When Does Teacher Incentive Pay Raise Student Achievement?: Evidence from Minnesota's Q-Comp Program

Aaron Sojourner*
University of Minnesota
Carlson School of Management

Kristine West
University of Minnesota
Dept. of Applied Economics

Elton Mykerezi
University of Minnesota
Dept. of Applied Economics

April 19, 2011

Abstract

Since 2005, dozens of Minnesota school districts have implemented pay for performance (P4P) plans as part of the state's Quality Compensation (Q-Comp) program. This paper performs the first systematic study of Q-Comp's impact on student achievement, exploiting variation across districts in the timing of participation as well as in the design of districts' P4P plans to study effects on achievement for grades 3 through 8. Results show a consistent zero average effect of Q-Comp participation on both reading and math achievement. However, effects on reading achievement differ depending on the design of the P4P plan. Specifically, districts offering greater rewards for teacher-centered actions or outcomes evidently experienced large gains in reading (0.11σ/\$1,000 bonus) while those offering rewards based on school-wide goals or formal subjective evaluations did not. Gains from specific P4P design features were not consistently evident in math. We also study effects on other outcomes, such as teacher characteristics and parent demand.

*Corresponding author: asojourn@umn.edu. Thanks to Avner Ben-Ner, John Budd, David Figlio, Paul Glewwe, Michael Lovenheim, Jason Shaw, Chris Taber, and Joel Waldfogel for comments and to Paul Kristapovich, Qianyun Xie, and Yingchun Wang for able research assistance. All errors are ours.

1 Introduction

Teachers vary widely in their ability to produce student achievement gains (Hanushek 1971, Hanushek and Rivkin 2010) but this ability is not predicted by educational degrees or experience beyond the first few years of a teacher's career (Hanushek 2003, Aaronson et al 2007). This has large economic consequences (Chetty et al 2010, Hanushek 2010), which motivates policy and research interest in pay for performance (P4P).

Advocates of P4P believe that tying teacher compensation to performance will support increased efforts from incumbent teachers and attract better potential teachers to the profession (Lazear 2003). Many school districts and states are experimenting with P4P plans, which set compensation criteria beyond the conventional ones: experience and education. Plans differ on many dimensions including whether teachers are rewarded individually or in teams, whether for test scores or other measures of teacher quality, and in the magnitude of incentive pay available.

Empirical evidence on the relative and absolute merit of these programs is decidedly mixed. While reviews of the literature point to some gains from P4P (Springer and Podgursky 2007, Neal 2011), evaluations of two large-scale P4P plans that were implemented as randomized trials found null or even negative effects on student achievement (Springer et al 2010, Fryer 2011). Whether plans implemented as long-term policies rather than short-term experiments or plans with other designs would produce better results remains an open question of great interest to policymakers, educators, and economists.

Absent a clear picture of what optimal compensation reform policy should look like and given the political challenges to adoption, the U.S. Department of Education has taken a decentralized approach. Its Race to the Top Fund and Teacher Incentive Fund are allocating billions of dollars for locally-designed reforms via competitive grants. The Department of Education sets guidelines, applicants submit proposals, and federal funds are distributed to support approved plans.

The State of Minnesota took a similar approach when it implemented the Quality Com-

pensation (Q-Comp) program in 2005 as the signature education initiative of Governor Tim Pawlenty. The Minnesota Department of Education (MDE) set general guidelines for acceptable programs and invited districts to propose specific P4P program designs that they would implement. If the proposal was approved, the state authorized up to $260 per student per year in additional funding to the district.

Districts designed plans that varied along many dimensions. Each district was required to specify the maximum incentive pay they would make available to teachers based on different types of criteria and there is great variation in what they chose. This allows us to construct continuous measures of each district's plan in terms of dollars at stake for: (1) individual teacher actions or outcomes, (2) school wide goals, or (3) through a subjective evaluation process. We exploit this variation in P4P plan designs, along with variation in when districts adopted Q-Comp, to provide evidence on the effect of plan design features on achievement scores and other outcomes.

For a number of reasons, Q-Comp provides an excellent opportunity to learn about the effects of P4P in a policy framework mirroring recent national efforts. First, the program has been in effect for over five years and was implemented as a permanent program rather than a time-limited experiment. Second, there is substantial variation in what criteria trigger P4P bonuses. Third, Minnesota has one of the longest lasting and most widely used inter-district open enrollment policies and a large number of charter schools, so parents have substantial choice in public schooling. This makes it possible to examine the effect P4P designs have on net student movements, which can reflect changes in parent demand. Overall, understanding the Minnesota P4P experience can provide valuable information to policy makers nationwide.

In addition to policy relevance, Q-Comp can provide new evidence on several issues of theoretical importance related to P4P contracts in education. For instance, it is not clear what the optimal team size for targeting bonuses should be. On one hand, incentives tied to school-level criteria may encourage efficient effort if there are positive externalities from cooperation (Weitzman and Kruse 1990) or variations in incentive strength across teachers (Ahn 2008). On the other hand, free riding may make individual or small team incentives

preferable (Kandel and Lazear 1992). Since Q-Comp districts experimented with a wide range of P4P contracts, we are able to investigate whether incentives offered at lower levels of aggregation (such as the individual teacher or grade) are more or less productive than those offered at higher levels of aggregation (such as the school or district level).

There are also important theoretical questions about how to measure teacher quality and performance. Measures based on principal or peer subjective evaluations have received some attention in the literature, especially since principals seem able to identify the best and worst teachers (Jacob and Lefgren 2008). However, high-stakes subjective evaluation processes may be captured and converted into de facto salary augmentations (Neal 2011). Minnesota's Q-comp offers a valuable opportunity to examine if a high-stakes P4P plan based on subjective evaluations affects educational outcomes.

Subjective evaluations have also received attention because they can provide a solution to multitasking problems (Holstrom and Milgrom 1991). Attaching high stakes to standardized test scores may induce teachers to narrow the curriculum, teach to the test, or worse. Subjective evaluations may alleviate this problem by harnessing richer information about a teacher's performance (Baker 1992, Rockoff et al 2010, Tyler et al 2010). Q-Comp's diverse P4P plans, along with the availability of net student movements in an environment of open enrollment as well as student attendance data, used as a proxy for family engagement with learning, allows us to examine if there are significant issues related to multitasking. If multitasking issues are important, incentives tied to test scores may have positive effects on test scores but adverse effects on other educational outcomes and incentives tied to broad subjective evaluation criteria may have adverse impacts on test scores but positive impacts on other measures of quality such as parent demand.

This study finds that districts that put higher stakes on individual teacher (or small team) level actions or outcomes experience bigger increases in student reading scores in standardized tests, on the order of 0.11 standard deviations per $1,000 of bonus offered. This is a shockingly large estimated effect for a relatively low price.[1] The finding appears quite robust within the

[1] Based on Hanushek (2010) and Chetty et al (2010), a lower bound on the social value of a 0.2σ achievement gain for a teacher's class each year is conservatively $200,000.

limits of our study design. Districts that link more rewards to outcomes measured at the school or district level do not see increases in reading achievement and districts that link rewards to subjective evaluations see small declines in reading scores. There is evidence of similar effects in math under some specifications but estimates are less precise and the overall pattern is suggestive of no effects on math.

The paper finds no evidence that multitasking provides an explanation for the small negative effect of dollars tied to subjective evaluations on test scores. These dollars are also found to reduce enrollment and do not have a statistically significant positive impact on net pupil movements or attendance rates. School P4P dollars, on the other hand, have no effects on student test score, to which they are almost always explicitly tied, but there is some suggestive evidence that they may increase net pupil inflows.

In addition to the inquiry into the heterogeneous effects of program design features, we investigate the average effect of the Q-Comp program overall. We find no average effects on reading or math achievement. While many districts chose programs that created gains, others chose programs that offset these to produce no effect on program participation on average.

The paper proceeds as follows. Section 2 provides more detail on the Q-Comp program and some of the challenges and opportunities it presents for study. Section 3 briefly reviews relevant theoretical and empirical literature. Section 4 introduces a model and discusses identification. In section 5 we present results on the relative success of different plan designs along with robustness checks and results on the effects of Q-Comp participation in the aggregate. Section 6 concludes with a discussion of how our results add to the existing literature on P4P in education as well as plans for future research.

2 Design of and selection into Q-Comp

2.1 Q-Comp participation

Q-Comp is sizable. Since its inception in 2005, over one million student-years have been taught in dozens of participating districts and charters and over $200 million of state funds

have been distributed to districts. As one of the nation's largest teacher P4P programs, Q-Comp has attracted significant policy and political attention. Yet little is known about the designs of the local P4P plans it funds or their effects.[2]

Selection into Q-Comp works as follows. The state promised additional annual funding to districts that would implement Q-Comp plans and defined guidelines regarding the content of these plans. Districts (and charters) decided whether to apply and proposed specific program designs. The state then decided whether to accept the proposal. Where teachers were unionized, teachers voted on whether to accept the proposed changes.[3] Districts that clear all these hurdles then participate in Q-Comp.

Since 2005, new districts joined the program each year. Table 1 on page 7 describes the number of districts, schools and students participating and not participating in Q-Comp each year. The population is all Minnesota public schools, including charters each constituting its own "district." In 2005, only eight of the state's 504 districts participated (1.6%). These included 59 of the 2,256 schools with 33,674 of the 838,997 students (4.0%). By the 2009-10 academic year, 14.1% of districts with 28.6% of students participated.[4] Most analysis will focus on grades 3 to 8 because, in these grades, all students took both math and reading tests. Participation statistics are provided for schools in this sample in the bottom panel.

