POINT/CounterPOINT:
The View from the Trenches of Education Policy Research

Dale Ballou
Peabody College
Vanderbilt University

Matthew G. Springer
Peabody College
Vanderbilt University

Daniel F. McCaffrey
RAND Corporation

J. R. Lockwood
RAND Corporation

Brian M. Stecher
RAND Corporation

Laura Hamilton
RAND Corporation

Matthew Pepper
Metropolitan Nashville Public Schools

Published in *Education Finance and Policy* (2012), 7(2)

This material is based on work supported by the National Center on Performance Incentives, which is funded by the U.S. Department of Education's Institute of Education Sciences (R305A06034).

Abstract

The Project on Incentives in Teaching (POINT) was a three-year study testing the hypothesis that rewarding teachers for improved student scores on standardized tests would cause scores to rise. Results, as described in Springer et al. (2010b), did not confirm this hypothesis. In this article we provide additional information on the POINT study that may be of particular interest to researchers contemplating their own studies of similar policies. Our discussion focuses on the policy environment in which POINT was launched, considerations that affected the design of POINT, and a variety of lessons learned from the implementation of the experiment.

1. Introduction

The Project on Incentives in Teaching (POINT) was a three-year study conducted by the National Center on Performance Incentives (NCPI) in the Metropolitan Nashville Public School (MNPS) system from 2006–7 through 2008–9. Middle school mathematics teachers who volunteered for the project were randomly assigned to treatment and control groups. Treatment teachers were eligible for financial rewards if their students made substantial gains on standardized tests. The experiment was intended to test the hypothesis that rewarding teachers for improved scores would cause scores to rise. It was up to participating teachers to decide what, if anything, they would do to raise student performance: participate in more professional development, seek coaching, collaborate with other teachers, focus their instruction on tested content, or simply reflect on their practices. Thus POINT was also a test of the broader thesis that American education suffers from an absence of appropriate financial incentives and that correcting the incentive structure would, in and of itself, constitute an effective intervention that improved student outcomes.

Results did not confirm this hypothesis. While the general trend in middle school mathematics performance was upward throughout the project, overall 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). Incentives appeared to have a positive effect in fifth grade during the second and third years of the experiment. However, the effect does not appear to have persisted after students left 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 fifth-grade teacher the year before was in the control group. Further details on the experimental results, including a large number of sensitivity tests, can be found in Springer et al. (2010b).

This article provides additional information on the POINT study that may be particularly interesting to researchers contemplating their own studies of this and similar policies. In section 2 we describe the policy environment at the time POINT began, which was characterized by a revival of interest in compensation plans that tied teacher pay to improvements in student test scores. In section 3 we discuss the design of POINT. Evaluating the impact of an educational intervention requires a counterfactual: what would have happened had the intervention not taken place. Finding a counterfactual becomes especially acute if schools or teachers are selected (or self-selected) for treatment based on expectations of their future performance with and without the treatment. Because researchers rarely possess information on all the factors that decision-makers take into account when making such decisions, it is unlikely that conventional methods such as regression analysis will yield estimates of the effect of the intervention that control for all other relevant factors. Even comparing post-intervention outcomes with pre-intervention outcomes in the same schools does not fully solve the problem, because it is rare for interventions to be undertaken in environments where everything else is being held constant. For these reasons, POINT was designed as a randomized controlled trial, with participating teachers assigned to treatment (eligible for bonuses) and control (not eligible) groups.

Notwithstanding our intentions, we were limited in our ability to create the conditions of a perfectly controlled experiment. We could randomize teachers into the two groups, but we had no control over student assignments. In addition, after the first year the equivalence of treatment and control groups (i.e., differing only by chance) could be upset by attrition of teachers from the study. Section 3 discusses these issues and other choices we faced in designing POINT.

In section 4, we review lessons learned from this study about designing and implementing experiments in education policy. Several of these lessons arise in connection with the limitations to which we have just alluded. While we do not advocate giving up on controlled experimentation in

this area, many of the problems that beset POINT are likely to arise in other randomized trials dealing with teacher compensation and other personnel policies. In calling attention to these issues, we hope that future researchers will develop superior research designs that ameliorate, even if they do not entirely avoid, the difficulties we encountered. The final section contains concluding thoughts and reflections on the response to POINT.

2. Policy Environment

In recent years, several factors have focused public attention on tying teacher compensation to teacher performance as reflected in student test scores. First is the frustration with the slow pace of progress. It is now nearly thirty years since the Reagan administration issued *A Nation at Risk* (NCEE 1983), yet improvement in public schools has been very slow, particularly at the secondary level. The United States continues to fare poorly in international comparisons (OECD 2010); the achievement gap between the affluent and the disadvantaged remains wide (Dillon 2009).

Second, state and district accountability systems, often adopted in response to federal legislation, have focused public attention on the use of standardized testing to evaluate school and teacher performance. Current interest focuses on deriving estimates of teacher effectiveness using value-added methods. Although there is still controversy about the validity of value-added measures of performance, it has been shown that it is feasible to develop these measures for teachers in some grades and subjects and that these measures correlate moderately with other indicators of student learning and teacher effectiveness.

Third, researchers estimating teacher value added have concluded that teachers' influence on student achievement is highly variable. Teachers appear to be the single most important schooling input, with educational outcomes depending more on teachers than on any other factor outside the home. However, only a small part of this variation is explained by teacher characteristics typically

found in administrative data, suggesting that more efficient compensation policies might be designed than the traditional salary schedules that reward teachers for experience and advanced degrees (Goldhaber and Brewer 1997; Rivkin, Hanushek, and Kain 2005; Aaronson, Barrow, and Sander 2007).

Taken together, these circumstances have renewed interest in the use of performance incentives in public education. The idea is promoted by political leaders at the federal, state, and district levels. Florida, Minnesota, and Texas allocate over $550 million to incentive pay programs that reward teacher performance. Funding for the federally sponsored Teacher Incentive Fund (TIF) quadrupled in 2010, and the Obama administration's 2011 budget request designated an additional $950 million for a new Teacher and Leader Innovation Fund to support the development and implementation of performance- oriented compensation as a viable tool for motivating teachers to higher levels of performance and for aligning teacher behaviors and interests with institutional goals.

This revival of interest was instrumental in launching POINT. Examples of performance pay plans in other localities, as reported in the *Wall Street Journal* and elsewhere,[1] drew the attention of some Nashville business leaders with an interest in public education. Many of these same individuals had supported a local public education foundation in the middle of the decade to support local school reform, and a performance pay demonstration project was viewed as an appropriate foundation activity. A school superintendent and a school board president open to compensation reform were among the other key players.

While some of the participants in this effort were drawn to the project out of a conviction that performance pay would improve educational out- comes, this was by no means true of all. The

[1] See, for example, Henninger, Daniel. "Give Top Teachers a Bonus." The *Wall Street Journal*, January 19, 2007; Gold, Russell. "Merit Based Teacher Bonuses Receive Higher Marks Than Hiring Incentives." The *Wall Street Journal*, February 21, 2001; "A Better Class of Teachers." The *Wall Street Journal,* July 5, 2002.

fact that POINT was designed as a randomized controlled trial was a critical factor in winning support. The then-mayor of Nashville had played a leading role in sponsoring the Tennessee Student/Teacher Achievement Ratio (STAR) experiment on class size reduction while serving in the state legislature in the 1980s and was eager to see Nashville become the site of another well-designed study of a salient education reform. With the mayor's assistance, district and state leadership figures of the Tennessee Education Association (TEA), the state's professional organization for teachers, were persuaded to endorse the plan. To obtain their buy-in, it was critical that the project had a rigorous design (a randomized controlled trial) and that the researchers conducting the project were seen as honest brokers, prepared to report whatever the experiment found, regardless of whether it supported the use of performance pay.

Two years of meetings and consultations were required before all the key stakeholders agreed to proceed with the project. Modifications were made to the research design to accommodate the concerns of various parties. The most important of these are discussed below as constraints imposed by the district or the TEA. Forging personal relationships was essential to winning the trust of the various players. Success was by no means foreordained; over the same period the research team was engaged in discussions with district and state officials and union representatives in other potential sites, none of which resulted in an agreement.

