WHEN DOES TEACHER PERFORMANCE PAY RAISE STUDENT ACHIEVEMENT?: EVIDENCE FROM MINNESOTA'S Q-COMP PROGRAM

Aaron J. Sojourner*

Kristine L. West

Elton Mykerezi

July 18, 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 on state-mandated tests. 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-level goals experienced large gains in reading (0.09σ/\$1,000 bonus) while those offering rewards based on school-wide goals or subjective evaluations did not. Gains from P4P design features were not consistently evident in math or for measures of parental demand. Drawing on alternative standardized tests available for some districts suggests that gains are not fully generalizable.

JEL Classifications: J33, I21, J45

*U. Minnesota, 321 19th Ave S 3-300, Minneapolis, MN 55409, asojourn@umn.edu. Thanks to Avner Ben-Ner, John Budd, David Figlio, Caroline Hoxby, Paul Glewwe, Karthik Muralidharan, Michael Lovenheim, Morris Kleiner, Colleen Manchester, Jason Shaw, Chris Taber, and Joel Waldfogel and participants at the NBER Spring 2011 Economics of Education meeting for comments and to Qihui Chen, Paul Kristapovich, Qianyun Xie, and Yingchun Wang for able research assistance. Thanks to Kristie Anderson of the Minnesota Department of Education, the U. Minnesota's Center for Urban and Regional Affairs, and the Kingsbury Center for support for data acquisition. All errors are ours.

1 Introduction

Many school districts are introducing pay for performance (P4P) plans, using teacher compensation criteria beyond just the conventional years of experience and education. Plans differ on many dimensions including whether teachers are rewarded individually or in teams, based on objective targets or subjective evaluations, and the size of incentives. Theory offers ambiguous guidance on the optimal plan and empirical evidence on the relative and absolute merit of different P4P plans is decidedly mixed. While reviews of the literature point to some gains from P4P (Podgursky and Springer, 2007; Neal, 2011), evaluations of two large-scale P4P plans that were implemented as randomized trials found null or even negative effects (Springer et al., 2010; Fryer, 2011).

In 2005, the State of Minnesota implemented the Quality Compensation program (Q-Comp) as the signature education initiative of Governor Tim Pawlenty. Q-Comp is a package of reforms including P4P. The Minnesota Department of Education set general guidelines for acceptable programs and invited districts to propose specific P4P plans that they would implement. If the proposal was approved, the state authorized up to $260 per student per year in additional funding. Q-Comp provides an excellent opportunity to learn about the merits of different kinds of P4P plans for many reasons.

First, districts that participated in Q-Comp 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 P4P plan in terms of dollars at stake based on: (1) individual teacher-level goals, (2) school-wide goals, or (3) subjective evaluations. We exploit this variation to provide evidence on the effect of P4P plan design features on achievement scores and other outcomes. This inquiry speaks to many issues at the heart of personnel economics, including objective versus subjective evaluations and individual versus team based incentives. Dixit (2002) urges empirical work that considers P4P plan heterogeneity and our investigation is certainly in this vein.

Second, the study has several properties that make non-experimental identification credible. Six different cohorts of districts adopted Q-Comp-funded P4P programs over a six year period. We exploit this variation in the timing of adoption with a generalized difference-in-difference approach. Further, not every district's application succeeded, which enables some important checks. Results are stable when analyzing alternative samples: adopters-only, applicants-only, or all schools in the state.

Third, Q-Comp programs are implemented as permanent changes starting in 2005, which has advantages. If teacher P4P works, it will do so through two primary mechanisms: supporting improved effort by incumbent teachers and attracting better potential teachers to the profession (Lazear, 2003). To operate fully, each mechanism requires an expectation that P4P is here to stay. Teachers may be less willing to alter behavior in response to a time-limited experiment or they may need a few years of trial and error to learn what to do to improve outcomes. And, few people will make career choices based on an incentive program that is not expected to last. Studying a policy change in the field increases external validity because teachers, administrators and families have incentives to adjust to the new policies.

Fourth, in Minnesota, various measures of education quality — standardized tests and parent demand — are available. The Minnesota Comprehensive Achievement Test Series Two (MCA-II) is a state-mandated standardized achievement test. Many districts also use the Northwest Evaluation Association's Measures of Academic Progress (NWEA). We use this to assess whether learning gains generalize to multiple assessments. Since test scores are an imperfect measure of education quality learning and parents may have a richer view of education quality, we also study effects of P4P programs on parent demand. Minnesota is an excellent setting for this. It has the nation's longest standing open enrollment legislation (1981) and charter school legislation (1991). State funding follows the child. Families can enroll their students in any available district or charter at only the cost of transportation, which is sometimes subsidized. These options are well known and widely used.

Lastly, Q-Comp's grantor-grantee structure for program design mirrors U.S. Department of Eduction efforts such as Race to the Top and the Teacher Incentive Fund. In all these programs, the funder sets out guidelines and asks local entities to design and propose plans within them. The grantor delegates some design decisions to take advantage of local knowledge about what will work and what is politically feasible. However, this comes with the risk that local grantees do not deliver. Q-Comp can provide evidence on the trade-offs involved with this approach.

This study finds that, on average, Q-Comp did not produce gains on any measure. However, P4P plan design matters and in interesting ways. While the grantor-grantee format did not lead to widespread improvements in student achievement, it does provide a unique opportunity to learn about P4P plans in education.

Q-Comp districts that tie performance bonuses to individual teacher-level or small-group criteria experience increases in average MCA-II reading scores of 0.09 standard deviations per $1,000 of bonus offered. The finding is quite robust within the limits of our study design and suggests a very large effect for a relatively low price.[1] In contrast, linking rewards to school-level criteria does not appear to cause increases in reading scores, nor does linking rewards to a subjective evaluation process. There is weak evidence that higher stakes on subjective evaluations may lead to declines in reading scores. For math, there are no apparent effects of the incentives tied to teacher- or school-level measures. In some specifications there is evidence of a negative effect of tying bonuses to subjective evaluations in math as well.

There is evidence that achievement gains are concentrated on, but not completely limited to, the high-stakes test. While we do not generally observe to which tests districts tie stakes, we can observe which districts purchase NWEA tests — a necessary condition for tying stakes to it. In these districts, we observe both MCA-II and NWEA scores. The impacts

[1]The social value of a 0.2σ achievement gain for a teacher's class each year has been recently estimated conservatively at $200,000 (Hanushek, 2010; Chetty et al., 2010). Our result would imply that a $1,000 bonus yields an average $90,000 in social value, if the value derives from reading only. There are many program elements accompanying the P4P reforms. However, this rate of return would be extremely large unless the other elements have costs that are a couple of orders of magnitude bigger than the direct cost of bonuses.

of teacher-level bonuses on MCA-II reading scores that we discussed above turn out to be mostly due to gains realized in districts that do not purchase NWEA tests. The impact of teacher-level incentives on MCA-II reading scores among districts using the NWEA is still positive, but of smaller magnitude and not significant at conventional levels. Turning to measures of parent demand, districts that put higher stakes on individual teacher-level criteria do not see an increase in demand. There is weak evidence that students move to districts that put higher stakes on school-level measures and leave districts that put higher stakes on subjective evaluations.