2.2 District P4P Design Features

Q-Comp is a package of reforms with P4P at its center. We focus our attention on the performance pay component because these are the most interesting theoretically, the best measured in the available data, and the most likely to constitute a real policy change from the pre-adoption period. Data on performance pay are collected primarily from letters sent by the MDE to each district upon approval of its Q-Comp application. The letters detail agreed-

[2] Neal (2011) summarizes U.S. and international empirical evaluations of performance pay plans and notes that there has been no previous independent study of Q-Comp. Nadler and Wiswall (2011) study selection into Q-Comp and do not address whether Q-Comp impacts student achievement.

[3] Almost all Minnesota districts are unionized though many charter schools are not. They usually informally negotiated the proposals with administrators in advance and officially ratified after state approval.

[4] A few participating districts dropped out of Q-Comp. These tables reflect stock given exit and entry flow.

Table 1: District and School Q-Comp Participation by Year

Year	Participants			Non-Participants		
	Districts	Schools	Students	Districts	Schools	Students
All schools						
2005-06	8	59	33,674	496	2,197	805,323
2006-07	50	322	183,216	458	1,922	657,346
2007-08	60	397	231,465	456	1,856	606,113
2008-09	70	429	252,716	457	1,786	583,218
2009-10	74	411	239,489	451	1,796	597,141
Schools including at least one grade in 3 to 8						
2005-06	7	52	23,131	404	1,511	567,202
2006-07	36	255	129,754	379	1,338	463,862
2007-08	43	309	162,499	379	1,278	462,980
2008-09	52	328	176,870	381	1,258	413,023
2009-10	56	315	166,697	375	1,256	427,549

upon features of the plan.[5] From these, we create three variables for each district measuring the maximum performance pay available to teachers for the following types of criteria:

1. Teacher P4P$: anything under a teacher's primary influence. This includes inputs related to professional development (e.g., attending meetings, taking classes, completing professional development plans and self-assessments) and outputs close to the teacher (e.g., student performance on teacher-created assessments, a teacher's own students' standardized test scores). It also includes analogous small team or grade-wide outcomes. The descriptions in the approval letters do not allow us to consistently distinguish between various elements within this domain.

[5]The stated policy, taken from the application guidelines available online, requires that each district enumerate the maximum incentive pay available to teachers for meeting various goals. These include the amounts of performance pay each teacher

1. is eligible to earn for meeting specified goals for student achievement measured at the teacher, team, or grade level by formative, summative or standardized tests
2. is eligible to earn for meeting specified goals for student achievement measured school-wide or district-wide
3. will earn through the teacher evaluation/observation process.

In creating variables for use in our analysis some judgments were necessary in interpreting the letters. Our results are robust to different interpretations of vague letters. These additional results are available upon request.

2. School P4P$: anything linked to school-wide or district-wide outcomes. These primarily involve hitting standardized test score targets.

3. Evaluation P4P$: anything linked to classroom observations or subjective evaluations performed by peers, administrators, or a district-sanctioned mentors including a formal annual review process.

Districts vary in the total levels of pay available across these three dimensions as well as the shares available through each dimension. The value of these variables is shared by all school-grades within a district-year. Table 2 summarizes the distribution of these measures across participating districts.[6] We prefer the summary statistics where district observations are weighted by the number of students tested because this approximates the average at the teacher level and because the regression results use these weights. Participating teachers can earn an average maximum of $815 a year in incentive pay through actions or outcomes closely connected to them (Teacher P4P$). They can earn an average maximum of nearly $1,000 by meeting criteria tied to annual evaluations and classroom observations (Evaluation P4P$). School or district level goals – usually based on student achievement test scores – can earn them an average maximum of around $250 (School P4P$). The triggers for paying out on these dimensions are set according to various locally-designed criteria within and across districts.

Table 2: Summary statistics for district Q-Comp program design variables measuring maximum pay available through each dimension, in thousands of dollars

	Unweighted		Weighted by students			
	Mean	Std. Dev.	Mean	Std. Dev.	Min.	Max.
Teacher P4P$	0.682	0.585	0.815	0.652	0	2.5
School P4P$	0.334	0.342	0.247	0.225	0	2.5
Evaluation P4P$	0.727	0.58	0.987	0.743	0	2.5
Number of participating districts			77			

Note: the 2010 cohort included additional districts but their plans are not coded.

[6] If only the share assigned to each dimension, and no dollar values, were listed in the approval letter nor in any available program documents, we assumed the modal total amount among observed districts: $2,000. Five letters were so ambiguous as to be impossible to code on these dimensions. Although districts and schools may change their designs over time, at this point we do not have measures of this, so we assume they stay constant at the initial levels. White noise measurement error would bias estimates to zero.

Table 3 describes how the maximums in the three dimensions are correlated. For Teacher P4P$, we observe a 0.44 correlation with School P4P$ and a -0.21 correlation with Evaluation P4P$. We also observe a -0.02 correlation between School P4P$ and Evaluation P4P$. So programs that reward teacher-centered actions or outcomes also tend to reward school- or district-level goals at the expense of rewards based on subjective evaluations.

Table 3: Correlation of districts' maximum pay available by dimension, weighted

	Teacher P4P$	School P4P$	Evaluation P4P$
Teacher P4P$	1.00		
School P4P$	0.44	1.00	
Evaluation P4P$	-0.21	-0.02	1.00

Figure 1 on page 10 displays the joint distribution in participating districts across program designs. Each point represents a district's (or a charter school's) Q-Comp P4P design. The size of each point represents the maximum total bonus available to teachers in that district, the sum of Teacher, School and Evaluation P4P$. Each district's share of awards tied to Teacher P4P$ criteria is graphed horizontally. The share tied to School P4P$ criteria is graphed vertically. The remaining share, tied to Evaluation P4P$ criteria, is represented by the distance to the frontier. For instance, the large dot appearing on the frontier represents a district with a plan offering each teacher over $4,000 in additional pay annually. Being on the frontier line means that none of the bonus is tied to subjective evaluation. Half the bonus is tied to Teacher P4P$ criteria and the other half to School P4P$ criteria. The small dot at the origin represents a district with a plan that awards between $1,000 and $2,000 based solely on subjective evaluations.

This graph makes some important points about the program designs clear. First, there is a lot of variation across districts in these four dimensions. They do not cluster around some general "optimal contract." Second, few districts offer less than $1,000 or more than $4,000 in total and most districts offer a mix across all three dimensions. Few lie on the edges of the triangle. Third, none offer more than half tied to School P4P$ criteria, although there is a lot variation in shares below half. Because of this variation in program design, analyzing

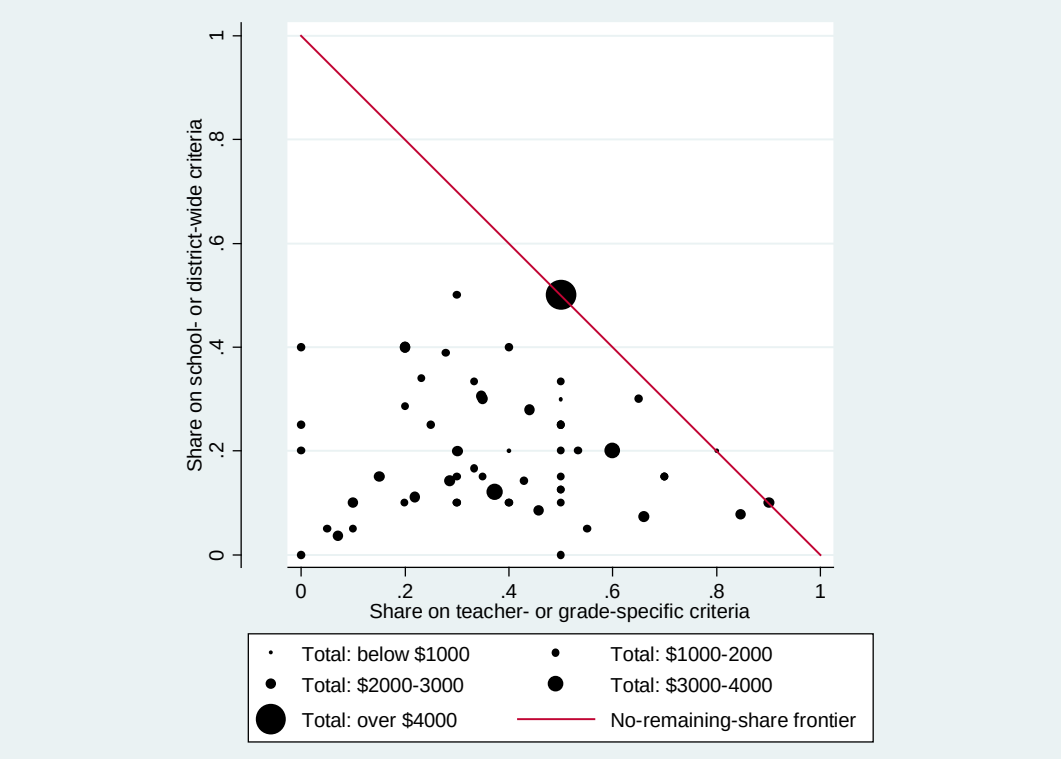


Figure 1: Joint distribution of P4P designs across Q-Comp districts

Q-Comp in the aggregate likely masks important differences across districts. Accounting for this heterogeneity in program design provides a unique opportunity to understand how P4P plan specifics impact educational outcomes.

One additional element of P4P plan designs is worthy of note. Many Q-Comp districts' approval letters also specified the academic subjects to which districts elected to tie their school-wide performance bonuses (School P4P$). These often varied by school within the same district. We coded whether these goals focus on reading or math. Schools were more likely to tie School P4P$ to reading rather than to math achievement. Three times more school-grades (15.6%) chose to focus exclusively on reading rather than exclusively on math (4.5%). The remainder divided their attention between math, reading, or other subjects. This provides an additional opportunity to examine a number of issues.

