian award-distribution plans when teachers have a role in designing school-level incentive systems (Springer et al., 2008; Taylor, Springer, and Ehlert, forthcoming).

This study also takes advantage of school-level data on institutional and organizational practices collected by the NYCDOE. The district collects survey data on student, parent, and

---

[3] A few pay-for-performance experiments are running concurrently in the U.S. public school system. The National Center on Performance Incentives has implemented an individual teacher incentive program in Nashville and a team-level incentive program in Round Rock, Texas. Mathematica Policy Research, Inc. is evaluating a five-year demonstration project examining the impact of the Teacher Advancement Program in the Chicago Public Schools.

[4] For more information on the role of teacher associations and collective bargaining agreements in teacher compensation reform, see Eberts (2007), Koppich (2008), Goldhaber (forthcoming), and Hannaway and Rotherham (2008).

[5] The NYCDOE secured funding from private sources to operate the SPBP during the first year of implementation. The district also appropriated public funding for year two of the program. Heinrich (2004) notes: "Districts and states rarely provide consistent funding for these programs, significantly reducing their motivational value."

[6] Sager (2009) contends that New York City should "take it up a notch" by implementing an individual-based incentive system.

teacher perceptions of the school learning environment, including items on academic expectations, communication, engagement, and safety. Teams of experienced educators also conduct two- to three-day on-site visits to review the quality of a school's institutional and instructional program. The school learning-environment survey and external quality reviews provide a means for studying the causal effect of the SPBP on intermediate outcomes. If significant differences in student achievement among schools assigned to the treatment and control conditions are detected, data on institutional and organizational practices could shed light on the types of changes that affected student achievement.

Our evaluation focuses on the impact of the SPBP on student achievement in mathematics during the first year of implementation. A series of analyses also uses school-level survey data on student, parent, and teacher perceptions of the school learning environment, as well as school-level data from enumerators' tests of how well a school is organized for the purpose of improving student learning. In addition, we explore the first-year impact on a variety of student and school characteristics.

Our sample includes 186 SPBP-eligible elementary, K–8, and middle schools and 137 control-condition schools in New York City over a two-year period. The 2006–07 school year is the baseline. The first year of implementation was the 2007–08 school year, though less than three months of school elapsed between the end of the period in which an eligible campus had to vote on whether to participate and the point at which New York State administered the high-stakes mathematics tests.[7] Test scores in mathematics for more than 100,000 students in grades three through eight were collected and reviewed. We do not include the English language-arts test because

[7] The SPBP was formally announced on October 23, 2007. The randomization of schools into treatment- and control-group conditions was announced in November and December of the same year. New York State's high-stakes English language-arts exams were administered from January 8 to 17, 2008. The high-stakes mathematics tests were implemented two months later (March 4–11, 2008).

it was administered a few weeks after the SPBP was implemented, and before the distributional rules of the SPBP reward system had been finalized by each school's compensation committee.

We found that the SPBP had no discernible effect on overall student achievement in mathematics during the first year of the program's implementation. The sign on the SPBP coefficient is negative in virtually all models, though the average treatment effect is always insignificant at any conventional level. There were no discernible impacts when adjusting estimates for SPBP-eligible schools that declined participation. The same holds true when using different achievement specifications.

An important question is whether any particular group of students or schools benefited from participating in the SPBP. Some previous studies of other pay-for-performance progams have found differential effects on student outcomes by student race (Ladd, 1996), prior student achievement (Lavy, 2008), family affluence (Muralidharan and Sundararaman, 2008), and parent education level (Lavy, 2002). Studies have also found no evidence of a significant difference attributable to student or teacher characteristics (Lavy, 2008; Muralidharan and Sundararaman, 2008; Lavy 2002). We found that neither student race nor initial student achievement produced statistically significant differences in the impact of the SPBP. We did not have access to data on other student characteristics such as free and reduced-price lunch status, parental education, or gender, or data on teacher characteristics such as years of experience, salary, or gender.

Organizational theory on group incentive programs suggests that social penalties and other strong forms of reciprocity can positively affect effort if the size of the team is not too large (Kandel and Lazear, 1992; Besley and Coate, 1995; Bowles and Gintis, 2002). Teachers in SPBP-eligible schools with large student enrollments may not respond to the SPBP because they feel less able to affect performance measures that establish qualification for bonuses. Contrary to previous theory, our findings suggest that mathematics achievement by students enrolled in schools with fewer

students remained static in response to the SPBP, while student achievement in schools with larger enrollments decreased. The potential moderating affect of school size on the direction and/or strength of the relationship between the SPBP and mathematics achievement will be revisited when data from the 2008-09 school year become available.

We also examined the impact of the SPBP on student, teacher, and parent perceptions of the school learning environment, as well as external enumerators' tests of a school's instructional program. We found no discernible differences in intermediate outcomes between SPBP schools and schools assigned to the control condition. Admittedly, estimates may be losing leverage. These data are aggregated at the school level, response rates on the survey vary considerably among schools, and enumerators' quality reviews are not available for all the schools in our sample. In addition, a positive and significant effect on teacher perceptions of the school learning environment would need to be interpreted cautiously, in light of the fact that scores count toward a school's overall Progress Report Card rating, which determines whether its teachers can qualify for a bonus award. We use a regression discontinuity design within the randomized evaluation to examine whether the difficulty of performance thresholds that SPBP schools needed to reach to earn a bonus contributed to the treatment effect. Schools had to meet different performance targets determined by their overall Progress Report Card score ranking to earn a bonus award. We use discrete cutoffs in performance target scores to identify these impacts. We found no evidence of a differential treatment impact among schools in response to the performance targets that they had to meet to earn a bonus award.

Note that this evaluation examines the impact of the SPBP after the program had been in operation for less than three months. Even though a randomized evaluation study of incentive programs in Andhra Pradesh, India observed a modest impact on student achievement after a single year, the governance structures in rural schools there are very different from the operational context

of New York City schools. The incentive structure facing teachers and schools in Andhra Pradesh is very weak compared to the accountability measures found in New York City (and the United States, more generally).[8] Furthermore, a series of educational reforms in New York City operating concurrently with SPBP potentially makes it difficult to distinguish the short-run effects that the SPBP generated. These reforms focus on the same outcome measures used in this study to assess the impact of the SPBP on student outcomes. Consequently, we will be able to offer a more comprehensive understanding than we have today of the impact of the SPBP on student outcomes, teacher behavior, and schooling practices as more years of data become available.

Finally, we evaluate the impact of the SPBP on the productivity of existing teachers and school personnel. A system of remunerating employees on the basis of individual or group performance is likely to do a better job of retaining such people and attracting new ones than systems that do not. The size of the sorting effect has been reported to be as large as the size of the incentive effect on the productivity of existing workers (Lazear, 2000).[9] The impact of the SPBP on teacher sorting and selection, as well as the teacher sorting and selection implications for student achievement and other outcomes of interest still await study.[10]

The remainder of this paper is organized as follows: Section 2 discusses the components of pay-for-performance programs. Section 3 reviews key findings from the relevant literature on teacher pay-for-performance programs. Section 4 provides a complete description of the SPBP.

---

[8] Kremer et al. (2004) reported the average absence rate for teachers in Andhra Pradesh, India was about 25 percent while only about half of the teachers in a nationally representative sample of government primary schools in India were actually teaching when external enumerators conducted unannounced visits. Teacher absenteeism in the United States is around five or six percent (Ehrenberg et al., 1991; Ballou, 1996; Podgursky, 2003; Clotfelter, Ladd, and Vigdor, 2007; Miller, Murnane, and Willett, 2007).

[9] In a case study of Safelite Glass Corporation, Lazear (2000) estimated that the compensation system's transition from hourly wages to piece rates was associated with a 44 percent increase in productivity (as measured by individual worker output per month). Interestingly, half of this effect was attributed to workers becoming more motivated, an incentive effect; the other half resulted from the sorting of more able workers largely through the hiring process.

[10] For more information on the relationship between teacher incentive programs and teacher mobility, see Taylor and Springer (2009), Springer et al. (2008), Springer et al. (2009), and Clotfelter, Glennie, Ladd, and Vigdor (2008).

Section 5 describes the data, sample, and random assignment of schools to treatment. Section 6 offers a description of the analysis plan, which sets the stage for Section 7, which is a discussion of overall results. Section 8 provides an analysis of potential differential treatment effects. Section 9 is the conclusion.

2. Understanding Components of Pay-for-Performance Programs

An organization's compensation system is arguably its most important human-resource management system (Ehrenberg and Milkovich, 1987; Lawler, 1981). Providing employees with financial incentives is believed to increase organizational productivity by strengthening employee motivation and attracting and retaining more effective individuals. However, in the public education sector, many contend that sufficient incentives reside in the work itself and that rewards can suppress teachers' intrinsic motivation (Johnson, 1986; Lortie, 1975). Social psychologists refer to the trade-off between intrinsic and extrinsic motivation as the "hidden cost of rewards" (Lepper and Greene, 1978) and "the corruption effect of intrinsic motivation" (Deci, 1975), or what behavioral economists have labeled the "crowding out" of intrinsic motivation (Frey, 1997).

However, numerous design components need to be understood before education reformers can conclude that teacher pay-for-performance programs are practical. For example, whose performance should determine bonus award eligibility? What performance indicators will monitor and appraise employee performance? Will the program reward school personnel according to a relative or an absolute performance standard? Who is part of the pay-for-performance system, and how will bonus awards be distributed to school personnel?

2.a. Forms of Teacher Pay-for-Performance Programs

The unit of accountability describes the entity whose performance determines award eligibility. It can be an individual teacher, a group or team of teachers (e.g., grade-level, department, interdisciplinary team, or school), or some combination thereof. Some literature indicates that pay-for-performance programs that are focused on the individual as the unit of accountability will achieve the best outcomes, particularly if output can be easily attributed to a single individual, the criteria for performance appraisals are observable and objective, and the work does not depend on the interdependence of employees (Deutsch, 1985; Milgrom and Roberts, 1990; Bowles and Gintis, 2002). A common critique of individual incentive programs rests on observation of dysfunctional behavior and system gaming (Prendergast, 1999; Murnane and Cohen, 1986).[11] Furthermore, in education and other sectors involving complex tasks and multiple goals, individuals have greater opportunity to maximize their own utility by reallocating effort to metered, rewarded activities (Holmstrom and Milgrom, 1991; Baker, 1992; Courty and Marschke, 2004).

Pay-for-performance programs that are focused on the group as the unit of accountability may contribute to greater productivity in organizations such as schools, where employees work interdependently. Group incentives can promote social cohesion and feelings of fairness and generate productivity norms (Lazear, 1998; Rosen, 1986; Pfeffer, 1995). A frequently cited threat to group incentive structures is free-riding or shirking, which suggests that some workers may underperform because they assume that others will take up the slack. However, the free-rider problem can be solved through mutual monitoring and the enforcement of social penalties if the team unit is not too large (Kandel and Lazear, 1992; Nalbantian and Schotter, 1997; Bowles and Gintis, 2002). Group structures may also create a perverse incentive by motivating effective teachers

[11] A growing body of education research documents dysfunctional behavior in response to high-stakes accountability programs, including systematically excluding low-scoring students from testing, reclassifying students assignment to particular student subgroups, altering student answer sheets, and focusing on marginally performing students. See, for example, Cullen and Reback (2006), Figlio and Getzler (2002), Figlio and Winicki (2002), Jacob and Levitt (2003), and Neal and Schanzenbach (forthcoming).

in low-performing schools to move to higher-performing schools, where their potential to earn a bonus award increases (Ladd, 2001; Clotfelter, Ladd, and Vigdor, 2005).

Pay-for-performance programs may use any number of performance indicators to monitor and appraise individual or group performance. Test scores are the most heavily weighted performance metric in most output-focused systems. These systems may also incorporate graduation or promotion rates, student or teacher attendance, a reduction in disciplinary referrals, increased test participation, and the like. On the other hand, some pay-for-performance programs may focus more heavily on input-based measures, particularly those that were developed and implemented prior to 2002 (e.g., teacher career ladder or knowledge- and skill-based pay programs).

Past compensation reforms in the education sector have been faulted for measuring what exists rather than proposing and testing what might be useful and important to measure. Today, most agree a pay-for-performance system must have multiple measures and can't be singularly focused on test scores. A structural misalignment between performance measures and a school's mission, or volatility in the outcome measure from one point in time to a later one, can create discontent among teachers and distort policy.

Incentive structure is another key component of teacher pay-for-performance programs. Programs can award a teacher, team of teachers, or an entire school contingent on the basis of how their performance compares with that of similarly situated individuals, groups, or schools using a rank-ordered tournament, or such programs can adopt a fixed performance standard by which any teacher or group of teachers meeting a predefined threshold wins (Lazear and Rosen, 1981; Green and Stokey, 1983).[12] Tournament incentive structures create competition among individuals or

[12] Neal (forthcoming) contends that it is important to come up with incentive pay designs specially suited to public education. He recommends rank-ordered tournaments of comparable schools that measure and reward school-wide performance. He identifies three challenges that the design of incentive-pay systems face: (1) defining the intended outcomes of public education; (2) the inability of existing assessment tools to identify and measure the contribution of specific teachers or schools to student learning; and (3) the lack of true market forces in the public education system.

groups to partake in a fixed pool of bonus awards, thus removing the financial risk inherent in operating a fixed performance–standard scheme. Incentive structures based on a relative performance standard may also be more practical when no obvious performance target exists or performance metrics are volatile. However, tournament incentive structures for teachers or teams of teachers have not received much support because schools have strong work interdependencies, though it is possible to design tournaments in which groups within schools are not competing against one another.

Bonus award distribution systems determine how evenly a pay-for-performance system distributes rewards to eligible employees. An egalitarian distribution plan distributes incentive money widely, in contrast to those plans that reward some individuals far more than others. Proponents argue that individualist reward plans help create a meritocracy able to retain an organization's highest performers, attract similar talent over the long run, send a clear signal to the lowest performers to improve or move elsewhere, and are more cost-effective (Milgrom and Roberts, 1992; Zenger, 1992; Ehrenberg and Smith, 1994; Pfeffer and Langston, 1993). At the same time, a growing body of research suggests that egalitarian pay distribution promotes cooperation and group performance, which are critical in participative organizations. Furthermore, Milgrom and Roberts (1992) suggest that greater pay dispersion may elevate the performance of the lowest performers, who also like receiving awards.

3. Review of Relevant Research and Experiments on Pay for Performance

This section offers a review of previous research studying the impact of teacher pay-for-performance programs on student outcomes, teacher behavior, and institutional dynamics. Our review focuses on evaluations of studies having experimental designs or those using regression

discontinuity (RD) designs in a quasi-experimental framework.[13] When implemented properly, such designs are ideal for assessing whether a specific intervention truly produces changes in outcomes under study or whether observed changes in outcomes are simply artifacts of pretreatment differences between two or more groups under study.