The paper proceeds as follows. Section 2 provides more detail on the Q-Comp program. Section 3 briefly reviews relevant theoretical and empirical literature. Section 4 introduces an empirical model and discusses identification. Section 5 presents results including survey evidence that Q-Comp's adoption led to real changes in district policies and programs, evidence on the relative success of different P4P plan design features, robustness checks, and tests for generalizability of the results to alternative outcomes. Lastly we report results on the average effects of Q-Comp participation. Section 6 concludes with a discussion of how our results add to the existing literature on P4P and 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 P4P plans it funds or their effects.[2]

Selection into Q-Comp works as follows. The state defined guidelines regarding the content of Q-Comp reform plans and promised additional annual funding to districts that implement approved plans. Districts (including charters) decide whether to apply and what specific P4P plans to propose. The state decides whether to accept the proposal. Where teachers were unionized, teachers vote on whether to accept the proposal.[3] Districts that clear all these hurdles participate in Q-Comp.

New districts have joined the program each year. Table I 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. A few participating districts dropped out of Q-Comp. These tables reflect stock given exit and entry flow.[4] Most analysis

[2]Neal (2011) summarizes U.S. and international empirical evaluations of P4P and notes that there has been no previous independent study of Q-Comp. A legislative auditor's report (Nobels, 2009) and a state-commissioned external report (Hezel Associates, 2009) provide evidence about Q-Comp's implementation but very little about resulting student achievement. Neither dealt with selection or covariates. Nadler and Wiswall (2011) use data on Q-Comp participation but do not address whether Q-Comp (or P4P design in general) impacts student achievement.

[3]Almost all Minnesota districts are unionized though many charters are not. Anecdotal evidence suggests districts informally negotiated proposals with unions in advance and teachers officially voted to ratify the contract after state approval.

[4]Districts' Q-Comp start date is based on the date of the approval letter sent by the state Department of Education. These dates differ slightly from those in a state's Legislative Auditor's report but results are robust to alternative coding of the start date.

will focus on grades 3 to 8 because in these grades all students took both math and reading MCA-II tests. Participation statistics are provided for schools in this sample in the bottom panel.

2.2 District P4P Design Features

Data on each Q-Comp district's P4P design are collected primarily from letters sent by the Minnesota Department of Education to each district upon approval of its Q-Comp application. In their applications to the state, districts had to specify the maximum bonus 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. can earn through the teacher evaluation/observation process.

The approval letters detail these agreed-upon features of the plan.[5]

We create three variables for each district measuring the maximum performance pay available to teachers for each of the three types of criteria outlined above. We label incentives under categories 1, 2 and 3 as Teacher P4P$, School P4P$ and Evaluation P4P$, respectively. A few clarifications on the details of each dimension are helpful.

Evaluation P4P$ are incentives tied to receiving a positive evaluation based on classroom observation. Depending on the district, the evaluator is the principal or other administrator, a peer, or a hired consultant (sometimes retired teachers). The state encouraged districts to use the Danielson evaluation framework (Danielson and McGreal, 2000) and conduct at least three observation sessions per year. Evaluations should be done by a trained evaluator and

[5]Each letter was coded by 3 independent coders. Our results are robust to different interpretations of vague letters and to dropping districts with vague letters from the analysis.

involve with pre and post observation conferences. Although the evaluations are "subjective," they rely heavily on a rubric and the state stresses the importance of inter-rater reliability, thus they may be very formalized "subjective" evaluations.

School P4P$ are incentives payable to all staff covered under the collective bargaining agreement for reaching a set target. These are primarily defined at the school level, with the exception of few small districts that set a district-wide goal. The goals are almost exclusively based on standardized test scores but vary on the targeted subject (e.g. math or reading) and assessment (e.g. MCA-II or NWEA). Most approval letters also specified the subject and assessment to which districts elected to tie their school-wide performance bonuses. Schools were more likely to tie School P4P$ to reading 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, other, or unspecified subjects.

Teacher P4P$ are incentives based on quantifiable targets defined at the teacher or small team level. The process of setting these targets was associated with rather significant complementary change which the Minnesota Department of Education refers to as "job embedded professional development." Specifically, with the support of their administration, teachers form Professional Learning Communities (PLCs) and meet regularly to analyze classroom practice, learn new instructional strategies and tactics, field-test them in the classroom, and report the results to each other (Hord and Hirsch, 2008; Darling-Hammond et al., 2009). Within a PLC, each teacher or small team must specify a target. They are not necessarily based on standardized test scores (though they can be), but they have to contribute to stated school-wide goal and have to be quantifiable. Data and assessment development teams were created to assist with the target setting and monitoring for each PLC. These are teams of teachers who meet together and analyze results for standardized tests or teacher-created assessments and use the evidence to determine teaching strategies that will improve student achievement.

This process is unique in several aspects. First, goals are set locally rather than externally. This can be effective if PLCs use local information to set goals that are more appropriate for each teacher. Also, teachers may be more inclined to pursue goals if they are actively involved in setting them. Perhaps most importantly, an investment was made in creating an infrastructure to monitor progress towards stated goals and provide support in achieving them. However, a process of setting goals locally would seem to have a higher risk of being captured and turned into defacto salary augmentations (Neal, 2011).

Q-Comp districts vary in the total levels of pay available across the three dimensions as well as the shares available through each dimension. The value of these variables is shared by all a district's school-grades in post-adoption years. Table III summarizes the cross-sectional distribution of these measures across participating districts.[6] Participating teachers can earn an average maximum of $872 a year in incentive pay through locally-set, individual or small team-level goals (Teacher P4P$), an average maximum of $247 for school or district-level goals (School P4P$), and an average maximum of $1,100 by meeting criteria tied to subjective evaluations (Evaluation P4P$). Table IV describes how the three dimensions are correlated. For Teacher P4P$, we observe a 0.12 correlation with School P4P$ and a -0.80 correlation with Evaluation P4P$, and a -0.15 correlation between School P4P$ and Evaluation P4P$. Teacher P4P$ and Evaluation P4P$ have a strong negative correlation. Figure I displays histograms for the marginal distributions of the three variables. The triggers for paying out on these dimensions are set according to various locally-designed, state-approved criteria within and across districts.

