Abstract Title Page
Not included in page count.

Title: Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT)

Author(s): Matthew G. Springer, Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, Brian M. Stecher

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

Despite the rocky history of merit pay in public schools, interest in tying teacher compensation to performance has revived, with the federal government taking a leading role in promoting compensation reform as a way to improve educational outcomes. With the expansion of standardized testing in systems of school accountability, the notion that teachers should be compensated (in part) on the basis of students' test score gains or more sophisticated measures of teacher value added has gained currency. However, the idea is controversial. Apart from debate over whether this is an appropriate way to measure what teachers do, it is not known how well this policy works in its own terms. If teachers are rewarded for an increase in student test scores, will test scores go up?

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

This paper presents the results of a rigorous experiment examining the impact of pay for performance on student achievement and instructional practice. This study, conducted by the National Center on Performance Incentives in partnership with the RAND Corporation examines an experimental pay for performance program administered via a randomized controlled study in the Metro Nashville Public Schools (MNPS) district.

The research questions are:

1: Does performance-pay alone improve student outcomes?

2: Does the opportunity to earn bonuses alter teachers' instructional practices and attitudes?

Setting:

Description of the research location.

The POINT experiment was conducted in MNPS district middle schools (grades 5-8) for three academic years from 2006-07 through 2008-09. MNPS includes 34 middle schools, enrolling roughly 20,000 students. The student population is approximately 47.4% African American, 15.9% Hispanic, 3.7% Asian and 32.7% White. 12.6% of students are English Language Learners and 75.9% are classified as economically disadvantaged. The consolidated city-county district covers Davidson County, an area of approximately 525 square miles and approximately 626,000 residents. MNPS district is the 48th largest urban school district in the nation.

Population / Participants / Subjects:

Description of the participants in the study: who, how many, key features or characteristics.

The POINT experiment includes middle school mathematics teachers in the MNPS district. All middle school mathematics teachers working in the district in 2006-2007 were eligible to participate, given that at least ten of their students per year completed the annual mathematics Tennessee Cumulative Assessment Program (TCAP) exam.

All teacher volunteers had to sign up in the first year of the experiment (2006-2007). Late enrollments were not permitted, nor were teachers who left the experiment permitted to re-enroll. Participating teachers could remain in the experiment even if they transferred schools as long as they continued to teach mathematics to at least one middle school grade in MNPS and remained above the ten-student threshold. Two-thirds of the district's eligible middle school mathematics teachers volunteered to participate. 296 teachers participated in the study in the beginning of the 2006-07 school year, though only 148 remained through the end of the third year (Tables 2 & 3.)

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.

Teachers who were randomly assigned to the treatment group were notified that they would be eligible for bonuses of up to $15,000 per year on the basis of their student test-score gains on the Tennessee Comprehensive Assessment Program (TCAP). Over the three years the experiment ran, POINT paid out more than $1.27 million in bonuses. A breakdown by year and bonus level appears in Table 1. 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.

In late summer of 2007, 2008, and 2009, NCPI calculated the performance measures and bonus awards. Confidential bonus reports were prepared for each teacher in the treatment group. Each report showed how the teacher's performance measure was calculated and whether that measure exceeded any of the thresholds entitling the teacher to a bonus. A roster of the student scores (without student names) used to calculate the teacher's performance measure was also provided. Bonus reports were mailed to treatment group teachers in September 2007, 2008, and 2009. Bonus awards were distributed to qualifying teachers in November paychecks.

Participating teachers were also surveyed annually during the experiment to gather information on how teachers respond to POINT how it affected teacher attitudes toward performance pay.

Research Design:
Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

The study was designed as a controlled, randomized experiment. Approximately half the teachers volunteering to participate were randomly assigned to a treatment group, in which they were eligible for bonuses of up to $15,000 per year on the basis of student test-score gains on the Tennessee Comprehensive Assessment Program (TCAP). The other half were assigned to a control group that was not eligible for these bonuses.

Participants were randomized into treatment and control groups using a two-stage process. First, schools were stratified into ten groups based on student TCAP scores in prior years.

