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

TEACHER PAY FOR PERFORMANCE:

Experimental Evidence from the Project on Incentives in Teaching

September 21, 2010

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

While permission to reprint is not necessary, the recommended citation is: Springer, M.G., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCaffrey, D., Pepper, M., and Stecher, B. (2010). Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

This report is also available on the NCPI website at www.performanceincentives.org

This page intentionally left blank.

ACKNOWLEDGEMENTS

This report was made possible by the efforts of many individuals and organizations over a five year period. We greatly appreciate the willingness of the Metropolitan Nashville School Board, the Metropolitan Nashville Public Schools, the Mayor's Office, the Metropolitan Nashville Education Association, the Nashville Alliance for Public Education, the Tennessee Education Association and the Tennessee Department of Education. Many individuals also contributed in meaningful ways to make this project a success, including Leonard Bradley, Tim Caboni, Paul Changas, Margaret Dolan, David Fox, Pedro Garcia, Pamela Garrett, Graham Greeson, James Guthrie, Janet Hansen, Chris Henson, Marc Hill, Keel Hunt, Erick Huth, June Keel, Julie Koppich, Meredith Libbey, Dan Long, Al Mance, Jayme Merritt, Melanie Moran, Bill Purcell, Jesse Register, Kay Simmons, and Gloria Towner. Without their involvement, this study would not have been feasible. We wish we could name the many other dedicated educators who cooperated with the study in many ways. These people were extraordinarily generous with their time and expertise.

We would also like to acknowledge the contributions from many other individuals at the National Center on Performance Incentives, RAND, and Vanderbilt University's Peabody College. We would like to acknowledge the input from Susan Burns, Christine DeMartini, Mark Ehlert, Cate Gardner, Bing Han, Robert Hickam, Rebekah Hutton, Kelly Fork, Jessica Lewis, Warren Langevin, Brian McInnis, Lawrence Painter, Art Peng, Michael Podgursky, John Smith, and Elias Walsh. Countless other colleagues at Vanderbilt University and RAND provided helpful comments and suggestions. The authors wish to thank Richard Colvin, Steve Glazerman, Carolyn Heinrich, Brian Jacob, Derek Neal, and Jeffrey Smith for their thoughtful review of the study.

This report was supported by the National Center on Performance Incentives, which is funded by the United States Department of Education's Institute of Education Sciences (R305A06034). The project would not have been possible without the generous support of an anonymous foundation that funded teacher bonuses.

The views expressed in this report do not necessarily reflect those of sponsoring agencies or individuals acknowledged. Any errors remain the sole responsibility of the authors.

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST

The research team for this evaluation consists of a prime grantee, Vanderbilt University's Peabody College; its subcontractors, RAND Corporation, University of Missouri – Columbia, and University of Michigan. None of these organizations or their key staff members has financial interests that could be affected by findings from the study. No one involved in the content of this report have financial interests that could be affected by findings from the study.

This page intentionally left blank.

FOREWORD

The Project on Incentives in Teaching (POINT) was a three-year study conducted in the Metro-Nashville Public Schools from 2006-07 through 2008-09. Middle school mathematics teachers voluntarily participated in a controlled experiment to assess the effect of offering financial rewards to teachers whose students showed unusual gains on standardized tests. This report contains a description of the project and a summary of the principal effects of the incentives on student achievement.

A longer, more comprehensive report will appear within the next few months. The longer report will contain an exhaustive description of data collection and a more elaborate analysis of teachers' responses to surveys that asked about their attitudes toward incentive pay, their perceptions of school climate, and changes in their behavior over the course of the experiment. We have made the decision to go forward with a shorter, more focused report at this time given the intense interest in this topic in education policy circles.

While this document is shorter than the full report to come, this should not be taken to mean that it is unduly simplified. The issues involved in analyzing the impact of incentives in POINT are complex, and much of the discussion is necessarily technical.

This page intentionally left blank.

TABLE OF CONTENTS

This page intentionally left blank.

LIST OF TABLES

This page intentionally left blank.

LIST OF FIGURES

This page intentionally left blank.

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.

This page intentionally left blank.

I. INTRODUCTION

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?

To test this proposition, the National Center on Performance Incentives (NCPI) partnered with the Metropolitan Nashville Public Schools (MNPS) to conduct the Project on Incentives in Teaching, or POINT. POINT was designed as a controlled 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. Because assignment to these conditions was random, there should be no systematic differences in the effectiveness of the teachers in the two groups apart from differences induced by the incentives. Better student outcomes in the treatment group would therefore be evidence that such incentives work: tying pay to an improvement in tests scores results in higher scores.

This page intentionally left blank.

II. DESIGN AND IMPLEMENTATION OF POINT

Several important considerations influenced the design of the experiment[1]

- Teachers would not compete against one another for bonuses.
- Awards would be made to individual teachers, not to teams or entire schools.
- Teachers would be evaluated on the basis of students' progress over the year and not their incoming level of achievement.
- The performance threshold for a teacher to earn a bonus award should not be so high that the goal appeared unattainable, nor so low that total bonuses paid out would exceed NCPI resources.
- Maximum bonuses should be large, providing strong motivation to improve performance.

The POINT experiment was open to middle school (grades 5, 6, 7, and 8) mathematics teachers working in the MNPS district during the fall of the 2006-07 school year. Teachers could teach other subjects than math, but they needed at least ten students taking the mathematics TCAP to participate. All teacher volunteers had to sign up in the first year of the experiment. Late enrollments were not permitted, nor were teachers who left the experiment permitted to re-enroll. Assignments to treatment (eligible for bonuses) and control (not eligible) groups were permanent for the duration of the project. 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.

To determine whether a teacher qualified for an award we used a relatively simple measure of teacher value-added. While more complicated and sophisticated measures could have been chosen (cf. Sanders, Saxton, and Horn, 1997, McCaffrey et al, 2004, Harris and Sass, 2006, Lockwood et al, 2007), simplicity and transparency seemed desirable. First, we needed to attract a sufficient number of volunteers to the program. Awarding bonuses on the basis of measures no one could understand struck us as unhelpful. Second, we felt a transparent measure of performance would give teachers the best opportunity to see why they had or had not received a bonus, and if they had not, by how much they fell short. This might in turn provide stronger motivation to improve than if we were to use a less transparent measure.

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

[1] Research has shown that teachers' responses to pay for performance are associated with the perceived fairness of their evaluations and with whether the targets are seen to be realistic (Kelley, Heneman & Milanowski, 2002; Milanowski, 2000). In addition, teachers and others have expressed concerns about negative effects of pay for performance on collegiality (Milanowski & Gallagher, 2000; Kellor, 2005), particularly in light of research that suggests the importance of collegiality and trust among school staff in promoting student learning (Rowan et al., 2002; Bryk & Schneider, 2002).

[2] Some smoothing of the state means was done to compensate for erratic patterns at the extremes of the distribution, where the number of scores can be quite small, even for the entire state.

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.[2] Benchmarking was simple: we subtracted the statewide average gain from a student's own gain to find out by how much his growth had exceeded the state average. Finally, we averaged these benchmarked scores over a teacher's class—more precisely, over students continuously enrolled in the teacher's class from the twentieth day of the school year to the spring TCAP administration, and for whom we had the prior year scores needed for benchmarking. This average was the value-added score used to determine whether the teacher qualified for a bonus.

To determine the thresholds that teachers' performance measures would need to reach to qualify for bonuses, we calculated the same performance measures for district teachers of middle school mathematics in the two years immediately prior to POINT, 2004-05 and 2005-06. We then set three thresholds based on the distribution of these measures: one at the 80th percentile, a second at the 85th percentile, and a third at the 95th percentile. Teachers whose performance during POINT reached the lowest of these thresholds were eligible for a $5,000 bonus. Those reaching the middle threshold were eligible for $10,000, and those reaching the highest threshold were eligible for $15,000.

It may be wondered whether we set the bar so high that few teachers would be motivated to change their instructional practices or raise their level of effort—that most teachers would regard the performance targets as unattainable no matter what they did, while others with strong prior performance would decide they did not need to make any changes in order to obtain bonuses. We have conducted an extensive analysis of this issue. In fact, neither statement appears to have been true of most teachers, to judge from performance in the pre-POINT years. Teachers' subjective probabilities of earning a bonus, as recorded on annual surveys given to POINT participants, strengthen this conclusion. Few thought they had no chance of winning a bonus or that it was a sure thing. (For complete analysis of this question, see Appendix A.)

Many MNPS middle school teachers, particularly in grades 5 and 6, teach subjects other than mathematics. Tying bonuses solely to mathematics test scores might encourage them to neglect other subjects. To safeguard against this, we calculated an analogous benchmarked performance measure for each teacher in all four tested subjects, including reading/English language arts, science, and social studies. To receive the full bonus for which a teacher qualified on the basis of the mathematics performance measure, it was necessary to match or exceed the district's mean benchmarked performance on the other measures in all the subjects for which the teacher provided instruction. Falling short of that goal cost the teacher a prorated portion of the mathematics bonus based on the proportion of her students tested in other subjects.

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 (*e.g.*, a disproportionate number of teachers in the highest performing schools being assigned to the treatment group by chance). 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 ad-

vanced 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. Because not every cluster appeared in every school, randomization occurred within 37 blocks. Slight deviations from this procedure were adopted to ensure that every school had at least one treatment teacher.[3]

IMPLEMENTATION

Two-thirds of the district's eligible middle school mathematics teachers volunteered to participate in POINT. Two hundred and ninety six teachers participated in the study in the beginning of the 2006-2007 school year though only 148 remained through the end of the third year. (See below for a discussion of attrition.) Each POINT teacher received a stipend of up to $750 for each year of participation in the experiment. This payment was to encourage even those volunteers assigned to the control condition to participate in various kinds of data-collection activities, as well as to mitigate negative reactions from being assigned to the control group. The stipend amount was reduced if teachers did not complete all of these activities. Teachers were notified of their stipend awards in letters sent out in the summer, with stipends paid in the late summer.

NCPI determined bonus awards and paid them to treatment teachers each year of the study. A careful audit of the rosters of treatment teachers was conducted at the beginning of each year to ensure that teachers were not held accountable for students not in their classes the requisite portion of the school year.[4] In late summer of 2007, 2008, and 2009, NCPI calculated the performance measures and bonus awards following the formula and methods described above. Confidential bonus reports were prepared for each treatment group teacher. 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.[5]

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. Sixteen teachers were one-time bonus winners, 17 repeated once, and 18 won bonuses in all three years. In all, 51 or 33.6 percent of the initial treatment group of 152 teachers received a bonus over the course of the experiment.

---

[3] We randomized by clusters to allow for an analysis of experimental outcomes that would be robust to efforts by treatment teachers to manipulate their student assignments (a form of system gaming). We do not present estimates based on cluster-level analyses in this paper and for that reason do not discuss it further here. For additional details and estimates that exploit this feature of the randomization scheme, please see the forthcoming longer report on POINT.

[5] This brief description of the project necessarily omits many details, including a full account of data collection activities and documentation of the variables used in the analysis below. For this additional information, see the longer forthcoming report.

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**

From an implementation standpoint, POINT was a success. This is not a trivial result, given the widespread perception that teachers are adamantly opposed to merit pay and will resist its implementation in any form. On surveys administered to participants each spring, both treatment and control teachers expressed moderately favorable views toward performance pay in general, though less so for POINT in particular. Although they became somewhat less positive over the course of the experiment, it was by no means the case that once they became familiar with the operation of the program, they turned against it en masse. The program ran smoothly. There were no complaints from teachers that they had not received the bonus they should have, and few questions about why they were not entitled to a bonus. Teachers did not question the fairness of the randomization process or the criteria used to determine bonus winners. There were no efforts to sabotage POINT that came to our attention. Names of bonus winners were not leaked to the media. Performance measures were not made public (a fear expressed by some teachers in pre-implementation focus groups).

No doubt some of the ease with which POINT ran was due to the understanding that this was an experiment intended to provide evidence on whether such performance incentives will raise achievement. Even teachers skeptical of the merits of the policy saw the worth in conducting the experiment. We believe there is an important lesson here: teachers may be more likely to cooperate with a performance pay plan if its purpose is to determine whether the policy is a sound idea than they are with plans being forced on them in the absence of such evidence and in the face of their skepticism and misgivings.

III. THREATS TO VALIDITY

In this section we examine two threats to the validity of our conclusions. First, though POINT was designed as a controlled experiment, for various reasons treatment and control groups may not have been equivalent on all relevant factors influencing student outcomes. Second, outcomes themselves are subject to manipulation, with the consequence that measured gains on standardized tests may not be valid indicators of how much students learned. We consider these in turn.

IMBALANCE BETWEEN TREATMENT AND CONTROL GROUPS

In an ideal experiment, the effect of treatment (in this case, being eligible for bonuses) can be inferred from a simple comparison of outcomes in the treatment and control groups. If the two groups are equivalent with respect to all relevant background factors, outcomes in the control group represent what would have been observed among treatment subjects in the absence of treatment. However, as with many studies involving human subjects, complicating factors can interfere with this equivalence. (1) Randomization may fail to achieve equivalence between treatment and control groups. (2) The assignment of students to teachers can be manipulated to improve treatment teachers' opportunity to earn bonuses, thus causing outcomes in the treatment group to differ from those in the control group for reasons other than instructional effectiveness. (3) Teacher attrition from the study can also make treatment and control groups non-equivalent. We begin by reviewing the extent to which each of these is a potential source of concern. We then review the evidence on whether problems of imbalance materialized.

The potential for imbalance

Problems with randomization. Though teachers were randomly assigned to treatment and control groups, imbalance can arise when the number of experimental subjects is small. The smaller the size of the groups, the greater the probability that chance can produce dissimilar groups. In POINT, treatment and control groups were reasonably well balanced overall on characteristics affecting student achievement. However, this was not the case for all subsets of participants (*e.g.*, teachers of students at a particular grade level), where small numbers become a greater problem. To estimate the effect of treatment in such subsets, it is necessarily to control for a variety of potentially confounding factors in order to remove the influence of pre-existing differences between treatment and control groups.

Assignment of students to teachers. POINT randomized participating teachers into treatment and control groups, but not their students. Because the assignment of students to teachers was controlled by the district, it is possible for principals and teachers to have manipulated the assignment process to produce classes for treatment teachers that enhanced their prospect of earning a bonus. This could involve changing the courses a teacher is assigned, if it is thought to be easier to produce gains in some courses than others. Or it might involve nothing more than removing a disruptive student from a class or transferring students out of courses in which they are not doing well. If principals received more requests of this kind from treatment teachers, or if they accommodated a greater per-

centage of requests from this group, systematic differences might have been introduced between treatment and control classes that would bias estimates of the effect of incentives.

Widespread system gaming on a scale that would reveal principals to be playing favorites among their staff seems to us unlikely, particularly on behalf of a set of teachers already benefiting from the chance to earn bonuses. For additional protection against this possibility, we took the following steps: (1) Principals were explicitly asked to run their schools during the POINT years just as they would have in the absence of an experiment; (2) Principals were not informed (by us) which of their faculty were participating in the experiment and whether they were treatment or control teachers; and (3) Participating teachers were required to sign a declaration that they would not reveal to other employees of the school system whether they had been assigned to the treatment or the control group. We also pointed out that by keeping this information to themselves, they could avoid having to answer potentially awkward questions about whether they had earned a bonus.

We are unsure how effective these efforts were. On a survey administered to POINT teachers in the spring of the experiment's third year, 72 percent of treatment teachers who were not themselves bonus winners, along with 81 percent of control teachers, indicated that they did not know whether anyone in their school won a bonus based on results in the previous year. This, in addition to limited anecdotal evidence that came our way, indicated that we were certainly not 100 percent effective in keeping the identities of treatment teachers secret. Even if principals did not know whether particular teachers were eligible for bonuses, they could have unwittingly abetted efforts to game the system by approving requests that treatment teachers were able to portray as educationally sound—for example, assigning a teacher to a course in which the teacher deemed herself more effective, or moving a struggling or disruptive student out of a particular class.