Figure II displays the joint distribution of Q-Comp districts across P4P design dimensions. Each point represents a district'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,

[6] In a few cases, only the share assigned to each dimension and no dollar values were listed in the approval letter nor in any available program documents. For these, we assumed the modal total amount among observed districts ($2,000) and applied the observed shares to this. Another five letters were so ambiguous as to be impossible to code. Although districts and schools may change their designs over time we assume they stay constant at the initial levels. A few districts filed change forms with the Minnesota Department of Education. These were small adjustments and they do not change the results reported here.

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 $4,000 in bonus pay annually. Being on the frontier line means that none of the bonus is tied to subjective evaluation. Half the bonus is tied to Teacher P4P$ 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 figure makes some important points about the Q-Comp designs clear. First, there is a lot of variation across districts. They do not cluster around some generally known, optimal contract. Second, almost all districts offer between $1,000 and $4,000 per year in total P4P$. Third, most districts offer a mix across all three dimensions; few lie on the edges of the triangle. Fourth, none offer more than half of their bonus to School P4P$ criteria, although there is a lot variation in shares below half. This variation in P4P plan design underscores our belief that analyzing Q-Comp in the aggregate likely masks important differences across districts. Accounting for this heterogeneity in design provides a unique opportunity to understand how P4P plan specifics impact educational outcomes.

3 Literature Review

Perhaps the most important issue facing P4P in education involves whether gains observed in response to P4P plans are only realized on the rewarded metric or whether these are generalizable to alternative measures of student learning (Koretz, 2002; Neal, 2011). Gains in the rewarded metric that are not generalizable to other measures of student achievement may result from unproductive hidden teacher action in the form of coaching (Jacob, 2005), socially wasteful gaming (Figlio and Winicki, 2005), or even cheating (Jacob and Levitt,

2003). Since no test score can fully capture teachers' effects on critical thinking, non-cognitive skills, and other unobserved yet valuable aspects of learning, high powered incentives tied to test scores may create a multitasking problem (Holmstrom and Milgrom, 1991; Baker, 1992, 2002). Teachers may 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 et al., 1994). Subjective evaluations are especially attractive because they can be used in non-tested subjects and research has shown that principals are able to distinguish effective from ineffective teachers (Jacob and Lefgren, 2008; Rockoff et al., 2011; Tyler et al., 2010). A recent study of a high quality teacher evaluation program in Cincinnati found immediate and medium-term student achievement gains (Taylor and Tyler, 2011). This Cincinnati program attached high-stakes, the possibility of firing, to the same Danielson framework commonly used among for Evaluation P4P$ among Q-Comp participants.

However, if principals are reluctant to use their knowledge of teacher effectiveness when making high-stakes decisions, such programs are subject to capture. Neal (2011) speculates that the failure of P4P programs in England (Atkinson et al., 2004) and Portugal (Martins, 2009) may be due 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 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."[7]

The theoretical literature on P4P more broadly also recognizes a trade-off between offering rewards based on individual versus team outcomes. Given complementarities in produc-

[7]There is evidence that almost all teachers in Q-Comp districts earn at least *some* performance-based pay, often through the subjective evaluation portion. (Johns, 2009) 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. 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.

tion, individual-level incentives can discourage productive cooperation (Alchian and Demsetz, 1972). On the other hand, team incentives 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).

We contribute empirical evidence to a growing body of studies on effectiveness of P4P in U.S. schools.[8] Neal (2011) thoroughly reviews the literature and concludes that P4P incentives seem able to shift performance targets but the gains are not necessarily generalizable to other measures of learning. However, two recent experiments find that P4P failed to move even the performance targets. First, in a randomized P4P trial in New York City, there were no positive impacts of school-wide bonuses on student achievement. The New York program provides rewards to each teacher if the school meets a specified target based on a composite measure that includes test scores, attendance and discipline. In fact, the school-level bonuses may even have decreased student achievement (Fryer, 2011). Second, a randomized P4P 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). Each of these two studies tested a single P4P design in a time-limited experiment and each found no boost in achievement on the performance targets.

Q-Comp's structure is based on the Milken Foundation's TAP model. Previous empirical research on TAP finds a mixed impact on achievement at best. Springer et al. (2008) find possible positive effects for elementary grades as measured by growth on NWEA exams but

[8]There is also a growing literature on P4P outside the U.S. This includes a recent large scale randomized trial in Andhra Pradesh, India. In this setting individual and small team rewards improve student achievement. Specifically, individual and small group rewards both had a positive impact on language and math tests 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 (Muralidharan and Sundararaman, 2011). There is also evidence in support of P4P from tournament structured P4P in Israel (Lavy, 2002, 2009) and school-wide bonuses in Kenya (Glewwe et al., 2010).

that program effects are negative and statistically significant for higher grades. Glazerman and Seifullah (2010) evaluate TAP in Chicago exploiting randomization in the timing of take-up and find no impact on student achievement growth as measured by average scores on the Illinois Standard Achievement Test.

Q-Comp implemented programmatic changes that were intended to complement the P4P. We focus on P4P because changes in it are most reliably measured but discuss our findings within the context of the full reform. Personnel economics has long recognized the importance of factors beyond compensation. Performance pay may be complemented by delegation of responsibility, monitoring, evaluation, and training. For instance, Prendergast (2002) argues that when an agent has local knowledge, such as a teacher's knowledge of the students' in his or her classroom, P4P should be paired with delegated responsibility. In the case of Q-Comp, individual teacher or small team-level bonuses were tied to goals set through a very structured professional development process. This process may be as important as, or even more important than, the pay increase. Marsden (2010) provides a similar discussion about P4P in British schools. He argues that P4P works primarily through an emphasis on goal setting. However, similar to the aforementioned concerns about tying bonuses to subjective evaluations, management literature is replete with warnings about the potential for individual goals to be corrupted and captured (Locke and Latham, 2002; Gerhart and Rynes, 2003).

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 measures are MCA-II average scores in math and reading. Since 2005-06 (coincidentally the first year of Q-Comp), these have been mandated for every student in third to eighth grade in both subjects.[9] As mentioned in the introduction, we also use NWEA tests, interdistrict movements and enrollment data as alternative outcomes.

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 $b \in \{M, R\}$.[10] NWEA outcomes are similarly defined and indexed. Interdistrict movements and enrollment are defined at the district-year level and thus are indexed by t and $d = 1, 2...D$ rather than s, g and b.

In explaining average student achievement, we use variants of this generalized difference-in-difference model:

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

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. Standard errors adjusted for correlation at the district level are provided throughout.