Agreement on POINT was obtained in large part because stakeholders in Nashville came to share one important common belief: more evidence was required on the effect of performance pay in education. This was true of both business leaders who favored this reform and union leaders who felt performance pay would not be shown to have a positive impact. The research team played an important role in convincing other parties that more study using more rigorous research designs was needed. While some previous investigations had found positive effects from the introduction of performance pay, the evidentiary base was weak. According to a recent review, the evaluation

literature on teacher incentive pay programs was "very slender" (Podgursky and Springer 2007). Looking for studies that used a conventional treatment and control evaluation design, with pretreatment benchmark data on student performance for both groups, the authors of this review found only four that dealt with incentive programs in the United States. None was a randomized controlled trial. Three (Clotfelter and Ladd 1996; Ladd 1999; Figlio and Kenny 2007) relied on cross-sectional comparisons of schools using incentives and a comparison group that did not. One (Winters et al. 2006) used a stronger difference-in-differences design, but the study was limited to two schools in which the intervention was tried and contained no information about why those schools had been selected. Unsurprisingly, the authors of the review article concluded that more research was needed, calling for further policy experimentation combined with rigorous evaluation. POINT was launched once key stakeholders in Nashville subscribed to this view.

3. Design of POINT

As noted in the introduction, POINT was designed as a randomized controlled trial. Such a design has distinct advantages over other research strategies. One of the simpler alternatives relies on a comparison of outcomes before and after the introduction of performance pay. However, from the mere fact that test scores rise after a district introduces performance pay, it does not follow that the policy is the reason. Compensation reforms often occur in environments with multiple reforms and accountability pressures that could create changes in student outcomes regardless of the changes to compensation, and it can be difficult to tease apart the separate contributions of the components to changes in outcomes over time.

For instance, compensation reform may be accompanied by other changes in personnel policy. Recent legislation in Florida eliminates tenure for newly hired teachers while replacing the traditional salary schedule with merit pay. Since 2002 (with the passage of No Child Left Behind

[NCLB]), and earlier in states with strong policies of their own, compensation reforms have been introduced in a context of enhanced accountability that has intensified pressure on schools to improve performance. Education reforms also often involve changes to a state's or district's testing regime. The tendency for scores to rise after the introduction of a new test, as teachers and students grow more familiar with it, can be confounded with the effect of other innovations undertaken around the same time.

Another strategy compares districts that have introduced performance pay with districts that have not. The difficulty of inferring the effect of compensation reform from pre- to post-reform comparisons within a district are typically magnified when the inference is based on comparisons across districts, where there will be still greater variation in potentially confounding factors. Yet such comparisons have generally been used to evaluate merit pay within the United States.

Given the complex environment in which compensation reform takes place, random assignment of teachers to treatment and control groups ap- pears to have decided advantages over less rigorous methods of evaluating the impact of the reform. In a controlled experiment, teachers exposed to the treatment (i.e., eligible for bonuses or merit awards) differ only by chance from those assigned to the control group (not eligible). The two groups are exposed to the same external influences (tests, accountability, other personnel policies). The counterfactual—what would have happened in the absence of the compensation reform—is therefore provided by outcomes in the control group.

While the design of POINT attempted to capitalize on the advantages of a randomized controlled trial, several limitations must be noted. First of all, two principal rationales for performance-based pay have been advanced: (1) the existing workforce will improve in response to incentives as teachers find ways to increase student learning that they do not now employ; (2) the use of performance incentives, particularly large incentives, will lead over time to an improvement in

the quality of the workforce as more capable individuals are attracted to careers in teaching. POINT was designed to test the first of these notions but not the second. Given that POINT ran for only three years, it was not possible to test effects on long-term career trajectories.

It should also be stressed that there are many different ways performance pay might be designed and implemented. POINT was able to test only one. The NCPI makes no claim that the incentives offered to teachers in POINT are the best or most cost-effective way to use incentives to improve teacher performance. Negative results in POINT by no means establish that other incentive plans would not be successful.

Finally, POINT's design should not be taken to imply an endorsement of standardized testing as the best way to measure what students learn and what teachers have contributed to their lives. The appropriate way to measure such things remains controversial. Positive findings in POINT would not settle these disputes: one might find that incentive pay of the type tested in POINT results in higher test scores but still regard it as a poor idea. POINT does not speak to this question. It merely sets out to test whether incentive pay tied to gains on standardized tests is successful at raising scores on those tests.[2]

In this section we review choices we confronted in designing POINT and our rationale for deciding as we did. In a few respects our choices were con- strained through negotiations with the district, which stipulated that teachers were not to compete with one another for bonuses and that no teacher would be forced to participate. The district also placed some restrictions on our data-collecting activities: NCPI was not given access to classrooms to observe treatment and control

[2] This leaves open the possibility that higher scores might be the result of narrowly "teaching to the test" or coaching students during the exam. Follow-up analyses of POINT outcomes were conducted to ascertain whether there was more evidence of these activities among treatment teachers than control teachers. There were no such indications. For further details, see Springer et al. 2010b.

teachers. Within this framework, however, NCPI was given broad latitude to design the experiment as it saw fit.

The Point Treatment

Teachers assigned to the treatment group in POINT received financial rewards if their students made unusually large gains on standardized tests. POINT involved no other incentives or systems of support for teachers in the treatment group. There was no requirement that teachers participate in professional development or that they alter their instructional practices in a particular way. What teachers did in response to these financial incentives was entirely up to them. We designed POINT in this manner not because we believed that an incentive system of this type is necessarily the most effective way to improve teaching performance but because the idea of rewarding teachers on the basis of student test scores had gained such currency. We sought a clean test of the proposition: if teachers are rewarded for an increase in student test scores, will test scores rise?

It would have been possible, of course, to test a more complicated intervention—say, one including bonuses plus some form of professional development. However, results would be more difficult to interpret unless the experimental design assigned some teachers to one or the other of the two treatment components and some teachers to both. Unfortunately we did not have the requisite sample size for such a design. In addition, the literature offers no consensus on the other components of a hybrid policy (e.g., what types of professional development are most promising). No choice would have satisfied criticisms that we had not tested the most promising intervention. Finally, there is a greater risk of contamination from treatment to control group when the intervention includes professional development. Ideas from such development would be easy for

teachers to share with their colleagues in their schools; sharing money is less likely or straightforward.

Although no single theory of action informed the design of POINT, it is clear that incentives alter student learning, if at all, through intermediate effects on teachers' perceptions, attitudes, and behaviors. Because we did not specify what teachers should do to raise student achievement, we monitored a wide variety of possible responses in order to learn how teachers viewed the intervention and what they actually did when they were eligible for bonuses. As part of POINT we gathered extensive data on these variables through surveys and interviews.

In addition, while we did not stipulate any particular set of activities that bonus-eligible teachers should undertake to improve performance, it is worth noting that the district provides opportunities for professional development that teachers can pursue on a voluntary basis. During POINT years, the district also offered peer coaching in mathematics to teachers who wished to take advantage of it. If test scores did not rise, it was not because teachers had no opportunity to improve.

Teacher Eligibility

Middle school teachers of mathematics were eligible to participate. A teacher could give instruction in other subjects but needed to teach at least ten students who would take the state's annual mathematics assessment in the spring in order to remain eligible. Teachers could transfer to other middle schools in the district and remain eligible, but teachers moving to elementary or high schools or who stopped teaching math were no longer in the experiment. As noted, only volunteers were included in the experiment.[3]

[3] Although this means we cannot generalize our results to the population of nonparticipants, the effect of incentives on volunteers still seems of considerable interest given the likelihood that many incentive plans would apply only to volunteers. Although one might think volunteers are drawn from the part of the teaching

Math was chosen because schools' and teachers' effects are more easily detected in mathematics than in other subjects (Schochet and Chiang 2010). We chose a design that made it easiest to find positive effects of incentives if there were any, a principle we followed in some other decisions as well. Middle school was chosen because all students take the same battery of standardized tests through grade 8. After middle school there is much more variability in what tests students take and when they take them. We also chose middle school because the number of math students per teacher was larger than in elementary schools, giving us more power to detect treatment effects, if any.[4]

Statistical reliability was one reason we required teachers to have at least ten math students to participate. We also believed that some teachers might object to the notion that someone with exceedingly few math students could win a bonus as an effective math teacher. Given that we needed teacher buy-in for the project to go forward, teachers' perceptions were an important consideration. While these factors argued for setting some floor for eligibility, the choice of ten was arbitrary.