Muralidharan and Sundararaman (2008) studied the impact of two output-based incentive systems (an individual teacher incentive program and a group-level teacher incentive program) and two input-based resource interventions (one providing an extra-paraprofessional teacher and another providing block grants). In what was known as the Andhra Pradesh Randomized Evaluation Study (AP RESt), 500 rural schools in Andhra Pradesh, India, were randomly selected to participate and then assigned to one of the four treatment conditions or to the control group. These schools had a weak incentive structure for teachers, with 90 percent of noncapital education spending going to regular teacher salary and benefits. The AP RESt intervention was developed in partnership with the government of Andhra Pradesh, a large nonprofit organization interested in education issues in India (the Azim Premji Foundation), and the World Bank.

The individual incentive program awarded bonus payments to teachers for every percentage point of improvement above five percentage points in their students' average test score. All recipients received the same bonus for every percentage point of improvement. The bonus award scheme was structured as a fixed performance standard, which means that awards were distributed to any teacher or school that was selected to be in the AP RESt intervention and that exceeded the performance threshold.

Muralidharan and Sundararaman (2008) reported that student test scores on high-stakes tests increased between 0.12 and 0.19 standard deviations in the first year of the program and between

[13] It is important to note that RD studies generate highly localized estimates of a treatment effect, and estimates tend to be low-powered in many applications because they are reliant on a subset of observations immediately above and below a cutoff point.

0.16 and 0.27 standard deviations in the second. Students enrolled in classrooms presided over by teachers eligible to receive a bonus award scored 0.11 to 0.18 standard deviations higher on low-stakes tests than those students whose teachers were not eligible to earn a bonus award. Students in treatment-condition classrooms also scored higher on a separate test that assessed high-order thinking which Muralidharan and Sundararaman (2008) indicate represents "genuine improvements" in learning, as opposed to better test-taking skills or perhaps other strategies employed by teachers to increase their chances of receiving a bonus award.

Muralidharan and Sundararaman (2008) also found that the schools assigned to the output-based intervention (i.e., individual- or group-incentive conditions) outperformed those schools assigned to the input-based resource interventions (i.e., paraprofessional or block grant conditions). Students enrolled in a classroom instructed by a teacher selected for the group incentive intervention also outperformed students in control-condition classrooms on the mathematics and language tests (0.28 and 0.16 standard deviations, respectively). At the same time, students enrolled in schools assigned to the individual incentive condition outperformed students in both the group incentive condition and the control condition following the second year of implementation.

Another interesting feature of the AP RESt study is that external evaluators collected data on intermediate outcomes in interviews and through classroom observation. Teacher interviews offered anecdotal evidence that teachers in the individual or group incentive intervention were more likely to assign homework, offer support outside of class time, have students complete practice tests, and focus attention on low-performing students. However, Muralidharan and Sundararaman (2008), using data collected by the observational protocol, found no significant differences between treatment- and control-condition classrooms.

Glewwe, Ilias, and Kremer (2008) studied the impact of the International Child Support Incentive Program (ICSIP), a group incentive intervention that randomly assigned 100 schools in

rural Kenya to either a treatment or a control condition. ICSIP's bonus scheme was structured as a rank-ordered tournament, and prizes ranged between 21 percent and 43 percent of average monthly base salary.[14] The ICSIP appraised school performance on the basis of student drop-out rates and test scores, with the twelve highest-performing and the twelve most-improved schools that were assigned to the ICSIP intervention receiving a prize.

Glewwe et al. (2008) found that students enrolled in schools participating in the ICSIP intervention had noticeably higher scores on high-stakes tests than students enrolled in schools assigned to the control condition. However, when comparing the performance of students enrolled in control- and treatment-group schools on a low-stakes test, Glewwe et al. (2008) found no differences in student test scores. It appeared that students enrolled in schools participating in the ICSIP intervention were coached in test-taking skills; an analysis of item-level test data revealed, for example, that treatment-condition students were significantly less likely to leave a test question blank.

Glewwe et al. (2008) also examined the impact of the ICSIP on teacher behavior. The authors found no differences in teacher attendance or pedagogy (behavior in classroom, instructional practices, number of homework assignments) among teachers in schools assigned to the ICSIP intervention and those working in a control-condition school. At the same time, teachers working in schools eligible for an ICSIP prize were 7.4 percentage points more likely to offer test-preparation sessions for students outside of normal school hours (typically when students were on vacation). In total, Glewwe et al. (2008) question the probability of the ICSIP program's improving long-run education outcomes, given the current state of schooling in the Busia and Teso districts of western Kenya.

---

[14] Unlike other incentive programs discussed in this section of the paper, ICSIP awarded teachers with prizes rather than cash bonuses. As noted by Glewwe, Ilias, and Kremer (2008), the ICSIP awarded prizes such as a suit worth about $50, plates, glasses and cutlery worth about $40, a tea set worth about $30, and bed linens and blankets worth about $25.

Unlike the above-mentioned controlled trials, in which teachers or schools were randomly assigned to research groups, the next several studies exploited the fact that teachers or schools assigned to intervention and control-group conditions differ solely with respect to a cutoff point along some pre-intervention assignment variable. When implemented properly, an RD design allows for unbiased comparison of average treatment effect on teachers or schools that fall just to the right or to the left of such selection cutoffs.[15] The remainder of this subsection presents an overview of major findings from three RD studies of education incentive interventions: two programs implemented in Israel and a program operating in Mexico since 1992.

Lavy (2002) evaluated a group incentive program that was implemented in sixty-two Israeli high schools and designed to reduce student drop-out rates and improve student achievement. The program rewarded school performance on the basis of three factors: mean test scores, mean number of credit hours, and school drop-out rate. The bonus scheme was designed as a rank-ordered tournament, with the schools in the top third of performers competing for $1.44 million in awards. Schools earning a bonus had to distribute to their teachers 75 percent of the school-level award funds in amounts proportional to their gross annual compensation, regardless of their performance during the school year; the remaining 25 percent was to be used for improving school facilities for teachers. Lavy (2002) reported that top-performing schools received between $13,000 and $105,000 during the first year of implementation, with teacher bonuses ranging from $250 to $1,000 per teacher.

Lavy (2002) found a positive and statistically significant effect on student outcomes. Following the second year of implementation, for example, the group incentive program was found to have had a positive effect on average credit hours earned, average science credits earned, average test scores, and proportion of students taking Israel's matriculation test. Estimates further indicated

[15] For a discussion of RD designs, see Thistlewaite and Campbell (1960); Hahn, Todd, and van der Klaauw (2001); and Lee and Lemieux (2009).

that the program affected particular groups of students more than others—for instance, students at the low end of the ability distribution performed much better than expected on Israel's exit tests.

Lavy (2002) also compared the effectiveness of Israel's group incentive intervention with an input-based intervention that had been implemented several years earlier. The input-based intervention provided twenty-two secondary schools with additional resources to implement professional training programs, reduce class size, and offer tutoring to below-average students. Although both programs improved student outcomes, Lavy (2002) concluded that the group incentive program is more cost-effective per marginal dollar spent. Muralidharan and Sundararaman (2008) similarly found that both the individual and group incentive programs were more cost-effective than either the "extra-paraprofessional" teacher or block-grant treatment conditions. The relative effectiveness of these interventions is particularly relevant to U.S. education policy because input-based reforms generally have been implemented more widely than output-based interventions such as New York City's SPBP.[16]

Lavy (2008) studied an individual incentive program in Israel that awarded bonuses to high school teachers in grades ten, eleven, and twelve based on their students' performance on national exit tests. The program was structured as a rank-ordered tournament and operated for a single semester (January–June 2001). Teachers in the intervention could earn a bonus for each class of students they prepared for the national exit tests, with awards ranging from $1,750 to $7,500 per class prepared. As reported by Lavy (2008), of the 302 teachers (48 percent of eligible teachers) awarded a bonus following the June 2001 exit tests, sixteen won bonuses for two of their classes.

Lavy (2008) creatively exploited two subtle features of the pay-for-performance program—measurement error in the assignment variable and a break along the pre-intervention assignment

[16] Hanushek (2003) provides a critical review of evidence on input-based schooling policies in the United States and abroad.

variable—to estimate the causal impact of the incentive program by using regression discontinuity design. Estimates of the net intervention effect indicated that the number of exit-exam credits earned by students instructed by a teacher in the incentive program increased by 18 percent in mathematics and 17 percent in English, while data from a survey of teacher attitudes and behaviors suggested positive changes in teaching practices, teacher effort, and instruction tailored to low-performing students. When investigating gaps in performance between the results of school tests and national tests taken by students enrolled in treatment and comparison schools, Lavy (2008) did not find evidence of opportunistic behavior or negative spillover effects.

Santibanez et al. (2008) used a RD design to estimate the impact of Mexico's Carrera Magisterial (CM) on student test scores. Implemented in 1992, CM is a teacher incentive program that was designed collaboratively by state and federal education departments and the national teachers' union. Teachers participating in the program can earn a financial bonus if they accumulate enough points on a variety of measures defined by CM guidelines, including input criteria such as years of experience, highest degree held, and professional development activities, as well as output criteria such as their performance on a subject-matter knowledge test and their students' test scores (Santibanez et al., 2008). Awards ranged from 24.5 to 197 percent of a teacher's annual earnings (McEwan and Santibanez, 2005; Ortiz-Jiminez, 2003).

Santibanez et al. (2008) take advantage of the financial incentive that individual teachers have to improve their students' test performance. Since the program appraises teachers on most performance measures before students take the high-stakes tests each school year, teachers participating in the CM program have a general sense of how many additional points they need to earn on the strength of their students' performance on the high-stakes test to receive an award. Santibanez et al. (2008) detected a negligible impact on test scores of students enrolled in elementary

Pay-for-Performance Experiments in the United States

Despite more than a quarter-century of sustained debate over teacher compensation reform, research on pay-for-performance programs in the U.S. have tended to be focused on short-run motivational effects and to be highly diverse in terms of methodology, population targeted, and programs evaluated (Podgursky and Springer, 2007). Indeed, the four pay-for-performance programs known to us to employ a random-assignment design, as this study does, are still being implemented or evaluated. Building a solid research base is necessary for making firm judgments about pay-for-performance programs generally and for deciding whether specific types of design features have promise.

In August 2006, the National Center on Performance Incentives (NCPI) implemented the Project on Incentives in Teaching (POINT) intervention in the Metropolitan Nashville Public Schools (MNPS) system.[17] The POINT experiment recruited 297 teachers of middle-school mathematics in grades five through eight and randomly assigned these teachers to the treatment or control condition. Teachers assigned to the intervention are eligible to receive bonuses of up to $15,000 per year for a three-year period on the basis of two factors: the progress of a teacher's math students over a year, as measured by their gains on the Tennessee Comprehensive Assessment Program (TCAP); and the progress of a teacher's nonmath students over a year, as measured by their gains on the TCAP as well.

The POINT experiment is designed as an individual incentive intervention in which performance is judged according to a fixed performance standard. Because this standard was determined at the beginning of the POINT experiment and will remain fixed for three years, all teachers have the opportunity to be rewarded for having improved over time. The experiment concludes following the 2008–09 school year, and preliminary results will be available sometime during the following year.

In October 2008, the NCPI implemented a demonstration project to study a group incentive intervention. Eighty-two grade-level teams of teachers in grades six, seven, or eight were randomly assigned to either the treatment or control conditions. A team is defined as a group of academic teachers who meet regularly to discuss a common set of students, performance goals, and outcomes for which they are collectively accountable. Teachers assigned to the incentive intervention are eligible to receive an award if their team is selected as one of the four highest-performing teams at their grade level, as measured by standardized achievement scores in reading, mathematics, science, and social studies. Treatment teachers are projected to earn a bonus of about $6,000 if their team qualifies for an award.

Glazerman et al. (2007) designed and implemented an impact evaluation of the Teacher Advancement Program (TAP), a program being implemented by the Chicago Public Schools using a federal Teacher Incentive Fund grant. The TAP is a comprehensive school-reform model consisting of four elements: (1) multiple career paths; (2) ongoing, applied professional growth; (3) instructionally focused accountability; and (4) performance-based compensation.[18] At the beginning of the 2007–08 school year, Glazerman and colleagues randomly assigned eight schools to receive the TAP intervention and eight schools to the control condition. The latter set of schools delayed implementation of TAP for a two-year period while serving as controls. Another sixteen schools were then recruited and randomly assigned to the TAP intervention or control conditions for the 2009-10 and 2010-11 school years.

---

[17] The NCPI, a state and local policy research and development center funded by the U.S. Department of Education's Institute of Education Sciences, was established in 2006 to conduct independent and scientific studies on the individual and institutional effects of pay-for-performance programs and other incentive policies. The NCPI is located at Vanderbilt University's Peabody College and core institutional partners include the RAND Corporation and the University of Missouri – Columbia. More information can be found at www.performanceincentives.org.

[18] More information on the TAP can be found at www.talentedteachers.org. For a recent, non-experimental evaluation of the TAP see Springer, Ballou, and Peng (2008). The Center for Educator Compensation reform also provides an overview of a related program in Chicago's Public Schools (http://www.cecr.ed.gov/initiatives/profiles/pdfs/Chicago.pdf).

school classrooms taught by teachers facing a strong incentive, while they detected small, positive effects at the secondary level. The authors note that their identification strategy relies on a factor in the CM program that may be worth too few points to motivate teachers to exert more effort to improve student test scores.

4. New York City's School-Wide Performance Bonus Program

The SPBP is a group incentive program developed collaboratively by the NYCDOE and the UFT. The SPBP sets expected incentive payments using fixed performance standards, not by constructing a rank-ordered tournament. The SPBP was conceived as a two-year pilot program, with the number of eligible schools increasing from approximately 216 to 400 in the second year. However, because of budgetary constraints, the number of SPBP-eligible schools did not grow in the 2008–09 school year. Stakeholders are currently exploring funding the SPBP for a third year by leveraging funding obtained from the Obama administration's American Recovery and Reinvestment Act (Hernandez, 2009).

Participating schools earn bonus awards if they meet performance targets established by the NYCDOE's Progress Report Card system, which is the primary accountability program in the school district.[19] The Progress Report Card system evaluates each public school on the basis of three factors: student attendance and student, parent, and teacher perceptions of the school learning environment (15 percent); student performance on New York State's high-stakes test in English language arts and mathematics (30 percent); and student progress in English language arts and mathematics (55 percent). All schools receive an overall Progress Report Card score and grade—

[19] Although performance targets were eliminated from the Progress Report Card system for the 2008–09 school year, the NYCDOE and the UFT elected to use the same metric from 2007-08 school year for schools participating in the second year of the SPBP. For an evaluation of the NYCDOE Progress Report Card system, see Rockoff and Turner (2008) and Winters (2008).

from A to F—which is based on how well they performed in these three areas in comparison with a set of schools serving a similar population of students.[20] The Progress Report Card system then assigns each public school a performance target for the subsequent school year based on the rank of its overall Progress Report Card score.

