NATIONAL CENTER ON
Performance Incentives

PROJECT ON INCENTIVES IN TEACHING

POINT

Teacher Pay for Performance

Experimental Evidence from the Project on Incentives in Teaching

Matthew G. Springer
Dale Ballou
Laura Hamilton
Vi-Nhuan Le
J.R. Lockwood
Daniel F. McCaffrey
Matthew Pepper
Brian M. Stecher

LED BY
VANDERBILT PEABODY COLLEGE

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

EXECUTIVE SUMMARY

The Project on Incentives in Teaching (POINT) was a three-year study conducted in the Metropolitan Nashville School System from 2006-07 through 2008-09, in which middle school mathematics teachers voluntarily participated in a controlled experiment to assess the effect of financial rewards for teachers whose students showed unusually large gains on standardized tests. The experiment was intended to test the notion that rewarding teachers for improved scores would cause scores to rise. It was up to participating teachers to decide what, if anything, they needed to do to raise student performance: participate in more professional development, seek coaching, collaborate with other teachers, or simply reflect on their practices. Thus, POINT was focused on the notion that a significant problem in American education is the absence of appropriate incentives, and that correcting the incentive structure would, in and of itself, constitute an effective intervention that improved student outcomes.

By and large, results did not confirm this hypothesis. While the general trend in middle school mathematics performance was upward over the period of the project, students of teachers randomly assigned to the treatment group (eligible for bonuses) did not outperform students whose teachers were assigned to the control group (not eligible for bonuses). The brightest spot was a positive effect of incentives detected in fifth grade during the second and third years of the experiment. This finding, which is robust to a variety of alternative estimation methods, is nonetheless of limited policy significance, for as yet this effect does not appear to persist after students leave fifth grade. Students whose fifth grade teacher was in the treatment group performed no better by the end of sixth grade than did sixth graders whose teacher the year before was in the control group. However, we will continue to investigate this finding as further data become available, and it may be that evidence of persistence will appear among later cohorts.

The report is divided into six sections. After a brief introduction, Section II describes the design and implementation of POINT. In POINT the maximum bonus an eligible teacher might earn was $15,000—a considerable increase over base pay in this system. To receive this bonus, a teacher's students had to perform at a level that historically had been reached by only the top five percent of middle school math teachers in a given year. Lesser amounts of $5,000 and $10,000 were awarded for performance at lower thresholds, corresponding to the 80th and 90th percentiles of the same historical distribution. Teachers were therefore striving to reach a fixed target rather than competing against one another—in principle, all participating teachers could have attained these thresholds.

It is unlikely that the bonus amounts were too small to motivate teachers assigned to the treatment group. Indeed, a guiding consideration in the design of POINT was our desire to avoid offering incentives so modest that at most a modest response would result. Instead, we sought to learn what would happen if incentives facing teachers were significantly altered. Was the bar set too high, discouraging teachers who felt the targets were out of reach? We devote considerable attention to this question in Appendix A, examining performance among teachers who were not eligible for bonuses (POINT participants prior to the implementation of the project, and control teachers during the project). We find that about half of these teachers could reach the lowest of the bonus thresholds if their students answered 2 to 3 more questions correctly on an exam of some 55 items. We conclude

that the bonus thresholds should have appeared within reach of most teachers and that an attempt to raise performance at the margin ought not to have been seen as wasted effort by all but a few teachers "on the bubble."

In Section III we consider other threats to the validity of our findings. We investigate whether randomization achieved balance between treatment and control groups with respect to factors affecting achievement other than the incentives that POINT introduced. While balance was achieved overall, there were differences between treatment and control groups within subsamples of interest (for example, among teachers within a single grade). Statistical adjustments through multiple regression analysis are required to estimate the effect of incentives in such subsamples. As always, this raises the possibility that different models will yield different findings. Thus, we place greatest confidence in estimates based on the overall sample, in which data are pooled across years and grades.