In order to boost power, our primary results pool across grades 3 to 8 and restrict program

[9] Prior to 2005, only grades 3, 5 and 7 were tested and on a different test, the MCA-I.

[10] 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. MCA-II data for the 2010-11 school year is not yet available.

effects to be the same across grades within subject, $\beta_{gb} \equiv \beta_b$.[11] To facilitate pooling, all scores are normalized to mean zero and standard deviation one across schools within grade-year-subject.

Interest centers on the effects of Q-Comp participation and of features of the P4P designs adopted. To allow β to capture the effect of Q-Comp participation on average, we define Q as a simple participation dummy. To measure the effects of various P4P design features, we use different definitions of Q. In most cases, we define Q as a vector measuring Teacher P4P$, School P4P$, and Evaluation P4P$ interacted with the post-adoption indicator.

To measure the effects of Q, we use two alternative comparison time periods. In specification (A), the reference category is all years prior to adoption. Specification (B) adds an indicator for academic years two or more years prior to adoption, 1(2+ pre-adoption). This conditions on and measures pre-adoption differences in achievement levels between adopters and non-adopters. The specification (B) reference category is the single year immediately prior to adoption (Lovenheim, 2009).

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 is negative it indicates districts do worse after leaving Q-Comp than they did in the year(s) prior to adoption.

It is worth noting that since our data start in the year that the first cohort adopted and that different-sized cohorts adopted during each year, there are 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 between 2007 and 2010.

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

[11] Results by grade are available in the appendix Table A2.

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}). These are listed in the top panel of Table II, 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, 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. These terms identify counter-factual year effects for each grade and subject (δ_{tgb}). 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.[12]

The model is identified by assuming that program variables (Q_{tsgb}) 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

[12]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.

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. Also, if administrators were able to forecast future achievement successfully and designed plans that differed based on these forecasts the difference-in-difference estimators above could produce biased results about the P4P plan features.[13]

Third, we estimate the 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,[14] and (3) just Q-Comp adopters who have not yet adopted. We refer to 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 non-adopters passed the first hurdle to participation; they choose to apply. Some even cleared the second hurdle (state approval). In this sense, interested non-adopters 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 III and IV present trends in average reading and math achievement levels among each adoption cohort, the cohort of never-adopters, and among interested non-participants. 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' scores

[13]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.

[14]Failed applications had to be obtained through a Freedom of Information Act to the Minnesota Department of Education.

are just below the never appliers. Among Q-Comp adopters, the 2005 and 2010 cohorts are most similar to the never-adopters and the interested non-adopters on average. 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 trends between cohorts, aside from the fluctuations in the very small 2009 cohort. Third, this foreshadows one of our 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.

Lastly, we estimate growth models that condition on lagged achievement. Because students move across schools, it would not be possible to get lagged achievement data for approximately one third of the sample if analysis were conducted at the school-grade-subject level, therefore growth models are estimated at the district-grade-subject level. At the district level we can obtain lagged scores based on all students, adding $\vec{y}_{(t-1)d(g-1)}$ as a covariate to explain y_{tdgb}. 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 Q-Comp adoption, P4P plan design and achievement growth. The results turn out to be quite similar qualitatively.

5 Results

5.1 Did Q-Comp Change Teacher Pay and Incentives?

We begin by asking whether Q-Comp program adoption actually changed teacher incentives as advertised. Grant recipients often elicit funds for activities they were already performing. In that case, our study design would find null effects. Was this the case with Minnesota school districts? Drawing on supplemental data, we present three pieces of evidence that Q-Comp actually did change the way teachers are paid.

First, adopting Q-Comp is significantly associated with districts starting to reward teachers for excellence, according to an analysis of data from the National Center for Educational Statistics' Schools and Staffing Survey (SASS). 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. Table V reports on the 55 Minnesota districts sampled in both the 2003-04 and 2007-08 waves of the SASS. Among districts not participating in Q-Comp at the time of the second SASS survey, 96% report no pay for excellence both before and after Q-Comp started. Among districts participating in Q-Comp in 2007-08, 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.

Second, in order to get more detail on the particular aspects of the P4P plans implemented in Q-Comp schools, we conducted an independent phone survey of district human resource professionals about their district's pay practices without mention of Q-Comp. It found that participating districts are vastly different from nonparticipants in how they compensate teachers. We obtained data from 92 districts (38% response rate), twenty-one of

whom participate in Q-Comp. Table VI summarizes our findings. 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. Participating districts are just as likely to pay for years of experience and educational credentials as are non-participants. Q-Comp P4P is clearly a supplement to, rather than a replacement of, traditional compensation criteria.

Lastly, introduction of Q-Comp is associated with a 2.5% increase in average teacher salaries when we use district log mean teacher pay as a dependent variable in our analysis. More pay comes in districts offering more Teacher P4P$ and Evaluation P4P$, not School P4P$.[15] 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

Next, we estimate the impact of program design features on standardized test scores. Table VII presents estimates for the effects on MCA-II 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.[16]

Schools which offer more Teacher P4P$ 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

[15] Detail is in the web appendix Table A1.

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

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 0.087 (0.025) to 0.112 (0.026) per $1,000 at stake. The stability of the results suggest that neither districts' application decision criteria nor the state's rejection criteria were correlated with time-varying unobservables.

The parameter estimates on School P4P$ are positive and estimated very imprecisely. We do not see evidence that school or district-level incentives increase reading test scores. This could be related to the fact that the average maximum bonus for School P4P$ is quite low.

Rewards for subjective evaluations (Evaluation P4P$) have a negative and statistically significant impact on reading achievement in the full sample. Parameter estimates are slightly smaller in the interested only and adopters only samples and standard errors are higher so the results are not statistically significant at conventional levels. The results suggest that districts that began attaching larger bonuses to the subjective evaluation process, if anything, did slightly worse on reading than they did prior to Q-Comp.

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-grade fixed effects pick up differences in stable growth trends for each grade across districts rather than differences in levels. Parameter estimates in Table VIII continue to indicate a significant impact of Teacher P4P$ on reading scores. School P4P$ are now large and negative but still imprecise. The estimated impact of Evaluation P4P$ is still negative but no longer significant.

The estimated impact of these same incentives on math scores is less clear. Estimates presented in Table IX indicate no statistically significant effect of Teacher or Evaluation P4P$ on achievement levels. School P4P$, on the other hand, show a large, marginally significant positive effect, but only on the specifications that use all pre-adoption years as a reference and not in those that use the single pre-adoption year and condition on differences

in prior years' achievement.[17] The district *growth* models for math, presented in Table X, indicate patterns more similar to those in reading for the Evaluation P4P$, but not for Teacher P4P$. Specifically, estimates imply a negative Evaluation P4P$ impact on math scores of similar magnitude to the impact on reading scores.