Table 1 displays descriptive information on the relationship between overall performance-score rankings and performance targets. For example, if a school's overall Progress Report Card score ranked it in the 75th percentile of schools—that is, in Category 2—its target improvement for the next year's score would have been 12.5 points for the 2007–08 school year. In other words, the school's overall performance target score for that school year was its overall performance score from the 2006–07 school year plus 12.5 points. Table 1 also displays the number and percentage of schools in our sample according to their Progress Report Card performance rankings and their target gains.

[INSERT TABLE 1 ABOUT HERE]

Schools participating in the SPBP that meet 100 percent of their performance target score receive $3,000 per UFT member in their school. Schools that meet 75 percent of their performance target score receive $1,500 per UFT member in their school.[21] As displayed in Table 2, of ninety-three SPBP-eligible schools meeting their performance target, sixty-five met 100 percent of their performance target and twenty-eight schools met 75 percent of their target. In total, $14.25 million

---

[20] A school can also earn bonus points, which are added to their overall Progress Report Card score when high-needs students make exemplary progress on New York State's high-stakes tests. The Progress Report Card system identifies five categories of high-needs students: (1) any student identified as having special needs; (2) any student identifieid as being limited English proficient; (3) Hispanic students in the bottom third of all NYCDOE students; (4) black students in the bottom third of all NYCDOE students; (5) all other students in the bottom third of all NYCDOE students. "Exemplary" gains are those in the highest 40 percent of all student gains per school type in the NYCDOE. For more information, see New York City Public Schools (2007).

[21] In June 2008, the NYCDOE and the UFT announced a third way that schools participating in the SPBP could earn a bonus award: by achieving two consecutive A-grades under the Progress Report Card system. Doing so entitles them to receive $1,500 per UFT member. However, this alternative does not have any bearing on our analysis of the first year because schools were unaware of the policy during the school year and thus could not have responded to it.

was awarded to these schools, with bonus awards ranging between $51,000 and $351,000 per school (with an average award of $160,095 per school).

[INSERT TABLE 2 ABOUT HERE]

Nearly all schools in our sample entered the lottery because of the challenges posed by the nature of their student bodies, not their previous achievement.[22] All schools in the lottery served students with difficult backgrounds. As illustrated in Table 1, some schools were identified as being more effective than others at improving student outcomes (and earned a high number of points under the Progress Report Card system or a high grade under No Child Left Behind [NCLB]). Furthermore, the percentage of schools in our sample in each accountability rating category is similar to the percentage of all schools in the NYCDOE in those categories. The impact of the SPBP should be generalizable to schools of varying productivity with high percentages of disadvantaged student populations.

Schools' receipt of bonuses under the SPBP does not necessarily indicate that program eligibility caused improvements according to the performance indicators: student attendance and school learning environment; performance and progress on high-stakes tests in mathematics and English language arts. The Progress Report Card system is going to identify high- and low-performing schools irrespective of the SPBP. We would expect some schools to meet their Progress Report Card targets and earn bonuses even if there were no treatment effects from the SPBP.

At the same time, since schools are assigned target gain scores on the basis of their overall Progress Report Card performance ranking, some schools may have a greater chance than others of meeting the bonus performance threshold. Table 2 reports the number and percentage of schools

[22] Middle schools are the exception. Middle schools were identified on the basis of their average proficiency ratings in mathematics and English language arts (ELA) in the fourth grade. Our sample contains fifty-five middle schools in the treatment sample and forty-one middle schools in the control sample. These schools make up 29.63 percent of schools in our sample.

assigned by lottery to an SPBP treatment group according to their Progress Report Card category and how many schools in each category met all, some, or none of their performance target during the 2007–08 school year. It is clear that the great majority of Category 4 and Category 5 schools met 100 percent of their performance target, while only about half of Category 2 and Category 3 schools met at least 75 percent of their performance target. Furthermore, even though 65 percent of Category 1 schools met at least part of their performance target, 70 percent of Category 1 schools received an award for earning two consecutive A-grades—a performance metric that was established after the first year of the program concluded.

We also estimated a simple binomial logit model to understand the relationship between Progress Report Card categories and the probability that a school met part of its performance target. Specifically, we estimate the odds of a school meeting at least 75 percent of its performance target when controlling for school level, breakdown of students by race/ethnicity, and peer index rating. We find that the odds of a Category 4 or 5 school's earning at least part of its performance bonus award are about ten times greater than the odds of a Category 3 or 2 school's meeting part of its performance target. Category 1 schools are about two to three times more likely to earn a performance bonus than Category 2 or 3 schools. The difference is explained by the two consecutive A-grades that the Category 1 school had earned.

The SPBP stipulates that schools participating in the SPBP establish a site-based compensation committee to determine how bonus awards will be distributed to school personnel. Compensation committees consist of the school principal, an individual appointed by the principal, and two staff people who are UFT members. A school's compensation committee "has complete discretion, without interference from either the [NYCDOE] or the UFT, to decide how to distribute the pool of bonus money available to the school. The compensation committee could choose to give every employee the same amount, give employees who did exceptional work more, give employees

in one title (for instance, teachers) more, give employees who only worked a partial year less, etc." (SPBP background document, August 1, 2008).

Table 3 provides descriptive statistics and Figure 1 illustrates the range of award amounts for the ninety-two elementary, middle, and K–8 schools that earned a performance-award bonus following the 2007–08 school year. Each vertical bar represents a single school, its lower end being the minimum distributed award (other than zero) and its upper end being the maximum award distributed. The mean bonus awarded to teachers was $2,417 at the school level. About three-quarters of all schools awarded a maximum individual bonus of $3,000 or less. When restricting the sample to only those school personnel classified as teachers, we find that the average bonus increases to $3,000, with more than 90 percent of all teachers receiving a bonus of between $2,500 and $3,500.

[INSERT TABLE 3 AND FIGURE 1 ABOUT HERE]

The average size of the bonus awards received by teachers in SPBP schools that met their performance target is around the size thought to be large enough to influence teacher behavior. For example, an average teacher bonus award of $3,000 is 45 percent of monthly base salary, or 5 percent of annual base salary, assuming a $60,000 average base salary and a nine-month pay period. Case studies suggest that bonus awards of 5–8 percent should be large enough to elicit a behavioral response (Odden, 2001). Furthermore, experimental studies that detected behavioral changes in response to teacher pay-for-performance interventions reported average bonus awards equivalent to about 40 percent of a single month's base salary.

5. Data, Sample, and Random Assignment

5.a. Data

The data for this study come from multiple sources. Student-level data were provided by the NYCDOE's Office of Accountability. The data set contains student demographic information, including race/ethnicity, special-education status, and English-language learner status. It also contains scores on New York State's mathematics and English language arts (ELA) tests administered during the 2006–07 and 2007–08 school years. Using data for the universe of students in the NYCDOE, we standardized student test scores in math and ELA by grade and school year. A negative z-score indicates that the score is below the mean for all tested students in that subject, grade, and year, while a positive z-score indicates that the score is above the distribution mean.

A second data set contained information on the SPBP. It identified eligible schools that voted in favor of, or against, participation. A separate annotated file provided details on both lotteries and documented any violations of the random assignment process between the first and second lotteries. The NYCDOE also provided a teacher-level file setting out the size of the actual bonus awards given in autumn 2008 to personnel who worked during the 2007–08 school year in an SPBP school that met at least 75 percent of its performance target or earned two consecutive A-grades.

We supplemented these files with school Progress Report Card data available on the NYCDOE website. Files contained aggregated data on student demographics, student attendance rates, and student enrollment, as well as information on the following accountability-system ratings: overall accountability, student performance, student progress, environment, engagement, communication, academic expectations, percentile rank, performance target score, and NCLB Adequate Yearly Progress (AYP) status.

We also obtained school-level data from a survey of students', teachers', and parents' perceptions of the school learning environment. The surveys were administered from April 30 to June 6, 2007, and from April 4 to April 18, 2008. Surveys were sent to all parents and teachers, and

to students in grades six through twelve. Response rates increased significantly from 2007 to 2008, with parents' response rate increasing from 26 percent to 40 percent, teachers' response rate increasing from 44 percent to 66 percent, and students' response rate increasing from 65 percent to 78 percent (NYCDOE, 2008).

Finally, we downloaded and keyed data from quality reviews completed for all the NYCDOE during the 2007–08 school year. The quality review process consists of trained teams of enumerators' conducting a prereview of a school and then visiting that school for two to three days. School site visits included a thirty-minute campus tour, ten to fifteen classroom observations lasting twenty minutes each, and structured and unstructured interviews with teachers and students (NYCDOE, 2008). Enumerators assess schools on the basis of five criteria indicating relative quality, each of which contains seven ratings, the lowest being "underdeveloped" and the highest being "outstanding."

5.b. Sample

Our sample includes 186 SPBP-eligible elementary, K–8, and middle schools and 137 control-condition schools over a two-year period comprising the baseline year (the 2006–07 school year) and the first treatment year (the 2007–08 school year). Student test scores are available in mathematics for more than 100,000 students in grades three through eight. We restrict the sample to schools identified on their Progress Report Cards as being an elementary, K–8, or middle school because test scores are unavailable in high school grades. We focus on student achievement in mathematics because the ELA test was administered weeks after the SPBP was implemented and before the distributional rules of the SPBP reward system had been finalized by each school's compensation committee.

Schools had to be elementary, middle, and high schools in the NYCDOE with the highest needs to qualify for the SPBP. The NYCDOE determines a school's "need" by resorting to a peer index ranking system in which: elementary and K–8 school rankings are based on a composite measure of student demographic factors such as the percentage of English-language learners, black students, Hispanic students, special-education students, and Title I free lunch-program students; and middle school and high school rankings are set in relationship to the average proficiency ratings in mathematics and ELA in a single grade (fourth grade for middle schools and eighth grade for high schools).

Table 4 displays descriptive statistics on demographic and performance measures for treatment schools, control-group schools, and all public schools in the NYCDOE. About 59 percent of schools eligible for SPBP were elementary schools, 11 percent K–8 schools, and 29 percent middle schools. The average school size is slightly under 600 students, with elementary schools being modestly smaller in size than K–8 and middle schools. On average, more than 95 percent of the school's students are identified as Hispanic (56 percent) or black (41 percent), 19 percent are identified as English-language learners, and 22 percent receive some level of special-education services. Standardized scores in mathematics and reading are, respectively, approximately 0.36 and 0.37 standard deviations below the mean test scores in the district.

[INSERT TABLE 4 ABOUT HERE]

More than half of SPBP eligible schools (53 percent) in our sample were in good standing, according to New York State's NCLB accountability plan. Another 27 percent of schools were restructuring, while approximately 19 percent attended schools that were either in need of improvement or under corrective action. Interestingly, Progress Report Card grades assigned to schools by the NYCDOE's accountability program suggest that schools in our sample are

distributed more evenly: 23 percent of schools received an A; 32 percent received a B; 27 percent received a C; 10 percent receive a D; and 9 percent received an F.

5.c. Random Assignment of Schools

The NYCDOE had it in mind that 200 schools (including high schools) would participate in the SPBP during the 2007–08 school year. How to arrive at this number was difficult to know. Schools randomly assigned (lotteried in) to the SPBP intervention had to vote in favor of participation. Schools not randomly assigned (lotteried out) to the SPBP intervention were assigned to the control group. Thus, the NYCDOE's Research and Policy Support Group was able to implement a two-stage clustered randomized trial, which is summarized in Figure 2.

[INSERT FIGURE 2 ABOUT HERE]

In early November 2007, the Research and Policy Support Group identified 429 schools meeting eligibility criteria for the SPBP. Almost all of them (404) were entered into the first lottery, from which the Research and Policy Support Group then randomly selected 233 schools, which it invited to participate in the SPBP. Schools had six weeks to vote for or against participation. Of the initial 233 schools lotteried into the SPBP, 195 voted to participate, 35 did not to participate, and 3 were excluded because of complicating factors.

In December 2007, the NYCDOE's Research and Policy Support Group held a second lottery. The second lottery included only the 189 schools that were not selected during the first lottery. Twenty-one schools were randomly selected and then invited to participate in the SPBP. Nineteen of these schools voted in favor of participation, and two schools declined participation. In total, 254 of 404 schools entered into the lottery were randomly selected to participate in SPBP. Thirty-seven schools lotteried into the SPBP declined participation.

Figure 2 indicates a few irregularities in the lottery process. To begin with, 25 schools were barred from the lottery even though these schools, on the basis of observable characteristics, met the selection criteria for entering the SPBP lottery. These schools also were similar to those included in the lottery on the basis of observable student and school characteristics (see Tables 4 and 5). In a conversation with the authors, the NYCDOE indicated that the 25 schools were ruled ineligible prior to the lottery process. While their exclusion could impair the external validity of our findings, it should not have any effect on their internal validity.

Noncompliance in the form of "no-shows" and "crossovers" may blur the contrast in outcomes between treatment groups by understating the average SPBP treatment effect (Bloom, 2006). Thirty-seven schools that lotteried into the treatment condition declined participation following a vote among school personnel (no-shows). Another eight schools were permitted to participate in the SPBP despite never having been lotteried into the SPBP (crossovers). The thirty-seven schools that were lotteried in but declined to participate were coded as having been deemed eligible for the policy but as not having participated. The eight schools that received treatment under special circumstances were coded as being ineligible for participation but to have received treatment.[23] We address noncompliance by using the local average treatment effect (LATE) framework developed by Angrist, Imbens, and Rubin (1996), which is a refinement of Bloom's (1984) average impact of treatment-on-the-treated strategy.

Our analytic strategy assumes that both the observed and unobserved characteristics of treatment and control schools are, on average, identical. Logically, we cannot attest to the identicalness of a school's unobserved characteristics. But we can establish whether there are observed differences between two categories of schools and then infer from a lack of difference

[23] Eight additional schools were offered treatment for some "special case." School personnel at six of these schools voted to participate in SPBP. Special-case schools were not entered into a lottery, so we removed them from our sample.

between them that they are identical in unobserved ways as well. We tested for differences on observables using a Kruskal-Wallis one-way analysis of variance, a nonparametric method for testing equality of population medians among groups (Kruskal and Wallis, 1952).

Table 4 displays descriptive statistics on demographic and performance measures by experimental status. We find that the sample of schools assigned to the SPBP treatment (column 1) are statistically indistinguishable from the schools assigned to the control condition (column 2), according to most demographic characteristics and performance measures in the baseline year (the 2006–07 school year). A slightly greater proportion of control-group schools received a D-grade under the NYCDOE's Progress Report Card system than were enrolled in SPBP-condition schools (0.17 vs. 0.10). We also find that a greater proportion of control-group schools than eligible schools were identified as being "in need of improvement" under NCLB (0.16 vs. 0.10).