The ten-student threshold made many, though not all, special education teachers ineligible. While we considered the idea of excluding all special education teachers, the fact that their classes differed from those of regular teachers did not appear to be a compelling reason, given the considerable heterogeneity across "regular" classes. Political considerations also played a role, in that

---

population more likely to respond to incentives and that our findings therefore represent an upper bound on the average treatment effect, that conclusion is speculative. It is possible that the teachers volunteering were those least bothered by the bonuses and that nonparticipants would have felt more keenly the pressures associated with knowing that their performance was being graded and rewarded, and therefore would have opted out.

[4] We assume readers are familiar with technical statistical concepts such as power, fixed and random effects, intent-to-treat analyses, etc. Background on these concepts can be found in standard texts such as *Statistical Methods* (Snedecor and Cochran 1980) or *Experimental and Quasi-experimental Designs for Generalized Causal Inference* (Shadish, Cook, and Campbell 2002).

we needed the support of the TEA leadership and, ultimately, its members to conduct POINT. Excluding teachers from participating did not seem wise from the standpoint of building support.

Teachers who changed schools during the experiment were permitted to continue, even though the change might have been sought to improve a teacher's chance of earning a bonus. While such behavior would pose a threat to our experiment, this was only one of several steps teachers might take for this purpose, most of which would not require changing buildings. As a result, we sought a more general solution to the problem (see the section "Randomization of Teachers to Treatment and Control Groups" below).

Although we anticipated considerable attrition from the experiment over the three years it ran, we restricted participation to teachers who were in the system at the beginning of the project and who volunteered at that time. We might have chosen instead to replenish the sample from new volunteers— either teachers new to the district or teachers already present who changed their minds about participating. However, it is unclear that this would have been helpful, given the likelihood that new volunteers would be less effective than the original set of volunteers, either because they were new teachers or because the initial decision to participate was not independent of teacher quality. Moreover, new teachers assigned to the treatment group would have had one or two years' less exposure to the treatment than initial volunteers. For these reasons, it is unlikely that a correctly specified model would pool data of initial and subsequent waves of participants as if there were no underlying differences. There was little to be gained by bringing in new waves of participants only to analyze their data separately.

Measurement of Teacher Performance and Criteria for Earning Bonuses

The Performance Measure

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 (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 sufficient volunteers to the program. Awarding bonuses on the basis of measures no one could understand struck us as unlikely to make participation in the experiment appealing to teachers. Second, we felt that a transparent measure of performance would give teachers an opportunity to see why they had or had not received a bonus and, if they had not, by how much they had fallen short in various parts of the achievement distribution. This might in turn provide better information and stronger motivation to improve than if we were to use a less transparent measure.

Insert Table 1 Here

Our value-added measure was based on students' year-to-year growth on the state test (the Tennessee Comprehensive Assessment Program, or TCAP). To control for the possibility that students at different points in the distribution of scores typically make different gains, we benchmarked each student's gain against the average gain statewide of all students taking the same test with the same prior year score. Benchmarking was simple: we subtracted the statewide average gain from a student's own gain to determine how much his or her growth exceeded the state average.[5] 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 administration of the TCAP 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

[5] It was a little more complicated than this statement implies. Because data are sparse at the extremes of the distribution (even using the population of test takers for the entire state), we smoothed average gains slightly in order to ensure a monotonic relationship between prior year scores and current year scores. The smoothing changed mean gains only at the extremes of the distribution, and the changes from the observed means were minimal.

qualified for a bonus. This calculation is illustrated in table 1 for a hypothetical teacher. The annual report that went to all teachers in the treatment group, explaining the bonus calculations, included a version of table 1 for each teacher's students, differing only in the fact that student names were masked.[6]

Not all students in a teacher's class(es) counted for determining bonuses. Teacher focus groups conducted prior to the start of POINT found that teachers were strongly opposed to counting students who moved in or out of a teacher's class in midyear, even on a prorated basis.[7] We ended up adopting the rule used by Tennessee under NCLB to identify students who count for purposes of determining whether a school has made adequate yearly progress (AYP)—a student had to be in a teacher's class by the twentieth day of the year and remain continuously enrolled through testing in the spring.

A second group of students had to be excluded from the performance measure: those for whom no prior year scores were available. This included most students new to the district as well as those who were absent during testing the previous year. For students transferring to the MNPS from other districts in Tennessee, we were frequently able to add prior year scores to our database, though this required hand entering test scores received by the MNPS in paper records. However, large numbers of students without prior year scores did not count for determining bonuses.

Approximately 20 percent of students were excluded by one or the other of these criteria—often by both. Of the 36,431 students who started a school year in the classroom of a participating teacher over the course of the experiment, test results for 7,960 did not affect bonuses. Four

---

[6] Names were masked because the data use agreement with the school district stipulated that the NCPI would not report scores for individual students.

[7] We conducted these focus groups in the summer of 2006 with mathematics teachers from a neighboring district. We went outside the district because we were not ready to roll out a fully specified program in Nashville, and we did not want to raise false expectations among teachers within the district.

thousand five hundred and sixty-seven students were missing test scores from the previous year, and 6,470 did not remain in the teacher's classroom through the rest of the year.[8]

Bonus Thresholds

There were fundamentally two ways we could determine bonus recipients using these performance measures. We might have awarded bonuses on the basis of relative performance, fixing the number of prizes, as in a tournament. However, this would put teachers in competition with one another, contrary to one of the district's stipulations.[9] The alternative was to compare teacher performance to a fixed target, which is the approach we adopted.

We also had to decide whether the same targets should apply to all teachers or whether teachers should be rewarded for improving on their own past performance. The former seemed advisable, for several reasons. First, teacher past performance is quite noisy. Many teachers would have been rewarded for having had abnormally poor performance in years prior to POINT. In addition, teachers who were already performing at the top of their ability might have felt the system was rigged against them, undermining support for POINT. Finally, this would have required separate rules for teachers who did not have past performance measures because they were new to teaching, new to the district, or new to teaching mathematics. Largely for these reasons, we also thought it unlikely that incentive schemes of this type would be implemented as frequently as plans that held all teachers to the same standard. Thus it made more sense to conduct a test of the latter.

We used historically determined targets to determine which teachers would earn bonuses. We calculated the benchmarked performance measures used in POINT for the district's middle

---

[8] Strictly speaking, these are counts of student-year observations, not of students. The same student can enter these totals more than once.

[9] We were opposed to this approach for our own reasons as well. The resulting competition could have had adverse consequences for student achievement. Given our commitment to testing incentives in conditions most conducive to their success, where possible, we also preferred to gauge teachers against an absolute threshold.

school mathematics teachers in the two school years immediately prior to POINT: 2004–5 and 2005–6. We then selected three thresholds from the distribution of these performance measures: one at the 80th percentile, a second at the 85th percentile, and a third at the 95th percentile. We sought thresholds that would strike teachers as attainable yet be high enough that there would be little question that winners were deserving. We also sought thresholds that made it unlikely that the sum of the bonus awards would grossly exceed available resources. It may be thought that these targets would have seemed out of reach to most teachers. For evidence to the contrary, see our initial report on POINT (Springer et al. 2010b).

Because these thresholds were historically determined and fixed through the duration of the experiment, it was theoretically possible for all POINT teachers assigned to the treatment group to qualify for bonuses. This represented an open-ended financial exposure for the NCPI, which was committed to funding the number of bonuses earned, regardless of the total amount.[10] While we could have limited the financial exposure by announcing that a fixed pool of money would be divided by the winners each year, this would have introduced uncertainty among teachers about bonus amounts, which seemed undesirable from the standpoint of incentive design. Instead we specified fixed awards for reaching each of these thresholds that also remained unchanged for the duration of the experiment.

Bonus Amounts

Previous incentive plans for teachers have generally offered modest rewards for improved performance. We regard it as unsurprising that responses by teachers have also been modest at best.

[10] We opted for this approach in part because we believed that if we ran out of money we could persuade interested private sector parties to provide additional funding. Overspending our budget would mean we had more bonus winners than we had anticipated, which would mean that incentives were having a pronounced effect on student achievement. It would be easy in such circumstances, we believed, to find private sector funding to get such a message out to the public. Events proved us wrong. While we had more bonus winners than expected, this was because achievement rose throughout the district, in the classrooms of control group teachers as well as treatment group teachers.

