
111
122
221

Does Whole-School Performance Pay Improve Student Learning?

Merit pay proponents argue that monetary incentives for better teaching can improve the quality of instruction in our nation's classrooms. Yet only a handful of studies have evaluated the impact of teacher merit pay on student achievement. These studies offer no conclusive recommendations regarding the optimal role of merit pay in U.S. school systems, leaving policymakers largely dependent on studies on other countries for information about how best to implement merit pay programs.

Recently, the New York City Department of Education (DOE) conducted a policy experiment to test whether merit pay given to all teachers at an effective school could increase student achievement. The city's School-Wide Performance Bonus Program, launched in 2007 and endorsed by both the DOE and the teachers union, was implemented in a randomly selected subset of the city's most disadvantaged schools. The randomized design of school selection makes it possible to separate out the causal effect of this form of merit pay from myriad other influences on student learning.

Our analysis is based on data from the first two years of the bonus program. In interpreting our findings, it is important to appreciate the key features of the program's structure. Teachers received bonuses based on the overall performance of all tested students in their school, rather

Evidence from the New York City schools

than just on the performance of students in their own classrooms. According to proponents of group incentives, this design can minimize conflicts and foster a spirit of cooperation among teachers at participating schools. However, under group incentive schemes, individual teachers may not have sufficient motivation to improve their own performance if they know that their success in attaining a bonus depends heavily on the efforts made by other teachers. Especially in schools with a large number of teachers, it may be difficult to sustain a school-wide push to mobilize the efforts of most teachers. The New York City bonus program thus provides valuable information on the effects of a school-wide bonus plan.

By SARENA GOODMAN and LESLEY TURNER

PHOTOGRAPH / AP IMAGES / RICARDO MAZALAN

Other specific characteristics of the bonus plan and the New York City context may also have influenced its effectiveness. If a school won a bonus, money was distributed among teachers and other school personnel by a committee consisting of two administrators and two teachers union representatives at the school. The bonus program was implemented alongside a new citywide accountability system that provided strong incentives to improve student achievement, regardless of whether a school was participating in the bonus program. Also, over the period we examine, all schools experienced increases in student achievement on the New York state test, leading some to suggest that the exam had grown easier (or at least easier to teach to). Roughly 90 percent of participating schools received a bonus in the second year of the program.

Schools had to gain the support of 55 percent of their full-time United Federation of Teachers (UFT) staff each year in order to participate.

Did the group bonus program operating in this policy environment have an impact on student achievement? We find very little effect overall, positive or negative. There is some evidence, however, that the program had a positive impact in schools where teachers were few in number, an environment in which it may be easier for teachers to cooperate in pursuit of a common reward. This study leaves open the question of whether a bonus program that rewards teachers for their own specific effectiveness would be more successful.

The Program

In November 2007, the New York City DOE launched the School-Wide Performance Bonus Program, randomly selecting 181 schools serving kindergarten through 8th grade to participate from a group of 309 high-need schools. (Disadvantaged high schools were also randomly selected into the program; we focus only on elementary and middle schools since these are the grades for which we can measure math and reading achievement.) The remaining 128 schools that were not selected serve as the control group for the purposes of our evaluation. The 309 schools included in the study differed from other city schools in the following ways: They had a higher proportion of English Language Learners (ELL), special education, minority students, and students eligible for the Title I free or reduced-price lunch program, as well as lower average math and reading scores. Teachers in these schools had slightly less experience and slightly more absences than teachers in other schools. The schools were smaller and had fewer teaching staff than other New York City schools.

The bonus program was the product of lengthy negotiations between district administrators and the teachers union. As a result of these negotiations, schools had to gain the support of 55 percent of their full-time United Federation of Teachers (UFT) staff each year in order to participate. Out of the 181 schools selected for the program, 25 schools voted not to participate in the first year of implementation or withdrew from the program following an initial vote of approval, and three more schools pulled out before the second year. Additionally, at the discretion of the DOE, two schools initially assigned to the treatment group were moved to the control group, and four schools initially designated as control schools were moved to the treatment group and subsequently voted to participate in the program. Of course, the schools that elected not to take part in the program and those moved by the DOE may differ in important ways from schools that chose to participate. We therefore consider the treatment group to include all 181 schools originally deemed eligible for bonus payments and take into account the fact that not all of them were actually participating in the program when interpreting our results.