Table 4 also displays the extent to which the group of eligible schools that voted in favor of participating in the SPBP differs from the group that chose not to participate. Interestingly, we find few differences between the observed characteristics of those eligible schools that voted to participate in the program and the characteristics of the schools that voted against the SPBP. Personnel in schools that voted to participate in the SPBP had slightly lower ELA scores in the 2006–07 school year (-0.37 vs. -0.27). Furthermore, schools that voted to participate in the SPBP were slightly more likely to have earned an F-grade on the Progress Report Card system (0.10 vs. 0.00) and to have been labeled "in need of improvement" under the NCLB accountability system (0.15 vs. 0.00) than schools in the nonparticipating group. All other observed demographic and performance characteristics of the participating and declining schools are statistically indistinguishable.

Table 5 displays descriptive statistics by experimental status for key constructs and response rates on the NYCDOE's school learning-environment survey. We find that schools eligible for the

SPBP (column 1) and those not lotteried into the SPBP (column 2) are similar in their scores on all characteristics during the baseline year (the 2006–07 school year). The same holds true for schools voting in favor of participating in the SPBP (column 3) and those schools that declined to participate (column 4). Furthermore, we do not detect any significant differences in student, parent, and teacher responses to the school learning-environment survey during the baseline year.

[INSERT TABLE 5 ABOUT HERE]

We used Hotelling's T-test to determine whether there were baseline imbalances between schools participating in the SPBP and control-condition schools (Hotelling, 1940). We say that the lottery is balanced if we cannot on statistical grounds, after examining all observable characteristics identified in Tables 4 and 5, dismiss the possibility that the treatment group and the control group are the same. Hotelling's T-test is the analog to a t-test when multiple variables are considered simultaneously. We fail to reject the hypothesis that the means of the treatment (column 1) and the control (column 2) conditions are different. We also found no significant differences in the means employed by the eligible participant sample of schools (column 3) and declining schools (column 4) as determined by Hotelling's T-test.

6. Analytic Strategy

6.a. Average Impact of Intention to Treat

We first estimate the average impact of the SPBP on student achievement using a standard intention-to-treat (ITT) approach. An ITT effect assumes that all schools lotteried into the SPBP elected to participate in the program, even though an approximate 14 percent of eligible schools in our sample did not participate. ITT estimates are relevant to policy because, by all accounts, if the SPBP is sustained in future years, it is likely that imperfect treatment implementation will continue to occur. Thus, to judge the overall impact of the SPBP as implemented, the combined effect of the

SPBP intervention and the effect of a school's decision not to comply with the policy can be expressed as:

(1) ITT = E[Y|Z=1]-E[Y|Z=0]

where Z=1 indicates a school's assignment to the SPBP intervention and Z=0 indicates a school's assignment to the control condition. Subscripts are suppressed for simplicity.

We also estimate a series of cross-sectional regression models to measure how student and school characteristics affected a student's math test score in the 2007–08 school year. A binary variable is set to equal one if a student was enrolled in a school that was lotteried into an SPBP treatment group and zero if a student was enrolled in a school that was not lotteried into a treatment condition. The average impact of the ITT effect is reported with and without regression adjustments. The most inclusive estimates control for a large number of observable student- and school-level covariates. Our most basic estimation strategy controls only for student grade.

Because SPBP eligibility was determined by lottery, and commonly used tests indicate balance across observable student and school characteristics, we interpret the relationship between SPBP intervention and student achievement in mathematics to be a direct consequence of the SPBP intervention. More formally, ITT estimates of the SPBP intervention are given by the ordinary least squares estimate for d4, which can be defined as:

(2) $$Y_{ist} = \delta_0 + \delta_1 f(Y_{ist-1}) + \delta_2 Student_{is} + \delta_3 School_s + \delta_4 Eligible_{st} + \phi_{ist} + \phi_s$$

where Yist represents the math test score of student i in school s at the end of program year *t* (April 2008); f(Yist-1) is a cubic function of the student's math test score in that subject at the end of year *t-1*; Student is a vector of observable student-level variables, including race/ethnicity, special-education status, limited English proficiency (LEP) status, and so forth; *School* is a vector of observable school-level attributes, including level of schooling (elementary, middle, or K–8) and percentage of students by race/ethnicity and borough; *Eligible* is an indicator variable that equals one

if student i is enrolled in a school that was lotteried into the SPBP intervention and zero if the school was not; $_{ist}$ is a stochastic error term; and $_{s}$ reminds us that this random error is clustered by school.[24]

We also tested for differential SPBP treatment effects by student and school characteristics. For example, previous research has documented system gaming and opportunistic behavior among school personnel in response to high-powered incentive policies.[25] School personnel may respond strategically to the SPBP intervention because the availability of bonus awards is determined by a school's overall Progress Report Card score, and schools can earn bonus points if high-needs students make exemplary progress on the high-stakes tests. We explore differential SPBP treatment effect by including in equation (2) a simple interaction term between *Eligible* and a particular student or school characteristic.

We typically report the average impact of ITT effects using a lagged achievement specification of (2), where the standardized form of a student's previous test score in mathematics at time t-1 is an explanatory variable. Controlling for lagged achievement helps to account for unobservable student attributes such as prior knowledge that students bring to the classroom. We also control for a cubic polynomial of a student's previous test score, which allows for the relationship between previous and current test scores to differ with reference to the student's previous score. Furthermore, the lagged achievement specification does not impose a specific assumption about the rate of decay in student achievement over time.

6.b. Impact of Treatment on the Treated

[24] When estimating equation (4) and all other equations at the student level, we calculate standard errors using the bootstrap method with 300 iterations. Among other advantages, the bootstrap method calculates consistent standard errors in light of potential autocorrelation in regression models, such as the value-added specification, that included a lagged dependent variable as a regressor (Cameron and Trivedi, 2006; Mackinnon, 2002).

[25] See footnote 17.

The ability of schools lotteried into the SPBP intervention group to vote against participation means that the ITT effect does not directly measure the intervention effect on schools that adopted the policy. A handful of schools that were never lotteried into the program received SPBP treatment, which complicates estimating the direct effect on student achievement. Specifically, the presence of forty-three noncompliant schools in the sample (thirty-five no-show and eight crossover schools) may be responsible for the understatement of the average SPBP treatment effect. We therefore estimate models according to a form of the treatment on the treated (TOT) framework developed by Bloom (1984) and advanced by Angrist, Imbens, and Rubin (1996) and others.

Bloom (1984) developed a strategy for estimating the average impact of TOT when no-shows are present in an experimental design (i.e., subjects are assigned to treatment but do not participate). A TOT approach isolates the impact of the SPBP intervention on the subset of schools lotteried into the SPBP condition that actually received the treatment, and it then compares the achievement scores of students enrolled in schools participating in the SPBP with those of the sample of students enrolled in schools that were not lotteried into the SPBP. In contrast to the basic ITT approach identified in equation (2), the TOT effect can be expressed as:

(3) $ITT = E[D|Z=1]TOT + [1-E(D|Z = 1)]0 = [E(D|Z=1)]TOT$

where Z=1 if the school was lotteried into the SPBP intervention and Z=0 otherwise, and D=1 for schools that receive the SPBP treatment and D=0 for those that do not.[26]

However, the TOT effect assumes that schools assigned to the control group did not participate in the SPBP treatment. Not accounting for crossovers may produce a downward bias in estimates of the average treatment effect. We therefore estimate the local average treatment effect

[26] As noted by Bloom (2006), equation (5) further shows that the average effect of ITT equals the weighted mean of TOT effect for schools that were lotteried into and participated in SPBP and it equals zero for the no-show schools, where weights are equal to the SPBP treatment receipt rate ([E(D|Z=1]) and the no-show rate (1-[E(D|Z=1). Equation (3) implies that: $TOT = ITT/[E(D|Z=1)]$.

(LATE) developed by Angrist, Imbens, and Rubin (1996). LATE not only accounts for the lack of participation among those randomly assigned to the SPBP treatment group (no-shows); it also adjusts for SPBP participation by schools that were not lotteried into the treatment group but are participants in the SPBP treatment nonetheless.[27] As noted in Bloom (2006), LATE can be expressed as:

(4)
$$LATE = \frac{ITT}{E(D \mid Z = 1) - E(D \mid Z = 0)} = \frac{\Delta Y}{\Delta D}$$

Equation (4) is equivalent to the Wald Estimator, a special case of an instrumental variables (IV) strategy. An IV strategy can be used to capture the effect of the SPBP intervention on compliers—that is, schools randomly assigned to the control condition but that participated in the SPBP intervention. The LATE is estimated using a two-stage ordinary least squares, by which an IV approach estimates the average treatment effect on the subset of schools that participated in the SPBP because of a lottery assignment, and the estimated probability of receiving the SPBP treatment is then used as an indicator variable in a second-stage regression model. More formally, to establish the probability that a school actually received the SPBP treatment, our first-stage regression model takes the form:

(5)
$$T_{st} = \pi_0 + \pi_1 School_{st} + \pi_2 Eligible_{st} + \omega_{st}$$

where *T*st indicates the school's actual participation in the SPBP during the first year of implementation (2007–08 school year), and all other variables are as previously defined. We then use the resulting coefficient estimates $\hat{\pi}_0, \hat{\pi}_1, \hat{\pi}_2$ to establish the probability that each school received the SPBP treatment.

[27] The LATE is also known as the complier-average causal effect of treatment (CACE).

The instrument in equation (5) is the variable *Eligible*, which indicates whether the school was lotteried into the SPBP treatment condition. The fact that program eligibility was determined randomly suggests that the *School* variables are relatively unnecessary, and the estimated probabilities of whether a school was actually treated resulting from equation (5) are nearly identical, whether or not we include these variables. The coefficient on *Eligible*, $\hat{\pi}_2$, is also very similar to the percentage of eligible schools that voted to participate in the policy. However, for the sake of completeness, we continue to include it in all estimates reported below. The first-stage (or instrumenting) model is performed at the school level because schools (not students) were randomly assigned to the treatment condition. Estimating the first stage at the student level would imply that individual students within a school with different observed characteristics had different probabilities of receiving treatment. The estimated probability that a school received treatment, $\hat{T}$, is merged on the student achievement data file. We then estimate the impact of the SPBP on student achievement using the estimated probability that the student's school received the treatment, which can be expressed as:

$$Y_{ist} = \theta_0 + \theta_1 f(Y_{ist-1}) + \theta_2 Student_{is} + \theta_3 School_s + \theta_4 \hat{T}_{st} + \kappa_{ist} + \kappa_s \quad (6)$$

where all variables are as previously defined in equation (5) and the coefficient on the probability of treatment, θ_4, provides a consistent estimate of the impact of the actual SPBP treatment on student mathematics proficiency.

7. Average Impact of the School-Wide Performance Bonus Program

Table 6 presents results for a series of estimates of the impact of the SPBP on student achievement in mathematics. Panel A reports ITT estimates with and without regression adjustments. Estimates of the ITT effect indicate no significant relationships between SPBP eligibility and student

performance in mathematics. The sign on coefficient estimates is always negative but never significant at conventional levels. Panel B indicates that the same holds true when we use an IV strategy to estimate the LATE, which means that the average treatment effect in the subpopulation of compliant schools is indistinguishable at a conventional level.

[INSERT TABLE 6 ABOUT HERE]

We are also interested in whether particular student subgroups benefit more from the SPBP. We test for differential effects by introducing the simple interaction term *Eligible* with a binary student demographic variable. The LATE effect is estimated by interacting the predicted treatment from equation (5), $\boldsymbol{T}$, with student demographic variables. NYC's Progress Report Card system also gives schools extra credit or bonus points if a high-needs student makes exemplary gains on the state's high-stakes tests. Using a basic $\mathbf{X}^2$ test, we also report whether estimates on these coefficients are jointly equal to zero.

Table 7 reports the average impact of the ITT effect (column 1) and the LATE (column 2), allowing for heterogeneous treatment effects by student race. Regression-adjusted estimates indicate no discernible differences among student race. Estimates are robust irrespective of the controls for student- and school-level covariates or whether we exclude a student's previous test score in mathematics from the regression equation. Furthermore, both the ITT effect and the LATE are robust if student attainment (rather than the lagged-achievement or value-added approach) is the dependent variable when comparing schools lotteried into the SPBP with those randomly assigned to the control condition.

[INSERT TABLE 7 ABOUT HERE]

Table 8 reports results when we allow the estimate of the SPBP treatment effect to vary by student ability, where ability is defined by the quartile of a student's previous test score in the tested subject. ITT estimates do not provide much evidence of the SPBP's benefiting students of a

particular ability group. We find no statistical difference in the performance of students according to the quartile of their baseline math score. However, students in the third quartile scored, on average, 0.0328 standard-deviation units below the typical student enrolled in a school participating in the SPBP. Students whose previous achievement scores were in the bottom performance quartile also performed worse than expected. Furthermore, we find that the LATE estimates in Panel B of Table 8 are qualitatively similar to estimates reported for the average impact of the ITT effect.

[INSERT TABLE 8 ABOUT HERE]

We also examined whether achievement scores in mathematics varied by school type. Our sample includes three types of schools: elementary, middle, and K–8 schools. Approximately 60 percent of students enrolled in SPBP-eligible schools attend an elementary school, while about 29 percent attend middle schools and 11 percent attend K–8 schools. Panel A of Table 9 displays estimates for the ITT effect. We do not find a significant difference in achievement between students enrolled in the SPBP treatment and those enrolled in schools assigned to the control condition. Furthermore, the TOT estimates find that students' achievement gains in mathematics at schools that actually received the treatment are not statistically different from those at untreated schools. Estimates reported in Table 9 are similar to estimates, with or without making adjustments for student or school characteristics, or for a student's previous achievement score.

[INSERT TABLE 9 ABOUT HERE]

In sum, we find no evidence of a significant SPBP treatment effect during the first partial year of implementation. Perhaps this is unsurprising. There was a limited window of opportunity for school personnel working in schools that were lotteried into the SPBP to respond to the SPBP intervention (less than three months), assuming that school personnel were disposed to respond to the program. We will repeat these analyses following the 2009–10 school year, when more years of data become available, including scores on student achievement in ELA.

8. Potential Mediators of the Treatment Effects

This section focuses on potential mediators of the SPBP treatment effect, including school size and the rigorousness of the performance target that a school has to meet to earn a performance bonus award. We also examine the association between institutional and organizational practices and student achievement in mathematics, as measured in surveys of student, teacher, and parent perceptions of the school learning environment. Finally, we examine data from independent appraisals of institutional practices conducted by an independent team of experienced educators.

8.a. Differential Impact by School Size

The SPBP is a group incentive program. School size may affect the strength of incentives offered school personnel in SPBP-eligible schools (Kandel and Lazear, 1992). Our intuition is that, in larger schools, the probability that social penalties can influence group performance diminishes. Further, an individual teacher has relatively little direct impact on the school's overall performance, which is the unit of accountability; in smaller schools, the impact of an individual teacher's performance is proportionately larger. It may also be easier for teachers in larger schools to free-ride, while a smaller school may contain incentives that shape teacher behavior in ways that an individual-level pay-for-performance program does.