At the same time, the discussion in the education policy community has increasingly focused not on marginal or incremental changes to teacher compensation but on dramatic changes (e.g., six-figure salaries for the best teachers). Our interest was therefore in seeing what teachers would do in response to substantial bonuses.

Teachers whose performance during POINT reached the lowest threshold (the 80th percentile of the historical distribution) 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. Our maximum bonus of $15,000 is one-third of the mean base pay among participating teachers—substantially greater than the 5–10 percent bonuses more commonly found in performance pay plans.[11] Although we did not expect most bonus recipients to reach that threshold, we suspected that this maximum amount would have greater salience than the smaller bonus amounts and would play a significant role in motivation.[12]

Many MNPS middle schoolteachers, particularly in grades 5 and 6, teach subjects other than mathematics. Tying bonuses solely to mathematics scores might lead them to neglect other subjects. This struck us as unwise educationally. It also seemed inadvisable politically. Parents upset that their child's teacher had neglected subjects other than mathematics could have cost POINT public support. To safeguard against this, we calculated an analogous bench- marked performance measure for each teacher in all four tested subjects, including reading/English language arts, science, and

---

[11] Using data from the Schools and Staffing Surveys, Taylor, Springer, and Ehlert (2008) find that approximately 13 percent of public school districts in the United States used some type of performance- based incentive in 2003–4. They estimate the average amount per recipient to be approximately $2,000, 4.6 percent of base pay. This is an average, not the maximum, but by contrast the average bonus paid in POINT exceeded $10,000.

[12] Anecdotal evidence suggests that this belief was at least partially correct. The large dollar amount was the focus of teachers' attention. In interviews with teachers in which the bonus amounts were mentioned, it was clearly the $15,000 figure that had grabbed their attention and to which they later referred. Whether this amount motivated a change in teaching practices is less clear.

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 median performance on the other measures in all the subjects for which the teacher provided instruction. Falling short of that goal would cost the teacher a prorated portion of the mathematics bonus based on the proportion of her students tested in other subjects.[13]

In addition to the bonuses earned by treatment teachers whose performance exceeded the aforementioned thresholds, stipends of $750 per year were paid to all participating teachers, both treatment and control. As a condition of receiving these stipends, teachers agreed to respond to written surveys sent out each spring and to participate when selected for oral interviews. Over the course of the experiment, these annual stipends totaled $2,250. The stipend was large for the amount of time the project required of teachers. The size of the stipend was an attempt to assuage any disappointment teachers might feel on being assigned to the control group rather than the treatment group. Such disappointment might have had a negative impact on teaching performance that could have been confounded with the effect of incentives.

Size and Duration of the Experiment

Because teachers might require time to respond fully to incentives, we decided to conduct POINT as a multiyear experiment. Teachers would be randomized into treatment and control groups at the start of the experiment. In order to see whether the long-term response differed from the short-term response, a teacher's status (treatment or control) was held fixed for the duration of

[13] The precise formula incorporating these adjustments follows: Let T equal the bonus for which a teacher qualified, based on the performance of her students in mathematics (either $5,000, $10,000, or $15,000). Let D_k equal 1 if the teacher fails to achieve the district's median value-added score (in the historical distribution) in subject k, where k = math (M), English (E), science (S), and social studies (SS); otherwise D_k is 0. Finally, let N_k be the number of students the teacher has in subject k (where, as noted above, students are counted only if they are continuously enrolled in the teacher's class from the twentieth day of the school year). Then $P_k = N_k/\Sigma_j N_j$, j = M, E, S, and SS. The teacher's bonus is then given by Bonus = $T\times [1 - P_E D_E - P_S D_S - P_{SS} D_{SS}]$.

the experiment.[14] The choice of three years was pragmatic. Funds were limited. A three-year experiment, followed by a year or more to analyze the data, would take us nearly to the end of the center's five-year life.

Preliminary power analyses indicated that we needed about 200 teachers enrolled in the experiment to detect an effect size of .12 to .17 standard deviations with a probability of 80 percent: 100 in the treatment group and 100 in the control.[15] Anticipating attrition during the experiment, we attempted to enlist more in the first year. As it turned out, 296 teachers agreed to participate, approximately 70 percent of those eligible.

Randomization of Teachers to Treatment and Control Groups

Two features of the study design had implications for randomization:

- Teachers would remain in the same experimental condition (treatment or control) for all three years of the study; and
- The district would retain control of student assignments to classes and teachers. Thus, while POINT could randomly assign teachers to treatment and control groups, we could not randomly assign students.

The first of these features meant that teachers would know whether they were in the treatment group and eligible for bonuses prior to receiving teaching assignments in the second and third years of the experiment. Suppose treatment teachers took advantage of that knowledge to influence the makeup of their classes. In that case, systematic differences could be introduced between treatment and control groups that might be confounded with the effect of bonus eligibility on teaching

[14]In addition, reassigning teachers at the start of each year to treatment and control would have blurred the distinction between the two groups: a current-year control teacher might have improved performance as the result of having been in the treatment group the previous year.

[15] This was on the assumption that effects would be estimated separately for each year.

performance (e.g., by reassigning struggling or disruptive students to the classroom of a teacher in the control group or a nonparticipating teacher).[16]

Given the potential for purposive assignments to bias estimated treatment effects, we developed a two-stage randomization scheme that would be robust to such threats.[17] If all teachers of a particular course in a particular school were assigned to the same experimental status (treatment or control), reassignment of students from one section of that course to another would not affect the overall equivalence of treatment and control groups. Of course, some transfers could occur outside this group. Students might move from a more advanced to an easier course, or vice versa, if their original placement was deemed a mistake. For broad groupings of courses, however, across-group transfers would be much less likely than within-group transfers and would be less apt to be made to accommodate a particular teacher's wishes than for educationally sound reasons.[18] Thus we created four course clusters within each school: grades 5 and 6 mathematics classes, grades 7 and 8 mathematics classes, special education mathematics classes, and algebra or more advanced mathematics classes. Each teacher was associated with one of these groups, based on the courses taken by the plurality of her students at the time of randomization. Each course cluster within a school (and the set of teachers associated with it) was then randomly assigned to treatment or control status. Because not all teachers participated in the experiment, a treatment cluster was not in

---

[16] Though the problem seemed most severe in years 2 and 3 of the experiment, even in the first year treatment teachers could attempt to influence the makeup of their classes by recommending students for transfer, for example.

[17] We also tried simpler measures. So that principals would not be in a position to assist treatment teachers (were they so inclined), we did not inform them which of their teachers were in the treatment group and which were controls. We also encouraged principals to run their schools as they would have in the absence of POINT. Participating teachers had to sign a pledge that they would not disclose their status to other district employees. Anecdotal evidence as well as teachers' responses to our surveys indicate that many teachers knew colleagues who were eligible for bonuses as well as the identities of at least some bonus winners.

[18] Students could transfer across schools, of course, but with rare exceptions these decisions would be made by parents and would not constitute the purposive assignment of students to improve a treatment teacher's chance of earning a bonus.

fact entirely made up of teachers assigned to treatment status. While this would dilute the treatment effect, it would not introduce bias provided the nonparticipating teachers in a cluster were not affected by the assignment. This would allow us to estimate responses to treatment at the level of the course cluster (analogous to an intent-to-treat estimate, in which nonparticipating teachers fail to take up the cluster's status) or to use a cluster's status as an instrument to estimate the effect of treatment on the treated.[19]

As implemented, however, clusters were not perfectly homogeneous. In- stead they could include both treatment and control teachers. Some teachers taught across clusters. For example, instructors who primarily taught algebra might have one class of seventh-grade mathematics. Because most of their instruction is in algebra, they would belong to the algebra cluster for their school, which might be assigned to treatment. Their seventh-grade mathematics students would therefore have a treatment teacher, whether or not that course was in a treatment or a control cluster in that school. Thus there were schools in which most seventh-grade mathematics students had teachers in the control group, yet some sections of the same course were taught by a teacher in the treatment group (or vice versa). While this further diluted the treatment effect, it did not introduce a potential bias. Even if the treatment teachers in such a case managed to shift onto control teachers their least promising students, as long as those students remained within the course cluster such behavior would not bias a cluster-level analysis of outcomes or an analysis in which cluster was used as an instrument for the treatment status of individual teachers.