3 Literature Review

The substantial variation in the plans' designs allows us to address at least two major issues where economic theory provides ambiguous guidance and existing evidence is thin. First, as described in the organization design literature on team production, there is a trade-off between offering rewards based on individual versus team outcomes. Given complementarities in production, individual-level incentives can discourage productive cooperation. This pressure pushes for incentives tied to criteria defined at the team level. On the other hand, team incentives also open the door for free riding, a problem that worsens in team size. In schools, grade levels are a natural grouping and additionally, many middle and high schools are organized into even smaller teams of core subject teachers. These groups may be small enough to exert sufficient peer pressure to overcome the free-rider problem (Kandel and Lazear 1992).

In our analysis of Q-Comp, we group together rewards based on criteria defined at the teacher level with those at the small-group level (i.e. subject teams or grade-level groups). We distinguish these Teacher P4P$ from School P4P$ tied to criteria defined at higher levels, such as whole schools or districts, where the free rider problem may be especially severe. Previous empirical literature on the effect of incentives at various levels is mixed.

A randomized trial in Nashville, Tennessee, in which teachers assigned to the treatment group could earn up to $15,000 based on their individual students' gains on state mathematics tests, found no significant treatment effect on student achievement (Springer et al 2010). We find that Q-Comp districts see increases in reading achievement on the order of 0.11 standard deviations per $1,000 Teacher P4P$ bonus available.[7]

Most Q-Comp plans include some school- or even district-wide measures of success in determining bonuses. P4P plans in North Carolina and New York reward exclusively at the school-level. Analysis of the North Carolina program finds no significant impact on

[7] In a randomized trial in Andhra Pradesh, India, Muralidharan and Sundararaman (2011) finds that individual rewards do impact student achievement. Their evidence suggests that individual and small group rewards both had a positive impact on language and math exams with effect sizes between 0.12 to 0.27 standard deviations. In the first year, individual and small group rewards were equally effective. In the second year, individual rewards were more effective. However, in their study, 92% of the treatment schools had between two and five teachers. This is a similar size to small groups of teachers defined at the team or grade level in Minnesota. Also, average teacher effort levels are much lower in India, as measured by high absenteeism.

student achievement resulting from school wide bonuses (Vigdor 2008). Fryer (2011) presents evidence from a randomized trial in New York City also showing no positive impact on student achievement. In fact, he writes that the school-level bonuses may have decreased student achievement, especially in larger schools. We do not find a significant impact on school- or district-level rewards.[8]

Second, theory provides mixed predictions on the role of subjective versus objective measures of teacher quality. Most P4P plans adjust pay based on objective measures of teaching quality such as test scores. These may not fully capture teachers' ability to affect critical thinking, non-cognitive skills, or other unobserved yet valuable aspects of learning. This is a multitasking problem (Holstrom and Milgrom 1991). High powered incentives tied to test scores may induce teachers to spend too much time on tested skills at the expense of other socially valuable skills, leading to "teaching to the test" or a "narrowing of the curriculum."

Adding subjective evaluation criteria may mitigate this problem (Baker, Gibbons and Murphy 1994). Subjective evaluations are especially attractive in schools because they can be used in non-tested subjects. Principals are able to distinguish effective from ineffective teachers as measured by value-added statistics (Jacob and Lefgren 2008, Rockoff et al 2010; Tyler et al 2010). Interestingly, the principal's overall rating of the teacher is a better predictor of future parent requests for the teacher than value-added statistics (Jacob and Lefgren 2008). This suggests principals' judgments about which teachers contribute to overall student learning agree with parents' judgements better than test scores do.

However, principals may be reluctant to use their knowledge of teacher effectiveness when making high-stakes decisions (Jacob 2010, Neal 2011). Neal (2011) attributes the failure of P4P programs in England (Atkinson et al 2009) and Portugal (Martins 2009) to the fact that they were largely based on subjective evaluations done by local staff. Such plans may not improve student achievement because evaluators lack incentives to assess teachers accurately. Neal asserts, with specific mention of Q-Comp, that plans which base pay on locally-defined

[8]Contrasting evidence from Israel indicates that school level incentives based on measures such as the average number of credits per student and the dropout rate had a positive and significant impact on student test scores (Lavy 2002).

goals and locally-conducted evaluations can become a "vehicle for raising base pay of most or all teachers whether or not these teachers improve their performance."[9]

We distinguish between Q-Comp plans that focus on subjective evaluations and those that tie bonuses to specific actions or student outcomes and find some support for Neal's prediction. Almost all Q-Comp contracts reward subjective evaluations but they differ in the amount of money at stake for a successful evaluation and in the details of the process or payout criteria.[10] While we are unable to say whether the evaluators correctly identify effective teachers, we can test how making the subjective evaluation process central to a P4P plan affects student achievement. Our evidence suggests that plans which tie bonuses primarily to evaluations may even be detrimental to student achievement on standardized tests. Districts with larger rewards linked to evaluation appear to score slightly worse in reading than they did before Q-Comp adoption.

Neal's concern that subjective evaluations may be made into de facto salary augmentations is premised on the assumption that schools do not face the competitive pressures that are at work in the private sector. In Minnesota, however, there are two mechanisms for subjecting districts to the discipline of the market. Minnesota has the nation's longest standing open enrollment and charter school laws. Parents can enroll their children in any public district or charter as long as there is space. Because of this, we also test the impact of Q-Comp on net pupil movements between regular districts and from districts to charters.

[9]There is evidence that almost all teachers in Q-Comp districts earn at least *some* performance-based pay, often through the subjective evaluation portion. A Minneapolis Star-Tribune investigation found that, in the 22 Q-Comp districts they researched, only 27 teachers got absolutely no performance payment out of the roughly 4,200 teachers eligible. (2/1/2009, Minneapolis Star Tribune). The article reports that "...many educators say [Q-Comp] is strengthening teacher evaluations and training. But others are questioning whether Q-Comp has just become a cash handout." However, not everyone earns the maximum evaluation payout nor meets the teacher-centered or school- or district-level standards based on student achievement. There are incentives unclaimed so this is not strictly a cash transfer program. In order to get teachers' assent for Q-Comp adoption, which creates uncertainty and variance in their future pay, teachers may demand some augmentation to base pay. Dollars tied to subjective evaluation may be the vehicle for it.

[10]The wording in many Q-Comp plans indicates rewards given simply based on the completion of a subjective evaluation cycle (i.e. a set number of observations coupled with pre- and post-observation meetings with the evaluator). Not all plans differentiate pay based on the evaluator's assessment of teacher effectiveness.

4 Model and Data

To learn about the impact of Q-Comp on student achievement, we analyze a panel of student achievement, demographic, and school characteristic data defined at the year-school-grade level using generalized difference-in-difference methods. Our primary achievement measure is the Minnesota Comprehensive Assessments Series II (MCA-II) average scores. Since 2005-06 (coincidentally the first year of Q-Comp), these have been mandated for every student in third to eighth grade in both reading and math.[11]

We study how schools' student achievement changes as their Q-Comp participation changes. The main outcome is average student achievement on MCA-II tests each academic year indexed $t = 2005, 2006, ...2009$, in each school indexed $s = 1, 2...S$, in each tested grade indexed $g = 3, 4, ...8$, and in either math or reading indexed $b \in \{M, R\}$.[12]

In explaining average student achievement, we use variants of this model:

$$y_{tsgb} = \beta_{gb} Q_{tsgb} + \alpha_{gb} w_{tsg} + \gamma_{sgb} + \delta_{tgb} + \epsilon_{tsgb} \tag{1}$$

Interest centers on the effects of Q-Comp participation and of features of the P4P designs adopted. These are captured by β_{gb}. In the simplest form, we will characterize Q-Comp participation as either

(A) a simple participation dummy, 1(Post-adoption) or

(B) the participation dummy paired with a dummy indicating academic years two or more years prior to adoption, 1(2+ pre-adoption) (reference category: the single year immediately prior to adoption).

Moving across specifications (A) and (B) the definition of Q changes but the model is otherwise the same. Specification (A) treats the whole pre-adoption period as the reference

[11] Unfortunately, Minnesota switched testing regimes in 2005. Prior to 2005, only grades 3, 5 and 7 were tested and on a different test.

[12] Before third grade, students are not tested. After eighth, tenth graders are tested only in reading and eleventh graders only in math. Estimated effects for these two series are also available on request.

category. Specifications (B) conditions on and measures pre-adoption differences in achievement levels between adopters and non-adopters using a dummy to indicate observations that come from years more than one year pre-adoption. The fact that our data start in the year that the first cohort adopted and that different-sized cohorts adopted during each year of our data generate imbalances in what data are available to identify various parameters. All observations from more than one year pre-adoption come from the smaller cohorts of districts that adopted in 2007, 2008 or 2009. Defining Q as a simple participation dummy allows β to measure the effect of Q-Comp participation on average.

To measure the effects of various aspects of Q-Comp P4P program design, we use different definitions of Q. We define Q to be a vector measuring P4P design features in years that they are participating and zeros otherwise. In particular, as discussed in section 2.2 we measure maximum P4P bonus available to teachers in a district for three kinds of criteria, measured as Teacher P4P$, School P4P$, and Evaluation P4P$.

Additionally, we include a variable to indicate district-years where the district once participated in Q-Comp but has since dropped out. This only affects a small number of districts. If the estimated coefficient on this dummy is negative it indicates districts do worse after leaving Q-Comp than they did in the year(s) prior to adoption.

Since Q-Comp participation is not randomly assigned, there may be systematic unobserved differences between districts that influence both Q-Comp adoption and our outcomes, which would bias estimates of program effects. We use four main strategies to guard against this threat.