To evaluate whether there is a differential SPBP treatment effect caused by school size, we interact the *Eligible* variable with school size. School size is defined as the number of unique students with a valid mathematics test score in the 2007–08 school year. The mean school size was slightly fewer than 600 students, with a standard deviation of approximately 260. The regression model is a

modified form of equation (4) and can be expressed as:

$$Y_{ist} = \pi_0 + \pi_1 f(Y_{ist-1}) + \pi_2 Student_{is} + \pi_3 School_s + \pi_4 Size_s + \pi_5 Eligible_{ist} + \pi_6 (Eligible_{ist} * Size_s) + \chi_{is} + \chi_s \quad (7)$$

where Size is the number of students in a school with a valid test score in mathematics during the baseline school year, and all other variables are as previously defined. The estimate on the interaction term, π_6, indicates the direction and strength of the association between school size and student achievement gains in mathematics.

We also estimate the differential effect of actually receiving the treatment by school size using a two-stage least squares regression. Following the four types of compliance behaviors identified by Angrist et al. (1996), we substitute the *Eligible* variable identified in (2) with estimates of a school's probability of participating in the SPBP that were generated from a linear probability model. We then run a second-stage regression model in which estimates from the first-stage participation model, T, become the instrument for estimating the relationship between school size and student achievement. Because a weak instrument can cause the precision of estimators to be low, we report regression-adjusted estimates.Panel A of Table 10 reports estimates of the differential effect of SPBP eligibility by school size. Model (1) suggests that student achievement gains in mathematics in SPBP-eligible schools tend to be inversely proportionate to school size. The coefficient on the average effect of SPBP eligibility is no longer significant at the 10 percent level in Model (3) of Panel B. Furthermore, there is a negative and significant relationship between school size and receipt of the SPBP treatment, which suggests that some schools participating in the SPBP may be large enough to have a negative effect on students' achievement gains in mathematics.

[INSERT TABLE 10 ABOUT HERE]

Table 10 also reports estimates of the association between school size and treatment status by school size as measured in quartiles. Models (2) and (4) compare student achievement gains in

Quartile 1, Quartile 2, or Quartile 4 schools with achievement gains of students enrolled in Quartile 3 schools using an interaction term between the SPBP indicator and a dummy variable for each quartile. Interestingly, estimates suggest that students who were enrolled in SPBP schools in the quartile containing the largest schools performed not as good as than students enrolled in control-group schools in the same quartile. Average estimates of both the ITT effect and the LATE are statistically significant at the 5 percent level.

We can also see this effect in the model incorporating an interaction between overall enrollment and treatment. Differentiating (7) with respect to *Eligible*, we can see that the overall impact of SPBP eligibility in this model is found by solving: $\hat{\pi}_5 + \hat{\pi}_6 * Size$. We can recover the school size at which treatment is no longer pointed in a positive direction by inputting the coefficient estimates from the regression $\hat{\pi}_5$ and $\hat{\pi}_6$, setting the resulting equation equal to zero, and solving for *Size*. Doing so for the ITT model in math yields a school size of 529 students, the point where the coefficient estimate for SPBP eligibility goes to zero and then turns negative with the enrollment of every additional student in a school.

We performed a series of X^2 tests to identify the points at which any positive effect (when school size is below 529) and any negative impact (when school size is greater than 529) are statistically different from zero. We find that school size must drop to under 120 students to produce a positive treatment effect that is statistically significant at the 10 percent level. No school in our sample is this small. We also find that the overall SPBP treatment effect becomes significantly negative in schools with more than 693 students, which represent about 30 percent of the schools in our sample.[28]

[28] We experimented with polynomials of school size and found that the relationship between school size and treatment was quite linear, so we keep the more parsimonious model here.

These results are inconsistent with economic theory, suggesting that the relationship between school size and treatment effect may be spurious. Previous theoretical models hypothesize an inverse relationship between school size and the positive effects of a group incentive program like the SPBP. However, there is no practical explanation for why larger groups would be negatively affected by the program. We expect that schools could be large enough to neutralize the positive effects of a group incentive program. As a consequence, teachers' behavior at large schools should simply return to its nontreatment norm (i.e., a treatment effect indistinguishable from zero). We plan to revisit these analyses when data from the 2008–09 school year become available.

8.b. Differences in Target Score to Earn Bonus Awards and Treatment Response

Variation in the performance target that an SPBP-eligible school must meet to receive a bonus award is an interesting design feature of the SPBP. Schools in the treatment condition receive bonus awards if they make significant improvements under the NYCDOE's Progress Report Card system. More specifically, schools with higher overall point totals at the end of 2006–07 school year than the rest of the city's schools and their peer group were required to make fewer point gains in the following year to receive a bonus than schools with lower overall scores (see Table 1). In effect, schools' target gain score affects the strength of the incentives acting upon them.

We take advantage of this variation to evaluate whether the targets set by the SPBP might send a signal to schools about the amount of effort they need to exert to raise their students' test scores. It is possible that schools that needed to make greater gains tried harder than schools with easier targets. However, there is also a chance that schools discouraged by targets that seemed unattainable would end up expending less effort than schools with easier targets.

Although schools were not randomly assigned improvement targets, we take advantage of the nonlinear structure of the performance targets reported in Table 1 to examine whether there is a

differential response attributable to the particular performance targets defined by the SPBP intervention. Discrete performance thresholds facilitate an RD design within the context of the randomized evaluation design. Under certain reasonable assumptions, we can estimate how the perceived rigorousness of performance targets affects student achievement in mathematics.[29]

Our analytic strategy follows the RD framework described in Rouse et al. (2007), and subsequently applied to a number of education-related studies (Winters, Greene, and Trivitt, 2008; Winters, 2008; Rockoff and Turner, 2008). We add a number of independent variables to equation (2), including a cubic function for the number of points earned by a student's school during the 2006–07 school year, dummy variables indicating the performance target category that a school needed to reach to earn a bonus award, and an interaction between school target score category and the SPBP treatment. The regression model can be expressed as:

$$(8)\quad Y_{ist} = \psi_0 + \psi_1 f(Y_{ist-1}) + \psi_2 Student_{is} + \psi_3 School_s + \psi_4 Eligible_{st} + \psi_5 g(Percentile_{st-1}) + \psi_5 Cat_{st-1} + \psi_6 (Cat_{st-1} * Eligible_{st}) + \omega_{ist} + \omega_s$$

where *g(Percentile)* is a cubic function for the percentile of the school's overall points (less the additional bonus points) relative to the rest of the city's schools in its type under the Progress Report Card system in the 2006–07 school year, which was used to put them into categories; *Cat* is a vector of binary variables indicating which of the five target levels of performance the school was required to meet in order to receive a bonus, and all other variables are as previously defined.

The estimated coefficients on the vector of interaction terms, $\hat{\psi}_6$, indicate any differential SPBP treatment effect between an included and an excluded category. In addition, we want to recover the respective estimated treatment impacts on students enrolled in schools in each individual

[29] McEwan and Santibanez (2005) and Santibanez et al. (2008) implemented a similar approach when evaluating Mexico's Carrera Magisterial.

category. Differentiating (8) with respect to *Eligible*, we see that the overall treatment effect on an eligible school in a particular category is the sum of the coefficient for eligibility and the coefficient on the interaction term for the particular category $\psi_4 + \psi_6$. We measure this relationship for each category and test its significance with a X^2 test.

The *Eligible* variable continues to be identified by random assignment. The identifying assumption for estimating the variables in the vector *Cat* is that there is no difference in school performance represented in the target-level category that is not conveyed in a cubic function of the percentile of the number of points that a school earned under the Progress Report Card system. We can then interpret the estimated effect of a school's being in a particular category (and thus facing a particular performance target) as the causal influence of assignment to that categorization (and consequently, the different performance target attached to the categorization) on student achievement. Furthermore, we can interpret the interaction of *Cat* and *Eligible* as a consistent estimate of the differential impact of the SPBP on schools facing varying performance targets.

The basic idea behind this technique is to take advantage of the cutoffs on either side of which schools are assigned scoring targets. In essence, this technique compares the performance of students in schools that just barely fell on either side of the benchmark cutoff. The cutoffs on the point scale that determine in which performance category a school is placed are set at somewhat arbitrary points. They convey little, if any, information about a school's effectiveness that is not already represented in an overall Progress Report Card score (nor do they convey the percentile of a school's overall Progress Report Card score). Though schools with similar point totals may be similar in their effectiveness, the rank that their overall performance score gives them determines the target that the school must meet in order to earn a bonus award under the SPBP.

Although RD designs are a powerful evaluation technique, several limitations are worth mentioning. RD designs focus on a highly localized impact of the SPBP—that is, on schools that are

very close to either side of the cutoffs. These estimates will not necessarily hold globally—that is, for all schools. Furthermore, RD designs require much larger sample sizes to produce impact estimates with sufficient statistical power (Cappelleri et al., 1994; Schochet, 2008a, 2008b; Bloom et al., 2005).

It is also worth emphasizing here that while all New York City schools are given performance targets, and thus could be affected by them, this analysis is not particularly concerned with the overall impact of the targets themselves on schools in our sample. We use the estimate on the interaction term to focus on the differential response of schools to performance target thresholds, not to recover the overall impact of the performance target scores. Even though it is plausible that both treatment and control schools would be affected by how they were categorized, only SPBP-eligible schools had the additional incentive of a performance bonus award if they met their performance target, which is what we focus on here.

The results from estimating various forms of (8) in mathematics are displayed in Table 11. None of the models finds that any kind of SPBP treatment makes a significant difference, regardless of a school's Progress Report Card target score.

[INSERT TABLE 11 ABOUT HERE]

8.c. Pay for Performance and School Learning Environment

The results above indicate that on average, the SPBP treatment had no effect on student achievement in mathematics. However, student attendance and student, parent, and teacher perceptions of the school learning environment account for 15 percent of a school's overall Progress Report Card score. Thus, schools may have sought to increase scores in ways unrelated to advancing student achievement.

We measure directly whether SPBP-eligible schools made larger improvements on the Progress Report Card system overall score, as well as on individual components of a school's overall

score. We use publicly available data at the school level to estimate regressions that explain the number of points a school earned on its 2007–08 Progress Report Card as a function of the SPBP treatment and observed school characteristics (including the number of points earned in 2006–07 school year). We estimate equations taking the form:

$$(9)\quad Point s_{st} = \theta_0 + \theta_1 Point s_{st-1} + \theta_2 School_s + \theta_3 Treat_{st} + \nu_{st} + \nu_s$$

where, depending on the specification, Points is a school's overall Progress Report Card score or its score is a component of the Progress Report Card system, including environment scores, progress score, or extra credit earned (bonus points). All other variables are as previously defined.

Table 12 displays results comparing SPBP-eligible and comparison-group schools on the basis of individual components that make up a school's overall Progress Report Card score. We find no relationship between SPBP eligibility and any component score of the school grading system. Nonetheless, the regression model evaluating a school's score on the performance score (Model 3 of Table 12) has particular interest. A school's performance score is determined by the percentage of its students meeting particular proficiency benchmarks on the New York State high-stakes mathematics and ELA tests. It might be thought that SPBP-eligible schools would respond to the importance of this component by focusing their efforts on students falling just short of the proficiency benchmarks.[30] However, the lack of statistical differences in the performance scores of eligible schools and those assigned to the SPBP intervention suggests that the latter have not responded in this way.

[INSERT TABLE 12 ABOUT HERE]

We also compare scores at the school level from the student, teacher, and parent school learning-environment surveys. Recall that a school's learning-environment survey score accounts for

[30] For studies on educational triage in response to high-stakes accountability systems, see Booher-Jennings (2005), Neal and Schanzenbach (forthcoming), Reback (forthcoming), Ballou and Springer (2008), and Springer (2007).

15 percent of its overall Progress Report Card score. Further, if schools have responded to the SPBP eligibility, it is possible for the school learning-environment survey to reflect some of these short-run outcomes. We also evaluated individual components of the student, teacher, and parent survey results, all of which are reported in Table 13. Once again, we find no difference in any of the components of the teacher, student, or parent surveys among SPBP-eligible and control-group schools.

[INSERT TABLE 13 ABOUT HERE]

9. Summary and Conclusion

In this paper, we present evidence on the impact of NYCDOE' SPBP during the program's first year of implementation. Because the number of schools meeting eligibility criteria under the SPBP guidelines required more than the amount of money budgeted for the program, NYCDOE' Research and Policy Support Group assigned schools to the SPBP intervention by random lottery. Our evaluation design takes advantage of the fact that schools were randomly lotteried into the SPBP intervention.

Our findings suggest that the SPBP has had negligible short-run effects on student achievement in mathematics. The same holds true for intermediate outcomes such as student, parent, and teacher perceptions of the school learning environment. We also find no evidence that the treatment effect differed on the basis of student or school characteristic. An exception is the differential effect of SPBP eligibility by school size, which suggests student performance in larger schools decreases when SPBP was implemented. The potential moderating effect of school size on the direction and/or strength of the relationship between the SPBP and mathematics achievement will be revisited when data from the 2008-09 school year become available.

Although a well-implemented experimental evaluation design would suggest that our estimates have strong internal validity, readers should interpret these initial findings with caution when considering the possible impact of this or any other program. First, the estimates presented here are of the short-run effects of the SPBP, which may limit our ability to identify any aspect or degree of the program's effectiveness. Schools learned that they were eligible for the program less than three months before New York State's high-stakes mathematics tests were administered. An evaluation of the SPBP's impact following the 2008-09 school year should provide much more reliable information.

Furthermore, readers should not lose sight of the fact that additional experimental and quasi-experimental evaluations of various forms of teacher compensation reform are needed. Pay-for-performance programs can exhibit various design components, including the unit of accountability, performance measurement, incentive structure, and bonus distribution. The education policy community needs to study a greater number of forward-thinking schools systems such as NYCDOE before it can construct a knowledge base sufficiently large to permit the making of sound policy decisions on the question of whether teacher pay-for-performance is a useful strategy for enhancing teacher effectiveness and school quality.

References

Angrist, J., G. Imbens, and D. Rubin. (1996). "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, 91 (434), 444–55.

Baker, G. (1992). "Incentive Contracts and Performance Measurement." *Journal of Political Economy*, 100 (3), 598–614.

Ballou, D. (1996). "Do Public Schools Hire the Best Applicants?." *Quarterly Journal of Economics*, 111 (1), 97–133.

———, and M.G Springer. (2008). "Achievement Trade-Offs and No Child Left Behind." Mimeograph. Nashville, TN: Vanderbilt University.

Besley, T., and S. Coate. (1995). "The Design of Income Maintenance Programs." *Review of Economic Studies*, 62 (1), 187–221.

Bloom, H. (1984). "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review*, 8 (2), 225–46.

———. (2006). "The Core Analytics of Randomized Experiments for Social Research." Working Paper. New York: Manpower Demonstration Research Corporation.

———, L. Richburg-Hayes, and A. Black. (2005). "Using Covariates to Improve Precision: Empirical Guidance for Studies That Randomize Schools to Measure the Impacts of Educational Interventions." Working Paper. New York: Manpower Demonstration Research Corporation.