5.3 Causality and Robustness

Next, we probe concerns about causality more deeply. If Q-Comp had been designed as an experiment, ideally each district would be randomly assigned the timing of adoption as well as the dollars at stake in each of the three categories. The program we study departs from this ideal experiment because the timing is not random but rather is driven by the district administration. Further, the plan designs are a function of the administration's preferences and teachers' preferences as represented by the union. Because of these concerns, we investigate sensitivity to various identification threats.

We begin by examining whether the timing of adoption is systematically related observable district characteristics using a hazard model of switching from nonparticipation to participation in Q-Comp next year.[18] We find that charter schools are more likely to adopt Q-Comp than are traditional public school districts. While interesting, this is no cause for concern since fixed effects deal with time-invariant characteristics such as charter status. The hazard model also describes time-varying observables associated with adoption. Year-over-year increases in the share of African American and/or Asian American students, the share of teachers with Masters degrees, and in parent demand all increase the likelihood of adoption. So adopting districts tend to be growing and attracting more students of color. Importantly though, changes in average math and reading scores on the MCA-II do not predict adoption. We do not find evidence that changes in student achievement drove Q-Comp adoption rather than vice-versa.

[17] The negative, significant estimate on "2+ yrs pre-adoption" indicates that the Q-Comp schools were improving math achievement leading up to adoption. Improvement did not continue after adoption.

[18] Full results are available in web appendix Table A5.

Because plan design is endogenously chosen by the districts rather than randomly assigned, the design feature "effects" really capture the combination of selection into P4P design and their effects. In order for the positive Teacher P4P$ effect to be explained by selection, districts able to forecast abnormal upward changes in reading test scores would have to be more likely to apply *and* the administration and union would have to be more likely to agree to load on Teacher P4P$. When gains in test scores are forecast, it seems more likely to load on School P4P$ that are explicitly tied to test scores and evenly distributed to all union members, rather than Teacher P4P$, which are subject to negotiation through PLCs.

To further investigate the time-varying determinants of program design, we predict Teacher P4P$, School P4P$ and Evaluation P4P$ in a seemingly unrelated regression framework (SUR) using one-year changes in observable district characteristics from the year leading into the application year. This tests for observable changes that predict adoption of a particular kind of program design. No patterns emerge.[19] Importantly, changes in math and reading achievement are not correlated with any particular plan design. There is no evidence that trends in MCA-II scores influenced what type of plan districts enacted. Further, if changes in unobservables are similarly uncorrelated with design characteristics, then our identifying assumption is valid.

If the timing of adoption for districts or a particular contract design happens to be correlated with other unobserved-time varying factors that also affect test scores, estimates could be biased. To assess this, we estimate the models dropping one adoption cohort each time (Jackson, 2010). This is useful for at least two reasons. First, identification depends on variation in the timing of adoption and the assumption that timing is not correlated with unobserved achievement trends. Dropping cohorts helps clarify if different cohorts are getting different effects from their designs. Second, because both the Q-Comp program and the outcome data start in 2005, no pre-adoption trends are available for the first two

[19]Results are in web appendix Table A6.

cohorts. Dropping each can reveal whether the results generalize to the later cohorts where pre-trends are available. Table XI reports the results with reading in the top panel and math in the bottom. The results are generally quite stable though when the biggest cohort, 2006, is dropped, the reading results weaken somewhat and become less precise. However, the results are qualitatively very similar. For math, the results are qualitatively stable, although, School P4P$ becomes large and significant when either the 2006 or 2007 cohorts are dropped.

To generate evidence about the lack of a mediating role played by student, teacher and district changes, we estimate the model with alternative sets of conditioning variables (w_{tsg}). Our primary analysis in Tables VII and IX uses student demographics and total enrollment at the school-grade-year level. Table XII shows that the results are robust to alternative conditioning sets. The first column shows the effect of P4P plan designs excluding student demographics and grade enrollment. Only fixed effects and a pre-trend are included. The second column reproduces the results from Tables VII and IX for comparison. The third column adds two teacher variables: average experience and percent with a Masters degree. The fourth column adds three district administrative variables: general reserve fund balance as a percent of previous year expenditures, net pupil inter-district movements, and log(average teacher salaries). All the results are quite stable. If unobservables are correlated with the P4P design variables similarly to these additional sets of observables, then the identifying assumption seems valid (Altonji et al., 2005).

These same exercises performed for the growth models produces similarly stable results. In the web appendix, Table A4 presents the analysis dropping cohorts and Table A3 presents the analysis with alternative conditioning sets.

Next, we introduce both linear and quadratic district-specific time trends as a check on the difference-in-difference specification (Angrist and Pischke, 2008) and supplement the data in various ways to help deal with the demands this creates. This allows each district to follow its own different trend and is more general than using the indicator for two or

more years prior to adoption. Table XIII reports these results, which fluctuate somewhat for reading. The math results are consistently null. The first column reproduces the baseline full sample, specification (B), results from Table VII and Table IX for ease of comparison. The second column introduces district-specific linear trends and the third column adds quadratic trends. With linear trends, the coefficient on Teacher P4P$ for reading falls from 0.087 with standard error (0.025) to 0.031 (0.026). With more general quadratic trends, it returns to 0.081 (0.035). However, with quadratic trends, the effect of Evaluation P4P$ on reading becomes positive and significant. This specification is very demanding given only 5 years of data, especially since one cannot estimate a pre-adoption trend for early adopters given that 2005 was the first year of both the outcome and the program.

To explore this further, we take advantage of two longer panels of test data though neither is complete in other dimensions. First, we bring in two prior years of data from the MCA-I. Because the MCA-I was given only to those in grades 3, 5, and 7, we focus only on these grades.[20] The results, presented in column 4, are very similar to those in the main analysis. For reading, the coefficient on Teacher P4P$ is 0.094 (0.013) and the coefficient on Evaluation P4P$ is -0.075 (0.026). Here, School P4P$ appear to be positively associated with reading achievement. Second, NWEA data are available for some districts as far back as 2002-03. However each district chooses whether to purchase access to the test in any given year and the numbers have grown each year, so this is an unbalanced, self-selected panel. As presented in column 5 of Table XIII, results indicate a positive, significant impact of Teacher P4P$ on NWEA reading scores, even larger than on the MCA-II.[21] The impact of School P4P$ and Evaluation P4P$ are negative but very imprecisely estimated. As with the MCA, none of the three plan dimensions appear to have an impact on NWEA math scores.

[20]We normalize the scores on both tests to mean 0 and standard deviation 1 but concern about whether these two tests are comparable remains.