First, since within any given school and grade, average student achievement may vary over time due to differences in student cohorts, we condition on a vector of year-school-grade student demographic characteristics and school-level variables (w_{tsg}).[13] These are listed in the top panel of Table 6 on page 33, which also provides summary statistics. These characteristics do not vary across subject, although their coefficients α_{gb} can.

Second, school-grade-subject fixed effects (1_{sgb}) are included to remove time-invariant,

[13]This alone improves over available evidence on Q-Comp's effects. Neither a legislative auditor's report (Nobles, 2009) nor a state-commissioned external report (Hezel Assoc., 2009) dealt with selection or covariates.

additive unobserved differences in achievement levels γ_{sgb} between schools. The model is identified from within-school-grade-subject, across-time variation.

Fixed effects for each year-grade-subject 1_{tgb} are also included.[14] These terms identify counter-factual year effects for each grade and subject (δ_{tgb}). The comparison group matters here because their experience across years defines these time effects. This is a generalization of difference-in-difference analysis that relies on differences in the timing of adoption across districts to separate time effects from program effects.[15]

The model is identified by assuming that program variables Q_{ts} are uncorrelated with unobserved influences ϵ_{tsgb} conditional on other observables, school-grade fixed effects, and year-grade fixed effects,

$$Cov[Q_{ts}, \epsilon_{tsgb} | (w_{tsg}, 1_{sgb}, 1_{tgb})] \equiv 0 \tag{2}$$

Within the restrictions of functional form, this model yields unbiased estimates of program effects even if selection into Q-Comp is based on stable differences in achievement levels. If, for instance, schools with higher achievement levels are more likely to adopt or to adopt earlier than schools with lower achievement levels, that is not a problem. The crucial assumption is that within-school, time-varying, unobserved influences on achievement levels are not systematically related to whether or when a school adopted Q-Comp or the features of the design it adopted. The estimates of β may be biased if districts select into participation or design based on fluctuations in achievement levels. For example, if a school is more likely to adopt in a year when levels would rise for other reasons than in a year when they would fall (perhaps, districts experimenting with Q-Comp are also experimenting with other reforms), this violates the identifying condition and would bias the estimated program effect upwards.

[14] Also, each year-grade-subject distribution of student scores is standardized to have mean zero and standard deviation one in order to facilitate pooling across grades.

[15] The first difference is the within-school comparison across time periods. The second difference is between the first-differences at adopting schools and those at non-adopting schools across the same time period. A within-school change between any two points in time is evaluated against changes across those same two years among other schools. With a simple participation dummy, β_{gb} measures the difference in average grade-g, subject-b achievement within adopting-schools in the years after adoption compared to the years prior to adoption conditional on changes in w_{tsg} and the average change experienced across these years by other schools. Lovenheim (2010) uses a similar approach to study the effects of teachers unionizing in different districts over time.

Third, we estimate these models with three different comparison groups. We compare the experience of participants to that of either (1) all other schools in the state, (2) districts that applied to Q-Comp but failed to adopt, due either to the state rejecting the proposal or their teachers voting against it,[16] and (3) just Q-Comp adopters who have not yet adopted. We refer to comparison group (2) as "interested" nonadopters and to the these three samples as the full, interested-only, and adopters-only samples, respectively. Excluding never-appliers from the analysis reduces precision because they contain information about the effect of observable characteristics w and the time effects δ. However, excluding them can reduce bias if they are fundamentally different from adopters or applicants in unobservable, time-varying ways. Also, unlike never-applying districts, interested nonadopters passed the first hurdle to participation; they choose to apply. Some even cleared the second hurdle (state approval). In this sense, interested nonadopters are more similar to adopters than the never-appliers are. Parameter estimates across all samples are provided for comparison and results turn out to be very stable.

Figures 2 and 3 on pages 18 and 19 present trends in average reading and math achievement among each adoption cohort, the cohort of never-adopters, and among interested non-participants. In each grade-year-subject, test scores are normalized to be mean zero and standard deviation one across schools weighted by students tested.

There are three points to make about these trends. First, there are differences in average achievement levels between cohorts. The never-applied cohort is the largest and hovers just below state mean achievement throughout the period. The interested non-adopters' trends are just below the never appliers generally. Among Q-Comp adopters, the 2005 and 2010 cohorts are most similar to the never-adopters and the interested non-adopters on average.[17] However, the 2006, 2007, and 2008 adoption cohorts were higher achieving than average. The 2009 adopters are lower achieving on average, around a half to a full standard deviation below the mean in math and reading. Second, there do not seem to be large differences in achievement

[16] Failed applications had to be obtained through a Freedom of Information Act to the state Dept. of Education.

[17] Achievement data for the 2010-11 year is not yet available.

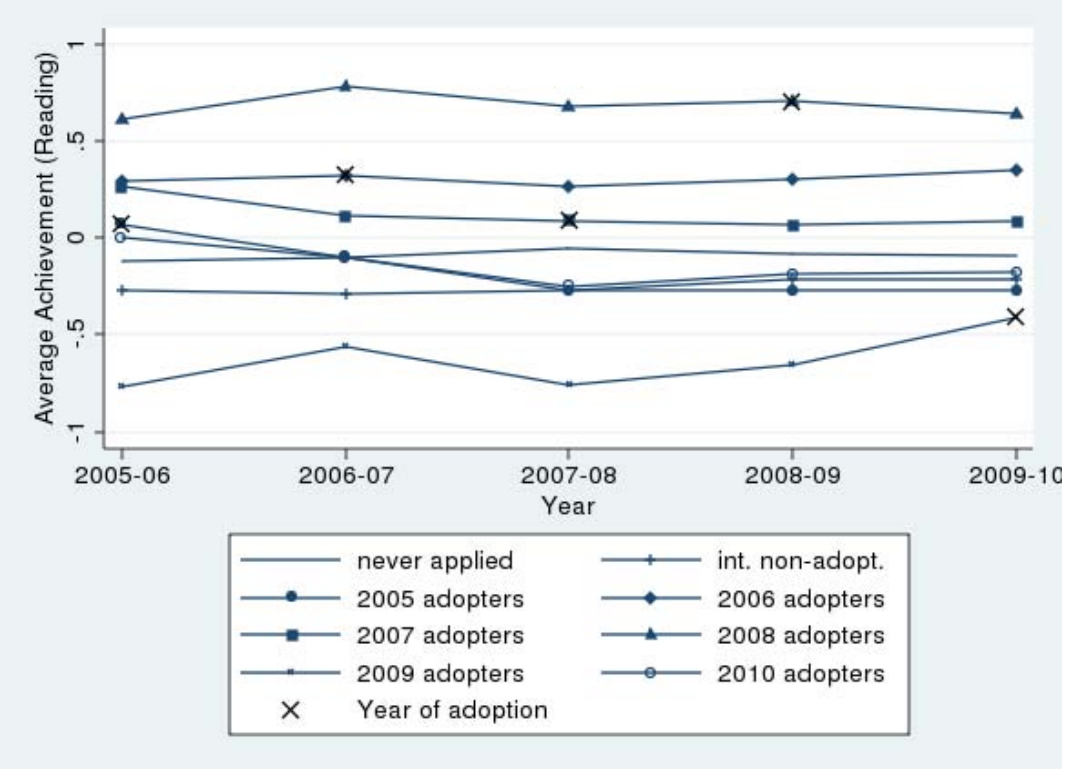


Figure 2: Trend in average reading achievement by Q-Comp adoption cohort

trends between cohorts, aside from the fluctuations in the very small 2009 cohort. Third, this foreshadows one of our main conclusions: the effects of Q-Comp participation appear to be null on average. Increases in achievement do not seem to follow Q-Comp adoption in the aggregate.

The above school-grade-subject level models can, however, still be susceptible to endogenous plan design based on time varying unobservables. If administrators were able to forecast future achievement and successfully and designed plans that differed based on these forecasts the difference in difference estimators above could produce biased results.[18]

So, lastly, we estimate growth models that condition on lagged achievement. We do this at the district level in order to avoid losing students as they transition between schools across grades. At the district level we can obtain lagged scores based on all students, adding

[18]We say "successfully" because if unobservable were simply correlated with application the "interested only" sample still produces unbiased results. So time-varying unobservables have to be correlated with actual implementation and/or design of P4P plans upon implementation.