In addition, we deliberately gave some teachers a different assignment from the rest of the instructors in their course cluster, if needed, to ensure that every school had at least one treatment

[19] Prior to assigning course clusters, schools were stratified into ten blocks based on student TCAP scores in prior years. Randomization then occurred by course cluster within block to ensure balance between treatment and control groups (e.g., so that the course clusters in the highest-performing schools were not all assigned to the treatment group by chance).

teacher. This was done to forestall a negative reaction should teachers discover that none of the participants in their school were eligible for bonuses.[20]

4. Lessons Learned

In this section, we review a variety of lessons learned during this experiment—lessons distinct from the central question addressed in the study: will incentive pay produce higher student test scores? (Interested readers should refer to Springer et al. 2010b for a summary of results.) We arrange them under three headings: lessons related to the design of POINT, lessons related to the implementation of POINT, and lessons related to the analysis of POINT outcomes. Though these categories are not airtight—implementation affected design (and vice versa), while both had a bearing on how we analyzed results—this division is broadly useful.

Lessons Related to the Design of Point

In hindsight, it appears we fretted over many details of the intervention that did not matter as much to teachers as we feared. While we worried about demoralization among control teachers, our surveys found that teachers in the control group were as supportive of the experiment as those in the treatment group. Few teachers in either group reported feeling any resentment due to the experiment.

Treatment teachers reported higher levels of teacher collegiality than control teachers, indicating that they understood they were not competing against one another for bonuses but against a fixed target. More generally, the majority of treatment and control teachers disagreed with

[20] We feared this would be perceived as evidence that we had violated our promise that all teachers would have an equal chance to be assigned to the treatment group. While this would not have been true, we were not confident that teachers unfamiliar with randomization by cluster within strata would recognize that fact.

the statement that "rewarding teachers based on performance destroys the collaborative culture of teaching."

Determination of Bonuses

It does not appear that teachers were as concerned as we feared about details of the bonus formula and what sort of value-added performance measure we used. However, our evidence on this issue is anecdotal. For example, NCPI staff who went into schools to recruit volunteers for the experiment noted that many teachers signed up without reading the multipage FAQs that provided these details. Staff who fielded complaints from teachers during the course of the experiment sometimes found that these complaints arose because teachers were unfamiliar with fundamental facts about how bonuses were determined.

Unfortunately we did not ask probing questions of teachers in the surveys to obtain more systematic information about these issues. The questions we asked instead asked teachers to rate their own knowledge of POINT's rules and formulas. While the majority professed to understand how bonuses were determined, we did not test their knowledge of details.

The teacher performance measure in POINT was substantially more transparent to teachers than the other value-added measure with which most were familiar: the Tennessee Value-Added Assessment System. However, we do not know whether transparency was important to teachers for either of the reasons that mattered to us—namely, that it made them more willing to participate and gave them greater motivation to improve. We did not ask if teachers had used the bonus calculation worksheets provided by POINT to assess their effectiveness, to gain insight into why they had failed to receive a bonus, or to plan future changes in their instructional approach.

Randomization by Cluster

Prior to POINT we worried a good deal about purposive assignments—that treatment teachers would game the system by seeking easier-to-teach classes. As explained in Springer et al.

(2010b), by a variety of indicators this does not appear to have been a significant problem. This was not something we could have known in advance, of course, and we designed POINT to provide a fallback method of analyzing the experiment that would be largely immune from this behavior—that is, the method of randomizing teachers into treatment and control groups described above. Unfortunately this method fails, for reasons we did not foresee when POINT was designed. While cluster assignment to treatment or control groups is a valid instrument in year 1 of POINT, given the high level of attrition and the fact that attrition appears to be related to treatment status, cluster assignment is unlikely to be a valid instrument in years 2 and 3. Even if the effectiveness of teachers leaving the experiment is unrelated to treatment status, the fact that more of them leave the control group than the treatment group has implications for average teacher quality at the cluster level. Clusters in the control group will have more new teachers and more teachers with new assignments, both of which are likely to impair performance, even if only temporarily. Thus we used a cumbersome randomization process that reduced statistical power while buying us nothing.

Lessons Related to Implementation

Data Quality

While the spread of state and district accountability systems has spurred the development of the capacity to link students to teachers, our experience in POINT suggests that such data systems fall short of what is required for high- stakes personnel decisions. For POINT to be credible in the eyes of teachers, it was important that teachers not miss out on rewards because the data management system held them accountable for students they did not teach or did not teach for the requisite portion of the school year, or failed to give them credit for students who were theirs.

Ensuring the accuracy of these linkages was not a simple matter, given that MNPS administrative records were not designed to provide a longitudinal tracking of student-teacher links.

Instead snapshots of student enrollment by course, section, and teacher were taken at specified intervals (up to six times per year). These were merged with daily school enrollment records that placed a student in a given school each day of the year (though these records were not free of inconsistencies and errors). From these two sources the NCPI had to determine which teacher a student had at every point over the course of the year. Intra-school transfers occurring between snapshots posed an obvious problem. Additional challenges arose from the student management system's incomplete tracking of special education pullout services.

Anticipating that our reconstruction of teacher-student links would contain errors, each May during the experiment we mailed rosters to all treatment teachers, indicating the students we would be using to determine bonuses. Teachers were required to verify that these rosters were accurate and to re- port discrepancies to the NCPI. Requested changes were cross-checked with alternative administrative data sources. In a few circumstances, personnel in the MNPS's Department of Assessment and Evaluation assisted in efforts to resolve roster discrepancies by discreetly contacting a school administrator or counselor.

Insert Table 2 Here

Table 2 shows statistics on rosters created and appeals submitted for the three years of the POINT experiment. The number of appeals declined over time. During the first year of the experiment the NCPI did not attempt to identify intra-school transfers within the year. This led to a larger number of appeals in the first year than in years 2 and 3. Improvements to the district's data system and the growing experience of school staffs with the system also contributed to the declining number of appeals. For instance, during the experiment's first two years, the district did not perfectly identify students who received special education pullout services. This issue was solved by the third year.

Checking class rosters was only one of the tasks the NCPI undertook to acquire accurate data for POINT. To deal with such issues, it embedded a staff member (paid by the NCPI) in the district office. In addition to securing for the NCPI the data required for POINT, this individual played a larger role in improving the district's data systems. This proved to be an extremely useful step in the implementation of POINT.

Finances

As explained in section 4, POINT involved an open-ended commitment to pay bonuses according to fixed criteria that, in principle, all participating teachers could have met. By year 3, it was clear that if the same proportion of teachers won bonuses as in previous years, we would exhaust our budget. Eventually we were able to secure funding from a combination of private and public sources and avoid the cancellation of the study's third year.

Publicity

Our preexperiment focus groups showed that teachers strongly preferred not to publicize the names of bonus winners (and nonwinners). Because we needed volunteers for POINT to go forward, we promised prospective participants that we would do all we could to preserve confidentiality. However, teachers' compensation is a matter of public record, and although bonus winners were not identified as such in any documents, an enterprising reporter could have identified many recipients by looking for otherwise unexplained compensation. We therefore tried to maintain a low profile, saying little about the experiment after it was launched and hoping that the press would forget about it. We were also concerned that the disclosure of interim results could cause the coalition of interest groups supporting POINT to unravel, given that the various stakeholders had different expectations of what the experiment would show.

While we were successful in maintaining a low profile, in retrospect we may have created an environment that weakened POINT incentives. There was no hoopla of any kind in connection with

the program. By requiring teachers to pledge that they would not reveal to other district employees whether they had been assigned to treatment or control groups, we discouraged a sense of camaraderie or shared purpose among teachers assigned to the treatment group. Though teachers preferred a quiet intervention, it is possible that such an incentive plan does not provide the same motivation as one conducted more publicly.

Lessons Related to Analysis of Results

Teacher Attrition

Teachers participating in POINT left the study at a very high rate, with only half remaining for all three years. Most of this attrition was teacher initiated, although as described above, teachers with fewer than ten eligible students were dropped from the study by the NCPI. Year-by-year attrition is shown in table 3. The entries in the table are the number of teachers who had left the experiment by the end of the school year in question. Thus column 1 shows the small number that dropped out before the end of the first year. Most attrition, as one would expect, occurred during the summer. Some of the spike in year 2 was due to a one-year reprieve granted teachers with fewer than ten math students in 2006–7. As a result, two years' worth of teachers failing to meet this requirement were dropped from the experiment at the beginning of 2007–8.