Randomization was done within strata to ensure balance between treatment and control groups. Second, clusters of teachers rather than individual teachers were assigned to treatment or control status. Clusters were based on four course-groups: grade 5 and 6 mathematics classes, grade 7 and 8 mathematics classes, special education mathematics classes, and algebra or more advanced mathematics classes. Each teacher was associated with one of these groups, based on the courses taken by most of her students. A cluster was the set of teachers in a given school in the same course group. Clusters of the same type from the various schools within each stratum were combined to create blocks and within each block half of the clusters were randomly selected to be part of the treatment group and the other half were assigned to the control group.

To determine whether a teacher in the treatment group qualified for an award, we used a relatively simple measure of teacher value-added. Our value-added measure was based on students' year-to-year growth on TCAP. To control for the possibility that students at different points in the distribution of scores are likely to make different gains, we benchmarked each student's gain against the average gain, statewide, of all students taking the same test with the same prior year score. This average was the value-added score used to determine whether the teacher qualified for a bonus.

Additionally, NCPI administered surveys to all teachers participating in the POINT experiment in the spring 2007, spring 2008, and spring 2009 semesters. The surveys included items on teacher attitudes, behavior and instructional practice, and school culture. Surveys asked teachers about their opportunities for professional growth and about their classroom practice—what resources they used related to curriculum standards and assessments (i.e., curriculum guides, assessment training manuals) and whether they used student achievement scores to tailor instruction to students' individual needs. Finally, surveys addressed contextual factors at school that may moderate the impact of a pay for performance program: the quality of collegial relations and school leadership, and the professional culture at the school (Figure 7.)

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

To estimate the treatment effects we used linear mixed models designed to account for features of the experimental design and randomization into treatment and control groups (Raudenbush and Bryk, 2002). Over the three years of the study, we have repeated measures on both students and teachers. These units are not nested, for students move across teachers as they progress through grades. Blocks combined clusters from schools with similar historic school-level value-added measures and study teachers were uniquely linked to randomization clusters based on their teaching assignments at the beginning of the study. The models account for the blocking and the cluster randomization by way of block fixed effects and cluster random effects.

Virtually all of the results we report were obtained from separate samples for each year. When data were pooled across years, the model also included block by year interactions and cluster by year random effects. Models included teacher random effects (or teacher by year effects, when data were pooled across years) as well as teacher by grade random effects. Students are observed more than once in the samples that pool data across years. In this case, within-student covariances over time are unrestricted. Finally the models included grade by year fixed effects to account for grade-level trends in the achievement scores.

To control for differences between treatment and control groups that might have arisen for reasons other than chance, we adjust for pre-experiment student characteristics including achievement in each of the four TCAP subjects, race/ethnicity, gender, English Language Learner (ELL) classification, special education participation, free and reduced price lunch participation, and the numbers of days of suspension and unexcused absences. Covariates were measured in the most recent year outside of the experimental frame of the 2006-07 to 2008-09 school years and grades 5-8.

In order not to distort the relative performance of treatment and control groups, we standardized the scores by grade and subject relative to the entire district in spring 2006, the testing period immediately prior to the experiment. Specifically, we used the district-wide TCAP data during 2005-2006 to create a mapping between scores and percentiles in the district, with separate mappings by grade and subject. For all other years, we assigned every scale score a percentile by locating it in the appropriate 2006 grade/subject distribution, using linear interpolation to estimate percentiles for scale scores that were not observed in 2006 (scores outside the observed 2006 range were assigned the percentile of the maximum or minimum 2006 score). The percentiles were then transformed by the standard normal inverse cumulative distribution function. We report results on this standardized scale.

NCPI administered annual teacher surveys. The dependent variable in the survey was measured on a 4-point Likert scale (strongly disagree, disagree, agree, strongly agree). We test for differences across grades and treatment status, and for changes over time, using an ordered probit model in which the regressors are randomization block, the proportion of a teacher's students at each grade level, year, and treatment status. The error structure includes a random effect for cluster.

Findings / Results:
Description of the main findings with specific details.

We find no significant difference overall between students whose teachers were assigned to the treatment group and those whose teachers were assigned to the control group (Figure 2) In addition, there were no significant differences in any single year, nor were there significant differences for students in grades 6-8 when separate effects were estimated for each grade level.