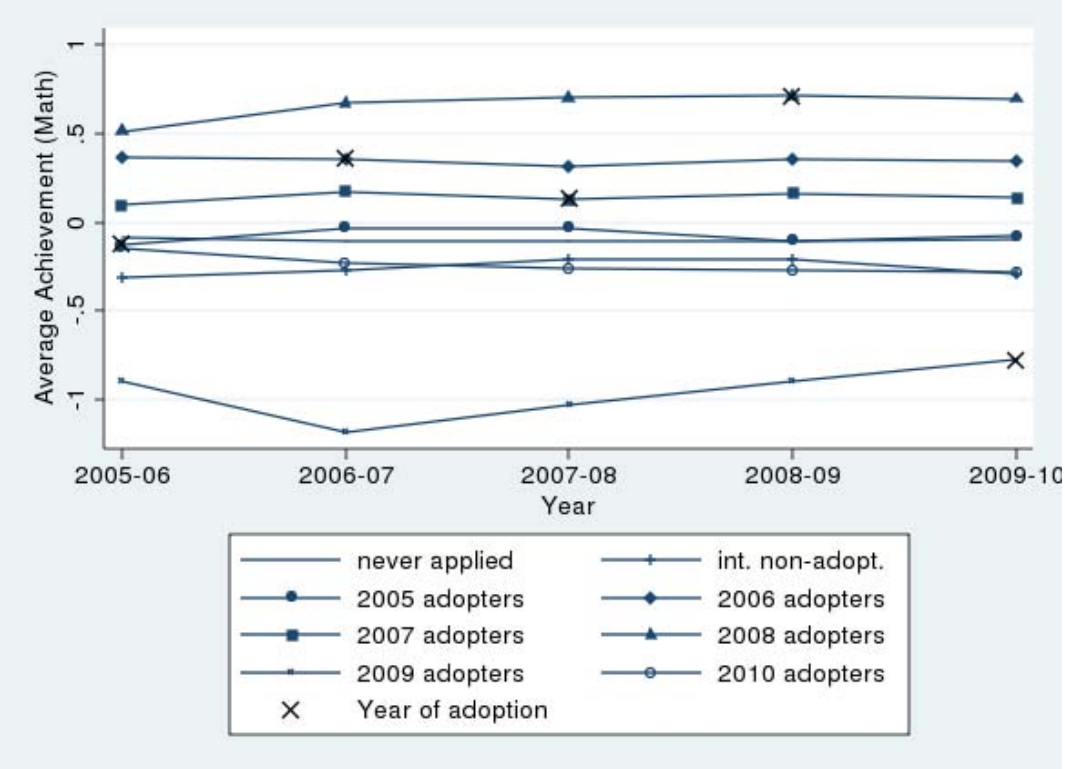


Figure 3: Trend in average math achievement by Q-Comp adoption cohort

$\vec{y}_{(t-1)d(g-1)}$ as a covariate to explain $y_{(t-1)d(g-1)b}$. These specifications do not use all the variation across grades and schools that the above models do but are more robust to omitted time varying variables that affect P4P plans and achievement growth. Here the results turn out to be even stronger than in the levels models.

In order to boost power, our primary results pool across grades 3 to 8 and restrict program effects to be the same across grades within subject, $\beta_{gb} \equiv \beta_b$ but some results by grade level are also provided.

Although most participation decisions are made at the district level, a few large districts allowed individual schools to participate in the program, so the participation decision was coded accordingly at the school level. Use of school-grade level data increases precision. However, standard errors adjusted for correlation at the district level are provided throughout.[19]

[19]Clustering at the school of school-grade level provides identical estimates and smaller standard errors. However, we present the most conservative estimates.

5 Results

5.1 Did Q-Comp Change Teacher Pay and Incentives?

Before examining the impact of Q-Comp P4P design features we first present three pieces of evidence on whether Q-Comp actually changed teacher incentives as advertised. It is often the case that innovative grant recipients find ways to elicit funds to accomplish tasks that they would perform even in the absence of the grant. Null effects would result. Was this the case with Minnesota school districts? By supplementing our data with information from the National Center for Educational Statistics' Schools and Staffing Survey (SASS) and by conducting a survey of district human resource professionals, we provide generate evidence that Q-Comp actually did change the way teachers were paid.

The SASS asks districts whether they use any pay incentives to "reward excellence in teaching." Q-Comp participation is significantly associated with switches from "No" before Q-Comp adoption to "Yes" after adoption among Minnesota districts. Table 4 reports on the 55 Minnesota districts sampled in both waves. Among districts not participating in Q-Comp in 2007-08, 96% report no pay for excellence both before and after Q-Comp started. Among districts participating in Q-Comp in 2007-2008, none reported paying for excellence before Q-Comp, in 2003-04. However, in contrast to the nonparticipants, 58% of participants report paying for excellence in the post-adoption SASS survey wave. This suggests that, many districts perceived something programmatic to have changed. The fact that 42% of surveyed Q-Comp districts still reported no pay for excellence also suggests that not all districts experienced deep changes or conceptualized Q-Comp in this way.

In order to get more detail on the particular aspects of the P4P programs implemented in Q-Comp schools, we conducted an independent phone survey of district human resource professionals about their district pay practices and without mentioning Q-Comp. It found that participating districts are significantly different from nonparticipants in how they compensate teachers. We obtained data from 92 districts (38% response rate). Twenty-one of these, we know from administrative data, participate in Q-Comp. Table 5 summarizes our findings.

Table 4: Evidence on change in "Pay for Excellence" among Minnesota districts by Q-Comp participation status from the Schools and Staffing Survey (SASS)

			Can teachers earn extra pay "for excellence"?				
District in Q-Comp in 2007-08	Districts in both waves	In 2003-04 In 2007-08	No No	Yes Yes	No Yes	Yes No	Total
Yes	12		42%	0%	58%	0%	100%
No	43		96%	2%	0%	2%	100%

Note: only these 55 districts appear in both the 2003-04 and 2007-08 waves of SASS. Q-Comp began in 2005.

Among Q-Comp participants, 86% report paying for student performance and 90% report paying for subjective evaluations. In stark contrast, none of the non Q-Comp districts report paying on either of these dimensions. Unsurprisingly, participating districts are just as likely to pay for years of experience and educational credentials as are non-participants. Q-Comp P4P is a supplement to, rather than a replacement of, traditional compensation criteria.

Table 5: Evidence on Q-Comp's impact on compensation from author survey in 2010

Districts in Q-Comp in 2010-11?	Percent of districts paying for: Student Perform.	Subjective Evaluation	Years of Experience	Education Credentials	N
Yes	86%	90%	95%	95%	21
No	0 %	0%	100%	100%	71

Lastly, using mean teacher pay as a dependent variable in our analysis shows that the introduction of Q-Comp is associated with a 2.5% increase in average teacher salaries. This is consistent with an average salary of $55,000 and an average Q-Comp bonus paid of $1,375 per year per teacher in participating districts.

5.2 Impact of P4P Design

We begin by estimating the impact of program design parameters on standardized test scores. Table 7 presents estimates for the effects on reading pooled across grades 3-8. As noted, specification (A) compares scores in post-adoption years to all pre-adoption years. Specification (B) compares post-adoption years to the single year prior to adopting. All specifications

condition on time-varying student demographics, school-grade effects and grade-year effects. The full sample includes 4,677 school-grades with multiple observations across years for each. Together they include 1,749,818 tested student-years. Each school-grade-year-subject observation is weighted by the number of students tested.

Schools which offer more performance pay based on factors over which teachers exert closer control produce large achievement gains in reading. This result is consistent across alternative comparison groups. Columns 1 and 2 present estimates using the full sample, columns 3 and 4 present estimates using only the sample of interested districts (those that ever applied for Q-Comp), and columns 5 and 6 present estimates using only districts that ever participate in the program at some point. The parameter estimates are positive and significant across specifications and samples, ranging from 0.112 (0.030) to 0.145 (0.032) per $1000 available.

The parameter estimates on School P4P$ are negative and estimated very imprecisely. So there is not evidence that school or district-level incentives increase test scores. This could be related to the fact that the average maximum bonus for School P4P$ is quite low, $247, though values up to $2,500 are observed in the data. Evaluation P4P$ in the full sample has a negative and statistically significant impact on achievement. Parameter estimates are almost identical in the interested sample and slightly smaller in the adopters only sample, but standard errors are higher so the results are not statistically significant at conventional levels. The results suggest that district adding high stakes subjective evaluations did worse on student achievement.

Because plan design is endogenously chosen by the districts rather than randomly assigned, these design feature "effects" really measure the combination of selection and the effect of the plan feature. If this were purely selection, however, to get oppositely signed effects on Teacher P4P$ and Evaluation P4P$ would require that districts anticipating abnormally *positive* (compared to non-adopters) time-varying growth trends be more likely to adopt incentives tied to teacher-based targets and those anticipating abnormally *negative* trends be more likely to adopt incentives tied to evaluation-based targets. This seems unlikely.

We also estimated the effect of P4P on achievement growth rather than on achievement

levels. The results are robust to this alternative specification. These models include lagged measures of achievement as predictors and, in this specification, district fixed effects pick up differences in stable growth trends across districts rather than differences in levels. Parameter estimates in Table 8 continue to indicate a sizable, significant impact of teacher P4P$ on reading scores with magnitudes slightly larger than in the levels models (Table 7). The estimated impact of teacher P4P$ is now 0.162 (0.072) in the interested only sample, specification (A). The estimated negative impact of evaluation P4P$ also has a somewhat higher magnitude than the school-level models but is estimated less precisely absent within-district variation.

The estimated impact of these same incentives on math scores is less clear. Estimates presented in Table 9 indicate no effect of any type of incentive on achievement levels.[20] However, the district level growth models presented in Table 10 indicate patterns more similar to those in reading. Specifically, estimates imply a positive impact of Teacher P4P$ of 0.115 standard deviations per $1000, marginally significant at the 10% level for the full sample but not in the applicants only or adopters only samples. Estimates also imply a negative Evaluation P4P$ impact in math scores of similar magnitude to the impact on reading scores (-0.049). These are significant at the 10% level for the full sample and the applicants-only sample, but not for adopters only.

The estimated negative impact of evaluation P4P$ is consistent with an environment where evaluators can observe a richer measure of teacher quality than is reflected on test scores. If so, rewarding this broad measure may focus teacher effort away from tested material resulting in lower test scores, but still benefit students and society through improved learning on other valuable dimensions.

In a state with a long standing and widely used open enrollment policy and many charter schools, net pupil movements between districts as well as total enrollment could reflect parent satisfaction with schools.[21] If parents observe a richer measure of the overall educational experience an examination of the impact that P4P plans had on pupil movements can shed

[20] The number of observations is slightly different for math because school-grade observations are not released by the state when there are fewer than ten students tested and this varies across subject.

[21] We analyze net pupil movements between and out of regular districts as a percentage of their enrollment. Alternative outcomes.

light on whether substantial multitasking issues are at play.