Schools that implemented the program could earn a lump-sum bonus for meeting school-wide goals. These goals were tied to the New York City accountability system and were mainly determined by student performance on state math and reading exams. Under this accountability system, schools receive scores and grades that summarize their overall performance on three sets of measures: school environment, student performance, and student progress. The school environment measure incorporates student attendance and the results from surveys of parents, teachers, and students. Student performance measures include average student achievement on reading and math exams, along with median proficiency and the percentage of students achieving proficiency. The student progress measure considers the average change in test scores from year to year and the percentage of students who made progress from one year to the next. The accountability system also gives "extra credit" for exemplary progress among high-need students. Schools received target scores based on

their accountability grades, and schools with lower accountability grades needed to make larger improvements to reach their targets.

Schools participating in the bonus program received awards based on their progress toward meeting target scores. Schools that achieved their goals received bonuses equal to $3,000 per union teacher. Schools that fell short but manage to meet 75 percent of their goal received $1,500 per union teacher. Schools that did not achieve their target faced no consequences from the bonus program beyond the absence of incentive pay. For a sense of the strength of the incentive provided by the bonuses, the full $3,000 award represents a 7 percent increase in the salary of teachers at the bottom of the pay scale and a 3 percent increase for the most experienced teachers. In other words, these bonuses provided a substantial monetary benefit to most recipients.

Each participating school was required to develop a plan for distributing any lump-sum bonus awarded to the school. In the first year of the program, plans had to be submitted to the DOE after students took the state math and reading exams but before exam results were released and, thus, before schools knew whether they would receive a bonus. In every school, a four-member compensation committee, consisting of the principal, a second administrator, and two teachers elected by the school's UFT members, determined how bonuses would be distributed. The DOE program guidelines placed only two restrictions on the schools' bonus distribution plans: all union teachers had to receive a portion of the bonus payment and bonuses could not be distributed based on seniority. Otherwise, the committees had full discretion over bonus amounts and over whether other school employees would also receive funds. About half of the school committees chose to divide the award roughly equally among all recipients. In these schools, the difference between the highest and lowest bonus payment was less than $100. In the rest of the schools, the difference between the highest and the lowest bonus ranged from a low of $200 to a high of $5,000.

Of the 158 schools that voted to participate in the first year of the program, 87 (55 percent) received bonus payments. The bonus pool totaled $14.0 million in the first year and averaged $160,500 per school. In the second year of the program, the 2008–09 school year, 139 participating schools (91 percent) earned bonus awards, averaging $195,100 per school and totaling $27.1 million.

Little Difference for Students

Before we get to the detailed findings of our study, it is important to make clear the nature of the incentives NYC teachers and administrators faced over the period we examine. First, the 2007–08 school year was the first year of both the bonus program *and* a new citywide accountability system. The accountability system provided strong incentives to improve student achievement, regardless of whether a school was participating in the bonus program. For example, schools that earned A or B accountability grades were eligible for principal bonuses and additional funds when students transferred from schools receiving a poor grade. Schools that received D and F grades faced potential consequences, including principal removal and school closure. With this in mind, we see the results of our study as representing the effect of group-based teacher merit pay for schools that are already under accountability pressure. However, given that all school districts in the United States are subject to No Child Left Behind and many states have implemented their own accountability systems, this may be the most appropriate context in which to study the consequences of merit pay.

The second thing to keep in mind is that the power of the bonus program incentives was likely muted in the first year because of the timing of the program announcement. Eligible schools were notified in November of 2007, leaving relatively little time for teachers and administrators to alter their educational plans before accountability exams were administered in January for reading and March for math. As noted above, the percentage of schools that hit their achievement targets increased between the first, truncated year of the program and the second, when schools had more time to respond to the program incentives. But we caution readers to remember that this leap in bonus payouts is not, by itself, evidence that merit pay worked. It may instead reflect citywide performance improvements or, more pessimistically, that the New York state tests decreased in difficulty over this period. The most important comparison to make is between

Treatment-group schools need to outpace their counterparts in the control group over [the] two years [of our study] for us to say that merit pay made a real difference for student achievement.

the treatment group schools eligible for the bonus program (most of which actually participated in the program) and the schools in the control group. Treatment-group schools need to at least outpace their counterparts in the control group over these two years for us to say that merit pay made a real difference for student achievement. It is this comparison that is at the heart of our analysis.