Insert Table 3 Here

Insert Table 4 Here

Teachers left for a variety of reasons, most prominently because they no longer worked in the district, moved to elementary or high schools in the MNPS, or stopped teaching middle school mathematics (table 4). While there were some differences between the reasons given by treatment and control teachers, they were not statistically significant.

The high rates of attrition are troubling for two reasons. First, by diminishing the overall size of the sample, they reduce statistical power. Second, year 2 attrition was 38 percent higher in the control group than in the treatment group, suggesting that attrition was endogenous to treatment assignment, with some treatment teachers avoiding employment changes that would have made them ineligible for bonuses. Although this pattern was reversed to some extent in year 3, this may reflect the fact that the experiment was nearing an end. Treatment teachers who had put off job changes in year 2 may have found that the benefit of continuing to do so no longer outweighed the cost, given that POINT had only one year left to run. On this reading of the evidence, the reversal in year 3 was a kind of catch-up. Be that as it may, it remains the case that over the course of the experiment attrition was greater from the control group. If attrition were also a function of how likely teachers were to earn a bonus, after the first year of the experiment treatment, control groups would no longer be equivalent to factors influencing student achievement.[21]

It is not evident that anything could have been done to prevent these high rates of attrition without drastically altering the design of POINT. Clearly the problem could have been avoided by re-randomizing teachers to treatment and control groups at the beginning of each year, but this would have sacrificed the experiment's longitudinal design and the possibility of detecting a change in treatment effects over time. Moreover, frequent changes in status would have blurred the distinction between treatment and control groups.

It may be thought that the problem posed by attrition could be made to disappear by defining the effect of treatment more broadly to include the impact of incentives on teacher

[21] Although the difference in attrition may reflect the choice of better teachers in the treatment group to remain eligible for bonuses, this is not a foregone conclusion. Subjective probabilities of winning a bonus, as reported on teacher surveys, bore little relationship to the actual incidence (Springer et al. 2010b). Thus it is possible that while the overall rate of attrition was lower in the treatment group, this was because the most optimistic treatment teachers—not the best treatment teachers— sought to preserve their eligibility for bonuses. However, if teacher optimism reflected unmeasured differences in students, the potential for bias remains.

turnover. Treatment would thus affect student outcomes through two channels: an effort effect (teachers work harder to earn bonuses) and a selection effect (the best of the treatment teachers are more likely to continue teaching middle school mathematics). The selection effect would become one of the ways incentives alter outcomes, not a source of bias.

The problem with this approach is not that such a redefinition of the treatment effect is unreasonable but that it does not solve the problem. The mean difference between outcomes in the treatment and control groups will not be an unbiased estimate of the treatment effect on student achievement, even under this broader definition, if the self-selection of treatment teachers in or out of POINT is a function of unobserved teacher *and* student characteristics. For example, suppose that in the absence of incentive pay, the probability that a teacher stops teaching middle school mathematics is a function of her ability plus the (unmeasured) engagement of her students, and that teachers with less engaged students are more likely to leave, ceteris paribus. Now suppose that incentive pay partially offsets this among teachers in the treatment group, depending on how effective the teacher is. Then more effective treatment teachers who continue teaching middle school math will tend to have less engaged students than the average control teacher making the same decision, and unobserved student characteristics become confounded with the effects of treatment.

To reduce the potential for attrition bias, we introduced a variety of covariates into the achievement equation to control for observable differences between treatment and control students and teachers. As a result, POINT in its second and third years resembles an observational study as opposed to a pure randomized experiment.

We also investigated whether the treatment group retained a higher proportion of its most effective teachers. Over time the treatment group came to have a higher proportion of students taught by female teachers and black teachers. There were also statistically significant but

substantively small differences in within-district experience, professional development credits, and days absent. No significant differences emerged in the variables most directly related to the experimental outcome—an estimate of teacher value added from the 2005–6 school year and students' pre-POINT scores in math and reading.

With hindsight we might have done more to correct for the effects of attrition on our estimates. On the annual surveys sent to participating teachers, we might have asked members of the treatment group whether eligibility for bonuses influenced their decisions with respect to the following: transferred to an elementary school or a high school, stopped teaching math, left the district, etc. Had teachers told us that bonuses were a factor in these decisions, we might have used their responses to reweight the sample of treatment teachers to resemble the group of non-attriting control teachers.[22] Given that we did not include such items on the surveys, this solution remains conjectural. Moreover, it is not clear that survey-based measures would provide accurate information about teachers' motives for career choices.

Randomization Inference

POINT used a complex randomization design, with teachers assigned to treatment and control groups by block and cluster. Because numerous factors contribute to student achievement, many of them unobserved, we estimated achievement equations in which some combination of block, school, cluster, teacher, and student effects was possibly present, as well as variation in each of those effects by year—block by year, school by year, etc. It was not apparent which of these effects was important a priori. A fully general model that incorporated all of them was intractable, but a misspecified model that incorrectly omitted some could yield incorrect standard errors and faulty inferences. Given the sensitivity of our inferences to model specification and our uncertainty

[22] The objective would have been to weight each treatment teacher by 1 minus the probability that that teacher would have left the experiment had she been assigned to the control group.

about which model was correct, we relied heavily on randomization inference, a simulation method whereby a large number of artificial experimental samples (e.g., five hundred, one thousand) are generated by reassigning teachers to treatment and control groups using the original randomization rules (Efron and Tibshirani 1994). Under the null hypothesis of no treatment effect, the distribution of the treatment-control differences over these samples mimics the distribution of the same statistic in repeated samples taken from the population from which the actual data were drawn. This in turn allowed us to ascertain whether our model-based inferences were accurate.

For example, we found fairly large preexperimental differences in the achievement of treatment and control teachers' classrooms, particularly in 2004–5. Because we randomly assigned teachers to treatment and control groups, we knew these differences arose from chance. However, in the models we initially favored, these differences were associated with quite small *p*-values, suggesting that we had made an improbable type I error. A randomization analysis, in which we artificially generated several hundred samples of treatment and control groups, revealed that such preexperimental differences occurred far more often than the reported *p*-values from our initial models, with a true probability exceeding the conventional thresholds of .05 or .10. This in turn led to the recognition that our initial model failed to capture all the relevant sources of variation in student test scores. As we introduced more richly specified models, reported standard errors corresponded to those obtained from the randomization analysis, and the *p*-values from the model closely matched those from the randomization inference.[23] The distribution of treatment and control group differences from the randomization analysis also revealed that the pretreatment differences were highly sensitive to the assignment of a single teacher, allowing us to test our main results for

[23] The models reported in Springer et al. 2010b included block fixed effects and cluster and teacher random effects. Because the data were not pooled across years, these were implicitly block-by-year, cluster-by-year, and teacher-by-year effects. In unpublished results using data pooled across years, these interactions are explicit. Individual students are observed more than once when data are pooled across years. In this case, within-student covariances over time are unrestricted.

sensitivity to this outlier. We found that our main results were robust, but the results suggested shortcomings in our models that allowed the outlier to be so influential.

In short, we found randomization inference a valuable check on our model specification and the validity of our inferences.

Differences in Treatment Effects by Year

POINT was designed as a longitudinal study in which teachers would remain in treatment or control groups for multiple years. Although the purpose was to see whether the impact of incentives increased with time, responses may have differed across years for other, more important reasons.

Teachers were told about POINT after the beginning of the 2006–7 school year. Recruitment took place in the fall, with teachers notified of their status (treatment or control) in early November. Given that the school year begins in mid-August and the state conducts testing in April, these delays meant that approximately three-eighths of the potential instructional time had passed before teachers knew whether they were eligible for bonuses.

However, the situation was worse than that because final approval of the project was not obtained until mid-January in a vote of teacher union members, three months before testing.[24] Although approval was widely anticipated, participating teachers may have postponed any effort to improve their instructional practices until they were certain that POINT was going forward.