Booher-Jennings, J. (2005). "Below the Bubble: 'Educational Triage' and the Texas Accountability System." *American Educational Research Journal*, 42 (2), 231–68.

Bowles, S, and H. Gintis. (2002). "Social Capital and Community Governance." *Economic Journal*, 112 (November), F419-F436.

Cameron, A., and P. Trivedi. (2006). *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Capelleri, J., R. Darlington, and W. Trochim. (1994). "Power Analysis of Cutoff-Based Randomized Clinical Trials." *Evaluation Review*, 18 (2), 141–52.

Clotfelter, C., E. Glennie, H. Ladd, and J. Vigdor. (2008). "Would Higher Salaries Keep Teachers in High-Poverty Schools? Evidence from a Policy Intervention in North Carolina." Journal of Public Economics, 92 (5–6), 1352–70.

Clotfelter, C., H. Ladd, and J. Vigdor. (2007). "Are Teacher Absences Worth Worrying About in the U.S.?." Working Paper 13648. Cambridge, Mass.: National Bureau of Economic Research.

———. (2005). "Who Teaches Whom? Race and the Distribution of Novice Teachers." *Economics of Education Review*, 24 (4), 377–92.

Courty, P., and G. Marschke. (2004). "An Empirical Investigation of Gaming Responses to Explicit Performance Incentives." *Journal of Labor Economics*, 22 (1), 23–56.

Cullen, J., and R. Reback. (2006). "Tinkering toward Accolades: School Gaming under a Performance Accountability System." Working Paper 12286. Cambridge, Mass.: National Bureau for Economic Research.

Deci, E. (1975). *Intrinsic Motivation*. New York: Plenum.

Deutsch, M. (1985). *Distributive Justice: A Social-Psychological Perspective.* New Haven, Conn.: Yale University Press.

Eberts, R. W. (2007). "Teacher Unions and Student Performance: Help or Hindrance?." *The Future of Children*, 17 (1), 175–200.

Ehrenberg, R., and G. Milkovich. (1987). "Compensation and Firm Performance." In *Human Resources and the Performance of Firms*, ed. M. Kleiner. Madison, Wisc.: Industrial Relations Research Association.

Ehrenberg, R., and R. Smith. (1994). Modern Labor Economics: Theory and Public Policy, 5th ed. New York: Harper Collins College Publishers.

Ehrenberg, R., D. Rees, and E. Ehrenberg. (1991). "School District Leave Policies, Teacher Absenteeism, and Student Achievement." *Journal of Human Resources*, 26 (1), 72–105.

Figlio, D., and L. Getzler. (2002). "Accountability, Ability and Disability: Gaming the System?." Working Paper 9307. Cambridge, Mass.: National Bureau for Economic Research.

Figlio, D., and J. Winicki. (2002). "Food for Thought? The Effects of School Accountability Plans on School Nutrition." Working Paper 9319. Cambridge, Mass.: National Bureau for Economic Research.

Frey, B. (1997). "A Constitution for Knaves Crowds Out Civic Virtues." *Economic Journal*, 107 (443), 1043–53.

Glazerman, S., et al. (2007). *Options for Studying Teacher Pay Reform Using Natural Experiments.* Washington, D.C.: Mathematica Policy Research.

Glewwe, P., N. Ilias, and M. Kremer. (2008). "Teacher Incentives." Mimeograph, Cambridge, Mass: Harvard University.

———. (2008). *Teacher Incentives in the Developing World.* Mimeograph. Cambridge, Mass.: Harvard University.

Goldhaber, D. (Forthcoming). "The Politics of Teacher Pay Reform." In *Performance Incentives: Their Growing Impact on American K–12 Education*, ed. M. G. Springer. Washington, D.C.: Brookings Institution Press.

Green, J., and N. Stokey. (1983). "A Comparison of Tournaments and Contracts." *Journal of Political Economy*, 91 (3), 349–64.

Hahn, J., P. Todd, and W. van der Klaauw. (2001). "Identification and Estimation of Treatment Effects with a Regression Discontinuity Design." *Econometrica*, 69, 201–9.

Hamilton, B. H., J. A. Nickerson, and H. Owan. (2003). "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation." *Journal of Political Economy*, 111 (3), 465–97.

Hannaway, J., and J. Rotherham. (2008). *Collective Bargaining in Education and Pay for Performance.* Nashville: National Center on Performance Incentives.

Hanushek, E. (2003). "The Failure of Input-Based Resource Policies." *Economic Journal*, 113, F64–68.

Heinrich, C. (2004). "Outcomes-Based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness." *Public Administration Review*, 62 (6), 712–25.

Hernandez, J.C. (2009). "New Education Secretary Visits Brooklyn School." Retrieved from *New York Times* at http://cityroom.blogs.nytimes.com/2009/02/19/new-education-secretary-visits-brooklyn-school/.

Holmstrom, B., and P. Milgrom. (1991). "Multitask Principal-Agent Analysis: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization*, 7, 24–52.

Hotelling, H. (1940). "The Selection of Variates for Use in Prediction with Some Comments on the General Problem of Nuisance Parameters." *Annals of Mathematical Statistics*, 11, 271–83.

Jacob, B., and S. Levitt. (2003). "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics*, 118, 843–77.

Johnson, S. (1986). "Incentives for Teachers: What Motivates, What Matters." *Education Administration Quarterly*, 22 (3), 54–79.

Kandel, E., and E. P. Lazear. (1992). "Peer Pressure and Partnerships." *Journal of Political Economy*, 100 (4), 801–17.

Kelley, C., and K. Finnigan. (2004). "Teacher Compensation and Teacher Workforce Development." Yearbook of the National Society for the Study of Education, 103, 253–73.

Koppich, J. (2008). "Toward a More Comprehensive Model of Teacher Pay." Nashville: National Center on Performance Incentives.

Kremer, M., et al. (2004). "Teacher Absence in India." Working Paper. Washington, D.C.: World Bank.

Kruskal, W., and W. Wallis. (1952). "Use of Ranks in One-Criterion Variance Analysis." *Journal of the American Statistical Association*, 47 (260), 583–621.

Ladd, H. F., ed. (1996). Holding Schools Accountable: *Performance-Based Reform in Education.* Washington, D.C.: Brookings Institution.

Ladd, H. (2001). "School-Based Education Accountability Systems: The Promise and the Pitfalls." *National Tax Journal*, 54 (2), 385–400.

Lavy, V. (2002). "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement." *Journal of Political Economy*, 110 (6), 1286–1317.

———. (Forthcoming). Performance Pay and Teachers' Effort, Productivity and Grading Ethics. *American Economic Review.*

Lawler, E. (1981). Pay and Organization Development. Reading, Mass.: Addison-Wesley.

Lazear, E. P. (2000). "Performance Pay and Productivity." American Economic Review, 90, 1346–61.

———. (1998). *Personnel Economics for Managers.* New York: Wiley.

———, and S. Rosen. (1981). "Rank-Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy*, 89 (5), 841–64.

Lee, D., and T. Lemieux. (2009). "Regression Discontinuity Designs in Economics." Working Paper 14723. Cambridge, Mass.: National Bureau of Economic Research.

Lepper, M., and D. Greene. (1978). *The Hidden Costs of Reward.* Hillsdale, N.J.: Erlbaum.

Lortie, D. (1975). *Schoolteacher.* Chicago: University of Chicago Press.

MacKinnon, J. (2002). "Bootstrap Inference in Econometrics." *Canadian Journal of Economics*, 35, 615–35.

McEwan, P., and L. Santibañez. (2005). "Teacher and Principal Incentives in Mexico." In *Incentives to Improve Teaching: Lessons from Latin America*, ed. E. Vegas, 213–53. Washington, D.C.: World Bank Press.

Milgrom, P., and J. Roberts. (1992). *Economics, Organization, and Management. Englewood Cliffs*, N.J.: Prentice-Hall.

———. (1990). "Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities." *Econometrica*, 58 (6), 1255–77.

Miller, R., J. Murnane, and J. Willett. (2007). "Do Teacher Absences Impact Student Achievement? Longitudinal Evidence from One Urban School District." Working Paper 13356. Cambridge, Mass.: National Bureau of Economic Research.

Mizala, A., and P. Romaguera. (2003). "Scholastic Performance and Performance Awards." Working Paper. Universidad de Chile. Centro de Economía Aplicada.

Muralidharan, K., and V. Sundararaman. (2008). "Teacher Incentives in Developing Countries: Experimental Evidence from India." Working Paper. Nashville: National Center on Performance Incentives.

Murnane, R. J., and D. K. Cohen. (1986). "Merit Pay and the Evaluation Problem: Why Most Merit Plans Fail and a Few Survive. Harvard Educational Review, 56 (1), 1-17.

Nalbantian, H., and A. Schotter. (1997). "Productivity under Group Incentives: An Experimental Study." American Economic Review, 87 (3), 314–41.

Neal, D. (Forthcoming). "Designing Incentive Systems for Schools." In *Performance Incentives: Their Growing Impact on American K–12 Education*, ed. M. G. Springer. Washington, D.C.: Brookings Institution Press.

———, and D. Schanzenbach. (forthcoming). "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics.*

New York City Public Schools. (2007). "Educator Guide: New York City Progress Report. Elementary/Middle School." Retrieved from http://schools.nyc.gov/NR/rdonlyres/DEFA8A3D-7BB8-4502-BEFC-F977FB206542/43571/ProgressReportEducatorGuide_EMS_091608.pdf.

———. (2008, July 1). "Parent, Teacher, Student Learning Environment Surveys: 2008 Citywide Results." Retrieved from http://schools.nyc.gov/NR/rdonlyres/4C0235D3-AE5A-4E9B-98F4-8B4F5697497F/40759/lesresults.pdf.

Odden, A. (2001). "Rewarding Expertise." *Education Matters*, 1 (1), 16–25.

Ortiz-Jiménez, M. (2003). "Carrera Magisterial: Un proyecto de desarrollo profesional." Cuadernos de Discusión 12. Mexico City: Secretaría de Educación Pública.

Pfeffer, J. (1995). *Competitive Advantage through People: Unleashing the Power of the Workforce.* Boston: Harvard Business School Press.

———, and N. Langston. (1993). "The Effect of Wage Dispersion on Satisfaction, Productivity, and Working Collaboratively: Evidence from College and University Faculty." *Administrative Science Quarterly*, 38, 382–407.

Podgursky, M. (2003). "Fringe Benefits." *Education Next*, 3 (3), 71–76.

———, and M. Springer. (2007). "Teacher Performance Pay: A Review." *Journal of Policy Analysis and Management*, 26 (4), 909–49.

Prendergast, C. (1999). "The Provision of Incentives in Firms." *Journal of Economic Literature*, 37, 7–63.

Reback, R. (Forthcoming). "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." *Journal of Public Economics*, 92 (5–6), 1394–1415.

Rockoff, J., and L. Turner. (2008). "Short-Run Impacts of Accountability on School Quality." Working Paper 14564. Cambridge, Mass.: National Bureau of Economic Research.

Rosen, S. (1986). "The Theory of Equalizing Differences." In *Handbook of Labor Economics*, ed. O. Ashenfelter and R. Layard, vol. 1. Amsterdam: North-Holland.

Rouse, C., et al. (2007). "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure." CALDER Working Paper: Urban Institute.

Sager, R. (2009). "Prez's Challenge to NYC Teachers." Retrieved March 31, 2009, from *New York Post* (March 2).

Santibañez, L., et al. (2007). "Breaking Ground: Analysis of the Assessment System and Impact of Mexico's Teacher Incentive Program 'Carrera Magisterial.' " RAND Corporation.

Schochet, P. (2008a). "Statistical Power for Random Assignment Evaluations of Education Programs." *Journal of Educational and Behavioral Statistics*, 33 (1), 62–87.

———. (2008b). "The Late Pretest Problem in Randomized Control Trials of Education Interventions." Jessup, Md.: National Center for Education Evaluation and Regional Assistance.

Springer, M. (2007). "Accountability Incentives: Do Failing Schools Practice Educational Triage?." *Education Next*, 8 (1), 74–79.

———, D. Ballou, and A. Peng. (2008). "Impact of the Teacher Advancement Program on Student Test Score Gains: Findings from an Independent Appraisal." Working Paper. Nashville: National Center on Performance Incentives.

Springer, M. et al. (2009). "Governor's Educator Excellence Grant (GEEG) Program: Year Two Evaluation Report." Nashville: National Center on Performance Incentives.

———. (2008). "Texas Educator Excellence Grant (TEEG) Program: Year Two Evaluation Report." Nashville: National Center on Performance Incentives.

Taylor, L., and M. Springer. (2009). "Optimal Incentives for Public Sector Workers: The Case of Teacher-Designed Incentive Pay in Texas." Working Paper. Nashville: National Center on Performance Incentives.

———, and M. Ehlert. (Forthcoming). "Characteristics and Determinants of Teacher-Designed Pay for Performance Plans: Evidence from Texas' Governor's Educator Excellence Grant Program." In *Performance Incentives: Their Growing Impact on American K–12 Education*, ed. M. G. Springer. Washington, D.C.: Brookings Institution Press.