How did bonus program schools fare compared to schools in the control group? Both groups of schools saw an increase in the average math and reading scores during the first two years of the bonus program; treatment-group schools, however, did not experience a statistically significant improvement in average test scores relative to the schools in the control group. Nor did these results change notably when we 1) made adjustments for the small differences in treatment and control school characteristics that existed despite randomization between treatment-group and control-group schools, or 2) took into account whether treatment-group schools elected to participate in the bonus program. It is possible, of course, that looking at average student achievement could divert our attention from changes for particular groups of students. Were teachers, we wanted to know, focusing their attention on either high-achieving or low-achieving students in an effort to meet target scores? We used statistical techniques similar to the one we employed to examine changes in average scores to assess the effect of the bonus program on the percentage of students achieving proficiency on math and reading exams. Once again, we found no evidence that the bonus program led to changes in this measure of student achievement. Participation in the bonus program did not, for example, boost the percentage of students who scored at or above the level designated as "proficient" under New York state accountability standards. Bonus-program schools fared no better than schools in the control group, and in the second year of the program, treatment schools experienced a statistically significant, although quite small, decrease in math proficiency.

This particular type of merit pay program, where bonuses are based on school-wide performance and teachers expect to receive bonus payments regardless of their effort, may not work in all schools.

On a related note, the New York City accountability system and, as a byproduct, the bonus program, contain incentives to focus on particular groups of students, since improvements for some student groups matter more in the calculations of a school's accountability grade. In addition to calculating overall achievement for all students in a school, components of the New York City accountability system take into account changes in the achievement of students who were in the lowest third of their grade in the prior year, those on the cusp of proficiency, and those close to the school's median score, along with students who are designated as ELL and students who are enrolled in special education programs. Again, we found no evidence to suggest that the bonus program led to achievement gains for any of these groups of students. On average, students in these groups fared just as well whether they attended a school that was participating in the bonus program or one in the control group.

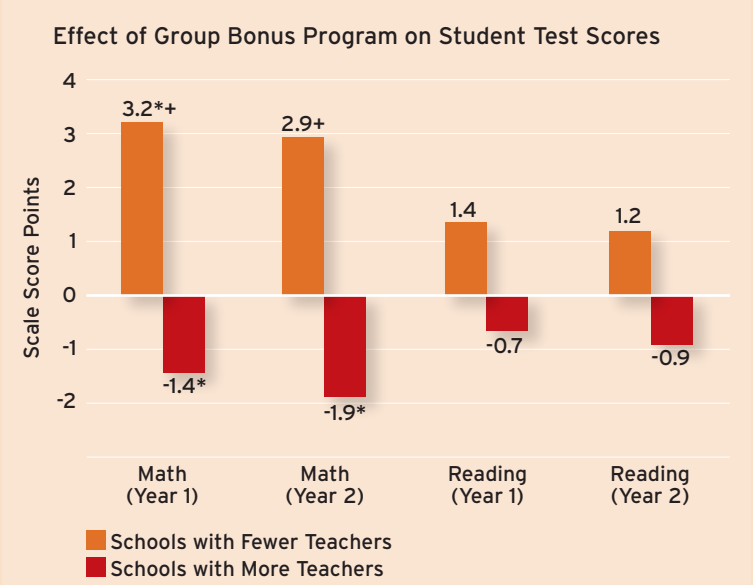
Limitations of Group Bonuses

Does evidence that the New York City bonus program did not lead to marked gains in student achievement, at least in the program's first two years, mean that merit pay for teachers in general does not work? That is certainly one possible conclusion to be drawn from our findings. Another possibility is that this particular type of merit pay program, where bonuses are based on school-wide performance and teachers expect to receive bonus payments regardless of their effort, does not work in all schools. Group bonuses may weaken the incentives for individual teachers to increase effort devoted to raising student achievement to the point that the programs become ineffective. And perhaps this problem would be mitigated in programs in which rewards are more tightly coupled to the effort an individual makes in the classroom.