Teacher attrition. Teachers participating in POINT left the study at a very high rate, with just more than half remaining through the third year. Most of this attrition was teacher initiated, although teachers with fewer than 10 math students were dropped from the experiment. Year-by-year attrition is shown in Table 3. Note the spike in year 2 of the experiment. Some (though certainly not all) of this spike is the result of granting teachers with fewer than 10 math students in 2006-07 a one-year reprieve, with the consequence that a disproportionate number of teachers who did not meet this requirement for a second year were dropped from the experiment at the beginning of 2007-08. Substantially more control than treatment teachers left in year 2, though that was reversed somewhat in the third year.

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

Teachers dropped out for a variety of reasons, most frequently because they left the district, stopped teaching middle school mathematics—although they remained teaching in the middle schools—or moved to elementary or high schools in the district. While there were some differences between the reasons given by treatment and control teachers, they were not statistically significant.

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

Teachers who left the study tended to differ from stayers on many of the baseline variables. Teachers who dropped out by the end of the second year of the experiment were more likely to be black, less likely to be white. They tended to be somewhat younger than teachers who remained in the study all three years. These dropouts were also hired more recently, on average. They had less experience (including less prior experience outside the district), and more of them were new teachers without tenure compared to teachers who remained in the study at the end of the second year. Dropouts were more likely to have alternative certification and less likely to have professional licensure. Their pre-POINT teaching performance (as measured by an estimate of 2005-06 value added) was lower than that of retained teachers, and they had more days absent. Dropouts completed significantly more mathematics professional development credits than the teachers who stayed. Dropouts also tended to teach classes with relatively more black students and fewer white students. They were more likely to be teaching special education students. A smaller percentage of their students were in math (as one would expect, given that teachers were required to have at least 10 mathematics students to remain in the study).

Teachers who dropped out in the third year of POINT were slightly more likely to be white than previous dropouts and somewhat less likely to hold alternative certification. They tended to teach somewhat greater percentages of white students. Differences between dropout and retained teachers on these dimensions therefore diminished from year 2 to year 3 of the study.

Evidence of imbalance

All three of the foregoing—randomization with small numbers of experimental subjects, purposive assignment of students to teachers, and attrition—are potential sources of imbalance between treatment and control groups. All could cause student achievement to differ for reasons other than the responses of bonus-eligible teachers to incentives. How great were the resulting imbalances? We consider two kinds of evidence: (1) Observable differences between the characteristics of students and teachers in the treatment and control groups during POINT operation, 2006-07 through 2008-09; (2) Differences in student outcomes during the two years prior to POINT, 2004-05 and 2005-06. Differences that appeared during POINT are the most immediately germane to the question: does the control group represent a valid counterfactual for the treatment teachers? Student assignments change; differences observed during the pre-POINT years would not necessarily have continued into the POINT period. However, pre-POINT discrepancies in achievement are still of interest, given that some of these discrepancies may be caused by persistent factors for which we are imperfectly able to control. The advantage of the pre-POINT comparison is that we are not limited to comparing treatment to control groups on observable factors believed to influence achievement. All factors that affect test scores are implicitly involved in such a contrast.

Differences between treatment and control groups during POINT. Table 4 below compares treatment to control groups on a range of teacher characteristics. Teacher means are weighted by the number of students assigned to the teacher at the start of the school year.[6] These weighted background variables are very similar for treatment and control group teachers at the start of the study. The only significant difference was in the percentage of English Language Learners (ELL): treatment teachers' classes contained somewhat greater proportions of ELL students than those of control teachers. Over time, as a result of attrition, the treatment group came to have a higher proportion of students taught by female teachers and black teachers. Weighted means for the treatment group with respect to year hired, professional development credits, and days absent were significantly greater than the corresponding means for the control group in years 2 and 3. However, the differences are substantively small: half a day more of absences, a third of year in year hired. Importantly, no significant differences emerge in the variables that are arguably the most directly related to the experimental outcome: the estimate of teacher value added from the 2005-06 school year, and mean prior-year student scores in math and reading.

[6] The adjusted group mean difference was estimated by a linear regression (or logistic regression model for dichotomous outcomes) that controlled for randomization block. The adjusted differences were standardized by the square root of the pooled within group variance. Standard errors for the adjusted differences account for clustered randomization of teachers.

TABLE 4.
Standardized Adjusted Treatment Versus Control Group Mean Differences Weighted by Number of Students Taught

	Year 1	Year 2	Year 3
Teacher Demographics			
Female	0.03	0.28†	0.35*
Race			
White	-0.03	-0.14	-0.11
Black	0.08	0.23†	0.21
Year of birth	-0.18	-0.10	-0.12
Preparation and Licensure			
Undergraduate mathematics major	0.03	0.12	0.01
Undergraduate mathematics major or minor	0.15	0.25	0.22
Undergraduate mathematics credits	0.10	0.10	0.08
Highest degree			
Bachelor's only	-0.03	-0.04	-0.17
Master's only	0.18	0.16	0.26
Master's plus 30 credits or advanced degree	-0.19	-0.16	-0.11
Alternatively certified	-0.18	-0.15	-0.11
Professional licensure	-0.06	-0.04	0.03
Teaching Experience			
Year hired	-0.15	-0.17	-0.34†
Years experience	0.10	0.07	0.07
New teacher	0.09	0.14	0.10
Tenured	-0.09	-0.08	-0.08
Professional Development			
Total credits, 2005-06	-0.17	0.01	-0.07
Core subject credits, 2005-06	-0.08	0.02	0.02
Mathematics credits, 2005-06	-0.15	-0.02	0.08
Teacher Performance			
Mathematics value added, 2005-06 school year	0.08	-0.02	-0.07
Days absent, 2005-06 school year	0.11	0.29†	0.45**
Teaching Assignment, Course Description			
Percentage of students in mathematics courses	0.08	0.09	0.22†
Teaching Assignment, Student Characteristics			
Percentage white students	-0.01	0.02	0.00
Percentage black students	-0.11	-0.18	-0.12
Percentage special education students	0.00	0.04	0.01
Percentage English Language Learner students	0.22*	0.30**	0.21†
Students' average prior year TCAP reading scores[c]	-0.03	0.03	0.06
Students' average prior year TCAP mathematics scores[c]	0.04	0.11	0.14

† $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$.

More signs of imbalance are evident in grade-level versions of Table 4 (see Appendix Tables B-1 – B-4). At the grade level differences between treatment and control groups are more pronounced and appear in variables that are arguably more central to our analysis. For example, grade 6 treatment teachers had higher pre-POINT value added than controls. The reverse was true in grade 7. Because these are observable differences between the groups, we can control for them when estimating the effect of treatment. Such controls are particularly important when the analysis is done at the grade level. However, that such discrepancies are evident in observable teacher characteristics raises the possibility that treatment and control groups differ with respect to unobservable determinants of achievement as well.

Table 5 compares the students in the treatment and control groups on their mathematics achievement in the last year before entering the POINT experiment (see Figure 1 for details on the years and grades of these measurements).[7] The differences were adjusted for the randomization block and the standard errors control for the cluster random design and the nesting of students within teachers (and, in column one, teacher by grade combinations). When the comparison is over all grades (column one), treatment and control groups have very similar levels of achievement before the study. Substantially greater differences are evident when the comparison is done at the grade level, with a difference of more than a quarter of a standard deviation in favor of the treatment group in grade 5 in 2007 and an equally large difference in favor of the control group in grade 7 in 2008. These differences underscore the importance of controlling for student characteristics like prior achievement when estimating treatment effects at the grade level.

TABLE 5.
Treatment vs. Control Group Differences in Math Achievement Prior to POINT

	Grade Level				
	All	5	6	7	8
Year 1	0.05	0.27*	-0.03	-0.07	-0.09
	(0.06)	(0.10)	(0.11)	(0.13)	(0.13)
Year 2	-0.11	-0.01	-0.11	-0.27†	-0.08
	(0.07)	(0.13)	(0.13)	(0.15)	(0.15)
Year 3	-0.03	-0.02	0.00	-0.08	-0.03
	(0.07)	(0.13)	(0.12)	(0.16)	(0.13)

† $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$.

[7] The comparisons in Table 5 differ from the comparisons of students' prior achievement in Table 4 because the data in Table 5 are student level whereas the data in Table 4 are teacher level, in which averages are calculated by teacher and then weighted by grade. Due to the way these weights are calculated, the results are not equivalent to averaging over all students.

Differences in achievement of students assigned to treatment and control teachers prior to POINT. Table 5 compares the pre-POINT achievement of students assigned to the classes of participating POINT teachers during the experiment. However, it is also of interest to compare the achievement of the students assigned to treatment and control teachers in the years before the experiment, given that such discrepancies may be caused by factors persisting into the POINT years. For this comparison we include only those students who were in a teacher's classroom from at least the twentieth day of the school year to the testing date. As we will be limiting our sample to this group when we analyze outcomes under POINT, it is reasonable to employ the same restriction when asking whether outcomes differed between treatment and control groups prior to the experiment. The labels treatment and control during these years reflect the status teachers will have when the experiment starts. Thus, they are literally "future treatment" and "future control" teachers. Not all POINT participants taught middle school mathematics during these years; however, there is no reason to expect any systematic differences between the subset of treatment teachers for whom we have data in those years and their counterparts among the control group. The comparison of pre-experimental outcomes is reassuring. The differences are small and statistically insignificant in both years (-.03 in 2005 and .06 in 2006).[8,9]

This is not true of comparisons at the grade level, particularly in 2005, when there were differences of at least .2 standard deviations between mean achievement in treatment and control groups in grades 5, 7, and 8. Once again, this contrast shows the importance of adjusting statistically for imbalances between groups. When we use the same adjustments on the pre-POINT data that we employ to analyze outcomes during POINT, these differences in mean achievement disappear.[10]

ADDITIONAL ANALYSES OF PURPOSIVE ASSIGNMENT AND ATTRITION

Comparisons of the samples of treatment and control teachers are not the only evidence we have on the extent to which attrition or purposive assignment pose threats to the validity of conclusions from POINT. We now summarize some of this additional evidence.

Intra-year movement of students. If treatment teachers shed more of their low performers throughout the year, the resulting differences in performance between treatment and control groups could be mistaken for differences in instructional quality.

We have estimated equations that predict the proportion of students that "switch out" of a teacher's

[8] These comparisons control for randomization block and for students' grade level, but for nothing else. Random effects were assumed at the cluster level and the teacher level, with an uncorrelated student-level residual.
[9] TCAP scale scores have been transformed to z-scores based on student's rank-order. To remove any influence POINT may have had on the distribution of scores, the distribution of scores in penultimate pre-POINT year, 2005-06, was used for this conversion. These z-scores have substantially smaller tails than the distribution of scale scores, conforming better to the assumption of normality used both in estimation and hypothesis testing. For details on this transformation, see Section IV below.
[10] The adjustments in question are fixed effects for randomization blocks and random effects for clusters, for teachers, and for teachers by grade.

class during the course of a year. A student switches out if his last day in a teacher's class occurs before TCAP administration in the spring. Such a student will not count for purposes of determining a teacher's bonus. We find no evidence that treatment teachers behave more strategically than control teachers in this respect—the difference in switching out rates between the two groups is less than one percentage point and is far from statistically significant ($p = 0.37$).[11,12]

Treatment teachers might also behave strategically by resisting the placement of new students in their classes during the school year. Even though these students won't count against a teacher for purposes of determining bonuses, they might be viewed as diluting a teacher's effort. To investigate this behavior, we estimate a model predicting the proportion of a teacher's math students that entered the class after the twentieth day of the academic year (and whose performance therefore does not count toward the bonus). The difference between treatment and control teachers was again less than one percentage point and statistically insignificant ($p = 0.74$ for math, $p = 0.68$ for non-math).

There remains the possibility that teachers behave strategically by requesting that struggling students be taken out of their classes. Note in this regard that a struggling student is not necessarily a student with low prior year scores. As we have already remarked, there are indications that treatment teachers would have preferred to instruct such students, expecting students with low prior scores to register the greatest gains. Moreover, when we estimate the effect of incentives, we can control for students' prior scores, so that even if teachers do attempt to screen students with a particular prior history from their classes, we can control for that student characteristic when comparing treatment to control group outcomes. More troubling would be evidence that treatment teachers attempt to shed students who are doing worse in the current year than one would expect on the basis of prior history. Systematically dropping them from the classes of treatment teachers introduces a bias in our estimate of the effect of incentives on outcomes that will be hard to correct, inasmuch as it is based on information known to the classroom instructor but not to the researcher.

Fortunately we are able to test this hypothesis using data from formative assessments in mathematics. These assessments, introduced on a limited basis in 2007-08, were given to nearly all students the following year, the third year of the experiment. Three assessments were administered, one in early fall, one in late fall, and one in the spring semester. Performance on these assessments gives us an opportunity to observe what the classroom instructor could see—a student whose mathematics performance was substantially below what would have been expected on the basis of prior TCAP scores. Using data from 2008-09, we have estimated a model in which performance on the first assessment is the dependent variable. Regressors include an indicator for students that switch out. This indicator is interacted with treatment status to see if those students leaving the classes of treatment teachers have lower scores on the first assessment than do those who leave the classes of control teachers. No significant difference was found ($p = 0.49$). Nor was there a significant difference when we added a control for the prior year TCAP mathematics score ($p = 0.27$). We then repeated

[11] All of the regressions described in this section included block effects to control for the fact that we randomized teachers to treatment and control status within blocks. They also included year and grade effects. Standard errors were corrected for clustering.

[12] An analogous test for non-mathematics students had a p-value of 0.69.

this analysis, using the score on the second formative assessment as the dependent variable and including the score on the first assessment as a regressor, thereby testing whether students that appear to be on a downward trend are more likely to leave treatment classrooms than control classrooms. Once again we found no difference ($p = 0.68$ without controls for the prior TCAP score, $p = 0.92$ with them).

Changes in teacher workload. Finally, we examined several workload indicators to determine whether there were significant differences in the jobs that treatment and control teachers were doing. First, we investigated whether either group taught a greater variety of subjects, involving more preparations. We constructed a Herfindahl index of subject concentration for each teacher. For this purpose we used four broad subject indicators interacted with the four grade levels to define subjects. Thus, fifth grade science was a "subject," as was seventh grade mathematics, etc.[13] We also considered whether treatment (or control) teachers simply had more students throughout the course of the year. We measured this in two ways: as a raw count of all students that showed up in their classes, and as a weighted count, where the weight represented the portion of the school year the student spent with that teacher. We looked for differences in the proportion of students in each of the four main subject areas, and in the proportion of students at each grade level. Finally, we calculated the proportion of the school year that a teacher's students spent, on average, in that teacher's classroom. Lower values mean more movement in and out, presumably making it more difficult for the teacher to do her job. With respect to none of these variables did we find significant differences at the 5 percent level between treatment and control teachers. Depending on the measure we use, treatment teachers have two to four fewer students than do control teachers, but this difference could easily arise by chance ($p = 0.14$). Differences are small even when marginally significant. For example, treatment teachers have about two percentage points fewer social studies students ($p = 0.08$).

We did, however, find that treatment teachers were less likely to switch from the school they had been teaching in at the start of the POINT experiment to another middle school. The difference in mobility rates is six percentage points ($p = 0.01$). To the extent that it helps teachers to remain in a familiar setting, we would expect this to enhance the performance of treatment teachers vis-à-vis controls. Because this difference appears to have been induced by assignment to the treatment group, any resulting difference in outcomes could be viewed as part of the treatment effect. That is the viewpoint we adopt here, though we recognize that this does not represent "improved performance" in the sense that most advocates of pay for performance in education have in mind.