We do find significant positive effects of being eligible for bonuses in the second and third years of the project in grade 5 (Figure 3.) The difference amounts to between one half and two-thirds of a year's typical growth in mathematics. However, for the 2007-08 fifth grade cohort (the only cohort we have been able to follow as yet as sixth graders), these effects are no longer evident the following year. That is, it makes no difference to grade 6 test scores whether a student's fifth grade teacher was in the treatment group or the control group.
There was also a significant difference between students of treatment and control teachers in fifth grade social studies (years 2 and 3 of the project) and fifth grade science (year 3). No differences for these subjects were found in other grades.

Based on survey responses, more than 80 percent of treatment teachers agreed that POINT "has not affected my work, because I was already working as effectively as I could before the

implementation of POINT." Fewer than a quarter agreed that they had altered their instructional practices as a result of the POINT experiment.

Treatment teachers were more likely to report that they collaborated with other teachers and were more likely to say that they aligned their instruction with the district's mathematics standards and spent classroom time on test preparation. When examining the relationship of these practices to student achievement, we do not find a positive, statistically significant association between the second set of activities and student achievement. Nor do we find evidence that the collaborative activities in which treatment teachers engaged were associated with higher test scores, with two exceptions: teachers that acted as mentors or coaches had better results, as did teachers that observed the work of others in the classroom. Because a teacher chosen to be a mentor or coach is likely a more effective teacher to begin with, the association may well be a selection effect.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

Given the limited scope of the effects and their apparent lack of persistence, we conclude that the POINT intervention did not lead overall to large, lasting change in student achievement as measured by TCAP. There is little evidence that POINT incentives induced teachers to make substantial changes to their instructional practices or their level of effort, and equally little evidence that the changes they did make were particularly well chosen to increase student achievement.

Possibly certain features of the project which were adopted in response to teachers' concerns ended up limiting its impact. The names of bonus winners were not publicized. Teachers were asked not to communicate to other district employees whether they received bonuses. A performance measure was used with which teachers were not familiar, and though it was easy to understand, nothing was done to show teachers how to raise their scores. Incentives were not coupled with any form of professional development, curricular innovations, or other pressure to improve performance.

The implications of these negative findings should not be overstated. That POINT did not have a strong and lasting effect on student achievement does not automatically mean another approach to performance pay 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.

Finally, we note that advocates of incentive pay often have in mind an entirely different goal from that tested by POINT. Their support rests on the view that over the long term, incentive pay will alter the makeup of the workforce for the better by affecting who enters teaching and how long they remain. POINT was not designed to test that hypothesis and has provided only limited information on retention decisions. A more carefully crafted study conducted over a much longer period of time is required to explore the relationship between compensation reform and professional quality that operates through these channels.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Raudenbush, S.W., & Bryk, A.S. (2002). Hierarchical linear models: Applications and data analysis methods. 2nd edition. Newbury Park, CA: Sage.

Appendix B. Tables and Figures

Not included in page count.

TABLE 1.
Bonus Awards by Year

	School Year		
	2006-07	2007-08	2008-09
# treatment teachers	143	105	84
# bonus recipients	41	40	44
# at $5,000	10	4	8
# at $10,000	17	15	23
# at $15,000	14	21	13
Average bonus award	$9,639	$11,370	$9,623
Total amount awarded	**$395,179**	**$454,655**	**$423,412**

TABLE 2.
Number of Teachers Who Dropped Out of the POINT Experiment by Treatment Status and School Year

	School Year		
Experimental Group	2006-07	2007-08	2008-09
Control	2	58	18
Treatment	3	42	23

TABLE 3.
Reasons for Attrition by Treatment Status

	Reason for Attrition						
	Change in Assignment					NCPI Initiated	
	In MNPS, Not Teaching	Retired	Moved to HS or ES*	Left MNPS	Still Teaching, not Math	Dropped from Experiment[a]	<10 Math Students
Control	8	0	14	27	18	1	10
Treatment	14	2	11	15	18	1	7

[a] One teacher declined to participate in the surveys and other aspects of the study and was dropped from the experiment; the other teacher was a long-term substitute who was not eligible and was dropped when status was revealed.
*HS - high school; ES - elementary school