[21]We standardize NWEA scores the same way we did MCA-II, to have mean 0 and standard deviation 1 across schools and within year-grade-subject. However, the sample is different. To increase comparability with the standardized MCA-II scores we have used thus far, we compute the standard deviation of standardized MCA-II scores in the sample of school-grade-years that have NWEA scores available — 0.84 in reading and 0.85 in math. We then report effects scaled by their reciprocals.

Analogous estimates to those in columns 4 and 5 which also including quadratic trends yield very similar results.

5.4 Generalizability and impact on alternative outcomes

It is not clear how to interpret the large positive effects of Teacher P4P$ on MCA-II and NWEA reading scores in terms of generalizability because, in any district, we do not observe whether Teacher P4P$ are tied to MCA-II, NWEA, or neither. Teacher P4P$ must be linked to measurable targets, but this could include teacher designed assessments rather than standardized tests. These targets are negotiated locally and not reported centrally. If the gains were being generated in districts that do not tie Teacher P4P$ to the outcome under study, then the gains would appear to be the result of generalized learning. However, if the MCA-II gains are being driven by districts tying Teacher P4P$ to MCA-II tests and the NWEA gains are being driven by those tying Teacher P4P$ to NWEA test, then the same pattern would appear to be the result of non-generalized learning.

To further investigate this issue, we exploit the fact that NWEA tests are not available in many district-years. We estimate the impact of P4P plan features on both the MCA-II and NWEA for district-years with data from both tests as well as the impact of P4P plan design on the MCA-II for the subset of districts where the NWEA is not available. Table XIV presents the results. The positive impact of Teacher P4P$ on MCA-II reading tests is disproportionately concentrated among districts that do not use the NWEA. However, the estimated impact of Teacher P4P$ on MCA-II scores in districts with the NWEA test is still positive, although about half the magnitude and estimated imprecisely. This suggests something intermediate between full and null generalizability.

Next, we turn our attention to measures of achievement other than test scores. Table XV shows the effect of different P4P plan dimensions on inter-district movement and log enrollment rates using data back to 2003. Summary statistics are in Table II. Teacher P4P$ appears to have no effect on enrollment or net student flow, suggesting that parents

do not respond to the induced achievement gains either because they do not value them highly or do not know about them. Of course, parent demand may respond to changes in quality only slowly. School P4P$ has a marginally significant positive effect on net pupil movements. Evaluation P4P$ has a negative impact on log enrollment, reinforcing results from the achievement outcomes.

Lastly, we investigate why the effects do not generalize across reading and math and find evidence against one plausible explanation. If a disproportionate number of teachers are rewarded for reading rather than math goals, this might generate the pattern of positive impacts from Teacher P4P$ on reading but not math. We investigate this using data on the subjects to which School P4P$ are tied, which can vary at the school-grade level within Q-Comp districts. Since teachers are encouraged to link their individual goals to the school-wide goals, it seems reasonable that Teacher P4P$ would follow the same subject bias as School P4P$. Table XVI presents additional models using an indicator of whether the subject in question is the only subject that School P4P$ are tied to: $1(\text{only high stakes goal})_{tsgb}$. 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 subject-year-school-grade. These results suggest that any differences in the P4P plan design effects between math and reading are not primarily due to differences in the incidence of goals set across subjects.

5.5 The Overall Effect of Q-Comp Participation

As discussed in the introduction, recent national efforts to spur education reform follow 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 and over $200 million in state funds? Table XVII presents estimated effects of program participation on reading (math) achievement on the MCA-II 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 -0.074 (0.038) estimated coefficient on 1(2+ yrs pre-adoption) implies that adopting districts were doing worse between 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. Other analysis shows that Q-Comp participation did not affect parent demand as measured by net pupil movements or log enrollment.

6 Discussion

Q-Comp P4P incentives tied to criteria defined at the teacher- or small team-level had a large, robust, positive impact on reading achievement. A 0.09 standard deviations increase in reading scores for a maximum bonus of $1,000 seems impressively inexpensive. Several factors unique to Q-Comp could explain why we find such a sizable effect of Teacher P4P$ on test scores. First, as noted, the prospective time-horizon for the program could have played a role; stake holders expected the program to last. Second, the Professional Learning Communities may be teams of teachers large enough to generate benefits from cooperation and small enough to overcome free rider problems. Third, the fact that teachers have a hand in setting their targets along with their peers and principal may increase the appropriateness of the goal and their ability and motivation to attain it. Fourth, there were significant investments in management practices built around the bonuses. Teachers were organized into teams, provided with time to consult, to enter mentoring relationships, to engage in low-stakes classroom observations, and to analyze student performance on assessments (standardized and teacher-created). The effects found here come as a result of the entire process involved in setting and helping teachers reach their targets, not only the $1,000 bonus.

Are these gains in reading strictly a result of hidden teacher action, such as coaching or teaching to the test? This question is closely related to what sorts of goals teachers set, which unfortunately we do not observe. If most teachers set goals that are not directly related to standardized tests then the gains do not result from unproductive hidden action. The Minnesota Department of Education requires that Teacher P4P$ be available to all staff covered by the collective bargaining agreement, so teachers that not responsible for teaching testable grades and/or subjects routinely pick goals that are not based on a standardized test. We have anecdotal evidence that these teachers often pick goals related to other metrics such as attendance, discipline and even AP classes (in higher grades). Teachers in tested grades and/or subjects can pick goals that relate to the MCA-II or NWEA but are not required to do so. In any, case we believe these teacher-level goals are not almost never based on

"value-added" statistics since most Minnesota districts do not currently have the capacity to compute teacher-level value-added. Finally, while the fact that MCA-II gains are found mostly in districts that don't have access to the NWEA and districts with NWEA access produce gains in NWEA reading does suggests some hidden action, we find some suggestive evidence that there are gains in MCA-II reading even in districts with access to the NWEA, although these are smaller in magnitude and imprecisely estimated.

We find no evidence that bonuses tied to larger groups, school- or district-level targets, led to achievement gains. This is consistent with recent evidence from New York City (Fryer, 2011) and with the idea that free riding may be an important problem for incentives defined at high levels of aggregation. We caution that our evidence on this point is statistically imprecise.

Subjective evaluations have been proposed as a potentially important component of P4P for teachers. This is largely based on studies such as Jacob and Lefgren (2008), Rockoff et al. (2011), and Tyler et al. (2010), which show that evaluations are correlated with value added measures of teacher quality. We test whether attaching bonuses to evaluations benefits student achievement, while remaining agnostic on whether evaluators are able or are choosing to distinguish teachers by quality. We find no evidence that bonuses tied to evaluations result in improvements in student achievement. If anything, we find that test scores may decrease in districts that attach bonuses to subjective evaluations.