Which kinds of teachers left the study? We have conducted an extensive variable selection analysis to identify the teacher characteristics that predicted attrition from the study, testing for interaction between these variables and treatment status.[14] There is little evidence that dropping out was mod-

[13] In principle it should be possible to construct a finer measure of concentration using course codes: thus, seventh grade algebra would not be treated as the same subject as seventh grade basic mathematics. However, discrepancies and anomalies in the coding of courses made this infeasible, with some teachers apparently assigned implausibly many subjects.

[14] We also tested for interaction with teachers' gender, as exploratory analyses suggested there was a strong interaction between treatment and gender even though gender was not a significant predictor of attrition. Exploratory analyses did not suggest any other omitted interactions.

erated by experimental treatment status. Of more than 20 variables examined—including teacher gender, teacher race, educational attainment, year hired, experience, tenure status, marital status, total and mathematics professional development credits (2005-06 school year), mathematics value-added (2005-06 school year), absences (2005-06 school year), proportion white students, proportion black students, proportion special education students, proportion English Language Learners, total number of students assigned to the teacher, number of mathematics students assigned to the teachers, and students' last pre-POINT mathematics and reading scores—only gender had a significant interaction with treatment. Treatment effects were much smaller (nearly null) for male teachers than for female teachers. In short, by none of these measures is there any indication that the higher retention rate among treatment teachers was a function of teacher characteristics related to the probability of winning a bonus (experience, pre-POINT value added) or to features of a teacher's job that might have made it easier to earn a bonus (student characteristics, workload).

Teachers' attitudes about performance-based compensation and the POINT experiment could influence how they respond to the intervention. Using data from surveys administered to participants each spring, we tested whether the size of the treatment effect on the likelihood of attrition varied with the following survey constructs:

- Negative effects of POINT
- Positive perceptions of POINT
- Support for performance pay
- Extra effort for bonus
- Hours worked outside of the school day
- The teacher's estimate of his or her likelihood of earning a bonus[15]

Again we found no evidence that attrition among treatment teachers relative to control teachers was sensitive to any of these teacher measures.

Although we have found no differences between treatment and control teachers that drop out (except for gender), it is possible that winning a bonus in the first or second year of POINT will encourage teachers to stay, an effect that is obviously only possible for teachers in the treatment group. Likewise, receiving a low rating on the performance measure used by POINT to determine bonus winners might encourage a teacher to consider an alternative assignment. We tested this conjecture using data from the treatment group teachers. These teachers received reports containing their performance measures and indicating whether they had won a bonus based on student achievement in 2006-07 in September of the second year of the study. This was too late to affect their decision to continue teaching in 2007-08, but this information could have influenced their decision for year 3 of the study. For the sample of treatment group teachers that remained in the study through year 2, we fit a series of logistic regression models to test for a relationship between their POINT performance measure, whether or not they won a bonus, and the probability that they remained in the study through year 3. The first models include only the performance measure or an indicator for winning

[15] Although the survey was administered to teachers after they began participating in the experiment, there were intervention effects on these measures. Hence, we believe there is limited risk of bias from modeling with post intervention variables.

a bonus, the next models include the performance measure or an indicator for winning a bonus plus baseline teacher background variables, the next set of models include the performance measure or bonus indicators interacted with sex, our survey based measures of the *Negative effects of POINT*, *Positive perceptions of POINT*, *Support for performance pay*, *Extra effort for bonus*, *Hours worked outside of the school day*, and *each teacher's estimate of his or her likelihood of earning a bonus.*

Neither the performance measure nor the bonus status was significantly associated with the probability of attrition between the end of year 2 and the end of year 3 in any of the models. However, our sample for these analyses is small, as it is restricted to the 107 treatment group teachers who remained in the study through the second school year. Of these only 23 (21 percent) dropped out the next year.

WERE IMPROVEMENTS IN TEST SCORES ILLUSORY?[16]

The possibility that improvements in student performance are illusory poses another threat to validity (Koretz, 2002). An obvious instance arises when the performance measured by the test is not the student's own—for example, when teachers alter answer sheets or coach students during an exam. But illusory gains can also be produced by less egregious behavior—such as narrowly teaching to the test, so that improvements do not generalize beyond a particular test instrument or fail to persist when the same students are re-tested the next year (Linn, 2000). Thus, even if we should find that students of treatment teachers have outperformed students of control teachers and that there appear to be no important confounding factors, we need to consider whether the difference was real—a permanent improvement in student mastery of the test domain—as opposed to a fleeting improvement on specific test items.

One potential indication that gains are illusory is a classroom in which student gains are high relative to how those same students tested in the previous year and relative to how they test the year following (Jacob and Levitt, 2003). In contrast, if large test score gains are due to a talented teacher, the student gains are likely to have a greater permanent component, even if some regression to the mean occurs. Hence, the first indicator of illusory gains is the extent to which a classroom's mean performance in year t is unexpectedly large and the same students' mean performance in year t+1 is unexpectedly small.

To create an indicator of whether a classroom's test performance in year t is unexpectedly good (or poor), we regress the mathematics score of student i in year t in classroom c in school s on measures of prior year achievement and a set of student and teacher-level covariates.[17] Separate regression for each grade/year in the analysis—*i.e.*, 6 total regressions: grades 5, 6 and 7 x years 2007 and 2008. Classroom mean residuals are multiplied by $\sqrt{N_{tcs}}$ as an approximate correction for sampling variability. Note that it is expected that large gains in one year will be followed by smaller gains the next (regression to the mean). Thus we will be looking for outliers with respect to this phenomenon:

[16] The analysis and discussion in this section was contributed by Brian Jacob and Elias Walsh.

exceptional swings from one year to the next for the same group of students.[18]

The second indication of illusory gains is based on the pattern of student item responses, on the assumption that teachers who intentionally manipulate student tests will generate unusual patterns in item responses. Consider, for example, a teacher that erases and fills in correct responses for the final 5 questions for the first half of the students in her class. In this case, there will be an unexpectedly high correlation between the student responses on these questions. We combine four different indicators of suspicious answer strings. The first is the probability, under the hypothesis that student answers within the same classroom are uncorrelated, of the most unlikely block of identical answers given by students in the same classroom on consecutive questions. The second and third measures capture the extent to which within-classroom deviations from the most likely answer to a given item (based on responses over the entire sample) are correlated. The first of these averages such correlations over items, reflecting the overall degree of correlation on the test. The second is a measure of the variability of such correlations across items. If a teacher changes answers for multiple students on some subset of questions, the within-classroom correlation on those particular items will be extremely high while the degree of within-classroom correlation on other questions will likely be typical. This will cause the cross-question variance in correlations to be unusually large.

The fourth indicator compares the answers that students in one classroom give to other students in the system who take the identical test and get the exact same score. Questions vary significantly in difficulty. The typical student will answer most of the easy questions correctly and get most of the hard questions wrong (where easy and hard are based on how well students of similar ability do on the question). If students in a class systematically miss the easy questions while correctly answering the hard questions, this may be an indication that answers have been altered. Our overall measure of suspicious answer strings is constructed in a manner parallel to our measure of unusual test score fluctuations. Within a given grade and year, we rank classrooms on each of these four indicators, and then take the sum of squared ranks across the four measures.[19]

---

[17] Student prior achievement measures include a quadratic in prior scores in for all four core subjects (a total of 8 variables), a quadratic in two years prior scores in all subjects (a total of 8 variables), and missing value indicators for each of the 8 test scores included in the regression (a total of 8 variables). Prior test scores that are missing are set to zero so that these observations are not dropped from the regression. The student demographics, X, include dummies for male, black, Hispanic, and other race, a cubic in age, a quadratic in days suspended, a quadratic in unexcused absences, a quadratic in excused absences, binary indicators for ELL eligible, free and reduced lunch, special education status, and having multiple addresses during the current school year. The "classroom" demographics, C, include fraction male, black, Hispanic, other race, free or reduced lunch, and special education in the class, and a quadratic in class size. These are defined at the year-school-grade-teacher-course level, as close to a true classroom as the data allow us to get.

[18] The statistic we employ is constructed by ranking each classroom's average test score gains relative to all other classrooms in that same subject, grade, and year, and then transforming these ranks as follows:

(3) $$SCORE_{cst} = (rank\_base_{cst})^2 + (1\text{-}rank\_post_{cst})^2$$

where rank_basecst is the percentile rank for class c in school s in year t and rank_postcst is the percentile rank for the same group of students in year t+1. Classes with relatively big gains on this year's test and relatively small gains on next year's test will have high values of SCORE. Squaring the individual terms gives more relatively more weight to big test score gains this year and big test score declines the following year.

We combine the two aggregate indicators—SCORE and STRING—to create a single indicator for each class by year combination. Classes with "high" values on both indicators are regarded as cases in which gains may be illusory (SUSPECT = 1). Of course, the definition of "high" is arbitrary. In this analysis, we consider classrooms that score above the 90th percentile on both SCORE and STRING.[20] In order to determine whether these suspect cases were more prevalent among treatment classes, we regress this binary indicator on teacher treatment status and several covariates: a measure of the teacher's value-added in the year prior to the experiment, the average incoming math score of students in the classroom, and fixed effects for the blocks within which random assigned occurred.[21] The sample was restricted to teachers that participated in the experiment and students in grades 5, 6, and 7 in years 2007 and 2008 so that all students remaining in MNPS would have the post-test observation needed to construct the SCORE variable.

Results are displayed in Table 6 below. Treatment classrooms were no more likely than control classrooms to be identified as suspect. Coefficients on the treatment indicator are both substantively and statistically insignificant. We do find that pre-POINT teacher value added has a strong positive relationship to the dependent variable, but this is expected. Value added is a measure of teacher quality, and classrooms of effective teachers should look different by both measures: strong gains during students' year with that teacher followed by smaller gains the next year, and a greater likelihood that students in these classrooms will answer more questions the same way (correctly). Separate regressions run for each grade also fail to detect any relationship between treatment status and SUSPECT.

It is possible, of course, that illusory gains could have resulted from behavior not picked up by the measures employed here. Nonetheless, it is reassuring that there was no difference between treatment and control classrooms with respect to measures that other research has shown to detect illusory test score gains.[22]

---

[19] Specifically, the statistic is constructed as

$$STRING_{cst} = (rank\_m1_{cst})^2 + (rank\_m2_{cst})^2 + (rank\_m3_{cst})^2 + (rank\_m4_{cst})^2$$

[20] Results were unchanged using alternative cutoffs corresponding to the 80th and 95th percentiles.

[21] The value added variable is set to zero if the teacher did not have a value-added score (for example, because the teacher was newly hired or newly assigned to teach math in 2006-07). Such cases were also distinguished by a binary indicator for missing value-added scores.

[22] See Jacob and Levitt (2003) for more detail. In particular, an audit study in which a random selection of classrooms suspected of cheating (based on the measures described in this memo) were re-tested under controlled conditions several weeks after the official testing. A random sample of other classrooms (not suspected of cheating) was also re-tested. Classrooms suspected of cheating scored substantially lower on the re-test than they had on the official exam only several weeks earlier while the other classrooms scored roughly equivalent on the re-test and official exam.

TABLE 6.
Estimates of the Treatment Effect on the SUSPECT Indicator

	Dependent Variable = SUSPECT Indicator (90th Percentile Cutoff)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment			0.00	-0.00	-0.00	-0.01	0.51
			(0.01)	(0.01)	(0.01)	(0.01)	(0.37)
Pre-experiment teacher value added					0.15**	0.18**	1922.81
					(0.04)	(0.05)	(4048.99)
Missing value added					-0.03**	-0.01	0.00
					(0.01)	(0.01)	(0.03)
Pre-experiment mean math score for students in a teacher's classes					-0.02**	-0.01	0.18
					(0.01)	(0.01)	(0.17)
Teacher fixed effects	Yes	No	No	No	No	No	No
School fixed effects	No	Yes	No	No	No	No	No
Block fixed effects	No	No	No	Yes	No	Yes	Yes
F-test of joint significance of fixed effects	0.76	1.50					
p-value from F-test	0.98	0.03					
Mean of dependent variable	0.03	0.03	0.03	0.03	0.03	0.03	0.06
Number of classrooms (observations)	500	498	500	500	500	500	228
R-squared	0.38	0.04	0.00	0.05	0.04	0.09	

† $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$.
Models (1)-(6) show fixed effect or OLS regression results. Model (7) shows odds ratios from a conditional logit regression.
Standard errors clustered by teacher are in parentheses.

IV. STUDENT ACHIEVEMENT

The ultimate purpose of changing teacher compensation is to improve outcomes for students in our nation's schools. Of course, standardized test scores are only one student outcome. Others, such as attainment or workplace productivity, may be of greater interest. However, student achievement on state tests is the currency of school evaluation and of great interest to policy makers and educators. It also is readily available and proximal to the intervention. Finally, growth in achievement as measured by the state test was the basis for the bonus awards: if incentives are going to have an effect, presumably it will be most evident in scores on these exams.

TREATMENT EFFECTS

POINT ran for three years. Each successive year provided teachers additional time to make adjustments to their teaching to improve their chances of earning a bonus. With each year, treatment teachers also received more information about their performance as measured by the award metric. Hence, there is potential for the effects of the intervention to vary across years.

Effects may also differ by grade level. Students in different grades take different tests and have varying amounts of exposure to teachers in the experiment. The majority of fifth and sixth grade students are in self-contained classrooms in which teachers provided instruction in multiple subjects. This was typically not the case in grades seven and eight, when mathematics instruction is generally provided by teachers specializing in math. Also, due to the way teachers were assigned to treatment and control groups, sixth and eighth grade students in treatment (control) classes in years 2 and 3 of the experiment were likely to have had a treatment (control) teacher in the preceding year. As a result, there is variation in total years of exposure to the intervention. Sixth and eighth grade students are apt to have had multiple years of exposure if they have had any, whereas students in grade 5 always had only one year of exposure, while about half of the treatment students in grade 7 had multiple years of exposure and half only a single year. Consequently, results at different grades might be measuring different degrees of exposure to teachers eligible for bonuses.

Given these various factors, we estimate not only overall treatment effects, but also separate effects by year and by grade within year.

MODELS

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). The models are complex. 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.

As described in Section II, the POINT experiment used cluster-randomization with clusters defined

by course-groups within schools. 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.[23] 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. To ensure the accuracy of standard errors, models included teacher random effects (or teacher by year effects, when data were pooled across years) as well as teacher by grade random effects.[24] 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 improve precision and to control for differences between treatment and control groups that might have arisen for reasons other than chance, we adjust for a variety of 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. For instance, the student-level covariates for an eighth grade student in year 1 (the 2006-07 school year) were measured when the student was in seventh grade in the 2005-06 school year whereas covariates for eighth grade students in year 2 (the 2007-08 school year) and year 3 were measured when the students were in grade 6 in the 2005-06 school year and grade 5 in the 2005-06 school year, respectively. See Figure 1 for details.

[23] We do not account for classes or sections taught by different teachers because this information was not included in our data. This omission should have limited effect on our estimates since we are accounting for the teacher. Also in years 2 and 3 some teachers left their original teaching assignments and are teaching in different randomization clusters. Because such changes could be endogenous, we use the cluster at the time of the initial randomization throughout our analyses. As noted above, a few teachers were assigned to the treatment group so that every school would have at least one treatment teacher. These teachers were assigned to separate clusters since they differed from other teachers in what would have been their cluster.

[24] For reasons of computational tractability, teacher by grade random effects were omitted when data were pooled across years. This likely results in a slight understatement of true standard errors for those estimates.