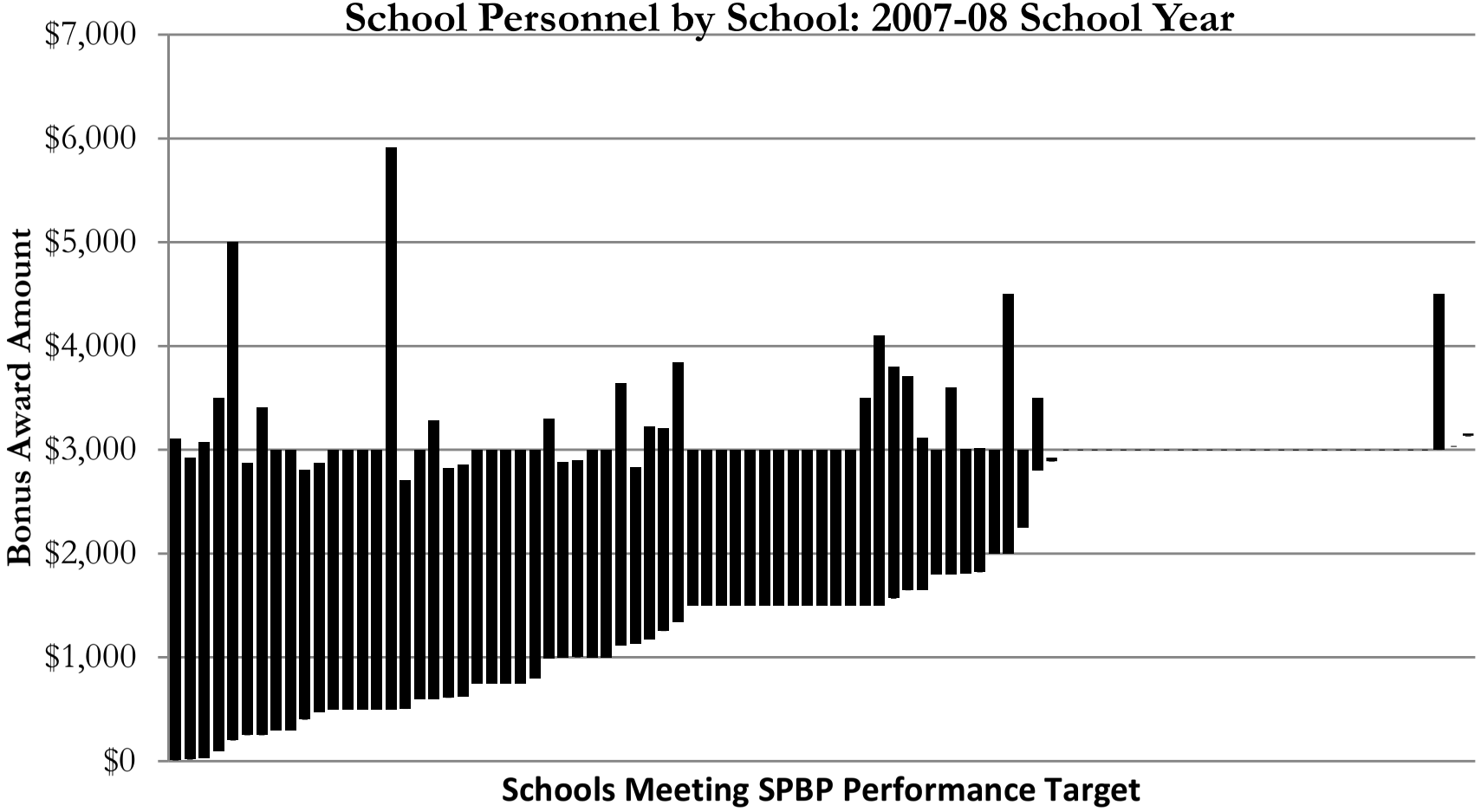
Thistlewaite, D., and D. Campbell. (1960). "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment." *Journal of Education Psychology*, 51, 309–17.

Winters, M. (2008). "Grading New York: An Evaluation of New York City's Progress Report Program." Manhattan Institute.

———, J. Greene, and J. Trivitt. (2008). "Building on the Basics: The Impact of High-Stakes Testing on Student Proficiency in Low-Stakes Subjects." Manuscript. Manhattan Institute.

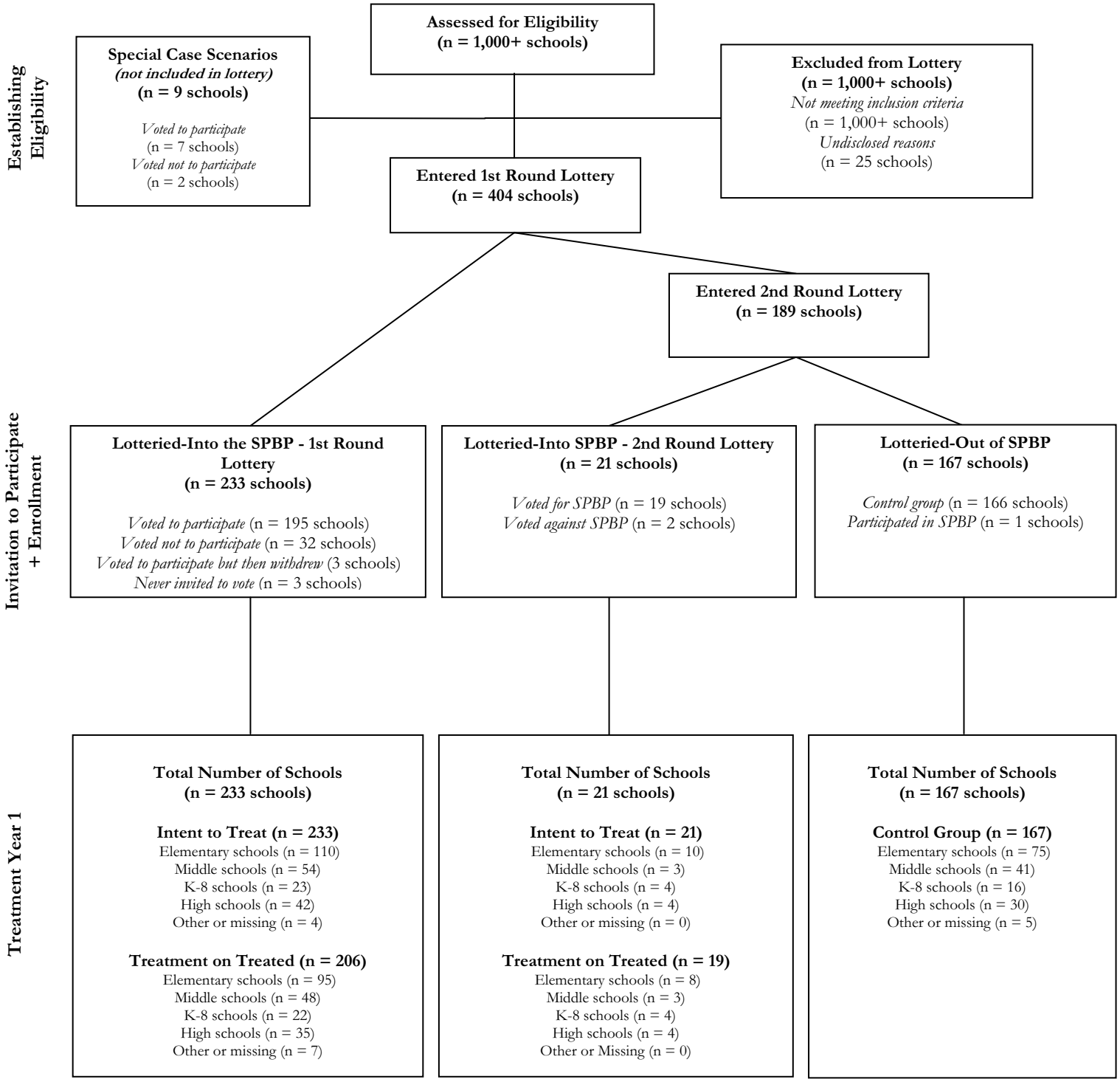
Zenger, T. (1992). "Why Do Employers Only Reward Extreme Performance? Examining the Relationships among Performance, Pay, and Turnover." *Administrative Science Quarterly*, 37 (2), 198–219.

Figure 1. Minimum and Maximum Bonus Amounts Awarded to School Personnel by School: 2007-08 School Year



Note: Authors' own calculations using Progress report Card system data obtained from NYCDOE.

Figure 2. Consort Diagram for New York City's School-Wide Performance Bonus Program*



** Seventeen schools associated with New York City's School-Wide Performance Bonus Program received a Progress Report Card score for their middle school grades and a second Progress Report Card score for their high school grades even though these students are enrolled in the same school. Of these seventeen schools, eight were lotteried-into the SPBP (6 participated, 2 did not), seven were lotteried-out of the SPBP, and 2 were excluded from the lottery for undisclosed reasons.*

Table 1. Performance Rankings and Target Gains for New York City's Progress Report Card System

	Performance Score Rank, 2006-07 school year	Target Gain, 2007-08 school year	Schools, 2006-07 school year	
			#	%
Category 1	>= 85th	7.5	41	12.69%
Category 2	< 85th and >= 45th	12.5	114	35.29%
Category 3	< 45th and >= 15th	15	99	30.65%
Category 4	< 15th and >= 5th	17.5	45	13.93%
Category 5	< 5th	20	24	7.43%
Total			323 schools	

Note: Authors' own calculations using Progress Report Card System data obtained from NYCDOE website.

Table 2. Schools Participating in School-Wide Performance Bonus Program by Percentage of Performance Target Met: 2007-08 School Year

	Performance Target						Total	
	Met 100%		Met 75%*		None			
Category (Target Gain)	*# schools*	*% schools*	*# schools*	*% schools*	*# schools*	*% schools*	*# schools*	*% schools*
Category 1 (7.5 points)	3	15.00%	10	50.00%	7	35.00%	20	12.12%
Category 2 (12.5 points)	20	31.75%	10	15.87%	33	52.38%	63	38.18%
Category 3 (15 points)	17	34.69%	5	10.20%	27	55.10%	49	29.70%
Category 4 (17.5 points)	11	68.75%	2	12.50%	3	18.75%	16	9.70%
Category 5 (20 points)	14	82.35%	1	5.88%	2	11.76%	17	10.30%
	65 schools (39.39%)		28 schools (16.97%)		72 schools (42.77%)		165 schools (100%)	

Notes: Authors' own calculations using Progress Report Card System data obtained from NYCDOE website. * Ten schools met target under A-A format and received $1,500 per UFT member in their school (nine schools in Category 1 and one school in Category 2).

Table 3. Descriptive Statistics for Bonus Award Distribution by School

	Mean	Minimum	Maximum	Std. Dev.	N
Bonus Amount	$2,417.19	$647.47	$3,750.00	$611.60	92
Minimum Bonus Amount	$1,700.75	$7.00	$3,134.00	$1,050.76	92
Maximum Bonus Amount	$3,156.41	$2,708.00	$5,914.00	$469.56	92
Range Bonus Amount	$1,462.74	$0.00	$5,414.00	$1,195.34	92
Number of Unique Bonus Award Amounts	3.51	1	19	3.18	92

Notes: Authors' own calculations using bonus award distribution data obtained from NYCDOE Office of Accountability.

Table 4. Descriptive Statistics for Treatment Schools, Control Group Schools, and All New York City Schools

	2006-07 School Year									
	Treatment Schools		*Control Schools*			*Participant Schools (among eligible)*		*Declining Schools (among eligible)*		*All NYC Schools*	
	(1)		(2)			(3)		(4)		(3)	
Type of School	Mean	Std. Dev.	Mean	Std. Dev.		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Elementary Schools	0.60	0.49	0.58	0.49		0.58	0.49	0.69	0.47	0.58	0.49
Middle Schools	0.29	0.46	0.31	0.46		0.30	0.46	0.23	0.43	0.29	0.46
K-8 Schools	0.11	0.32	0.11	0.31		0.12	0.32	0.08	0.27	0.12	0.33
Enrollment	592.89	257.28	595.02	262.53		592.50	255.96	595.31	270.45	671.32	329.06
NYC Progess Report											
Overall Score	52.03	16.18	51.63	14.73		51.41	16.19	55.88	15.90	53.88	14.19
Environment Score	0.46	0.20	0.47	0.18		0.46	0.20	0.48	0.21	7.63	2.80
Performance Score	0.49	0.17	0.48	0.18		0.48	0.17	0.53	0.16	13.64	4.51
Progress Score	0.51	0.21	0.50	0.19		0.50	0.21	0.55	0.21	30.10	11.24
Extra-Credit Score	2.58	2.48	2.46	2.50		2.57	2.52	2.63	2.28	2.29	2.29
Peer Index	55.89	34.09	54.46	34.57		55.23	34.48	59.97	31.96	42.58	30.64
Progress Report Grade											
A	0.23	0.42	0.20	0.40		0.21	0.41	0.35	0.49	0.23	0.42
B	0.32	0.47	0.32	0.47		0.33	0.47	0.27	0.45	0.38	0.48
C	0.27	0.44	0.28	0.45		0.28	0.45	0.23	0.43	0.26	0.44
D	0.10	0.30	0.16	0.37	*	0.09	0.28	0.15	0.37	0.09	0.28
F	0.09	0.28	0.04	0.21		0.10	0.30	0.00	0.00	0.04	0.20
NCLB Accountability Status											
Good Standing	0.53	0.50	0.53	0.50		0.51	0.50	0.65	0.49	0.69	0.46
In Need of Improvement	0.13	0.34	0.23	0.42	**	0.15	0.36	0.00	0.00	** 0.09	0.28
Corrective Action	0.06	0.24	0.04	0.19		0.07	0.25	0.00	0.00	0.02	0.15
Restructuring	0.27	0.45	0.20	0.40		0.26	0.44	0.35	0.49	0.13	0.34
Student Demographics											
% Asian/Pacific Islander	0.02	0.02	0.02	0.03		0.01	0.02	0.03	0.04	0.11	0.17
% Black	0.41	0.28	0.43	0.28		0.41	0.28	0.40	0.29	0.34	0.30
% Hispanic	0.56	0.28	0.54	0.28		0.56	0.28	0.56	0.29	0.40	0.27
% Native American	0.00	0.01	0.00	0.01		0.00	0.01	0.00	0.00	0.00	0.01
% White	0.01	0.02	0.01	0.02		0.01	0.02	0.01	0.02	0.13	0.21
% English Language Learners	0.19	0.13	0.19	0.13		0.19	0.13	0.18	0.12	0.15	0.14

% Special Education	0.22	0.10	0.22	0.10	0.22	0.11	0.20	0.06		0.15	0.08
Student Test Scores											
ELA Scale Score (normalized)	-0.36	0.24	-0.38	0.22	-0.37	0.24	-0.27	0.23	**	-0.02	0.47
Math Score Score (normalized)	-0.37	0.27	-0.40	0.25	-0.38	0.27	-0.29	0.27		-0.03	0.47
Number of Schools	186		137		160		26			1002	
Hotelling T-Test (p-value)	0.7246				0.3359						

Note: Test for signifcant differneces with a Kruskal-Wallis equality-of-populations rank test. * significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Table 5. Descriptive Statistics for New York City's School Environment Survey by Treatment Schools and Control Schools

	2006-07 School Year									
	Treatment Schools		*Control Schools*		*Participant Schools (among eligible)*		*Declining Schools (among eligible)*		*All Schools*	
	(1)		(2)		(3)		(4)		(3)	
Student Survey	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Response Rate	0.70	0.22	0.68	0.24	0.70	0.22	0.67	0.25	0.68	0.23
Academic Score	6.96	0.42	6.97	0.37	6.96	0.44	6.95	0.19	7.03	0.48
Engagement Score	6.51	0.52	6.57	0.47	6.53	0.53	6.34	0.43	6.51	0.61
Communication Score	5.77	0.44	5.75	0.44	5.78	0.46	5.61	0.17	5.69	0.61
Safety Score	5.72	0.61	5.76	0.62	5.76	0.61	5.30	0.43	6.11	0.82
Number of Schools	85		60		75		9		832	
Parent Survey										
Response Rate	0.23	0.08	0.23	0.09	0.23	0.08	0.22	0.08	0.27	0.14
Academic Score	7.28	0.57	7.25	0.60	7.28	0.56	7.24	0.62	7.31	0.64
Enagagement Score	6.53	0.54	6.49	0.51	6.53	0.53	6.49	0.62	6.26	0.65
Communication Score	7.15	0.56	7.15	0.53	7.15	0.56	7.20	0.55	7.16	0.66
Safety Score	7.56	0.64	7.51	0.62	7.54	0.64	7.66	0.66	7.70	0.74
Number of Schools	186		130		160		27		1002	
Teacher Survey										
Response Rate	0.42	0.17	0.43	0.19	0.42	0.17	0.39	0.18	0.46	0.19
Academic Score	6.38	1.07	6.57	1.00	6.36	1.07	6.50	1.06	6.84	1.09
Enagagement Score	5.37	1.19	5.53	1.09	5.37	1.20	5.38	1.16	5.81	1.23
Communication Score	5.75	1.00	5.93	0.88	5.77	0.99	5.65	1.07	6.00	1.03
Safety Score	6.03	1.08	6.17	0.91	6.00	1.04	6.20	1.28	6.66	1.14
Number of Schools	186		132		160		26		1002	