Table 11 shows the effect of different P4P plan dimensions on several teacher attributes and student outcomes.[22] Teacher P4P$ appears to have no effect on enrollment, net student flow, or student attendance rates. School level incentives, on the other hand, may influence net flows positively, despite having no impact on test scores. Finally, the only statistically significant effect of high stakes subjective evaluations appears to be on enrollments and it is negative. These results do not suggest an environment where subjective evaluations promote unobserved quality valued by parents. Of course, this may also reflect the fact that our alternative educational outcome measures are imperfect. Parent demand may respond to changes in quality only slowly.

In Table 11 we also examine the impact P4P design features had on teacher average salaries, experience and education levels in order to examine whether P4P plans induced teacher sorting. There is evidence of an impact of both Teacher and Evaluation P4P$ on salaries, but not of School P4P$. The estimated parameters are consistent with average salaries and average bonuses paid out by Q-Comp and, thus, provide good evidence that Q-Comp is associated with real changes to compensation. We also find that the percentage of teachers with masters degrees increases slightly following Q-Comp adoption, which could indicate sorting of more educated teachers into Q-Comp districts or incentives within Q-Comp contracts to obtain graduate degrees (even though education is not supposed to be explicitly rewarded by Q-Comp funds). In any case, teacher sorting with respect to teacher education or experience is not an important channel through which these P4P contracts generated changes in student test scores. When the percentage of teachers with a masters and average teacher experience are entered into the student achievement models as control variables, they are generally not significant and do not mediate the relationship between Q-Comp and student achievement.

We also investigate possible multitasking by using data on what subject School P4P$ are tied to. Table 12 presents additional models using an indicator of whether the subject in

[22] Thid analysis uses data back to 2003. Summary statistics are in the bottom panels of Table 6.

question is the only subject that School P4P$ are tied to: 1(only high stakes goal)$_{tsgb}$. For each subject, we estimate the main effect of the three P4P$ dimensions as well as interacting them with the only-goal indicator. None of the P4P$ dimensions are significantly more effective when applied in a high-stakes year-school-grade-subject. While imprecise, the magnitude of the School P4P$ interactions are positive and large, consistent with what one would expect from theory. So this is weak evidence against multitasking as a big concern. Further, these results suggest that any differences in the effect of P4P$ effects between math and reading are not primarily due to differences in the incidence of goals set across subjects.

Finally we take a look at whether the different P4P designs examined in this paper have disproportionate impacts by grade. Table 13 indicates that the positive impact of P4P incentives at the teacher level is highly concentrated in higher grades (seven and eight), while the negative impact of high stakes peer evaluations is spread across most grades. In math there is more fluctuation, with high stakes evaluations having a statistically significant negative impact only on grade seven and P4P incentives at the teacher level having a positive and significant impact only in grade eight. Against pattern, the estimated impact of teacher-level incentives is negative and somewhat significant in grade 5.

5.3 The Overall Effect of Q-Comp Participation

Recent national efforts to reform teacher compensation are following a similar general approach as Q-Comp in that they set guidelines and accept proposals from districts. How did Minnesota's program fare overall with this flexible approach? What was the average effect of the program after six years?

Table 14 presents estimated effects of program participation on reading (math) achievement pooled across grades 3 to 8 in the upper (lower) panel. Across all samples and in both subjects, we see evidence of a null effect. In math, specification (B) reveals evidence that participating districts may have been already improving in the years prior to adoption. The omitted category here is the year immediately prior to adoption. Therefore, the -0.072 (0.040) estimated coefficient on 1(Pre-adoption) implies that adopting districts were doing worse be-

tween four and two years prior to adoption than they were in the year immediately prior to adoption. However, once they adopted, the progress did not continue.[23]

6 Conclusion

Incentives tied to criteria defined at the teacher- or small team-level had a large, robust, positive impact on reading achievement. We found some evidence of Teacher P4P$ on math but this results was more fragile. It is worth recalling that the Teacher P4P$ were, sometimes but not usually, tied to measure of student achievement analyzed here. Often they were tied to completion of professional development plans or student achievement on teacher-developed assessments. Getting such large effects from incentives of this size is somewhat surprising. Springer et al (2010) report experimental results with much larger incentives and null effects on math achievement. This difference could be due to the difference in subject, in the prospective time-horizon for the programs, in the attainability of the targets, or the research designs. In any case, this result warrants further scrutiny.

We find no evidence that rewards tied to larger groups (i.e. school- or district-level) led to achievement gains. This is consistent with evidence North Carolina (Vigdor, 2008) and New York City (Fryer, 2011). Taken with the prior result, this evidence is consistent with the idea that free riding may be an important problem for incentives defined at high levels of aggregation in education.

Subjective evaluations have been proposed as a potentially important component of performance pay for teachers. This is largely based on studies such as Jacob and Lefgren (2008), Rockoff et al (2010) and Tyler et al (2010) which show that evaluations are correlated with value-added measures of teacher effectiveness. These evaluations, however, are conducted in low-stakes environments rather than being attached to bonuses.

We test whether attaching bonuses to these evaluations benefits achievement, while re-

[23] Allowing heterogeneous effects by grade reveals that the estimated impact of participation on reading is positive and significant in grades 3 and 8 while it is positive and significant only for grade 8 in math. Given the disproportionate effect on 8th grade we also tested to see if there is also a disproportionately large program effect in high school. We have reading scores for grade 10 and mathematics scores for grade 11. We find no large program effect on these scores. Results by grade are available from the authors upon request.

maining agnostic on whether evaluators are able or are choosing to distinguish teachers by quality. There is no evidence here that high stakes evaluations result in improvements in student achievement. If anything, we find that reading test scores decrease in districts that attach high stakes to subjective evaluation. The fact that high stakes evaluations may decrease test scores does not necessarily mean that they are ineffective. Subjective evaluations may be doing their job and solving the multitasking problem, discouraging teachers from teaching to the test. However, our results using measures of family demand for education — namely attendance, enrollment and net pupil movements — do not support this hypothesis. If tying dollars to evaluations led teachers to produce more engaging and desirable lessons then we might expect to see increases in these alternative measures of educational quality. We do not.

A pessimistic interpretation of these results would be that high stakes evaluations do not elicit productive effort, perhaps because of the capture issues discussed in Neal (2011), and may even divert effort from productive activities. There may be a *dog-and-pony show* effect, where teachers divert effort towards developing observational experiences evaluators value but that do not benefit measured student achievement or parent-assessed education quality. This interpretation seems overly strong on the basis of this evidence alone, but similar evidence may turn up in other settings.

The experience in Minnesota adds to our understanding of locally-designed P4P plans. The grantor-grantee relationship between education authorities and districts has advantages because it allows use of local information and experimentation in finding appropriate, feasible designs. Our findings suggest that if a granting authority proposes a range of reforms and allows districts to design plans locally, many districts (in cooperation with local teachers' unions) will design plans that base rewards largely on subjective evaluations and this does not seem to benefit student achievement. On the other hand, some districts (in cooperation with their local teachers' unions) will weight rewards to more specific teacher-centered criteria and this appears beneficial for reading achievement.

The fact that, despite large gains in some areas of the program, Minnesota spent $200,000,000

to get a net effect of zero also points out risks associated with too much local control over the plans. Some plans will operate to extract rents from the state more than to improve education. State and federal governments can, however, use the experiences of early adopters, such as Minnesota, to chose more appropriate program guidelines. These tradeoffs have been explored in other contexts through principal-agent models and there is great potential to apply them in this setting.

7 Works Cited

1. Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and Student Achievement in Chicago Public High Schools. *Journal of Labor Economics*, 25(1).

2. Ahn, Tom. 2008. The Missing Link: Estimating the Impact of Incentives on Effort and Effort on Production Using Teacher Accountability Legislation. Duke University Working Paper.

3. Atkinson, Adele, Burgess, Simon, Croxson, Bronwyn, Gregg, Paul, Propper, Carol, Slater, Helen and Deborah Wilson. 2009. Evaluating the Impact of Performance-related Pay for Teachers in England. *Labour Economics*, 16(3): 251–61.

4. Baker, George. 1992. Incentive Contracts and Performance Measurement. *Journal of Political Economy*, 100: 598–614.

5. Baker, George, Robert Gibbons, and Kevin J. Murphy. 1994. Subjective Performance Measures in Optimal Incentive Contracts. *Quarterly Journal of Economics*, 109, 1125–56.

6. Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2010. How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR. National Bureau of Economic Research Working Paper No. 16381.

7. Fryer, Roland. 2011. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. NBER Working Paper 16850.

8. Glazerman, Steven and Allison Siefullah. 2010. An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year Two Impact Report. Mathematica Policy Research, Inc.

9. Hanushek, Eric. 1971. Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data. *The American Economic Review.* 61(2): 280-288.

10. Hanushek, Eric. 2003. The Failure of Input-Based Resource Policies. *The Economic Journal*. 113(485): F64-F98.

11. Hanushek, Eric. 2010. The Economic Value of Higher Teacher Quality. National Bureau of Economic Research Working Paper 16606.

12. Hanushek, Eric. and Steven Rivkin. 2010. Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review*. 100(May 2010):267-271.

13. Hezel Associates, LLC. 2009. Quality Compensation for Teachers Summative Evaluation.

14. Holmstrom, Bengt, and Paul Milgrom. 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, and Organization*, 7: 24-52.

15. Jacob, Brian and Lars Lefgren. 2008. Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, 26(1):101-36.

16. Jacob, Brian. 2010. Do Principals Fire the Worst Teachers? National Bureau of Economic Research Working Paper 15715.

17. Johns, Emily. 2009. Is It "Merit Pay" If Nearly All Teachers Get It? *Minneapolis Star Tribune*.

18. Kandel, Eugene and Edward Lazear. 1992. Peer Pressure and Partnerships. *Journal of Political Economy*, 100(4): 801–17.

19. Lavy, Victor. 2002. Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement. *Journal of Political Economy*, 110(6): 1286–1317.