These problems could have been avoided had the NCPI been allowed a longer lead-in period to launch POINT. However, we were required to launch POINT in the 2006–7 school year,

[24] The January vote arose through circumstances that might best be described as a fluke. In the same year that POINT was launched, a competing pay-for-performance plan sponsored by another Nashville group was proposed for a small number of schools, contingent on a vote of teachers in the affected schools. They voted it down. This led other members of the union to ask why our proposal was not also required to clear the same hurdle. Although the union leadership had already approved POINT, they felt that the procedures applied to one proposal had to be applied to the other, and at a late date (fall) it was decided that the district's participation in POINT had to be put to a vote of the members. With the endorsement of the union leadership, the proposal passed.

only a few months after the award of the center by the Institute for Education Sciences. This haste is doubly regrettable in that the first year of the experiment, before significant levels of attrition took place, offered the cleanest test of our hypothesis with the greatest statistical power.

POINT was conceived as a three-year experiment. With our limited budget and the need to bring closure to the study within the NCPI's five-year lifetime, it would have been difficult to have made it longer. In hindsight, however, the fact that the experiment had a very visible termination date may have contributed to our failure to find a positive effect of incentives. With each successive year there was less time to amortize the cost to teachers of improving their performance. Had the incentives been put in place permanently, different results might have been obtained.[25]

Why Did Incentives Have No Effect?

As noted in the introduction (and widely reported in the news media), there was no general detectable influence of incentives on student achievement in POINT. As is true of most experiments, there are competing reasons for this (all, some, or none of which may be true). Did teachers not make much effort to improve (and why not)? Did they try the wrong approaches? Our ability to answer these questions is limited to what we can learn from self-reported responses to teacher questionnaires and interviews, to a small amount of administrative data on participation in professional development activities and absenteeism, and to surveys of principals and math mentors. A better understanding of what was taking place in classrooms might have been obtained through classroom observations, but our agreement with the district did not permit the NCPI into classrooms.

---

[25] We do not want to overstate this point. There are no permanent innovations in education policy. Many compensation reforms last no longer than our experiment. Merit pay plans are notoriously short lived, and teachers are rightly dubious that the money promised for better results will be forthcoming. Thus the fact that POINT ended after three years does not mean it is of no relevance to understanding how teachers would respond to pay for performance in the real world.

Taken at face value, teachers' responses to surveys suggest that treatment teachers made few changes to instructional practices over the course of POINT. More than 80 percent indicated that they were not changing what they did in order to earn a bonus, because they were already teaching as effectively as they could. Yet within the period of the experiment, mathematics test scores rose within the district in grades 4–8, calling into question the notion that teachers were doing nothing different. Something was changing, but whatever it was affected the classrooms of treatment and control teachers (and nonparticipants) alike.

Perhaps the most obvious explanation is that the district was under pressure to improve test scores to avoid NCLB sanctions. Conceivably, the response to NCLB dampened the effect of POINT incentives. If there is only so much a teacher can do to improve in a given year, and all the district's math teachers were already doing that to make AYP, there were no actions left for teachers to take in response to the new incentives. The pressure to meet AYP is strong and motivational to teachers. In a study of pay-for-performance in New York City, teachers reported that meeting accountability goals was more salient than earning a bonus (Marsh et al. 2011). NCLB created strong incentives to improve student performance, and POINT merely added to that signal.

In a larger, more heterogeneous district, one might put this hypothesis to the test by considering whether incentive effects were stronger in middle schools with a history of making AYP (and therefore under less pressure than other schools to respond to NCLB). There are too few schools of this kind in the MNPS to conduct such a test. Moreover, it is not clear that such results would represent the appropriate counterfactual for harder-pressed schools. Without direct evidence that incentives raise achievement in struggling inner-city schools, results from an experiment like POINT are bound to be disappointing, given current policy priorities.

5. Concluding Remarks

The Role of Experimentation in Educational Policy Research

The challenges POINT faced are common in educational experiments and highlight the limitations of such studies. The political climate of the public school system—including issues of fairness, the need for results in the short run, and the competing priorities of multiple stakeholders—constrained the intervention and the study design. The history of POINT also shows how difficult it is to maintain the integrity of experimentally controlled groups in education settings, particularly in longitudinal designs needed to study treatments whose effects might take time to manifest. Unlike textbook examples of experiments with single units that can be independently randomized into treatment and control groups and consistently measured, education settings are multilevel organizations of units (students, classes, teachers, schools) in which the population of experimental subjects and their relations to one an- other are changing both within and across years. For example, the POINT study units were teachers, but their outcome measures were provided only indirectly through the students they taught. The students assigned to teachers in the treatment and control groups were changing across years and had opportunities to evolve differently over the course of the experiment due to teacher-course, student-course, and student-teacher assignment mechanisms and exposure to past study teachers that were outside our control. Teacher attrition from the study also created potential confounds to the estimated intervention effect so that in years 2 and 3 POINT was effectively an observational study rather than a textbook random assignment study.

We believe that POINT was not uniquely vulnerable to this challenge. To avoid these problems, studies of long-term education interventions would probably require a degree of control over educational settings that few if any educators would be willing to provide. Nonetheless, we are not persuaded that randomized experiments have no role in education. Because POINT was a controlled trial, teachers participated in a well-defined intervention al- lowing us to precisely frame

the policy conclusions. Quasi-experiments or observational studies of policies that districts implement on their own accord rarely have such well-defined interventions or support such clear conclusions. Random assignment can be useful even when there are some breakdowns of groups or implementation. For example, had teacher attrition not been a significant problem, the cluster-based randomization scheme used in POINT would have furnished a method for dealing with nonrandom selection of students into treatment classes—something worth bearing in mind for studies of interventions limited to a single year in which teacher attrition is not an issue. Randomization of volunteers also removes the potential for preexisting differences among teachers who would or would not volunteer for a controversial intervention—differences that can be very strong—to bias estimated effects.

Finally, we stress that POINT was only one experiment—in effect, one observation—and that conclusions about the merits of experimentation in education policy research need to be based on broader evidence. Future re- searchers may discover ways to circumvent some of the problems we encountered. Indeed, with that hope, we present this account of POINT's implementation.

The Public Response to POINT

The NCPI released a report summarizing findings from POINT in the fall of 2010. While the report received considerable attention from the media, it is difficult to gauge what effect it has had, if any, on public policy. Efforts continue to reform teacher compensation by creating financial incentives for increasing student test scores, as one can see from recent issues of *Education Week* selected virtually at random.

> *Gov. Otter [of Idaho] also signed a pay-for-performance plan that will allow teachers to earn higher salaries if they improve students' achievement. (Cavanaugh 2011, p. 4)*

The first bill Mr. Scott signed as governor [of Florida] ended tenure for new teachers and created a merit pay system tied to test scores. (Associated Press 2011a, p. 5)

Gov. Robert F. McDonnell of Virginia has invited 57 districts with struggling schools to apply for $3 million in state funding for merit pay. . . At least 40 percent of each evaluation would be tied to student academic performance, including improvements in standardized-test scores. (Associated Press 2011b, p. 5)

Apparently it will take more than one experiment to convince advocates of this model of compensation that it is not likely to produce the results they desire. This is only fair. One experiment is not definitive. More experiments should be conducted to see whether POINT's findings generalize to other settings. Unfortunately we are not optimistic about the prospects for additional tests of incentives of the type tried in POINT. In part this reflects the sheer difficulty of mounting an experiment, particularly in an area as sensitive and contentious as teacher compensation. It is difficult to obtain the agreement of stakeholders. It is hard to secure funding. Moreover, any attempt to redo POINT in another setting is apt to run into the objection that, for example, those incentives were tried in Nashville and they didn't work. Opponents of performance-based pay will argue that another test is not needed. Proponents will be reluctant to throw good money after bad.

The history of research on smaller classes cautions us not to expect too much. Three decades after the Tennessee STAR experiment, it remains the only rigorous experimental evaluation of reductions in class size, despite concerns frequently raised about its generalizability. More than in the natural sciences, experiments in the social sciences need to be replicated, for many background factors will differ from one setting to the next. The annals of education policy furnish many examples of programs that worked in some places but not in others. Replication should therefore be the rule, but it is not.