FIGURE 2.
Math Achievement Trends Overall

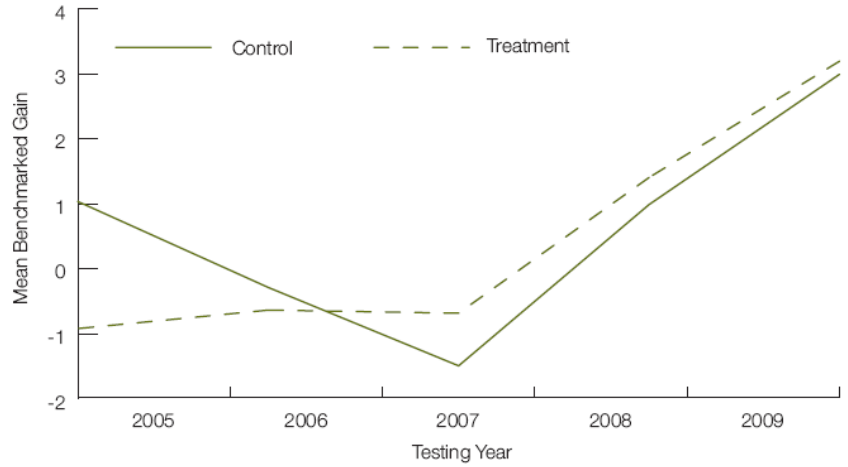

FIGURE 7.
Survey Items on Teacher Effort and Instructional Practices

Category: MNPS standards

I analyze students' work to identify the MNPS mathematics standards students have or have not yet mastered.
I design my mathematics lessons to be aligned with specific MNPS academic standards.
[All items answered: Never (1), once or twice a year (2), once or twice a semester (3), once or twice a month (4), once or twice a week (5), or almost daily (6)]

Category: Use of instructional time

Aligning my mathematics instruction with the MNPS standards.
Focusing on the mathematics content covered by TCAP.
Administering mathematics tests or quizzes.
Re-teaching topics or skills based on students' performance on classroom tests.
Reviewing test results with students.
Reviewing student test results with other teachers.
[All items answered: Much less than last year (1), a little less than last year (2), the same as last year (3), a little more than last year (4), or much more than last year (5)]

Category: Practicing test-taking skills

Increasing instruction targeted to state or district standards that are known to be assessed by the TCAP.
Having students answer items similar to those on the TCAP (*e.g.*, released items from prior TCAP administrations).
Using other TCAP-specific preparation materials.
[All items answered: No importance (1), low importance (2), moderate importance (3), or high importance (4)]

FIGURE 7. Cont.
Survey Items on Teacher Effort and Instructional Practices

Category: Time devoted to particular teaching methods in mathematics
Math students spending more time on: Engaging in hands-on learning activities (*e.g.*, working with manipulative aids). Working in groups. *[All items answered: Much less than last year (1), a little less than last year (2), the same as last year (3), a little more than last year (4), or much more than last year (5)]*
Category: Time outside regular school hours
During a typical week, approximately how many hours do you devote to school-work outside of formal school hours (*e.g.*, in the evenings, before the school day, and on weekends)?
Category: Level of instructional focus
I focus more effort on students who are not quite proficient in mathematics, but close. I focus more effort on students who are far below proficient in mathematics. *[All items answered: Never or almost never (1), occasionally (2), frequently (3), or always or almost always (4)]*
Category: Use of test scores
Use test scores for the following purposes: Identify individual students who need remedial assistance. Set learning goals for individual students. Tailor instruction to individual students' needs. Develop recommendations for tutoring or other educational service for students. Assign or reassign students to groups. Identify and correct gaps in the curriculum for all students. *[All items answered: Not used in this way (1), used minimally (2), used moderately (3), or used extensively (4)]*
Category: Collaborative activities with other mathematics teachers
Analyzed student work with other teachers at my school. Met with other teachers at my school to discuss instructional planning. Observed lesson taught by another teacher at my school. Had my lessons observed by another teacher at my school. Acted as a coach or mentor to other teaches or staff in my school. Received coaching or mentoring from another teacher at my school or from a district math specialist. *[All items answered: Never (1), once or twice a year (2), once or twice a semester (3), once or twice a month (4), once or twice a week (5), or almost daily (6)]*