The fact that high-stakes evaluations may decrease test scores does not necessarily mean that they are undesirable. Subjective evaluations may be solving the multitasking problem, discouraging teachers from teaching to the test. In this case a decline in test scores may be offset by gains in non-tested aspects of learning. However, our results using measures of parental demand for education – namely enrollment and net pupil movements – do not support this interpretation. If tying bonuses to evaluations led teachers to produce more engaging and desirable lessons, then we might see increases in these alternative measures of educational quality. We do not. A pessimistic interpretation of the negative effect of

Evaluation P4P$ is that high stakes evaluations do not elicit productive effort, perhaps because of the capture issues discussed in Neal (2011). There may even 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.

Lastly, the experience in Minnesota adds to our understanding of locally-designed education reform. The grantor-grantee relationship between education authorities and districts has advantages because it allows use of local information and experimentation in finding appropriate, feasible P4P designs. Minnesota's experience suggests that if a granting authority proposes a range of reforms and allows districts to design P4P plans, 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 teacher-level outcomes and this appears beneficial, at least 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 findings from Q-Comp, to chose more appropriate program guidelines. These findings suggest encouraging districts to tie rewards to locally set teacher or small team-level goals supported by PLCs, rather than school or district level goals or subjective evaluations.

Authors' affiliations. Aaron Sojourner: University of Minnesota's Carlson School of Management and IZA. Kristine West: University of Minnesota's Dept. of Applied Economics. Elton Mykerezi: University of Minnesota's Dept. of Applied Economics.

References

A.A. Alchian and H. Demsetz. Production, information costs, and economic organization. *The American Economic Review*, 62(5):777–795, 1972.

J.G. Altonji, T.E. Elder, and C.R. Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113(1):151–184, 2005.

J.D. Angrist and J.S. Pischke. *Mostly harmless econometrics: an empiricist's companion*. Princeton Univ Pr, 2008.

A. Atkinson, S. Burgess, B. Croxson, P. Gregg, C. Propper, H. Slater, and D. Wilson. Evaluating the impact of performance-related pay for teachers in england. *Department of Economics, University of Bristol, UK, The Centre for Market and Public Organisation*, 60, 2004.

G. Baker. Distortion and risk in optimal incentive contracts. *The Journal of Human Resources*, 37(4):728–751, 2002.

G. Baker, R. Gibbons, and K.J. Murphy. Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics*, 109(4):1125–1156, 1994.

G.P. Baker. Incentive contracts and performance measurement. *The Journal of Political Economy*, 100(3):598–614, 1992.

R. Chetty, J.N. Friedman, N. Hilger, E. Saez, D.W. Schanzenbach, and D. Yagan. How does your kindergarten classroom affect your earnings? evidence from project star. Technical report, National Bureau of Economic Research Working Paper No. 16381, 2010.

C. Danielson and T.L. McGreal. *Teacher evaluation to enhance professional practice*. Ascd, 2000.

L. Darling-Hammond, R.C. Wei, A. Andree, N. Richardson, and S. Orphanos. Professional learning in the learning profession, 2009.

A. Dixit. Incentives and organizations in the public sector: An interpretative review. *The Journal of Human Resources*, 37(4):696–727, 2002.

D.N. Figlio and J. Winicki. Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics*, 89(2-3):381–394, 2005.

R.G. Fryer. Teacher incentives and student achievement: Evidence from new york city public schools. Technical report, National Bureau of Economic Research Working Paper No. 16850, 2011.

B.A. Gerhart and S. Rynes. *Compensation: Theory, evidence, and strategic implications*. Sage Publications, Inc, 2003.

S. Glazerman and A. Seifullah. An evaluation of the teacher advancement program (tap) in chicago: Year two impact report. *Mathematica Policy Research, Inc.*, page 62, 2010.

P. Glewwe, N. Ilias, and M. Kremer. Teacher incentives. *American Economic Journal: Applied Economics*, 2(3):205–227, 2010.

E.A. Hanushek. The economic value of higher teacher quality. *Economics of Education Review*, 2010.

Hezel Associates. Quality compensation for teachers summative evaluation. Technical report, Syracuse, NY, 2009.

B. Holmstrom and P. Milgrom. Multitask principal–agent analyses: incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and organization*, 7(special issue):24, 1991.

S.M. Hord and S.A. Hirsch. Making the promise a reality. *Sustaining professional learning communities*, page 23, 2008.

C.K. Jackson. A little now for a lot later: A look at a texas advanced placement incentive program. *Journal of Human Resources*, 45(3):591, 2010.

B.A. Jacob. Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of Public Economics*, 89(5-6):761–796, 2005.

B.A. Jacob and L. Lefgren. Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1):101, 2008.

B.A. Jacob and S.D. Levitt. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, 118(3):843, 2003.

Emily Johns. Is it "merit pay" if nearly all teachers get it? *Minneapolis Star-Tribune*, Feb. 2009.

E. Kandel and E.P. Lazear. Peer pressure and partnerships. *The Journal of Political Economy*, 100(4):801–817, 1992.

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

V. Lavy. Evaluating the effect of teachers performance incentives on pupils achievements. *Journal of Political Economy*, 110(6):1286–1317, 2002.

V. Lavy. Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review*, 99(5):1979–2011, 2009.

E.P. Lazear. Teacher incentives. *Swedish Economic Policy Review*, 10(2):179–214, 2003.

E.A. Locke and G.P. Latham. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9):705, 2002.

M. Lovenheim. The effect of teachers' unions on education production: Evidence from union election certifications in three midwestern states. *Journal of Labor Economics*, 2009.

D. Marsden. The paradox of performance-related pay systems. *Paradoxes of Modernization*, 1(9):185–203, 2010.

P.S. Martins. *Individual teacher incentives, student achievement and grade inflation*. IZA Discussion Paper No. 4051, 2009.

K. Muralidharan and V. Sundararaman. Teacher performance pay: Experimental evidence from india. *The Journal of Political Economy*, 119(1):39–77, 2011.

C. Nadler and M. Wiswall. Risk aversion and support for merit pay: Theory and evidence from minnesota's q comp program. *Education Finance and Policy*, pages 1–31, 2011.

D. Neal. The design of performance pay in education. Technical report, National Bureau of Economic Research Working Paper No. 16710, 2011.

James Nobels. Evaluation report: Q comp quality compensation. Technical report, Minnesota Office of the Legislative Auditor, 2009.

M.J. Podgursky and M.G. Springer. Teacher performance pay: A review. *Journal of Policy Analysis and Management*, 26(4):909–950, 2007.

C. Prendergast. The tenuous trade-off between risk and incentives. *Journal of Political Economy*, 110(5), 2002.

J.E. Rockoff, B.A. Jacob, T.J. Kane, and D.O. Staiger. Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, 6(1):43–74, 2011.