Prospects are probably better for testing other incentive designs, though re- cent evidence suggests that other designs may yield results similar to POINT: small effects on student achievement that cannot be distinguished from null. Fryer (2011) found small negative effects of New York City's School-Wide Performance Bonus Program, which paid teachers, counselors, and support staff bonuses if their school met performance targets. Springer and coauthors (2010a) also reported a null effect for a team-level bonus program in the Round Rock Independent School District that was conducted as a tournament in which teams with the highest (value-added) impact on student achievement received financial rewards. Like POINT, these studies pro- vided incentives unlinked to specific professional development or training. Together with POINT, these results provide evidence that bonuses alone do not appear to improve student learning as measured by standardized tests.

Given the challenge of conducting experiments, quasi-experimental evaluations of alternative compensation plans may be more common, particularly in the context of federal programs to support education reform (e.g., Race to the Top, Investing in Innovation). However, quasi-experiments run an in- creased risk of drawing false conclusions. Suppose POINT had been conducted using only an interrupted time series design (one of the options suggested for evaluating projects supported under the Investing in Innovation program). In that case, the researchers might have deemed performance-based pay a substantial success in Nashville. More sophisticated variants, such as the multiple baseline design, do not fully allay the concerns raised by the absence of a true control group. In this design, a varying time schedule is used to phase in the program—all teachers are eventually made eligible for bonuses, but not immediately, so those who are not immediately eligible serve as a control group for those who are. Such a design has obvious appeal for administrators who do not want to treat some teachers substantially differently from others. But the potential for contamination is great. Teachers who know they will be eligible for performance-based pay in a year or two may begin taking steps now to improve. Some critics of POINT have argued

that the major (even the only) benefit from incentive pay is through the impact on teacher recruitment and retention. The premise underlying this criticism—that there is little prospect of getting better results from current teachers, so the best hope of substantial improvement is to replace the weakest part of the present workforce—is by no means universally shared. It is contrary to the assumptions underlying the account- ability movement. It was certainly not the only hypothesis of interest in the studies of performance incentives cited above. Nor does the widespread use of incentives in business rest on the notion that incentives work primarily (let alone solely) by attracting more productive workers instead of altering current employees' performance.

Nonetheless, this is an important question about which POINT says very little. A test of the hypothesis that financial incentives tied to student performance would change teacher recruitment and retention for the better seems feasible. Clearly, such an experiment would not look like POINT. It would need to run for many years, as do some longitudinal trials in medical research. Credible promises would have to be made to participants in the treatment group (e.g., by placing funds in escrow) that if they enter teaching and per- form at a particular level, bonuses will be forthcoming. A large experiment would probably span multiple districts and would need to avail itself of teacher performance measures already being produced by state systems (such as the value-added assessment systems of Tennessee and Florida). But all this appears doable, if those involved in making education policy are willing to wait ten years for an answer.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and student achievement in Chicago high schools. *Journal of Labor Economics* 25(1): 95–136.

Associated Press. 2011a. Florida commissioner stepping down in June. *Education Week* 30(26): 4–5.

Associated Press. 2011b. Virginia governor rolls out teacher merit pay plan. *Education Week* 30(29): 5.

Cavanaugh, Sean. 2011. Idaho gov. signs merit pay, collective bargaining bills. *Education Week* 30(26): 4.

Clotfelter, Charles, and Helen Ladd. 1996. Recognizing and rewarding success in public schools. In *Holding schools accountable: Performance-related reform in education*, edited by Helen Ladd, pp. 23–64. Washington, DC: Brookings Institution.

Dillon, Sam. 2009. "No Child" law is not closing a racial gap. New York Times, 28 April.

Efron, Bradley, and Robert Tibshirani. 1994. An introduction to the bootstrap. Boca Raton, FL: Chapman & Hall/CRC.

Figlio, David, and Lawrence Kenny. 2007. Individual teacher incentives and student performance. Journal of Public Economics 91(5–6): 901–14.

Fryer, Roland G. 2011. Teacher incentives and student achievement: Evidence from New York City public schools. NBER Working Paper No. 16850.

Goldhaber, Dan, and Dominic Brewer. 1997. Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources* 32(3): 505–23.

Harris, Douglas, and Tim Sass. 2006. Value-added models and the measurement of teacher quality. Unpublished manuscript, Florida State University.

Ladd, Helen. 1999. The Dallas school accountability and incentive program: An evaluation of its impacts on student outcomes. *Economics of Education Review* 18(1): 1–16.

Lockwood, J. R., Daniel F. McCaffrey, Louis T. Mariano, and Claude Setodji. 2007. Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics* 32(2): 125–50.

Marsh, Julie A., Matthew G. Springer, Daniel F. McCaffrey, Kun Yuan, Scott Epstein, Julia Koppich, Nidhi Kalra, Catherine DiMartino, and Art (Xiao) Peng. 2011. *A big apple for educators: New York City's experiment with schoolwide performance bonuses*. Santa Monica, CA: RAND Corporation.

McCaffrey, Daniel F., J. R. Lockwood, Daniel M. Koretz, Thomas S. Louis, and Laura S. Hamilton. 2004. Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics* 29(1): 67–101.

National Commission on Excellence in Education (NCEE). 1983. *A nation at risk: The imperative for educational reform*. Available www2.ed.gov/pubs/NatAtRisk/index.html. Accessed 15 November 2011.

Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA). 2010. *PISA 2009 results: What students know and can do—student performance in reading, mathematics and science*. Vol. 1. Paris: OECD.

Podgursky, Michael, and Matthew G. Springer. 2007. Teacher performance pay: A review. *Journal of Policy Analysis and Management* 26(4): 909–49.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73(2): 417–58.

Sanders, William L., Arnold M. Saxton, and Sandra P. Horn. 1997. The Tennessee Value-Added Assessment System: A quantitative, outcomes-based approach to educational assessment. In *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* edited by Jason Millman, pp. 137–62. Thousand Oaks, CA: Corwin Press.

Schochet, Peter Z., and Hanley S. Chiang. 2010. *Error rates in measuring teacher and school performance based on student test score gains*. Washington, DC: U.S. Department of Education.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.

Snedecor, George W., and William G. Cochran. 1980. *Statistical methods*. 8th ed. Ames, IA: Iowa State University Press.

Springer, Matthew G., Susan Burns, Daniel F. McCaffrey, John Pane, and Ann Haas. 2010a. Teacher pay for performance: Experimental evidence from Round Rock's Project on Incentives in Teaching. Paper presented at the 32nd Annual APPAM Research Conference, 6 November, Boston.

Springer, Matthew G., Laura Hamilton, Daniel F. McCaffrey, Dale Ballou, Vi-Nhuan Le, Matthew Pepper, J. R. Lockwood, and Brian M. Stecher. 2010b. *Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching (POINT)*. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Taylor, L. L., Matthew G. Springer, and Mark Ehlert. 2008. Characteristics and determinants of teacher-designed pay for performance plans: Evidence from Texas' Governor's Educator Excellence Grant (GEEG) Program. National Center on Performance Incentives Working Paper No. 2008–26, Nashville.

Winters, Marcus A., Gary W. Ritter, Joshua H. Barnett, and Jay P. Greene. 2006. An evaluation of teacher performance pay in Arkansas. Unpublished paper, University of Arkansas.

Table 1. Measurement of Teacher Performance

Student	Individual 2006 Math TCAP Score	State Math Benchmarks for 2007	Individual 2007 Math TCAP Score	Individual Difference from State Benchmark
J. Smith	250	270	285	+15
M. King	260	279	277	-2
F. Esposito	265	284	302	+18
L. Davis	255	273	267	-6
A. Aziz	230	255	258	+3
E. Jones	261	280	297	+17
…	…	…	…	…
T. Sawyer	237	260	271	+11
P. Morel	244	265	262	-3
V. Fleming	251	270	285	+15
I. Petrovitch	269	282	285	+3
L. Belkin	253	273	280	+7
		Class average difference		+7

Table 2. POINT Rosters and Appeals

School Year	Rosters Created	Total Number of Appeals	Number of Appeals Approved or Partially Approved	Requested Number of Student Changes	Number of Students Changed
2006-7	143	55	48	188	153
2007-8	107	35	30	83	70
2008-9	84	9	5	16	7

Table 3. Number of Teachers Who Dropped out of the POINT Experiment by Experimental Group and School Year

	School Year		
Experimental Group	**2006-7**	**2007-8**	**2008-9**
Control	2	58	18
Treatment	3	42	23

Table 4. Reason for Attrition by Experimental Group

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