FIGURE 1.
Grade Level and School Year of Covariate Measurements by Grade and Year of Study Participation and Outcome Measurements

Grade and Year of Covariate Measurement		School Year and Grade of Outcome Measurement											
		Year 1				Year 2				Year 3			
Year	Grade	5	6	7	8	5	6	7	8	5	6	7	8
2005-06	4	X					X					X	
	5		X					X					X
	6			X					X				
	7				X								
2006-07	4					X					X		
2007-08	4									X			

To account for missing covariates (*e.g.*, due to missed testing or students being new to the district) we used a pattern mixture approach where we assigned students to one of four common observation patterns of covariates and included pattern indicators and separate coefficients in the model for the covariates within each pattern.[25] All of these terms were interacted with both grade and year to account for potentially different associations between the covariates and the test score outcomes from different grades or years. The variances of the residual errors are not held constant but are allowed to vary by covariate observation pattern. This is important: residual errors of students without pre-experiment test scores are substantially more variable than those of students with such scores.

The models also included adjustment for three teacher-level covariates: an estimate of the teacher's value-added in mathematics from the year prior to the experiment, an indicator for this quantity being unobserved, and the average pre-POINT mathematics score of the students taught by each teacher in each year.

Finally, the models included teacher treatment status in one of three ways: 1) a single treatment indicator to provide an overall intervention effect; 2) treatment effects by year; and 3) treatment effects for each of the twelve grade by year cells. Separate models were fit for each of these three cases using REML estimation with the *lme* routine available in the R environment.

[25] This is a generalization of a commonly used method of dealing with missing data, in which the missing covariate is set to an arbitrary value (say, zero or the sample mean) and a dummy variable for observations with missing values is added to the model. Here a dummy variable is defined for each pattern of missing values and interacted with the covariates that determined these patterns. Observations that did not fit one of the four most common patterns of missing data were made to fit by setting some covariates to missing. A small amount of data was lost in this way at a considerable gain in computational tractability.

ANALYSIS SAMPLE

Using data from the MNPS student information system we identified all the students enrolled in middle schools during the experimental years. We also identified all the courses each of these students was enrolled in and the teacher(s) who instructed them in each course. The database for mathematics courses taught by POINT participants comprised 38,577 records and 37,130 unique student-year combinations from 25,656 unique students across the four grades and three years of the study, with data from 289 unique teachers.[26]

Some student-years occur more than once in this dataset because the student switched schools or switched mathematics teachers within the same school during the year. We restricted the data to the first record for each student in each year reflecting either their beginning-of-year assigned mathematics teacher, or their first mathematics teacher upon entering the district mid-year. This restriction left 35,625 records from 35,625 unique student-year combinations from 25,001 unique students.

Furthermore, we identified student-years where students were taught by a single mathematics teacher for 90 percent or more of the school year. We refer to these student-years as having a "stable" mathematics enrollment. Attribution of achievement outcomes to responsible instructors is clearly easier in the stable cases, compared to situations in which a student has had multiple teachers for significant portions of the year. Of all student-years linked to treatment teachers, 80.9 percent had stable enrollments, compared to 82.5 percent for control teachers. This difference was not statistically significant.[27]

Only students who took the TCAP mathematics test can be included in the estimation of the intervention effects on mathematics achievement. More than 95 percent of the student-year observations in participating teachers' classes had mathematics scores. The percentages were 95.5 percent for control teachers and 95.2 percent for treatment teachers. A small number of students tested outside their grade level were excluded. After restricting to records with on-grade mathematics test scores, our analysis dataset had 33,955 observations of 33,955 unique student-year combinations from 23,784 unique students and 288 unique teachers.

OUTCOME VARIABLES

The test score outcomes for all models are students' TCAP criterion referenced test (CRCT) scores during the experiment time period. On their natural scale, these scores have heavy tails that may invalidate normal approximations made in interpreting the model results. We therefore transformed

[26] Only 289 teachers are part of the outcomes analysis file because five teachers dropped out of the study during year 1 before student outcomes were measured.

[27] We fit a Generalized Linear Mixed Model (Raudenbush and Bryk, 2002) to test for differences between the intervention and control groups on the proportion of "stable" students. The model predicted the probability of a student being classified as stable as a function of treatment assignment and other terms to control for features of the design and clustering including random effects for the teacher and cluster.

the scores using "rank-based z-scores" to improve the plausibility of the assumption that residual disturbances are distributed normally. 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 CRCT data during 2005-2006 to create a mapping between CRCT scale 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.

Because the intervention awarded mathematics teachers bonuses primarily on the basis of their students' mathematics achievement, our primary outcome is student achievement in mathematics. Students were also tested in reading, science, and social studies. As described in Section II, these scores were factored into bonus calculations when mathematics teachers also taught these other subjects. We thus analyzed achievement in these other subjects to study possible positive or negative "spillover" effects from the primary intervention. We used rank-based z-scores for all tests, regardless of subject.

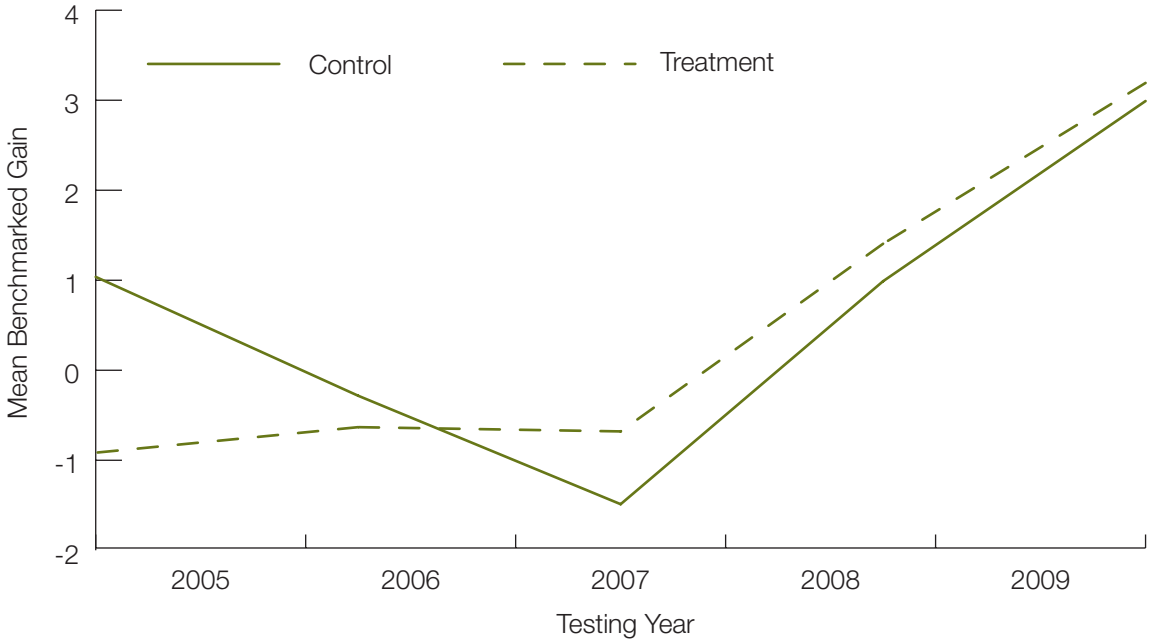
RESULTS

Before we present the estimated treatment effects described above, we present in graphical form information on achievement trends in the district. The graphs are both easy to understand and illuminating. In several respects, they prefigure findings from our more sophisticated analyses.

Figure 2 presents mean achievement from spring 2005 through spring 2009. The achievement measure is a student's year-to-year gain on the math TCAP, benchmarked against the state mean gain for students with the same previous year score. In the pre-POINT years, "treatment" and "control" refer to the status their teachers will have when the experiment starts.[28] Achievement is higher in the control group in 2005, but the gap is almost completely gone in 2006. The difference in 2007, the first year of POINT, is neither large nor statistically significant. Thereafter both groups trend upward. This may be a function of growing familiarity with a new set of tests introduced in 2004, or a response to pressures the district faced under No Child Left Behind. (A similar upward trend, not displayed in this figure, is evident among students of teachers that did not participate in POINT.) This trend also illustrates why we cannot take the large number of bonus winners in POINT as evidence that incentives worked. There were more bonus winners than expected on the basis of the district's historical performance, but this was because performance overall was rising, not because teachers in the treatment group were doing better than teachers in the control group.

[28] The mix of teachers changes over these years, but very similar patterns are obtained when the sample is restricted to teachers who taught middle school math in all five years.

FIGURE 2.
Math Achievement Trends Overall



Figures 3-6 show trends by grade level. The general upward trend is also evident at each of these grade levels. The pre-POINT differences between treatment and control groups are greater, particularly in 2005, than they were in Figure 2, where a positive difference in grade 6 partly offset negative differences in the other grades. We also note that these gaps between treatment and control groups can be quite unstable. They can vary considerably even within the pre-POINT period, suggesting that we should be wary of taking the pre-POINT gap as an indication of what would have followed in the absence of incentives. Consistent evidence of a treatment effect is evident only in grade 5: a small gap in favor of the treatment group in the first year of the experiment, widening considerably in the second year.