20. Lazear, Edward P. 2003. Teacher Incentives. *Swedish Economic Policy Review*, 10.

21. Lovenheim, Michael. 2009. The Effect of Teachers' Unions on Education Production: Evidence from Union Election Certifications in Three Midwestern States.*Journal of Labor Economics*, 27(4):525-87.

22. Martins, Pedro. 2009. Individual Teacher Incentives, Student Achievement and Grade Inflation. IZA Discussion Paper No. 4051.

23. Muralidharan, Karthik and Venkatesh Sundararaman. 2011. Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy.* 119(1).

24. Nadler, Carl and Matthew Wiswall. 2011. Risk Aversion an Support for Merit Pay: Theory and Evidence from Minnesota's Q Comp Program. *Education Finance and Policy*, 6(1):75-104.

25. Neal, Derek. 2011. The Design of Performance Pay in Education. NBER Working Paper No. 16710.

26. Nobels, James. 2009. Evaluation Report: Q Comp Quality Compensation. Minnesota Office of the Legislative Auditor.

27. Podgursky, Michael and Matthew Springer. 2007. Teacher Performance Pay: A Review. *Journal of Policy Analysis and Management*, 26(4):909-50.

28. Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2008. Can You Recognize An Effective Teacher When You Recruit One? NBER Working Paper 14485.

29. Springer, M., D. Ballou and Peng. 2008. Impact of the Teacher Advancement Program on Student Test Score Gains: Findings for an Independent Appraisal. National Center on Performance Incentives Working Paper 2008-19.

30. Springer, Matthew, Laura Hamilton, Daniel McCaffrey, Dale Ballou, Vi-Nhuan Le, Matthew Pepper, J.R. Lockwood and Brian Stecher. 2010. Teacher Pay for Perfor-

mance: Experimental Evidence from the Project on Incentives in Teaching. National Center on Performance Incentives.

31. Taylor, Eric and John H. Tyler. 2011. The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-Career Teachers. National Bureau of Economic Research Working paper 16877.

32. Taylor, Lori and Matthew Springer. 2008. Optimal Incentives for Public Sector Workers: The Case of Teacher-Designed Incentive Pay in Texas. National Center on Performance Incentives Working Paper 2009-05.

33. Tyler, John, Eric Taylor, Thomas Kane, and Amy Wooten. 2010. Using Student Performance Data to Identify Effective Classroom Practices. *American Economic Review*, 100(2): 256-60.

34. Vigdor, Jacob L. 2008. Teacher Salary Bonuses in North Carolina. Conference paper, National Center on Performance Incentives.

35. Weitzman, Martin L. and Douglas Kruse. 1990. Profit sharing and productivity. In *Paying for Productivity*. Edited by A. Blinder. Brookings Institution: Washington, D.C.

8 Tables

Table 6: Descriptive statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
Student: school-grade-year					
Total enrollment	194.3	172.5	1	871	2,695,237
Share male	0.513	0.064	0	1	2,695,237
Share free lunch	0.241	0.197	0	1	2,695,237
Share reduced price	0.080	0.052	0	1	2,695,237
Share special educ.	0.134	0.077	0	1	2,695,237
Share Afr.-American	0.086	0.141	0	1	2,695,237
Share Hispanic	0.056	0.084	0	1	2,695,237
Share Asian-American	0.058	0.096	0	1	2,695,237
Share Native American	0.021	0.072	0	1	2,695,237
Teacher: district-year					
Mean salary ($_{td}$)	56,511	8,334	10,279	118,398	356,946
% teachers with masters	19.0	10.126	0	49.8	357,261
Mean years of experience	14.5	2.8	1	27.1	357,261
District					
Inter-district flow rate†	-2.536	9.113	-51.9	225.2	5,749,030
Attendance rate	94.3	5.6	3.1	100	3,239,630

†Measures net flow for regular districts as percentage of total enrollment. Does not average to 0 due to transfers to charters.

Table 7: Reading achievement effect of maximum performance pay available by dimension - pooled across grades 3 to 8 and academic years 2005-06 to 2009-10

Sample	Full		Interested Only		Adopters Only	
Specification	(A)	(B)	(A)	(B)	(A)	(B)
Teacher P4P$	0.117***	0.118***	0.112***	0.114***	0.138***	0.145***
	(0.028)	(0.028)	(0.03)	(0.029)	(0.034)	(0.032)
School P4P$	-.053	-.059	-.050	-.055	-.067	-.081
	(0.095)	(0.088)	(0.104)	(0.095)	(0.1)	(0.092)
Evaluation P4P$	-.049**	-.051**	-.050	-.050	-.033	-.033
	(0.024)	(0.025)	(0.031)	(0.032)	(0.033)	(0.032)
2+ pre-adoption		-.013		-.012		-.033
		(0.046)		(0.045)		(0.044)
1(Dropped Q-Comp)	-.088**	-.089**	-.092**	-.091**	-.080*	-.075
	(0.041)	(0.041)	(0.047)	(0.046)	(0.046)	(0.046)
Enrollment, 1,000s	0.042*	0.042*	0.075***	0.074***	0.045**	0.045**
	(0.023)	(0.023)	(0.012)	(0.012)	(0.019)	(0.019)
Share free lunch	-1.222***	-1.223***	-1.344***	-1.345***	-1.366***	-1.367***
	(0.125)	(0.125)	(0.193)	(0.194)	(0.176)	(0.176)
Share red. price	-.802***	-.802***	-.786***	-.787***	-1.072***	-1.075***
	(0.118)	(0.117)	(0.234)	(0.233)	(0.296)	(0.293)
Share special Ed.	-1.850***	-1.851***	-1.794***	-1.794***	-1.679***	-1.682***
	(0.101)	(0.101)	(0.18)	(0.18)	(0.215)	(0.215)
Share Male	-.483***	-.482***	-.405***	-.405***	-.444***	-.442***
	(0.064)	(0.064)	(0.105)	(0.105)	(0.131)	(0.131)
Share Afr.-American	-1.598***	-1.597***	-1.890***	-1.885***	-1.745***	-1.721***
	(0.284)	(0.284)	(0.181)	(0.179)	(0.252)	(0.247)
Share Hispanic	-1.288***	-1.287***	-1.080***	-1.075***	-1.178***	-1.167***
	(0.179)	(0.178)	(0.266)	(0.264)	(0.325)	(0.322)
Share Asian-American	-.793***	-.790***	-.644***	-.635***	-.541**	-.506**
	(0.266)	(0.268)	(0.212)	(0.215)	(0.249)	(0.252)
Share Native American	-.725***	-.725***	-1.202***	-1.202***	-.700*	-.703*
	(0.255)	(0.255)	(0.381)	(0.38)	(0.408)	(0.411)
School-grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-grade FE	Yes	Yes	Yes	Yes	Yes	Yes
N districts	471	471	134	134	101	101
N school-years	4677	4677	1785	1785	1335	1335
N tested students	1749818	1749818	755801	755801	607067	607067
Adj. R^2	0.887	0.887	0.917	0.917	0.911	0.911

Coefficient (within-district-correlation corrected SE). Significance: $*$: 10% $**$: 5% $***$: 1%.
The single year immediately prior to adoption is always omitted.

Table 8: Program design effects on student achievement *growth* - reading

DV: Reading average achievement for district-grade-year						
Sample	Full		Interested Only		Adopters Only	
Specification	(A)	(B)	(A)	(B)	(A)	(B)
Teacher P4P$	0.152**	0.155**	0.162**	0.164**	0.205***	0.211***
	(0.072)	(0.072)	(0.072)	(0.072)	(0.07)	(0.069)
School P4P$	-.233	-.266	-.279	-.295	-.267	-.301
	(0.228)	(0.229)	(0.228)	(0.229)	(0.221)	(0.224)
Evaluation P4P$	-.071	-.072	-.068	-.069	-.042	-.042
	(0.059)	(0.058)	(0.056)	(0.056)	(0.053)	(0.053)
Lagged reading	0.32***	0.319***	0.322***	0.32***	0.299***	0.295***
	(0.019)	(0.019)	(0.038)	(0.038)	(0.041)	(0.041)
Lagged math	0.153***	0.153***	0.148***	0.149***	0.159***	0.16***
	(0.02)	(0.02)	(0.04)	(0.04)	(0.048)	(0.048)
2+ pre-adoption		-.050		-.027		-.057
		(0.06)		(0.064)		(0.068)
Student observables	Yes	Yes	Yes	Yes	Yes	Yes
District-grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-grade FE	Yes	Yes	Yes	Yes	Yes	Yes
N districts	415	415	128	128	94	94
N students	1,038,784	1,038,784	450,964	450,964	348,156	348,156
Adjusted R^2	0.94	0.94	0.968	0.968	0.958	0.958

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.

Variables are year-district-grade averages.

Lags are prior year, prior grade $(t-1)d(g-1)b$.