POINT randomized participating teachers into treatment and control groups. It did not randomize students. Because the assignment of students to teachers was controlled by the district, it is possible that principals and teachers manipulated the assignment process in order to produce classes for treatment teachers that enhanced their prospect of earning a bonus. In addition, attrition of teachers from POINT was high. By the end of the project, half of the initial participants had left the experiment. Such high rates of attrition raise the possibility that our findings could reflect differential selection (for example, more effective teachers might remain in the treatment group than in the control group).

We conducted a variety of analyses to ascertain whether differential attrition or the manipulation of student assignments biased our results. We conclude that neither produced significant differences between treatment and control groups and that experimental estimates of the incentive effect are free of substantial bias. In addition, to remove the impact of differences between the teachers and students assigned to treatment and control that arose by chance, we estimate treatment effects using models in which we control for student and teacher characteristics. Our conclusions about the overall effect of incentives are robust to the omission of such controls: a straightforward comparison of mean outcomes in the treatment and control groups and estimates from the more complicated model both show no overall treatment effect. This is not true of estimates based on subsets of the full sample—for example, outcomes by grade level. At the grade level there were substantial imbalances between treatment and control groups whose influence on achievement must be controlled for.

It is also possible that test score gains were illusory rather than proof of genuine achievement. This would obviously be the case if treatment teachers engaged in flagrant forms of cheating to promote their chances of earning a bonus. But it might also result from the adoption of instructional strategies intended to produce short-term gains on specific test instruments. Our investigation (including a statistical analysis of item-level responses) does not reveal this to have been a problem, though we have not had access to test forms in order to look for suspicious patterns of erasures.

In Section IV we present our findings. As already noted, we find no effect of incentives on test scores overall (pooling across all years and grades). We do find a positive effect among fifth graders whose teachers were eligible for bonuses. We have explored a variety of hypotheses that might account for

a positive effect in grade 5 but not the other grades. Only one seems to have played an appreciable role: fifth grade teachers are more likely to instruct the same set of students in multiple subjects. This appears to confer an advantage, though it is unclear precisely what the advantage consists of—whether it is the opportunity to increase time on mathematics at the expense of other subjects, or the fact that these teachers know their students better, or something else. And even this is at best a partial explanation of the fifth grade response.

POINT participants (both treatment and control teachers) completed surveys each spring over the course of the project. In Section V we summarize some of the findings, focusing on two issues: (1) how teachers' attitudes toward performance pay were affected by POINT; and (2) why we found no overall response to incentives.

Participating teachers generally favored extra pay for better teachers, in principle. They did not come away from their experience in POINT thinking the project had harmed their schools. But by and large, they did not endorse the notion that bonus recipients in POINT were better teachers or that failing to earn a bonus meant a teacher needed to improve. Most participants did not appear to buy in to the criteria used by POINT to determine who was teaching effectively. Perhaps it should not be surprising, then, that treatment teachers differed little from control teachers on a wide range of measures of effort and instructional practices. Where there were differences, they were not associated with higher achievement. By and large, POINT had little effect on what these teachers did in the classroom.

In the concluding section, we summarize our main findings and explore their implications for education policy. The introduction of performance incentives in MNPS middle schools did not set off significant negative reactions of the kind that have attended the introduction of merit pay elsewhere. But neither did it yield consistent and lasting gains in test scores. It simply did not do much of anything. While it might be tempting to conclude that the middle school math teachers in MNPS lacked the capacity to raise test scores, this is belied by the upward trend in scores over the period of the project, a trend that is probably due to some combination of increasing familiarity with a criterion-referenced test introduced in 2004 and to an intense, high-profile effort to improve test scores to avoid NCLB sanctions.

It should be kept in mind that POINT tested a particular model of incentive pay. Our negative findings do not mean that another approach would not be successful. It might be more productive to reward teachers in teams, or to combine incentives with coaching or professional development. However, our experience with POINT underscores the importance of putting such alternatives to the test.