FIGURE 3.
Math Achievement Trends in Grade 5

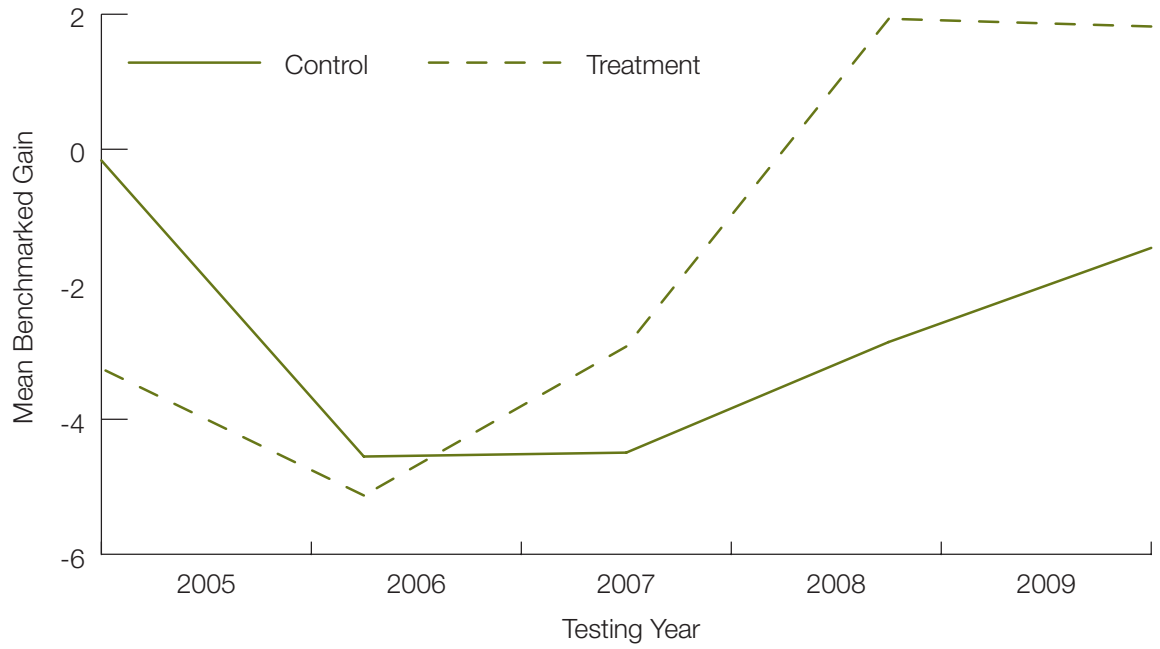


FIGURE 4.
Math Achievement Trends in Grade 6

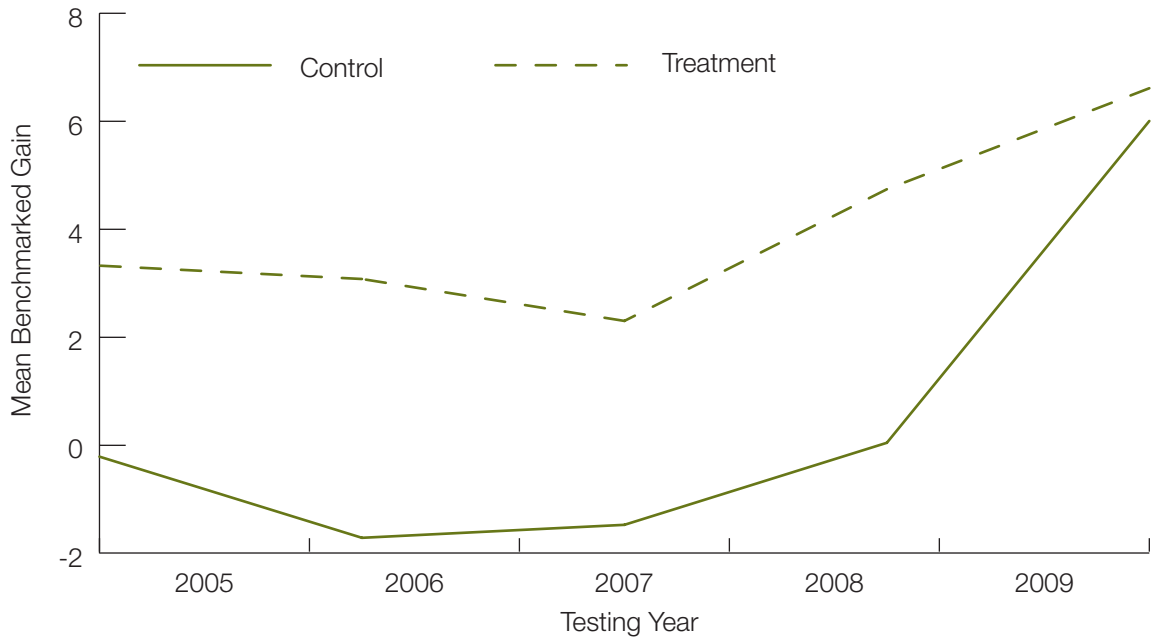


FIGURE 5.
Math Achievement Trends in Grade 7

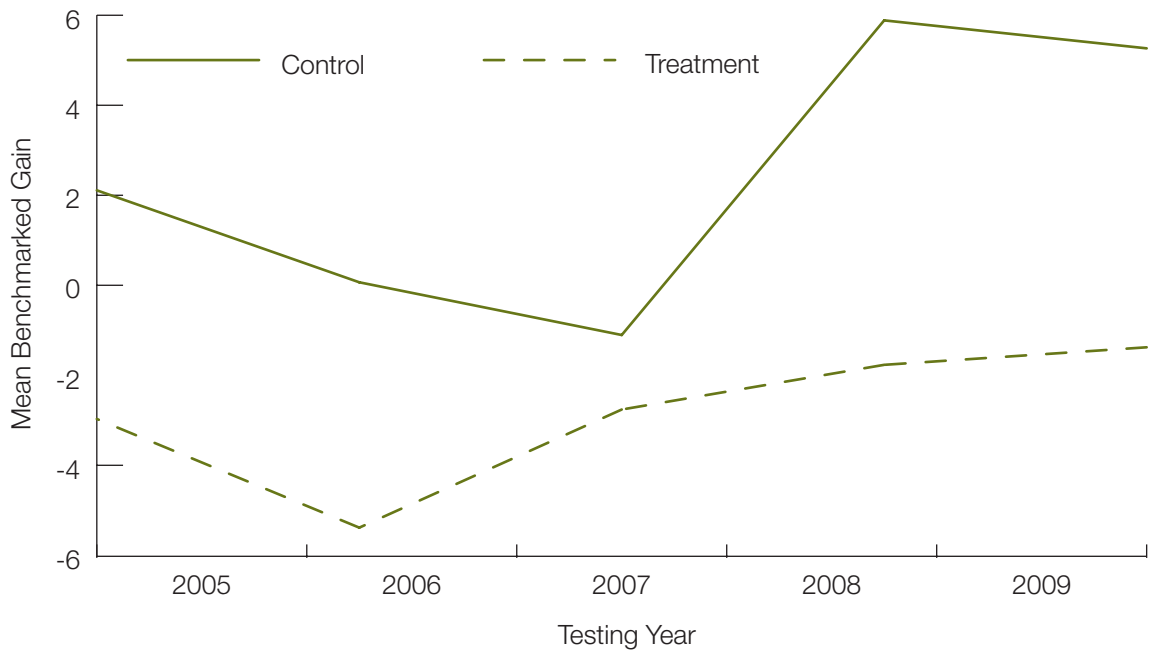
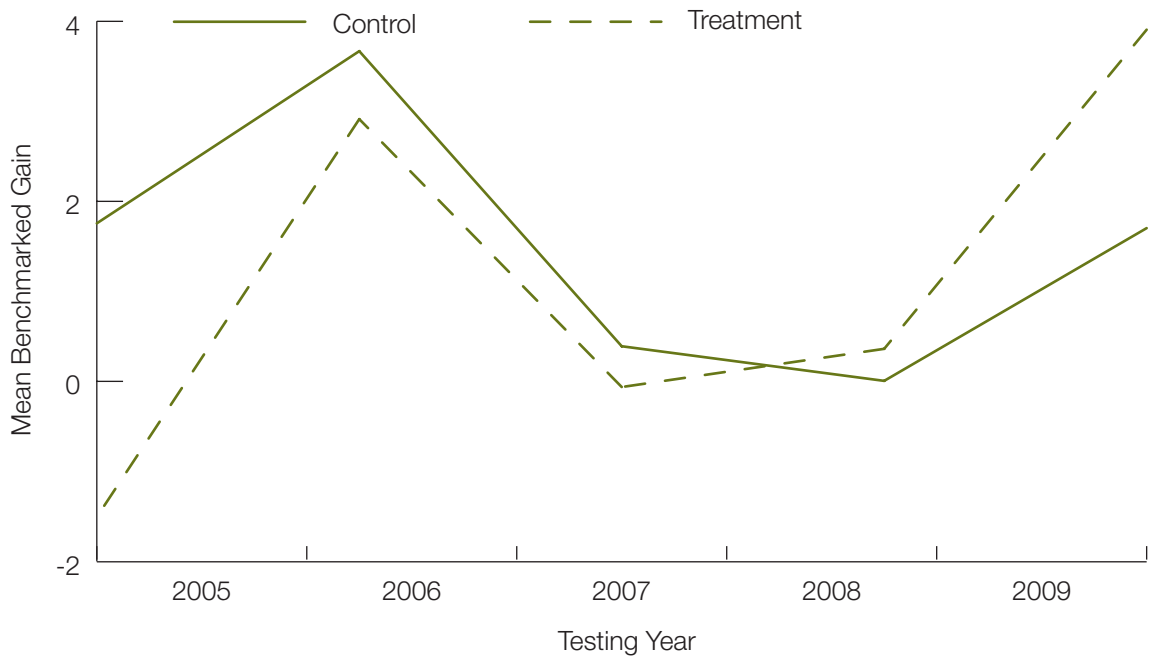


FIGURE 6.
Math Achievement Trends in Grade 8



Could spillover from the treatment group be responsible for improved performance in the control group? We find little support in the data for this hypothesis. First, it is implausible that such spillover would increase achievement as much in the control group as among teachers who were eligible for bonuses. A finer look at the evidence also argues against such a conclusion. There was variation from school to school and from grade to grade in the same school in the number of teachers in the treatment group. However, gains were no greater for control teachers who were exposed to a higher proportion of treatment teachers as colleagues. In addition, the same upward trend in mathematics scores shown in Figures 2-6 occurred in elementary schools, where the great majority of teachers had no day-to-day contact with teachers in the experiment.

Turning to our statistical analysis, we estimate an overall treatment effect across all years and grades of 0.04 with a standard error of 0.02—a small and statistically insignificant result. While this estimate is derived from the model described above, it is replicated in model-free comparisons of treatment and control group outcomes that control only for student grade level and randomization block, with random effects for clusters and teachers to ensure the accuracy of the standard errors. The difference between treatment and control groups remains small and statistically insignificant. The fact that we obtain the same results with or without the extensive set of controls for student and teacher characteristics suggests that neither attrition nor attempts to game the system disturbed the balance between treatment and control groups enough to impart a substantial upward bias to estimated treatment effects.

However, there are differences by grade level, as shown in Table 7. Results in grades 6, 7, and 8 are not significant, but those in grade 5 are, with positive effects in the second two years of the experiment amounting to 0.18 and 0.20 units on the transformed CRCT scale. Since the variance of the transformed scores is roughly one, these values are similar to effect sizes. These grade 5 treatment effects are equivalent to between one-half and two-thirds of a typical year's growth in scores on this exam. These differences are significant even if we use a Bonferroni adjustment to control for testing of multiple hypotheses on math outcomes (Steele, Torrie, and Dickey, 1997).

TABLE 7.
Estimated Treatment Effects in Mathematics

	Grade Level					
Year	All	5	6	7	8	N
1	0.03	0.06	0.01	-0.02	0.03	12311
	(0.02)	(0.04)	(0.04)	(0.05)	(0.05)	
2	0.04	0.18**	0.05	-0.01	-0.10	8878
	(0.04)	(0.06)	(0.06)	(0.07)	(0.07)	
3	0.05	0.20**	0.03	-0.05	-0.01	7812
	(0.04)	(0.08)	(0.07)	(0.09)	(0.08)	

† $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$.

Appendix Tables C-1 to C-3 present estimates of treatment effects on student achievement in reading, science, and social studies. There are no significant effects for reading. However, there are significant differences between treatment and control group students in grade 5 for both science and social studies. For both subjects, treatment group students scored significantly higher than students in the control group in year 3, with effects of 0.180 and 0.171 for science and social studies, respectively. There was also a marginally significant effect for social studies in year 2 of 0.131.

A response on the part of teachers to financial incentives is of little long-term value if their students' gains are not sustained into the future. A failure to sustain these gains may also indicate that teachers achieved these results by teaching narrowly to the test, so that the gains evaporated when students were re-tested using a different instrument. Because our only positive findings concern fifth grade teachers in years two and three of the experiment, we are able to examine longer-term effects for one cohort only: those students who were in fifth grade during the second year of the study and in sixth grade during the third year. (In the future, we will have follow-up data permitting us to extend this analysis to the 2008-09 fifth grade cohort.) To look for evidence of sustained effects, we restricted the data to the sample of students contributing to the grade 5, year 2 effect and examined the grade 6 test scores for the approximately 88 percent of these students who remained in the district and were tested during year 3. We fit a model analogous to our main outcomes model, but using grade 6 rather than grade 5 test scores. We considered models with and without controls for the grade 6 teacher status (treatment, control, and study non-participant), and considered several restrictions on the student population (all students linked to the sixth grade teacher to whom they were assigned at the beginning of the year, students that remained with the same sixth grade teacher from the twentieth day of the school year onward—"stable" students—and stable students whose sixth grade teacher was a POINT participant). Across all of these configurations and across all subjects, there were no statistically significant effects of grade 5 teacher treatment status on grade 6 outcomes. The largest estimated effect was 0.08.

To summarize, we find no overall effect, pooling across years and grades, of teacher incentive pay on mathematics achievement. Likewise, we find no overall effect by year, pooling across grades. Our only positive findings are in grade 5 in the second and third years of the experiment. These grade 5 results are also found in science and social studies in at least some years. However, the grade 5 gains do not persist into the future, at least in the cohort we have been able to check. By the end of sixth grade, it does not matter whether a student had a treatment teacher in grade 5.

SENSITIVITY TESTS

We have explored several alternative approaches to estimating treatment effects to see whether these findings hold up. To guard against model misspecification, we have re-estimated the achievement equations with an expanded set of covariates that includes the square and cross-products of all regressors. Results are virtually unchanged.

Our outcome measure for testing whether incentives raised achievement—the rank-based z-score described above—is not the same as the performance measure that determined whether teachers

qualified for a bonus. That measure was the TCAP scale score benchmarked to the average score statewide among students with the same prior year score (literally, the difference between the two, averaged over a teacher's class). Moreover, the set of students for whom we have estimated treatment effects is not precisely the set for whom teachers were accountable in POINT. Our analysis sample has included some students who are missing prior year scores and who did not count in POINT because we could not compute their benchmarked score, and it excludes some students who did count because they entered a teacher's class by the twentieth day of the school year, although they were not there from the start. Teachers were informed of these rules, and it is possible that they influenced decisions about how much assistance to give particular students. Given all this, it may be that another analysis, using the performance measure that determined bonuses and including only those students whose scores mattered, would reveal a different pattern of effects.

We have conducted an extensive set of such analyses, using three samples of students: all students that started the year with a given teacher, the set of stable students (the sample used in Table 7), and the set of students whose performance counted towards the bonus. We have estimated models with and without the set of student covariates for which we controlled in Table 7, as such covariates were not used when evaluating teacher performance for bonus purposes. We would note, however, that grade-level estimates without these controls are apt to be misleading, given that the randomization of teachers into treatment and control groups left imbalances on multiple dimensions, for which benchmarking to a single prior score is not a sufficient remedy.

Broadly speaking, results are consistent with those in Table 7. There are no significant treatment effects overall when pooling across grades and years or when estimating separate effects by year but pooling grades. We continue to find strong positive treatment effects in the second and third years of the experiment in grade 5, though not in the sample that includes students who left a treatment teacher's class in mid-year. There is also a significant positive effect in grade 5 in the first year of the experiment ($p = 0.09$) and a negative point estimate in grade 7 in the third year ($p = 0.09$), though these appear only when background controls are omitted.

WHY WAS FIFTH GRADE DIFFERENT?

In our baseline models as well as the additional models we have estimated as sensitivity tests, we have consistently found significant effects for grade 5 students in years 2 and 3. This is true of no other grade or year. Are these results spurious, or are there reasons why incentives worked in grade 5 but only in that grade?

Model misspecification. In our main analyses we have controlled for students' prior achievement, using their last pre-POINT score from the year before they entered grades taught by teachers in the study. As shown in Figure 1, for students in grades 6 to 8 in years 2 and 3, these scores date from two or three years prior to the study year, raising the possibility that the information they contain is dated, failing to capture systematic differences in student assignments to teachers reflected in more recent achievement results.

Accordingly, we have re-estimated our achievement models including the immediate prior year math score as a covariate.[29] The results (below) are qualitatively similar to those of Table 7. There are large effects for grade 5 in years 2 and 3 but not for other grades and years. While these estimates are difficult to interpret because the prior year score is a post-treatment outcome for some students and therefore endogenous, it is clear that controlling for prior achievement does not change our finding that the positive treatment effects were limited to grade 5 in the second and third years of the experiment.

TABLE 8.
Estimated Intervention Effects from Models Including Prior Year Mathematics Scores as a Covariate and Using Separate Models Per Year

	Grade Level					
Year	All	5	6	7	8	N
1	0.03	0.06	0.01	0.02	0.02	12311
	(0.02)	(0.04)	(0.04)	(0.05)	(0.05)	
2	0.05	0.17**	0.06	-0.02	-0.07	8878
	(0.04)	(0.06)	(0.06)	(0.07)	(0.06)	
3	0.04	0.18**	-0.02	-0.03	0.03	7812
	(0.04)	(0.07)	(0.06)	(0.08)	(0.07)	

† $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$.

Advantages of teaching multiple subjects in a self-contained classroom. Although housed in the middle schools, many grade 5 classes are self-contained where the teacher provides students all their instruction in core subjects and spends much of the day with these students. In some instances, the teacher will provide core instruction in two or three of the core subject areas, while students rotate to other teachers for the others. As shown in Table 9, 10 percent of grade 5 students received only mathematics instruction from the teacher who taught them mathematics; 28 percent received all of their core instruction from their mathematics teacher and an additional 30 percent received instruction in all but one core subject. The core subject most likely not to be taught by students' mathematics teachers was reading/English language arts.

The assignment of students to teachers for core instruction is very different in grades 7 and 8. By grades 7 and 8, instruction is nearly fully departmentalized with over 90 percent of students receiving no core instruction other than mathematics from their mathematics teacher. Special education students account for a sizeable fraction of the students receiving core instruction for other subjects from their mathematics teacher. Grade 6 occupies an intermediate ground: nearly a third of students receive no core instruction other than mathematics from their mathematics teachers and only 6 percent receive all their instruction in core subjects from their mathematics teachers.

TABLE 9.
Proportion of Students Taught 1, 2, 3, or 4 Core Courses by Their Mathematics Teacher, by Grade Level

	Number of Core Subjects Taught			
Grade Level	1	2	3	4
5	0.10	0.32	0.30	0.28
6	0.32	0.37	0.24	0.06
7	0.91	0.06	0.01	0.02
8	0.90	0.07	0.01	0.02

Do these differences account for the fact that we see treatment effects in grade 5 but not the other grades? When students have the same instructor for multiple subjects, that teacher has the opportunity to reallocate time from other subjects to mathematics. Two of the items on the surveys administered to POINT teachers each spring deal with instructional time in math. One asks whether a teacher increased math time for all her students, the other whether she increased math time for low-achieving students. After converting the responses to a binary indicator, we have run logistic regressions in which treatment status was interacted with the proportion of a teacher's students in grade 5, grade 6, etc. Because the focus here is on the comparison of treatment to control teachers, these equations included indicators of randomization block. The model also included random effects for cluster and teacher. The sample comprised all responses from treatment and control teachers pooled over the three POINT years. Thus teachers remaining in the experiment all three years responded three times.

There were no significant interactions of treatment with the proportion of grade 5 students (or any of the grade-level proportions) for either dependent variable. As an indirect test of this hypothesis, we replaced the grade-level proportions with three variables measuring the proportion of a teacher's math students to whom the teacher gave instruction in one other core subject, two other core subjects, and three other core subjects. Interactions with treatment were small and insignificant when the dependent variable was time for all students. However, when the dependent variable was time for low achievers, the interactions with treatment were actually negative and statistically significant for two of the three regressors.

The instructional time variable is self-reported, and it may be that these data are not of high quality. As an alternative we create a binary indicator of whether a student's math instructor also had the student for at least two other core subjects and introduce this into our student achievement model, both as a stand-alone variable and interacted with treatment status. Separate equations were estimated for each POINT year. The multiple subject indicator had a significant main effect only in 2007. The grade 5 treatment effect was unaffected by the inclusion of this variable. However, when we estimate a model in which the multiple subject indicator is interacted with treatment status by

grade, we find a significant positive coefficient on the interaction for grade 5 treatment teachers in 2008—in that year, students whose math teacher also provided instruction in at least two other core subjects had higher math scores. The approximate effect size is 0.15 ($p = 0.01$). If we take this estimate at face value, it largely accounts for the positive grade 5 treatment effect. The coefficient on the grade 5 treatment effect for students whose math teacher does not provide instruction in at least two other core subjects falls to 0.11 and is not significant ($p = 0.13$). Qualitatively similar, though weaker, effects are seen in 2009. The interaction of the multiple subjects indicator with grade 5 treatment teachers is insignificant in 2009, but the grade 5 treatment effect for students whose math teachers do not provide instruction in multiple subjects drops to 0.17 and loses some significance ($p = 0.07$). In addition, we find weak evidence that having the same math teacher in multiple subjects raises grade 6 achievement in treatment classes compared to control classes in 2007 ($p = 0.09$).