Note: Test for significant differences with a Kruskal-Wallis equality-of-populations rank test. * significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Table 6. Impact of New York City's School-Wide Performance Bonus Program on Mathematics Scores

	Panel A: Intention-to-Treat				Panel B: Impact-on-the-Treated			
(model)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.0261	-0.0288	-0.0279	-0.0263	-0.0463	-0.0391	-0.0324	-0.0322
	[0.0275]	[0.0252]	[0.0251]	[0.0166]	[0.0327]	[0.0305]	[0.0306]	[0.0203]
Sample								
Number of Students	110502	110502	110502	83966	110502	110502	110502	83966
Number of Schools	323	323	323	323	323	323	323	323
R-squared	0.002	0.162	0.178	0.563	0.002	0.162	0.178	0.563
Student-Level Covariates		√	√	√		√	√	√
School-Level Covariates			√	√			√	√
Prior Scale Score (quartic)				√				√

Notes: Student-level controls include race/ethnicity, English language learner status, and whether the student has an IEP. School-level covariates include school peer index, percent of students with certain race/ethnic status, percent English language learner, percent IEP, borough and total school enrollment. Standard errors in brackets adjust for clustering at the school level using bootstrap techniques with 300 iterations. * significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Table 7. Impact of New York City's Schoolwide Performance Bonus Program on Mathematics by Race/Ethnicity

	Panel A: Intention-to-Treat	Panel B: Impact-on-the-Treated
(model)	(1)	(2)
Treatment	-0.0101 [0.0505]	-0.0424 [0.0588]
Treatment * American Indian	-0.05 [0.0783]	-0.0269 [0.0964]
Treatment * Asian	-0.012 [0.0531]	0.00342 [0.0578]
Treatment * Black	-0.021 [0.0531]	0.000956 [0.0614]
Treatment * Hispanic	-0.0128 [0.0497]	0.018 [0.0585]
Treatment * White	… …	… …
Chi-Squared Test (p-value)		
*Eligible + (Eligible * American Indian) = 0*	0.3779	0.4048
*Eligible + (Eligible * Asian) = 0*	0.5824	0.4079
*Eligible + (Eligible * Black) = 0*	0.135	0.1058
*Eligible + (Eligible * Hispanic) = 0*	0.2158	0.2803
Sample		
Number of Students	83966	83966
Number of Schools	323	323
R-squared	0.563	0.563
Student-Level Covariates	√	√
School-Level Covariates	√	√
Prior Scale Score (quartic)	√	√

Notes: Student-level controls include race/ethnicity, English language learner status, and whether the student has an IEP. School-level covariates include school peer index, percent of students with certain race/ethnic status, percent English language learner, percent IEP, borough and total school enrollment. Standard errors in brackets adjust for clustering at the school level using bootstrap techniques with 300 iterations. * significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Table 8. Distributional Impact of New York City's Schoolwide Performance Bonus Program on Mathematics Test Scores

	Panel A: Intention-to-Treat	Panel B: Impact-on-the-Treated
(model)	(1)	(2)
Treatment	-0.0328** [0.0166]	-0.0373* [0.0202]
Treatment * Prior Scale Score in Lower 25th Percentile	-0.00803 [0.0183]	-0.0137 [0.0218]
Treatment * Prior Scale Score in 25th - 50th Percentiles	0.014 [0.00950]	0.0152 [0.0113]
Treatment * Prior Scale Score in 51st - 75th Percentiles	… …	… …
Treatment * Prior Scale Score in Top 25th Percentile	0.0195 [0.0180]	0.0188 [0.0219]
Chi-Squared Test (p-value)		
*Eligible + (Eligible * Lower 25th) = 0*	0.0601	0.0513
*Eligible + (Eligible * 25th - 50th) = 0*	0.2962	0.3103
*Eligible + (Eligible * Top 25th) = 0*	0.5716	0.5172
Sample		
Number of Students	83966	83966
Number of Schools	323	323
R-squared	0.564	0.564
Student-Level Covariates	√	√
School-Level Covariates	√	√
Prior Scale Score (quartic)	√	√

Notes: Student-level controls include race/ethnicity, English language learner status, and whether the student has an IEP. School-level covariates include school peer index, percent of students with certain race/ethnic status, percent English language learner, percent IEP, borough and total school enrollment. Standard errors in brackets adjust for clustering at the school level using bootstrap techniques with 300 iterations. * significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Table 9. Impact of New York City's Schoolwide Performance Bonus Program on Mathematics and English Language Arts Test Scores by Type of School

	Panel A: Intention-to-Treat	Panel B: Impact-on-the-Treated
(model)	(1)	(2)
Treatment	-0.0237 [0.0554]	-0.0331 [0.0665]
Treatment * Elementary School	0.0145 [0.0611]	0.0223 [0.0733]
Treatment * Middle School	-0.0223 [0.0656]	-0.0237 [0.0801]
Treatment * K-8 School	… …	… …
Chi-Squared Test (p-value)		
*Eligible + (Eligible * Elementary) = 0*	0.6673	0.6796
*Eligible + (Eligible * Middle = 0*	0.1216	0.1337
Sample		
Number of Students	83966	83966
Number of Schools	323	323
R-squared	0.563	0.563
Student-Level Covariates	√	√
School-Level Covariates	√	√
Prior Scale Score (quartic)	√	√

Notes: Student-level controls include race/ethnicity, English language learner status, and whether the student has an IEP. School-level covariates include school peer index, percent of students with certain race/ethnic status, percent English language learner, percent IEP, borough and total school enrollment. Standard errors in brackets adjust for clustering at the school level using bootstrap techniques with 300 iterations. * significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Table 10. Impact of New York City's Schoolwide Performance Bonus Program on Mathematics Test Scores by School Size

	Panel A: Intention-to-Treat		Panel B: Impact-on-the-Treated	
(model)	(1)	(2)	(3)	(4)
Treatment	0.0873* [0.0490]	-0.0327 [0.0303]	0.0941 [0.0595]	-0.0363 [0.0372]
Treatment * School Size	-0.000165** [7.07e-05]		-0.000185** [8.73e-05]	
Treatment * School Size in Lower 25th Percentile		0.0383 [0.0488]		0.0304 [0.0612]
Treatment * School Size in 25th - 50th Percentiles		0.0770* [0.0448]		0.0844 [0.0557]
Treatment * School Size in 51st - 75th Percentiles		… …		… …
Treatment * School Size in Top 25th Percentile		-0.0409 [0.0475]		-0.0514 [0.0589]
Chi-Squared Test (p-value)				
*Eligible + (Eligible * Lower 25th) = 0*		0.8828		0.9008
*Eligible + (Eligible * 25th - 50th) = 0*		0.2297		0.2852
*Eligible + (Eligible * Top 25th) = 0*		0.0297		0.0404
Sample				
Number of Students	84821	84821	84821	84821
Number of Schools	323	323	323	323
R-squared	0.56	0.56	0.56	0.56
Student-Level Covariates	√	√	√	√
School-Level Covariates	√	√	√	√
Prior Scale Score (quartic)	√	√	√	√

Notes: Student-level controls include race/ethnicity, English language learner status, and whether the student has an IEP. School-level covariates include school peer index, percent of students with certain race/ethnic status, percent English language learner, percent IEP, borough and total school enrollment. Standard errors in brackets adjust for clustering at the school level using bootstrap techniques with 300 iterations. * significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Table 11. Differential Impact of the Schoolwide Performance Bonus Program's Performance Targets on Mathematics Test Scores

	Panel A: Intention-to-Treat			Panel B: Impact-on-the-Treated		
(model)	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.0905 [0.0708]	-0.0691 [0.0614]	-0.0656 [0.0480]	-0.108 [0.0839]	-0.0859 [0.0790]	-0.082 [0.0597]
Treatment * Performance Category 1	… …	… …	… …	… …	… …	… …
Treatment * Performance Category 2	0.102 [0.0840]	0.073 [0.0710]	0.0659 [0.0545]	0.121 [0.101]	0.0915 [0.0915]	0.0829 [0.0682]
Treatment * Performance Category 3	0.0913 [0.0810]	0.065 [0.0721]	0.0484 [0.0546]	0.107 [0.0951]	0.0835 [0.0923]	0.0627 [0.0685]
Treatment * Performance Category 4	0.0355 [0.0818]	0.00888 [0.0716]	0.0434 [0.0544]	0.0336 [0.0957]	-0.000263 [0.0900]	0.0508 [0.0666]
Treatment * Performance Category 5	0.127 [0.106]	0.106 [0.106]	0.0562 [0.0924]	0.138 [0.128]	0.109 [0.133]	0.0357 [0.116]
Chi-Squared Test (p-value)						
*Eligible + (Eligible * Category 2) = 0*	0.7642	0.9106	0.991	0.7785	0.8977	0.9775
*Eligible + (Eligible * Category 3) = 0*	0.9827	0.9083	0.5365	0.981	0.9556	0.5690
*Eligible + (Eligible * Category 4) = 0*	0.2362	0.1831	0.4756	0.1519	0.1060	0.4026
*Eligible + (Eligible * Category 5) = 0*	0.6317	0.6763	0.9092	0.7545	0.8368	0.657
Sample						
Number of Students	110502	110502	83966	110502	110502	83966
Number of Schools	323	323	323	323	323	323
R-squared	0.18	0.192	0.567	0.18	0.192	0.567
Student-Level Covariates	√	√	√	√	√	√
School-Level Covariates		√	√		√	√
Prior Scale Score (quartic)			√			√

Notes: Student-level controls include race/ethnicity, English language learner status, and whether the student has an IEP. School-level covariates include school peer index, percent of students with certain race/ethnic status, percent English language learner, percent IEP, borough and total school enrollment. Standard errors in brackets adjust for clustering at the school level using bootstrap techniques with 300 iterations. * significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Table 12. Impact of New York City's Schoolwide Performance Bonus Program on Other Aspects of Progress Report Card System

	2007-08 School Year					
	Overall Score	**Environment Score**	**Performance Score**	**Progress Score**	**Bonus Points**	**Quality Review**
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.01405 [1.3930]	-.0994 [0.2182]	0.2949 [0.4024]	0.1566 [1.0751]	0.0907 [0.1996]	0.3094 [0.2338]
Sample	315	315	315	315	315	315
R-squared	0.2409	0.4689	0.2468	0.1171	…	…
School-Level Covariates	√	√	√	√	√	√
Prior Score (cubic)	√				√	

Notes: Each cell contains an estimate from a seperate regression that controls for school peer index, percent of students with certain race/ethnic status, percent English language learner, percent IEP, burough, and total school enrollment. * significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Table 13. Impact of New York City's Schoolwide Performance Bonus Program on Student, Parent, and Teacher Perceptions of School Environment by Four Domain Scores on Survey

	Panel A: Student Survey			
	Academic Score	*Engagement Score*	*Communication Score*	*Safety Score*
	(1)	(3)	(5)	(7)
Treatment	-.0547	-.2203	0.0760	-.1614
	[0.2943]	[0.2973]	[0.2974]	[0.2966]
Sample	145	145	145	145
	Panel B: Parent Survey			
	Academic Score	*Engagement Score*	*Communication Score*	*Safety Score*
	(9)	(11)	(13)	(15)
Treatment	0.0594	-1051	-.0854	-.0278
	[0.1986]	[0.1987]	[0.1987]	[0.1988]
Sample	316	316	316	316
	Panel C: Teacher Survey			
	Academic Score	*Engagement Score*	*Communication Score*	*Safety Score*
	(17)	(19)	(21)	(23)
Treatment	-.0189	0.1232	-0.0016	-.0557
	[0.0198]	[0.1980]	[0.1979]	[0.1980]
Sample	317	317	317	317
School Characteristics	√	√	√	√
Prior Score (cubic)	√	√	√	√
Response Rate (cubic)	√	√	√	√

Notes: Each cell contains an estimate from a seperate regression that controls for school peer index, percent of students with certain race/ethnic status, percent English language learner, percent IEP, burough, and total school enrollment. * significant at 10% level; ** significant at 5% level; *** significant at 1% level.

Faculty and Research Affiliates

Matthew G. Springer
Director
National Center on Performance Incentives

Assistant Professor of Public Policy and Education
Vanderbilt University's Peabody College

Dale Ballou
Associate Professor of Public Policy and Education
Vanderbilt University's Peabody College

Leonard Bradley
Lecturer in Education
Vanderbilt University's Peabody College

Timothy C. Caboni
Associate Dean for Professional Education and External Relations
Associate Professor of the Practice in Public Policy and Higher Education
Vanderbilt University's Peabody College

Mark Ehlert
Research Assistant Professor
University of Missouri – Columbia

Bonnie Ghosh-Dastidar
Statistician
The RAND Corporation

Timothy J. Gronberg
Professor of Economics
Texas A&M University

James W. Guthrie
Senior Fellow
George W. Bush Institute

Professor
Southern Methodist University

Laura Hamilton
Senior Behavioral Scientist
RAND Corporation

Janet S. Hansen
Vice President and Director of Education Studies
Committee for Economic Development

Chris Hulleman
Assistant Professor
James Madison University

Brian A. Jacob
Walter H. Annenberg Professor of Education Policy
Gerald R. Ford School of Public Policy
University of Michigan

Dennis W. Jansen
Professor of Economics
Texas A&M University

Cory Koedel
Assistant Professor of Economics
University of Missouri-Columbia

Vi-Nhuan Le
Behavioral Scientist
RAND Corporation

Jessica L. Lewis
Research Associate
National Center on Performance Incentives

J.R. Lockwood
Senior Statistician
RAND Corporation

Daniel F. McCaffrey
Senior Statistician
PNC Chair in Policy Analysis
RAND Corporation

Patrick J. McEwan
Associate Professor of Economics
Whitehead Associate Professor of Critical Thought
Wellesley College

Shawn Ni
Professor of Economics and Adjunct Professor of Statistics
University of Missouri-Columbia

Michael J. Podgursky
Professor of Economics
University of Missouri-Columbia

Brian M. Stecher
Senior Social Scientist
RAND Corporation

Lori L. Taylor
Associate Professor
Texas A&M University

NATIONAL CENTER ON

Performance Incentives

EXAMINING PERFORMANCE INCENTIVES IN EDUCATION

National Center on Performance Incentives
Vanderbilt University Peabody College

Peabody #43
230 Appleton Place
Nashville, TN 37203

(615) 322-5538
www.performanceincentives.org

VANDERBILT
PEABODY COLLEGE