Think about two schools, one with many more teachers than the other, both participating in a school-wide merit pay program. In each school, the impact of an individual teacher's effort on the expected bonus is determined by the number of other teachers with tested students, since bonus receipt is primarily based on student performance on math and reading exams. Because of this, a very good teacher with a large number of teaching colleagues can do less to raise school-wide student performance than a teacher of the same quality in a school with fewer teachers. In the school with more teachers, the diffusion of responsibility for test-score gains across many teachers may erode the incentive that any individual

Small Group Benefits (Figure 1)

In schools with a small number of teachers, the NYC School-Wide Performance Bonus Program boosted student learning in math, but it may have had a negative effect in schools with a larger number of teachers.



* Indicates that the effect is statistically significant at the 90 percent level.

\+ Indicates that the effect for schools with few teachers is statistically significantly different from the effect for other schools at the 90 percent confidence level.

Note: Schools with few teachers are those in the bottom quartile of schools in terms of the number of teachers with tested students; these schools had 10 or fewer such teachers in elementary/K-8 schools and five or fewer in middle schools.

SOURCE: Keeping Pace with K-12 Online Learning, 2010

teacher has to increase effort in the classroom. Some teachers may conclude that exerting additional effort will produce little difference in the overall performance of the school. The central idea here is that teachers could face relatively strong or weak incentives under the same merit pay program as a result of the number of teachers at their school. With this logic in mind, we examined the effect of the New York City school-wide merit pay program at schools with different numbers of teachers with test-taking students. Did schools with fewer teachers show signs that teachers were responding to merit pay incentives?

We conducted a statistical analysis similar to our method for estimating the average effect of the bonus program across all New York City schools in the experiment. But this time, we looked for different effects on math scores in schools with more and fewer math teachers and different effects on reading scores on schools with larger and smaller cohorts of reading teachers.

It turns out that the effectiveness of school-wide bonus programs may, in fact, depend on the number of teachers with tested students in a school (see Figure 1). For schools in the bottom quartile of the number of teachers with tested students, that is, schools with approximately 10 or fewer such teachers in elementary and K–8 schools and five or fewer in middle schools, school-wide merit pay *did* lead to improved student achievement. We estimate that the New York City bonus program had a positive effect on student math achievement in these schools in both program years, although the estimated effect in the second year fell just short of conventional levels of statistical significance. Conversely, this analysis also indicates that the program may have slightly lowered student achievement in schools with larger teaching staffs. Math achievement gains attributable to the bonus program in schools with smaller teaching staffs were modest in size but meaningful. In the first year of the program, the bonus program boost to math scores was, by our estimates, 3.2 points on the New York state test, or 0.08 student-level standard deviations. To benchmark this effect against the magnitude of other familiar results, it is slightly smaller than the estimated 0.1 standard deviation gain in achievement that results from being assigned to a teacher at the 85th percentile of the effectiveness distribution rather than a teacher at the median.

The Devil in the Details

The New York City bonus-pay program provides us with a valuable opportunity to study the effect of merit pay for teachers in an experimental setting. We are a long way from amassing a convincing body of research on either side of the debate over merit pay in education, but what this experiment makes frustratingly clear for merit pay proponents is that the structure of the payment scheme can make a large difference. For merit pay to improve student outcomes, teachers must face strong incentives to improve their performance. Our study indicates that school-wide bonus programs may be able to provide those incentives in schools with relatively small teaching staffs. They may also be appropriate for schools characterized by a high degree of staff cohesion, in which teachers work collaboratively to improve student learning and it is difficult to isolate the performance of a single teacher. The early experience with the New York City School-Wide Performance Bonus Program suggests, however, that a heavy reliance on school-wide rewards may hamper the effectiveness of merit pay programs in schools with large teaching staffs that are not highly collaborative.

Sarena Goodman and Lesley Turner are PhD candidates in Columbia University's Department of Economics.