In summary, the evidence on time reallocation is mixed. According to teachers' own reports, reallocation of time to mathematics from other subjects is not the reason we have found a treatment effect in grade 5 but not in other grades. However, it appears that having the same teacher for at least three core subjects can help mathematics achievement, though the evidence is spotty. We note also that this hypothesis is not consistent with the finding that achievement in science and social studies also rose in fifth grade but not in other grades (though there may have been some spillover between mathematics instruction and student performance in other subjects involving measurement, map-reading skills, and the like).

Even if teachers did not reallocate time from other subjects to math, a self-contained class in which the same instructor is responsible for multiple subjects could be advantageous in other ways. The teacher may also know his or her students better and be better able to adapt instruction to meet the students' learning styles and needs. However, most sixth grade mathematics teachers also teach at least one other subject to their math students, affording them some of the same opportunities to get to know their students better and to reallocate time from other subjects to mathematics that fifth grade teachers enjoy. Yet estimated treatment effects in grade 6 are quite small and far from statistically significant. We conclude that while teaching largely self-contained classes may be a contributing factor to the positive response to treatment found in grade 5, it appears to be far from the entire explanation.

Attrition. In Section IV we also found several differences between treatment and control groups in teacher characteristics. Most were evident in the baseline year but others grew more pronounced as the result of teacher attrition. Across all grades these characteristics included gender, advanced degrees, and number of days absent. Among teachers who taught grade 5, there were more differences. In years 2 and 3 the treatment group tended to have a greater proportion of white teachers and a smaller share of black teachers. Treatment teachers were also more likely to hold alternative certification than teachers in the control group. Treatment teachers in grade 5 also had more years of experience on average than their control group counterparts.

To reduce the scope for attrition bias, we include additional teacher characteristics shown to be related to attrition in the model. These models yield nearly the same estimates as the models without the additional covariates, suggesting that differences on observed variables between groups due to

teacher attrition did not contribute to the observed intervention effect on fifth grade students.

We have also run analyses restricting the sample to the 148 teachers who remained in the study for all three years, again using separate models by year. These are presented in Table 10. As a comparison with Table 7 shows, restricting the sample to teachers who remained to the end of the study ("non-attriters") does not change the pattern of results across time and leads to minimal changes overall.

Table 10. Estimated Treatment Effects from Sample Restricted to Teachers Remaining in the Study for Three Years Using Separate Models Per Year

	Grade Level					
School Year	All	5	6	7	8	N
1	0.04	0.07	0.03	0.03	0.00	9349
	(0.03)	(0.05)	(0.05)	(0.06)	(0.05)	
2	0.05	0.22**	0.04	0.00	-0.09	7875
	(0.05)	(0.07)	(0.07)	(0.07)	(0.07)	
3	0.05	0.20**	0.03	-0.05	-0.01	7812
	(0.04)	(0.08)	(0.07)	(0.09)	(0.08)	

† $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$.

This analysis does not guarantee that attrition bias is not present in our estimates. If non-attriting treatment teachers systematically differ from non-attriting control teachers, the resulting selection bias will certainly affect the estimates in Table 10. However, in this case we would expect to see evidence of a systematic difference in teacher quality in every year, as there is no change over time in the sample of teachers. This is not the case. In fact, restricting to the completers has almost no effect on the year 1 grade 5 intervention effect, which continue to be small and statistically insignificant.

Changes in teacher assignments. We also investigated whether changes in teacher assignments during the study could explain the grade 5 effects. The mix of grade levels taught by individual teachers changes over time. If treatment teachers believed that teaching grade 5 students would increase their chances of earning a bonus, they may have attempted to change their teaching assignments to grade 5 in years 2 and 3 of the study, which could result in differences between the treatment and control groups. Overall 64 of the 148 stable study teachers taught at least one grade 5 student over the course of the study. There was not strong evidence of a systematic shift of treatment teachers to grade 5 over the course of the study. The percentages of control teachers who taught any grade 5 students were 34 percent, 36 percent, and 31 percent for years 1-3, respectively. The corresponding percentages for treatment teachers were 39 percent, 33 percent, and 39 percent. We also conducted

a sensitivity analysis where we removed from the sample of 148 teachers 21 teachers whose pattern of grade 5 teaching was not consistent over the course of the study, where consistency was defined as grade 5 students comprising either less than 20 percent or more than 80 percent of a teachers' mathematics students in every year. We fit annual models analogous to those used to produce Table 7 but using the restricted sample of 127 teachers. The estimated grade 5 treatment effects by year were 0.12 ($p = 0.06$), 0.17 ($p = 0.06$), and 0.12 ($p = 0.18$). Although the results do not attain the same level of statistical significance as before, this is not surprising given that the analysis removed about 1/3 of all the teachers contributing to the grade 5 effects. The grade 5 treatment effect is higher in 2007 and lower in 2009 when the sample is restricted in this way, suggesting that over the course of the experiment, somewhat less effective teachers exited from fifth grade classrooms while stronger teachers entered. However, these changes are imprecisely estimated. The other grade-level treatment effects remain insignificant.

Other hypotheses. We have considered several other explanations of the grade 5 effect. Without presenting evidence, we mention them here for the sake of completeness.[30] (1) For whatever reason (say, better alignment between the MNPS curriculum and the TCAP), grade 5 teachers start out closer to the performance threshold at which they qualify for a bonus. This encourages more of them to make an effort to improve; (2) Teacher performance, as measured by POINT, is more variable in grade 5 than in other grades. This means that simply by chance, an average grade 5 teacher is likely to get closer to the performance threshold than the average teacher in higher grades, and this in turn encourages them to put in the effort to make marginal improvements in their performance; (3) For unspecified reasons, grade 5 teachers made a greater effort to earn a bonus; and (4) The activities in which grade 5 teachers engaged in an effort to earn a bonus (professional development, collaborative instructional practices, etc.) happen to have been a more effective mix than that pursued by teachers in other grades. We examined these hypotheses using achievement data, administrative records, and surveys of POINT participants and district math mentors. None accounted, even in part, for the grade 5 difference.

SUMMARY

Overall we find no effect of teacher incentives on student achievement. Grade-level analyses show positive effects in the second and third years of the experiment, but only in grade 5. Most of the explanations we have considered for why effects would be limited to grade 5 have been rejected. One, the advantage of teaching multiple subjects in a self-contained class appears to be a factor, but accounts for only part of the grade 5 difference. Changes to teacher assignments may also have played a minor role.

[30] Contact lead author for evidence related to discussion.

V. TEACHER ATTITUDES AND EFFORT

NCPI administered surveys to all teachers participating in the POINT experiment in the spring 2007, spring 2008, and spring 2009 semesters.[31] The surveys included items on teacher attitudes, behavior and instructional practice, and school culture. Surveys asked teachers about their opportunities for professional growth—whether they sought professional development/training beyond that which was required; the content, frequency, and format of training opportunities; and whether they participated in informal learning opportunities at school (i.e., teacher networks, mentoring relationships).

Surveys also asked teachers 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.

In this report, we turn to the surveys for information on two issues: (1) how teachers' attitudes toward performance pay were affected by POINT; and (2) why we found no overall response to incentives. The first of these questions is motivated by the controversial history of merit pay in public schooling and the common perception that where it has been tried, it hasn't worked. If this is the case, one would expect that teachers' attitudes will sour over time as they observe an incentive plan in operation. The second of these questions is clearly driven by the failure of incentives in POINT to produce a systematic improvement in achievement.[32]

ATTITUDES TOWARD PERFORMANCE PAY AND POINT

POINT participants were generally supportive of the idea that more effective teachers should be paid more than less effective teachers. In this connection, it should be remembered that all participants were volunteers. A majority (64 percent) agreed with the statement: "Teachers should receive additional compensation if their students show outstanding achievement gains" in spring of 2007. Two years later this figure was virtually unchanged (66 percent). There were no significant differences across grades or between treatment and control groups.[33]

This does not mean, however, that teachers thought highly of POINT. On the whole they did not put a great deal of stock in the criteria used to determine who received bonuses. This may reflect dis-

[31] Survey response rates were extremely high, ranging from 96 to 98 percent for control teachers and from 93 to 100 percent for treatment teachers. For the most part, teachers responded to all applicable survey items.
[32] A much more extensive analysis of the survey data appears in the forthcoming longer report.
[33] The dependent variable is 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.

satisfaction with TCAP and with standardized testing more generally. In spring of 2007, before any bonus winners had been announced, 69 percent of participants disagreed with the statement: "The POINT experiment will do a good job of distinguishing effective from ineffective teachers in the treatment group." There were no significant differences between treatment and control groups or by grade. Though responses tended to become more favorable over time, in 2009 64 percent still disagreed.

Participants were evenly divided on the question of whether the method used to award bonuses was fair to all treatment teachers. Treatment teachers were somewhat more likely to agree ($p = 0.08$). However, many of those that believed the method was fair still did not think it was particularly good at identifying deserving teachers. In 2007 80 percent agreed that "The POINT experiment ignores important aspects of my performance that are not measured by test scores." This percentage was even higher (85 percent) two years later. Among treatment teachers denied a bonus, more than 80 percent disagreed with the statement: "The fact that I did not earn a bonus means I need to improve my effectiveness as a teacher."

Merit pay has often been criticized for lowering morale and reducing cooperation among teachers (Chamberlin, et al, 2002). We did not find this to be the case in POINT. In each of the three surveys, more than 80 percent of participants disagreed with the statement: "The prospect that teachers in the POINT treatment group can earn a bonus discourages staff in the school from working together." In 2007 90 percent disagreed with the statement: "I have noticed increased resentment among teachers since the start of the POINT experiment." The proportion of teachers agreeing rose over time, but only slightly: in 2009, the percentage in disagreement was still 84 percent. On both items, teachers in the treatment group were somewhat more likely to disagree than teachers in the control group.

To summarize, participating teachers were generally supportive of the concept of extra pay for better teachers. They did not come away from their experience in POINT thinking it had harmed their schools. But by and large, they did not endorse the notion that bonus recipients were better teachers or that failing to earn a bonus ought to lead one to consider way to improve performance. In short, most participants did not appear to buy in to the criteria used by POINT to determine who was teaching effectively. This should be kept in mind when we consider why performance incentives failed to produce greater learning gains.

HOW TEACHERS RESPONDED TO POINT

If we accept at face value teachers' survey responses, it should not be a surprise that mathematics achievement did not increase among students of teachers eligible for bonuses. Most teachers claim to have made few if any changes in response to POINT. In each year, more than 80 percent of treatment group teachers agreed with the statement: "I was already working as effectively as I could before the implementation of POINT, so the experiment will not affect my work." Most disagreed with the statement: "I have altered my instructional practices as a result of the POINT experiment," though there was some change over time, with the percentage in disagreement falling from 87 percent in 2007 to 76 percent in 2009.

Some caution is required in interpreting these responses. Teachers may have been reluctant to agree with the first of these statements, as it carries the implication that they were not working as effectively as they could before the experiment. Some teachers who said POINT had no effect on their work nevertheless made changes to their classroom practices over the course of the project, though these may have been changes that would have occurred anyway. The surveys asked POINT participants about a wide range of teacher behavior and instructional practices. What do they tell us?

The survey items we examined are shown in Figure 7 below. They fall into the following categories: (1) Alignment of instructional with MNPS standards; (2) Use of instructional time; (3) Development of test-taking skills; (4) Use of particular teaching methods; (5) Use of test scores to inform and shape instruction; and (6) Collaboration with other math teachers.

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)]

In addition to teacher surveys, we turned to two other sources of data. From administrative records, we obtained various indicators of teacher involvement in professional development: (1) Total professional development credit hours earned during the year; (2) Professional development credits in core academic subjects; (3) Math professional development credits; (4) How frequently a teacher was a 'no-show' in a professional development workshop for which she had registered; (5) How frequently a teacher was a late drop from a professional development workshop; and (6) The number of times a teacher logged into Edusoft, the platform through which the district administers formative assessments (with the number of logins an indicator of the frequency with which an instructor used the assessment tools and reports available on the Edusoft website). Finally, using surveys of the district's math mentors, we constructed an index of the frequency and duration of teachers' contacts with mentors.[36]

We regressed each of these variables on the proportion of a teacher's students at each grade level and on treatment status. We used OLS when the dependent variable was continuous, probit when it was binary, and ordered probit in the remaining cases. All models included randomization block fixed effects and random effects at the level of the randomization cluster.

There are few survey items on which we have found a significant difference between the responses of treatment teachers and control teachers. (We note all contrasts with p values less than 0.15.) Treatment teachers were more likely to respond that they aligned their mathematics instruction with MNPS standards ($p = 0.11$). They spent less time re-teaching topics or skills based on students' performance on classroom tests ($p = 0.04$). They spent more time having students answer items similar to those on the TCAP ($p = 0.09$) and using other TCAP-specific preparation materials ($p = 0.02$). The only other significant differences were in collaborative activities, with treatment teachers replying that they collaborated more on virtually every measured dimension. Data from administrative records and from surveys administered to the district's math mentors also show few differences between treatment and control groups. Although treatment teachers completed more hours of professional development in core academic subjects, the difference was small (0.14 credit hours when the sample mean was 28) and only marginally significant ($p = 0.12$). Moreover, there was no discernible difference in professional development completed in mathematics. Likewise, treatment teachers had no more overall contact with the district's math mentors than teachers in the control group.

Finally, where treatment teachers did differ from controls, we do not find the differences for the most part associated with higher levels of student achievement. We have introduced each of the preceding dependent variables into the student achievement equations as an additional explanatory

[36] Mentors were asked how frequently they had worked with a teacher in each of six skill areas. Responses were never, once or twice a semester, once or twice a month (plus indicators of more frequent contact that were never or almost never selected). They were also asked the average duration of sessions: < 15 minutes, 15 minutes, 30 minutes, 45 minutes, 1 hour, more than 1 hour. To construct the index we treated once or twice a semester as a baseline (=1). Relative to this, a response of once or twice a month (or still more often) was assigned a value of 3. "Never" was 0, of course. We treated <15 minutes as equal to 15 minutes and >1 hour as equal to 1 hour, and multiplied the revised duration values by the three frequency values (0, 1, or 3). We then summed this over the 6 skill areas and across all mentors that worked with a given teacher to obtain a crude index of how much contact a teacher had with the math mentors.

variable. Virtually none had any discernible relationship to mathematics achievement. The only exceptions were two of the collaborative activities: teachers that acted as mentors or coaches had better results, as did teachers that observed the work of others in the classroom, though the latter is only marginally significant ($p = 0.14$). 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.

In summary, 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 appears to have had little effect on what these teachers did.

VI. SUMMARY AND DISCUSSION OF FINDINGS

Implementation. In terms of implementation, POINT was a success. At the district's request, participation was voluntary. Given the controversial history of performance incentives in education, we had some concern that a sufficient number of teachers would choose to participate. More than 70 percent of eligible teachers volunteered, exceeding our target. Only one teacher asked to be removed from the study. Responses to teacher surveys administered in the spring of each year ranged between 92 percent and 100 percent. Through the three years that the project ran, it enjoyed the support of the district, the teachers union, and community groups. Bonuses were paid as promised. Because focus groups conducted prior to the project indicated that teachers were concerned about adverse consequences if the list of bonus winners were publicized, we promised that to the extent possible we would maintain confidentiality about who participated and who earned bonuses. We were able to keep this promise, despite paying out nearly $1.3 million in bonuses. POINT enjoyed a relatively low profile in the community. In contrast to the experience with performance pay elsewhere, no list of winners appeared in the local press, nor did irate teachers seek outlets in the media to express dissatisfaction with their treatment.