Table 9: Math achievement effect of maximum performance pay available by dimension - pooled across grades 3 to 8 and academic years 2005-06 to 2009-10

Sample	Full		Interested Only		Adopters Only	
Specification	(A)	(B)	(A)	(B)	(A)	(B)
Teacher P4P$	-.015	-.008	-.035	-.023	-.030	-.017
	(0.034)	(0.034)	(0.035)	(0.036)	(0.038)	(0.039)
School P4P$	0.047	0.012	0.059	0.025	0.033	0.004
	(0.139)	(0.136)	(0.151)	(0.146)	(0.15)	(0.145)
Evaluation P4P$	-.002	-.009	-.010	-.013	-.011	-.011
	(0.025)	(0.021)	(0.028)	(0.025)	(0.027)	(0.026)
2+ pre-adoption		-.075*		-.072*		-.067
		(0.04)		(0.041)		(0.043)
1(Dropped Q-Comp)	0.02	0.016	0.025	0.028	0.001	0.011
	(0.139)	(0.136)	(0.144)	(0.141)	(0.135)	(0.132)
Enrollment, 1,000s	0.035**	0.035**	0.054***	0.052***	0.025	0.025
	(0.015)	(0.015)	(0.013)	(0.013)	(0.024)	(0.024)
Share free lunch	-1.075***	-1.077***	-1.182***	-1.184***	-1.240***	-1.243***
	(0.14)	(0.14)	(0.227)	(0.228)	(0.253)	(0.253)
Share red. price	-.578***	-.582***	-.745**	-.751**	-1.059***	-1.064***
	(0.135)	(0.136)	(0.291)	(0.292)	(0.358)	(0.359)
Share special Ed.	-1.900***	-1.902***	-1.996***	-2.001***	-1.811***	-1.820***
	(0.121)	(0.12)	(0.213)	(0.212)	(0.232)	(0.231)
Share Male	-.007	-.005	0.125	0.129	0.101	0.105
	(0.078)	(0.078)	(0.145)	(0.144)	(0.173)	(0.173)
Share Afr.-American	-1.650***	-1.641***	-2.065***	-2.036***	-1.929***	-1.882***
	(0.355)	(0.356)	(0.22)	(0.226)	(0.302)	(0.31)
Share Hispanic	-.875***	-.870***	-.683***	-.656***	-.566*	-.541*
	(0.171)	(0.172)	(0.242)	(0.246)	(0.293)	(0.297)
Share Asian-American	0.1	0.121	0.114	0.169	0.093	0.164
	(0.205)	(0.206)	(0.222)	(0.225)	(0.265)	(0.267)
Share Native American	-.671***	-.669***	-1.144***	-1.144***	-.585	-.597
	(0.258)	(0.258)	(0.399)	(0.397)	(0.387)	(0.39)
School-grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-grade FE	Yes	Yes	Yes	Yes	Yes	Yes
N districts	469	469	134	134	101	101
N school-years	4666	4666	1779	1779	1329	1329
N tested students	1698331	1698331	729520	729520	586667	586667
Adj. R^2	0.859	0.859	0.89	0.89	0.884	0.884

Coefficient (within-district-correlation corrected SE). Significance: $*$: 10% $**$: 5% $***$: 1%.
The single year immediately prior to adoption is always omitted.

Table 10: Program design effects on student achievement *growth* - math

DV: Math average achievement for district-grade-year						
Sample	Full		Interested Only		Adopters Only	
Specification	(A)	(B)	(A)	(B)	(A)	(B)
Teacher P4P$	0.112	0.115*	0.091	0.096	0.098	0.105
	(0.069)	(0.069)	(0.072)	(0.071)	(0.073)	(0.072)
School P4P$	-.238	-.277	-.258	-.290	-.261	-.297
	(0.206)	(0.207)	(0.221)	(0.218)	(0.214)	(0.21)
Evaluation P4P$	-.046*	-.047*	-.049*	-.049*	-.043	-.043
	(0.025)	(0.024)	(0.029)	(0.028)	(0.029)	(0.029)
Lagged reading	0.177***	0.176***	0.187***	0.185***	0.172***	0.167***
	(0.02)	(0.02)	(0.039)	(0.04)	(0.046)	(0.046)
Lagged math	0.323***	0.323***	0.347***	0.349***	0.365***	0.366***
	(0.021)	(0.021)	(0.041)	(0.041)	(0.047)	(0.047)
2+ pre-adoption		-.058		-.052		-.060
		(0.047)		(0.049)		(0.049)
Student observables	Yes	Yes	Yes	Yes	Yes	Yes
District-grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-grade FE	Yes	Yes	Yes	Yes	Yes	Yes
N districts	414	414	127	127	93	93
N students	1038147	1038147	450596	450596	347832	347832
Adjusted R^2	0.936	0.936	0.963	0.963	0.957	0.957

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.

Variables are year-district-grade averages.

Lags are prior year, prior grade $(t-1)d(g-1)b$.

Table 11: Program design effects on alternative outcomes using academic years from 2003-2004 to 2009-2010 and all available grades (K-12)

	Teacher			Student		
	Log (mean pay)	Mean yrs. exper.	% M.A.	Log (Enrlmt.)	Inter-dist. net flow %	Attend. %
Teacher P4P$	0.025**	-.266	1.621*	0.022	-.392	-.062
	(0.012)	(0.211)	(0.887)	(0.04)	(0.766)	(0.074)
School P4P$	-.097	0.211	-2.806	0.139	3.301*	-.082
	(0.069)	(1.005)	(3.529)	(0.113)	(1.793)	(0.375)
Evaluation P4P$	0.025***	0.139	0.393	-.059**	0.418	-.058
	(0.01)	(0.151)	(0.383)	(0.024)	(0.33)	(0.089)
2+ pre-adoption	0.00007	0.279	-.141	-.087***	-1.136*	0.05
	(0.007)	(0.224)	(0.441)	(0.027)	(0.587)	(0.152)
Excludes:					Charters	'06,'09
N district-years	498	500	500	558	345	543
Weighted by	FTE	FTE	FTE	Grade	Student	Student
N	356992	357307	357307	3974	5749030	4132170
Adj R^2	0.921	0.887	0.948	0.986	0.907	0.884

Significance: $*$: 10% $**$: 5% $***$: 1%.

Coefficient (within-district SE). Year effects and district effects included. All use district-level variables, except enrollment (district-grade) & attendance rate (school-grade).

Table 12: Effects of design features on achievement allowing for differential effects by whether School P4P$ are tied to student achievement exclusively in a single subject - pooled across grades 3 to 8 and academic years 2005-06 to 2009-10

Dep. Variable:	Reading	Math
Teacher P4P $	0.122***	-.009
	(0.035)	(0.037)
1(Goal for this subject only) · Teacher P4P$	-.022	-.067
	(0.068)	(0.077)
School P4P$	-.127	0.007
	(0.136)	(0.141)
1(Goal for this subject only) · School P4P$	0.231	0.494
	(0.314)	(0.512)
Evaluation P4P$	-.039	-.016
	(0.029)	(0.024)
1(Goal for this subject only) · Evaluation P4P$	-.051	0.06
	(0.033)	(0.089)
2+ pre-adoption	-.009	-.075*
	(0.047)	(0.04)
N districts	471	469
N school-grades	4,677	4,666
N tested students	1,749,818	1,698,331
Adj. R^2	0.887	0.86

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.

Each column is a separate regression of specification B in full sample as in second column of Tables 7 and 9.

Table 13: Design features on achievement with heterogeneous effects by grade-subject

Grade:	3	4	5	6	7	8
			Reading			
Teacher P4P$	0.098	-.032	0.041	0.115	0.224***	0.272***
	(0.061)	(0.042)	(0.032)	(0.095)	(0.064)	(0.058)
School P4P$	0.129	0.245*	0.034	-.097	-.173	-.623***
	(0.192)	(0.144)	(0.131)	(0.251)	(0.2)	(0.209)
Evaluation P4P$	-.081**	-.066**	-.039	-.050	-.104**	0.049
	(0.04)	(0.026)	(0.036)	(0.041)	(0.042)	(0.05)
2+ pre-adoption	0.016	0.028	-.033	0.125	-.156**	-.060
	(0.074)	(0.053)	(0.059)	(0.093)	(0.07)	(0.08)
			Math			
Teacher P4P$	-.060	-.029	-.074*	0.014	-.017	0.117**
	(0.055)	(0.068)	(0.044)	(0.1)	(0.057)	(0.057)
School P4P$	0.241	-.003	-.148	-.166	0.232	-.095
	(0.212)	(0.213)	(0.203)	(0.289)	(0.268)	(0.313)
Evaluation P4P$	0.005	0.017	0.07	-.039	-.090***	-.012
	(0.049)	(0.035)	(0.043)	(0.052)	(0.028)	(0.037)
2+ pre-adoption	-.122**	-.048	-.032	-.011	-.160***	-.080
	(0.056)	(0.055)	(0.095)	(0.071)	(0.061)	(0.091)

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.
In each sample, estimates from a single regression with separate effects by grade from specification B in full sample as in second column of Tables 7 and 9.

Table 14: Participation effects on achievement - pooled across grades 3 to 8 and academic years 2005-06 to 2009-10

Sample	Full		Adopters Only		Interested Only	
Specification	(A)	(B)	(A)	(B)	(A)	(B)
			Reading			
Post-adoption	0.017	0.019	0.014	0.016	0.051	0.051
	(0.031)	(0.032)	(0.036)	(0.037)	(0.044)	(0.044)
2+ yrs. pre-adoption		0.013		0.017		0.002
		(0.054)		(0.054)		(0.055)
N districts	471	471	134	134	101	101
N school-grade-years	4677	4677	1785	1785	1335	1335
N tested students	1749818	1749818	755801	755801	607067	607067
Adj. R^2	0.886	0.886	0.916	0.916	0.909	0.909
			Math			
Post-adoption	0.011	-.003	-.002	-.009	-.008	-.008
	(0.025)	(0.024)	(0.031)	(0.03)	(0.034)	(0.036)
2+ yrs. pre-adoption		-.072*		-.074*		-.070
		(0.04)		(0.041)		(0.043)
N districts	469	469	134	134	101	101
N school-grade-years	4666	4666	1779	1779	1329	1329
N tested students	1698331	1698331	729520	729520	586667	586667
Adj. R^2	0.859	0.859	0.89	0.89	0.883	0.884

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.

Specification as in Table 7, except 1(Post-adoption)$_{tsg}$ substituted for the three P4P$ variables.

The single year immediately prior to adoption is always omitted.