Probably the greatest problem from the standpoint of implementation was the high rate of attrition from the project. POINT began with 296 participating teachers. By the end of the third year, only 148 remained. Attrition occurred for a variety of reasons: teachers left the district, they switched to administrative jobs, they took positions in elementary schools or high schools, they ceased teaching math, or the number of math students they had fell below the threshold of ten. Cumulative attrition by the end of the project was higher among control teachers than treatment teachers (55 percent versus 45 percent), though the difference was only marginally statistically significant ($p = 0.12$). The experiment therefore provides weak evidence that the opportunity to earn a bonus reduces teacher attrition, though attrition from the study is not necessarily the kind of attrition that concerns policy makers. However, there is no evidence that being eligible for a bonus had a differential impact by teacher quality, as would be the case if being assigned to the treatment group made more effective teachers particularly likely to stay.

Outcomes. Of greatest interest is the impact of performance incentives on student achievement, the central question the study was designed to address. Our principal findings can be summarized as follows:

- With respect to test scores in mathematics, 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.

- 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. 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.

- 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 changes in student achievement as measured by TCAP.

These findings raise further questions. Why did we find no effect on most students? Why was there an effect in grade 5?

We have considered three explanations for the absence of an effect: (1) The incentives were poorly designed. Bonuses were either too small or the prospect of obtaining a bonus too remote for teachers to change their instructional practices; (2) Teachers made little or no attempt to improve, either because they believed they were already doing the best job of which they were capable, or because they did not know what else to try; and (3) Teachers did attempt to improve their performance, but the measures they took were not effective.

The first explanation does not appear to be credible. Most treatment teachers were within range of a bonus, in the sense that they would have qualified for a bonus had their students answered correctly 2-3 more questions (on a mathematics test of approximately 55 items). A third of the teachers assigned to the treatment group actually did earn a bonus at some point during the project—despite the fact that 45 percent of treatment teachers limited their opportunity to do so by dropping out before the experiment ended. Responses to teacher surveys confirmed that the POINT bonuses got their attention. More than 70 percent of treatment teachers agreed that they had a strong desire to earn a bonus. The size of the bonuses—$5,000, $10,000, and $15,000—relative to base salaries in the district makes it extremely unlikely that teachers viewed them as not worth the bother.

These surveys contain much stronger evidence in support of the second explanation. 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. Teachers' responses to such questions are not perfectly reliable indicators of their behavior: there may have been some reluctance to disagree with the first statement, thereby indicating that a teacher was not already working as effectively as she could. And indeed, responses to survey items dealing with specific in-

structional methods reveal that some teachers claiming to have done nothing different in response to POINT did change classroom practices over the course of the project. Nonetheless, on the whole the availability of bonuses does not appear to have inspired participating teachers to have done very much that they would not have done otherwise. On a wide range of questions about teaching practices, there are few to which treatment and control teachers gave consistently different answers in all years of the project. Nor were there significant differences between the two groups in the number of teachers reporting that they increased time spent on mathematics, either for all students or for low achievers in particular.

The conclusion that eligibility for bonuses did not induce teachers to make substantial changes to their instructional practices or their effort is corroborated by data from administrative records and surveys administered to the district's math mentors. Although treatment teachers completed more hours of professional development in core academic subjects, the difference was small (0.14 credit hours when the sample mean was 28). Moreover, there was no discernible difference in professional development in mathematics. Likewise, treatment teachers had no more overall contact with the district's math mentors than teachers in the control group.

We are not able to say as much about the third hypothesis. Analysis of survey data on instructional methods is problematic. First are the obvious limitations of self-reported data. Second, while information was sought on practices that have been deemed ways of improving instructional effectiveness (with varying degrees of supporting evidence), choices of teaching method are affected by teachers' perceptions of student needs and their own strengths and weaknesses. That a given teacher does or does not adopt a particular practice tells us little about whether that teacher is making the right instructional decisions for her circumstances. Finally, success in using any teaching method depends on implementation. We cannot tell from survey responses whether teachers using particular methods did so in a way that would enhance their effectiveness.

With these caveats in mind, what can we say about the way treatment teachers responded? Treatment teachers differed from control in two major respects: (1) they were more likely to report that they collaborated with other teachers (planning, reviewing student test results, coaching and being coached or observed); and (2) they were more likely to say that they aligned their instruction with the district's mathematics standards and spent classroom time on test preparation, including the taking of practice exams modeled on the TCAP. When we examine 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, though the latter is only marginally significant ($p = 0.14$). 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.

To conclude, 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, though the

latter inference must be carefully qualified for the reasons indicated above. This might not be disturbing if it were in fact true, as 80 percent of project participants claimed, that they were already teaching as effectively as they could. However, that claim is called into question by the substantial improvement in mathematics achievement across all middle school classrooms over the duration of the project, particularly in the final year when the district faced the threat of state takeover under NCLB. Under that threat, test scores improved. Yet they did not in response to monetary incentives.

The overall negative conclusion is tempered by the finding of a positive response in fifth grade during the second and third years of the experiment. What made fifth grade the exception? It might be explained by the fact that math teachers in fifth grade normally have the same set of students for multiple subjects, giving them the opportunity to increase time spent on math at the expense of other subjects in a way that is not possible in grades 7 and 8, where math teachers typically specialize. While we found limited support for this hypothesis, it did not appear to be a factor in all years. Nor did tests scores fall in other subjects; in fact, they rose in fifth grade science and social studies. Other possibilities remain conjectural. Because fifth grade teachers have fewer students for longer periods, it may be that they achieve better understanding of their students and enjoy greater rapport with them, both of which might contribute to higher achievement when the stakes are raised for teachers. Fifth graders are the youngest students in middle school. Not yet adolescents, they may have been more responsive to attempts by their teachers to cajole from them greater effort.

Finally, while the positive fifth grade effect might seem to be "good news," the effect did not last. By the end of sixth grade it did not matter whether a student's fifth grade math teacher had been in the treatment group or the control group. If not spurious, the fifth grade effect seems at best short-lived (though we have not yet been able to test this hypothesis for the third year of the project), possibly a sign that it was achieved by narrowly teaching to the test or test-prep activities that had no enduring impact on achievement.

Teacher surveys obtained information about teachers' perceptions and attitudes as well as their instructional practices. Some of what we learned is encouraging (if one believes there is a role for performance incentives in education). Teachers on the whole had a moderately positive attitude toward POINT, though it declined slightly over time. Failing to win a bonus did not sour treatment teachers; if anything, they seemed to put forth somewhat greater effort the following year, as measured by the time they put in outside regular school hours. Perceptions of teacher collegiality were not adversely affected by the experiment. The generally positive view of POINT may be due to the fact that teachers were not competing with one another for bonuses. It may also reflect the fact that the project was clearly understood to be an experiment in which even teachers opposed to incentives of this kind could see value.

In sum, 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. 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. All of these may have contributed to a tendency for POINT to fade into the background. By contrast, an intense, high-profile effort to improve test scores to avoid NCLB sanctions appears to have accomplished considerably more. This is not to say that performance incentives would yield greater results if introduced in a similarly stressful manner. Certainly we would expect adverse consequences to multiply. Yet POINT provides little support for the view that it is sufficient to tie teacher compensation to test scores, stand back, and wait for good things to happen.

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.

This page intentionally left blank.

REFERENCES

Bryk, A.S., & Schneider, B. (2002). Trust in schools: A core resource for improvement. New York: Russell Sage Foundation.

Chamberlin, R., Wragg, T., Haynes, G., & Wragg, C. (2002). Performance-related pay and the teaching profession: a review of the literature. Research Papers in Education, 17(1), 31-49.

Harris, D.N., & Sass, T.R. (2006). "Value-Added Models and the Measurement of Teacher Quality." Unpublished. Tallahassee, FL: Florida State University.

Jacob, B., & Levitt, S. (2003). "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." Quarterly Journal of Economics, 118(3), 843-877.

Kelley, C., Heneman III, H., & Milanowski, A. (2002). Teacher motivation and school-based performance awards. Educational Administration Quarterly, 38(3): 372-401.

Kellor, E. M. (2003). Catching up with the Vaughn express: Four years of performance pay and standards-based teacher evaluation. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research, Consortium for Policy Research in Education.

Koretz, D.M. (2002).Limitations in the use of achievement tests as measures of educators' productivity. Journal of Human Resources, 37(4), 752-777.

Linn, R.L. (2000). Assessments and accountability. Educational Researcher, 29(2), 4–16.

Lockwood, J.R., McCaffrey, D.F., Mariano, L.T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. Journal of Educational and Behavioral Statistics, 32(2), 125-150.

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. Journal of Educational and Behavioral Statistics, 29(1), 67-101.

Milanowski, A. (2000). School-based performance award programs and teacher motivation, Journal of Education Finance, 25(4), 517–544.

Milanowski, A., & Gallagher, A. (2000). Vaughn next century learning center performance pay survey school report. Unpublished manuscript, University of Wisconsin-Madison.

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

Rowan, B., Correnti, R., & Miller, R.J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects Study of Elementary Schools. Teachers College Record, 104(8), 1525-1567.

Sanders, W., Saxton, A., & Horn, B. (1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to educational assessment, In Millman, J. (Ed.), Grading teachers, grading schools: Is student achievement a valid evaluation measure? (pp. 137-162). Thousand Oaks, CA: Corwin Press, Inc.

Steel, R.G.D., Torrie, J.H. & Dickey, D.A. (1997) Principles and Procedures of Statistics, 3rd. ed. McGraw Hill.

APPENDIX A:
WERE POINT PERFORMANCE TARGETS UNREALISTIC FOR MOST TEACHERS?

This page intentionally left blank.

POINT tests whether large bonuses linked to student test scores motivate teachers in some unspecified set of ways to raise those scores. There are, of course, other ways to design incentives. Teachers might have been offered smaller amounts for incremental improvements over their own past results. In designing POINT as we did, we sought to test one model for radical reform of teacher compensation, in which high rewards are offered for excellent teaching, rather than a set of modest incentives that would yield at best modest results.

However, it may be wondered whether we set the bar at a height where few teachers would be motivated to change their instructional practices or raise their level of effort—that most teachers would regard the performance targets as unattainable no matter what they did, while a smaller number with strong past performance would also have little reason to make changes, but for the opposite reason: they could win a bonus without doing anything different. If the great majority of teachers fall into one of these two groups, only a few on the margin (or "the bubble") have much incentive to do anything differently.

To address this concern, we examine achievement in the two years immediately before POINT, asking how many of the teachers that participated in POINT would have earned a bonus in one of those years had the same rules been in effect then. Focusing on the teachers for whom we have results in both years, we find 25 were "winners" in 2005 but not 2006, 18 were "winners" in 2006 but not 2005, and 23 would have won in both years, for a total of 66 who won in at least one year, compared to 94 that won in neither. Clearly it is not the case that only a small minority of teachers had a realistic chance of winning, as 41 percent of the teachers observed in both years actually did qualify at least once.

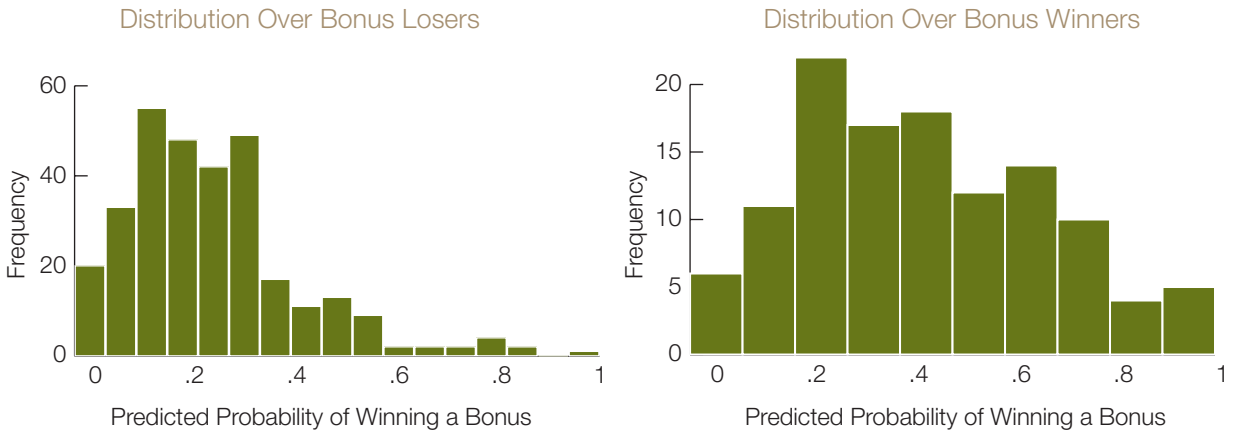
We conduct the same calculation for teachers in the control group during POINT. (Like teachers during the pre-POINT years, control teachers were not eligible for bonuses, so that this tabulation gives us the incidence of rewards assuming a "historical" level of effort.) 30 of the teachers observed in both years "won" at least once, compared to 59 that did not. Of those 59, an additional 8 were "winners" in 2009. Thus, among control teachers that remained in POINT through the final year of the experiment, 38 met the bonus performance target at least once, versus 51 that did not, or 43 percent versus 57 percent.

These tabulations overlook those who failed to qualify but came close. For a more nuanced examination of this question, we employ the mean benchmarked score, which, as described above, determined whether a teacher qualified for a bonus. Using a sample of all future participants in the pre-POINT years and the control teachers during the POINT years, we regress this performance measure on its lagged value, obtaining a predicted performance measure (EXPECTED PERFORMANCE)—what a teacher might reasonably have expected her students to do in the coming year, based on the year just completed.[37] We then use this prediction as the independent variable in a logistic regression in which the dependent variable is a binary indicator for whether the teacher

[37] Note that this prediction incorporates regression to the mean.

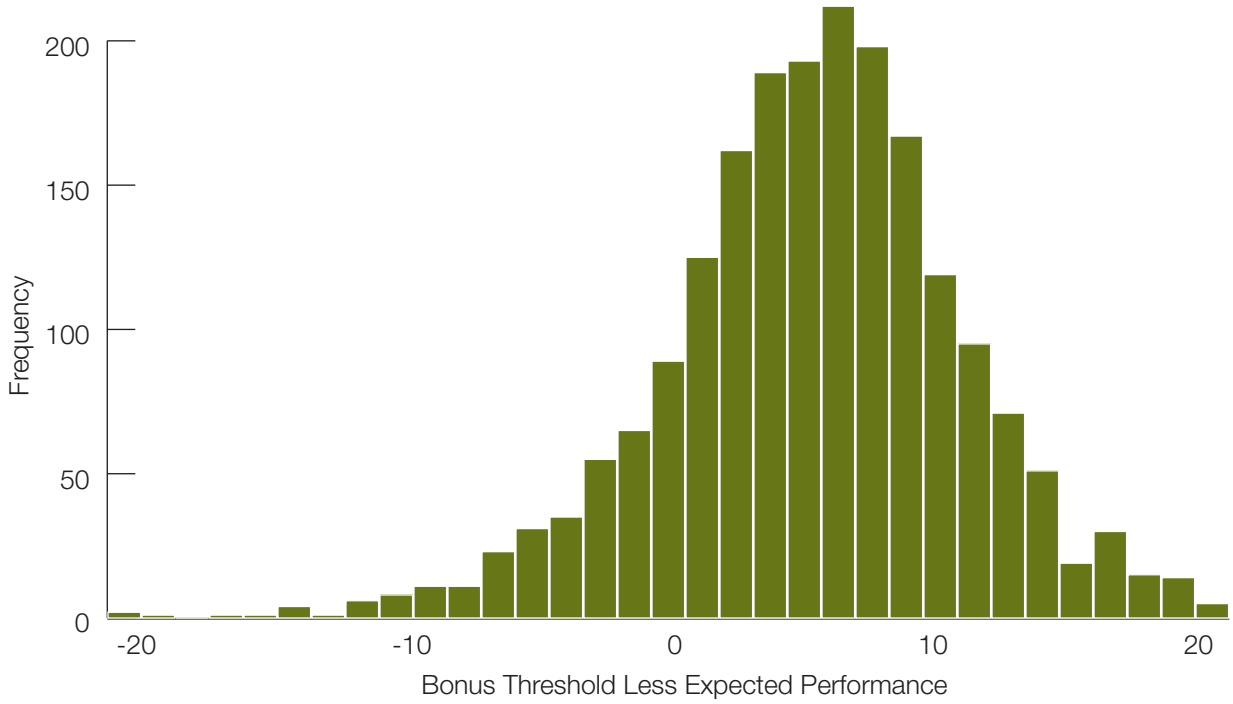
qualifies for a bonus in the coming year. Not surprisingly, EXPECTED PERFORMANCE is a strongly significant predictor of the probability of earning a bonus in the coming year, as teachers that have done well in the past tend to do well in the future. Figure A-1 contains histograms of the predicted probability of winning a bonus—the probabilities predicted from the logistic regression. There are substantial differences between losers and winners in the predicted probability of winning a bonus. Virtually all of the losers have predicted probabilities below 50 percent; only about half of the winners are this low. However, there are very few winners whose predicted probability of earning a bonus was so high that a marginal improvement in performance would have had no payoff.

FIGURE A-1.
Probability of Winning a Bonus



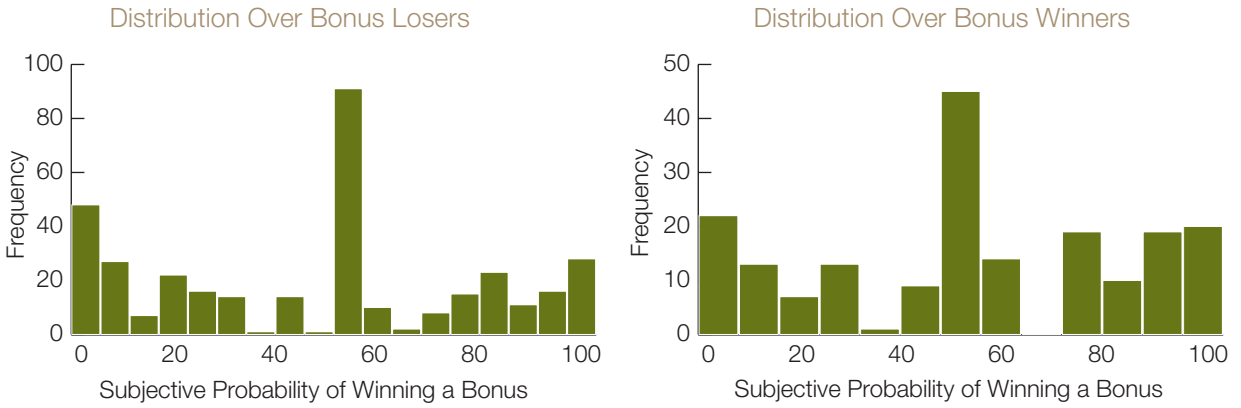
How much did teachers with low probabilities in Figure A-1 have to improve to obtain a bonus? One way to assess whether bonus thresholds appeared out of reach is by the improvement in student scores needed for a teacher to reach the minimum bonus level of 3.6. This is calculated as 3.6 minus EXPECTED PERFORMANCE. The distribution of the resulting values is shown in Figure A-2 (a small number of teachers with values below -20 or above 20 are omitted from the graph). Negative values represent teachers whose EXPECTED PERFORMANCE already exceeded the minimum threshold for earning a bonus. Most teachers are in the positive range. Of this group, half would qualify for a bonus if they could raise their students' performance by 6 scale score points—that is, if on average students could answer 2-3 more test questions correctly (on a test of approximately 55 items in total). If this improvement is more than most teachers could effect on their own, it would appear that some combination of greater effort and good luck was often required to reach the bonus level. However, such combinations were not unusual—as Figure A-1 shows.

FIGURE A-2.
Required Improvement to Earn a Bonus



The preceding analysis has used data on teachers' performance measures to calculate how likely teachers were to win bonuses as a function of EXPECTED PERFORMANCE. As an alternative, we can use teachers' subjective probabilities of winning bonuses, as reported in surveys conducted each spring during POINT. Arguably, teachers' beliefs are more important than a statistical analysis of historical data in understanding whether the design of POINT provided them with sufficient incentive to modify their practices. Figure A-3 depicts the distribution of these subjective probabilities over bonus losers and winners. Compared to the previous graphs, losers and winners look remarkably similar. Subjective probabilities bear almost no relationship to whether teachers actually won or lost bonuses. Teachers that thought they had almost no chance of earning a bonus are represented about equally in both groups, as are teachers that believed they were a sure thing. In both the modal value is 50 percent.

FIGURE A-3.
Subjective Probabilities of Winning a Bonus



To conclude, it is not the case that teachers mainly fell into two groups: those for whom the bonus thresholds were hopelessly out of reach, and those who were assured of reaching them without doing anything extra. Chance appears to have had a lot to do in determining who qualified for a bonus. Many bonus "winners" had predicted probabilities between .2 and .4. (Recall that this is an analysis of notional winners who were not actually responding to incentives, so these are not individuals with low ex ante probabilities who worked their way to a higher level in order to earn a bonus.) Thus, bonus thresholds should have appeared within reach of most teachers, as long as they understood that luck was going to play a role in determining whether they actually got there.

APPENDIX B:
GRADE-LEVEL COMPARISONS OF TREATMENT AND CONTROL GROUPS

This page intentionally left blank.

TABLE B-1.
Standardized Adjusted Treatment Versus Control Group Mean Differences Weighted by Number of Grade 5 Students Taught

	Year 1	Year 2	Year 3
Teacher Demographics			
Female	-0.11	0.29	0.14
Race			
White	0.12	0.57*	0.51
Black	-0.04	-0.49†	-0.42
Year of birth	0.04	-0.11	0.03
Preparation and Licensure			
Undergraduate mathematics major	-0.31†	-0.41	-0.40
Undergraduate mathematics major or minor	-0.17	-0.39	-0.35
Undergraduate mathematics credits	-0.05	-0.16	-0.14
Highest degree			
Bachelor's only	-0.11	-0.32	-0.39
Master's only	-0.14	-0.23	-0.14
Master's plus 30 credits or advanced degree	0.32†	0.65*	0.64
Alternatively certified	0.17	0.38†	0.10
Professional licensure	-0.00	-0.01	0.22
Teaching Experience			
Year hired	-0.05	-0.04	0.05
Years experience	0.14	0.48†	-0.01
New teacher	0.11	-0.20	-0.29
Tenured	0.05	-0.03	0.15
Professional Development			
Total credits, 2005-06	-0.03	-0.07	-0.21
Core subject credits, 2005-06	0.03	0.03	-0.08
Mathematics credits, 2005-06	0.13	0.10	0.12
Teacher Performance			
Mathematics value added, 2005-06 school year	-0.34	0.23	0.10
Days absent, 2005-06 school year	0.00	0.33†	0.08
Teaching Assignment, Course Description			
Percentage of students in mathematics courses	0.19	0.39†	0.64*
Teaching Assignment, Student Characteristics			
Percentage white students	0.34	0.45	0.20
Percentage black students	-0.52*	-0.58*	-0.53†
Percentage special education students	-0.21**	-0.26**	-0.14
Percentage English Language Learner students	0.31	0.25	0.48
Students' average prior year TCAP reading scores[c]	0.20	0.30	0.06
Students' average prior year TCAP mathematics scores[c]	0.25	0.34	0.09

† $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$.

TABLE B-2.
Standardized Adjusted Treatment Versus Control Group Mean Differences Weighted by Number of Grade 6 Students Taught

	Year 1	Year 2	Year 3
Teacher Demographics			
Female	-0.31	0.06	0.31
Race			
White	0.00	-0.04	-0.14
Black	-0.00	0.04	0.14
Year of birth	-0.30	-0.16	-0.11
Preparation and Licensure			
Undergraduate mathematics major	-0.40†	-0.62†	-0.00
Undergraduate mathematics major or minor	-0.42†	-0.62†	-0.00
Undergraduate mathematics credits	-0.00	-0.37	0.24
Highest degree			
Bachelor's only	-0.54*	-0.48	-0.77*
Master's only	0.14	0.30	0.48
Master's plus 30 credits or advanced degree	0.73**	0.34	0.45
Alternatively certified	-0.17	-0.19	-0.36
Professional licensure	0.08	-0.25	-0.02
Teaching Experience			
Year hired	-0.15	0.03	-0.13
Years experience	0.32	0.07	0.24
New teacher	-0.31	0.01	-0.05
Tenured	0.14	-0.10	0.04
Professional Development			
Total credits, 2005-06	-0.16	-0.09	-0.16
Core subject credits, 2005-06	0.01	0.03	-0.01
Mathematics credits, 2005-06	-0.15	0.17	0.02
Teacher Performance			
Mathematics value added, 2005-06 school year	0.60**	0.22	0.30
Days absent, 2005-06 school year	0.05	0.38	0.66*
Teaching Assignment, Course Description			
Percentage of students in mathematics courses	0.07	0.44*	0.57*
Teaching Assignment, Student Characteristics			
Percentage white students	0.19	0.33	0.27
Percentage black students	-0.21	-0.48†	-0.14
Percentage special education students	0.09	0.06	0.11
Percentage English Language Learner students	0.21	0.29†	-0.23
Students' average prior year TCAP reading scores[c]	-0.05	-0.03	0.07
Students' average prior year TCAP mathematics scores[c]	0.00	0.12	0.16

† $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$.

TABLE B-3.
Standardized Adjusted Treatment Versus Control Group Mean Differences Weighted by Number of Grade 7 Students Taught

	Year 1	Year 2	Year 3
Teacher Demographics			
Female	-0.28	-0.34	-0.35
Race			
White	-0.28	-1.00**	-0.80
Black	0.40†	1.48**	1.44*
Year of birth	-0.28	0.11	-0.01
Preparation and Licensure			
Undergraduate mathematics major	-0.13	-0.05	-0.27
Undergraduate mathematics major or minor	-0.06	0.11	0.03
Undergraduate mathematics credits	-0.13	-0.19	-0.71†
Highest degree			
Bachelor's only	0.20	0.37*	-0.00
Master's only	0.31	0.00	0.99*
Master's plus 30 credits or advanced degree	-0.58*	-0.44	-1.22*
Alternatively certified	-0.30	0.05	-0.59
Professional licensure	-0.29	-0.49†	-0.44
Teaching Experience			
Year hired	-0.18	-0.25	-1.21*
Years experience	-0.21	-0.47	-0.17
New teacher	0.34	0.64*	0.73
Tenured	-0.14	-0.65*	-0.50
Professional Development			
Total credits, 2005-06	-0.70*	-0.07	-0.78
Core subject credits, 2005-06	-0.82**	-0.47	-0.37
Mathematics credits, 2005-06	-0.94**	-0.34	-0.16
Teacher Performance			
Mathematics value added, 2005-06 school year	-0.34	-0.96**	-0.78†
Days absent, 2005-06 school year	0.35	0.47	1.00†
Teaching Assignment, Course Description			
Percentage of students in mathematics courses	-0.21	-0.51	-0.68
Teaching Assignment, Student Characteristics			
Percentage white students	-0.38†	-0.36	-0.91†
Percentage black students	0.27	0.32	0.65†
Percentage special education students	-0.00	0.04	0.10†
Percentage English Language Learner students	0.30	0.54*	0.19
Students' average prior year TCAP reading scores[c]	-0.22	-0.13	0.30
Students' average prior year TCAP mathematics scores[c]	-0.10	-0.04	0.21

† $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$.

TABLE B-4.
Standardized Adjusted Treatment Versus Control Group Mean Differences Weighted by Number of Grade 8 Students Taught

	Year 1	Year 2	Year 3
Teacher Demographics			
Female	0.64*	0.92**	0.80†
Race			
White	0.12	0.06	-0.00
Black	-0.10	-0.06	0.00
Year of birth	-0.21	-0.18	0.04
Preparation and Licensure			
Undergraduate mathematics major	0.41	0.59*	0.79*
Undergraduate mathematics major or minor	0.95**	1.10**	1.21**
Undergraduate mathematics credits	0.36	0.53	0.45
Highest degree			
Bachelor's only	0.28	0.15	0.32
Master's only	0.42	0.42	0.08
Master's plus 30 credits or advanced degree	-0.96**	-0.82**	-0.67†
Alternatively certified	-0.45†	-0.59†	-0.60
Professional licensure	0.11	0.18	0.38
Teaching Experience			
Year hired	-0.45†	-0.51†	-0.25
Years experience	0.12	0.23	0.01
New teacher	0.37	0.33	0.11
Tenured	-0.26	0.02	0.04
Professional Development			
Total credits, 2005-06	-0.10	0.19	0.44
Core subject credits, 2005-06	-0.02	-0.02	0.26
Mathematics credits, 2005-06	0.02	-0.02	0.43*
Teacher Performance			
Mathematics value added, 2005-06 school year	0.06	0.01	-0.32
Days absent, 2005-06 school year	0.15	0.30	0.57†
Teaching Assignment, Course Description			
Percentage of students in mathematics courses	-0.09	-0.02	-0.29
Teaching Assignment, Student Characteristics			
Percentage white students	-0.29*	-0.36*	-0.31
Percentage black students	-0.01	-0.02	-0.16
Percentage special education students	-0.00	0.07	0.08
Percentage English Language Learner students	0.29†	0.36*	0.57**
Students' average prior year TCAP reading scores[c]	-0.08	-0.06	-0.19
Students' average prior year TCAP mathematics scores[c]	0.04	0.04	0.00

† $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$.

APPENDIX C:
ESTIMATES OF TREATMENT EFFECTS ON STUDENT ACHIEVEMENT IN READING, SCIENCE, AND SOCIAL STUDIES

This page intentionally left blank.

TABLE C-1.
Reading

School Year	Grade Level				
	All	5	6	7	8
1	-0.01	0.02	-0.03	-0.03	-0.01
	(0.02)	(0.04)	(0.04)	(0.04)	(0.04)
2	-0.03	0.01	-0.03	-0.01	-0.08
	(0.03)	(0.05)	(0.05)	(0.05)	(0.05)
3	0.01	0.02	0.00	0.02	-0.01
	(0.02)	(0.05)	(0.04)	(0.06)	(0.05)

† p < 0.10, * p < 0.05, and ** p < 0.01.

TABLE C-2.
Science

School Year	Grade Level				
	All	5	6	7	8
1	0.01	0.03	0.04	0.04	-0.07
	(0.03)	(0.05)	(0.05)	(0.06)	(0.06)
2	-0.02	0.06	-0.02	0.02	-0.13
	(0.05)	(0.07)	(0.07)	(0.08)	(0.08)
3	0.08†	0.18*	-0.00	0.12	0.06
	(0.04)	(0.08)	(0.07)	(0.09)	(0.08)

† p < 0.10, * p < 0.05, and ** p < 0.01.

TABLE C-3.
Social Studies

School Year	Grade Level				
	All	5	6	7	8
1	0.02	0.07	0.01	-0.02	-0.00
	(0.03)	(0.05)	(0.05)	(0.05)	(0.05)
2	0.02	0.13†	0.01	-0.05	-0.03
	(0.04)	(0.07)	(0.07)	(0.08)	(0.07)
3	0.07†	0.17*	0.02	0.04	0.06
	(0.04)	(0.07)	(0.06)	(0.08)	(0.07)

† $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$.

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 Public Policy
Vanderbilt University's Peabody College

Timothy C. Caboni
Associate Dean for External Relations;
Lecturer 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
Statistician
RAND Corporation

Daniel F. McCaffrey
Head of Statistics
Senior Statistician
RAND Corporation

Patrick J. McEwan
Associate Professor of Economics
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