M.G. Springer, D. Ballou, A. Peng, and National Center for Performance Incentives (US). *Impact of the Teacher Advancement Program on student test score gains: Findings from an independent appraisal*. National Center on Performance Incentives, Vanderbilt University, Peabody College, 2008.

M.G. Springer, L. Hamilton, D.F. McCaffrey, D. Ballou, V.N. Le, M. Pepper, JR Lockwood, and B.M. Stecher. Teacher pay for performance: Experimental evidence from the project on incentives in teaching. 2010.

E.S. Taylor and J.H. Tyler. The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers. Technical report, National Bureau of Economic Research Working Paper No. 16877, 2011.

J.H. Tyler, E.S. Taylor, T.J. Kane, and A.L. Wooten. Using student performance data to identify effective classroom practices. *American Economic Review*, 100(2):256–60, 2010.

7 Tables

Table I: 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

Table II: Descriptive statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
Student: school-grade-year, weighted by enrollment					
Total enrollment	167.8	139.3	1	826	1,826,036
Share male	0.513	0.064	0	1	1,826,036
Share free lunch	0.256	0.202	0	1	1,826,036
Share reduced price	0.083	0.052	0	1	1,826,036
Share special educ.	0.139	0.08	0	1	1,826,036
Share Afr.-American	0.092	0.146	0	1	1,826,036
Share Hispanic	0.063	0.092	0	1	1,826,036
Share Asian-American	0.061	0.099	0	1	1,826,036
Share Native American	0.021	0.075	0	1	1,826,036
Teacher: school-year					
% teachers with masters	11.0	13.1	0	92.2	3,373
Mean years of experience	13.3	4.4	0	34.0	3,372
District: district-year					
Inter-district flow	-0.36	498.6	-11,037	2,599	3,244
General Reserve Fund/Expend.	12.4	10.7	-54.7	174.0	3,199
Log(Average teacher salary)	10.8	0.19	9.2	11.7	3,120

Student and teacher characteristics restricted to grades 3-8 and years 2005-2009.
District characteristics not restricted by grade and include data from 2002-2009.

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

	Mean	Std. Dev.	Min.	Max.
Teacher P4P$	0.872	0.692	0	2.5
School P4P$	0.247	0.214	0	2.5
Evaluation P4P$	1.1	0.694	0	2.5
Number of participating districts		77		

Note: weighted by numbers of tested students. The 2010-11 cohort included additional districts but their plans are not coded.

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

	Teacher P4P$	School P4P$	Evaluation P4P$
Teacher P4P$	1.00		
School P4P$	0.12	1.00	
Evaluation P4P$	-0.80	-0.15	1.00

Table V: 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	Districts	In 2003-04	No	Yes	No	Yes	
in 2007-08	in both waves	In 2007-08	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.

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

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

Table VII: Program design effects on student achievement *levels* - reading

Sample	Full		Interested Only		Adopters Only	
Specification	(A)	(B)	(A)	(B)	(A)	(B)
Teacher P4P$	0.087***	0.087***	0.096***	0.097***	0.108***	0.112***
	(0.025)	(0.025)	(0.03)	(0.03)	(0.027)	(0.026)
School P4P$	0.036	0.034	0.037	0.033	0.032	0.024
	(0.08)	(0.08)	(0.075)	(0.074)	(0.077)	(0.076)
Evaluation P4P$	-.051**	-.051**	-.044	-.045	-.035	-.034
	(0.021)	(0.023)	(0.029)	(0.029)	(0.031)	(0.031)
2+ pre-adoption		-.007		-.013		-.024
		(0.046)		(0.047)		(0.042)
1(Dropped Q-Comp)	-.030	-.032	-.022	-.022	0.0005	0.005
	(0.068)	(0.069)	(0.09)	(0.091)	(0.094)	(0.096)
Enrollment, 1,000s	-.174	-.175	-.242	-.249	-.270	-.288
	(0.222)	(0.224)	(0.345)	(0.353)	(0.355)	(0.361)
Share free lunch	-1.211***	-1.211***	-1.304***	-1.305***	-1.366***	-1.367***
	(0.118)	(0.118)	(0.156)	(0.156)	(0.17)	(0.169)
Share red. price	-.763***	-.763***	-.693**	-.695**	-1.052***	-1.055***
	(0.132)	(0.131)	(0.285)	(0.283)	(0.325)	(0.323)
Share special Ed.	-1.855***	-1.855***	-1.837***	-1.837***	-1.698***	-1.700***
	(0.099)	(0.099)	(0.17)	(0.17)	(0.207)	(0.207)
Share Male	-.484***	-.483***	-.391***	-.390***	-.439***	-.437***
	(0.064)	(0.064)	(0.106)	(0.106)	(0.132)	(0.132)
Share Afr.-American	-1.589***	-1.588***	-1.815***	-1.809***	-1.690***	-1.671***
	(0.279)	(0.279)	(0.166)	(0.167)	(0.245)	(0.242)
Share Hispanic	-1.311***	-1.311***	-1.129***	-1.124***	-1.191***	-1.181***
	(0.188)	(0.188)	(0.29)	(0.285)	(0.315)	(0.312)
Share Asian-American	-.723**	-.721**	-.460*	-.451*	-.456*	-.432
	(0.291)	(0.293)	(0.255)	(0.266)	(0.259)	(0.266)
Share Native American	-.738***	-.738***	-1.267***	-1.267***	-.738*	-.741*
	(0.261)	(0.261)	(0.38)	(0.38)	(0.396)	(0.399)
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	0.886	0.886	0.916	0.916	0.91	0.91

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

Table VIII: 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$	0.074**	0.073**	0.083**	0.081**	0.081**	0.076**
	(0.036)	(0.035)	(0.038)	(0.037)	(0.04)	(0.036)
School P4P$	-.128	-.121	-.163	-.153	-.126	-.115
	(0.106)	(0.103)	(0.108)	(0.104)	(0.107)	(0.101)
Evaluation P4P$	-.030	-.028	-.029	-.027	-.028	-.027
	(0.033)	(0.037)	(0.039)	(0.041)	(0.038)	(0.039)
Lagged reading	0.309***	0.309***	0.281***	0.282***	0.291***	0.292***
	(0.02)	(0.02)	(0.037)	(0.038)	(0.035)	(0.036)
Lagged math	0.143***	0.143***	0.161***	0.16***	0.149***	0.148***
	(0.016)	(0.016)	(0.031)	(0.03)	(0.035)	(0.034)
2+ pre-adoption		0.02		0.031		0.033
		(0.051)		(0.046)		(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	446	446	132	132	98	98
N school-years	1989	1989	584	584	442	442
N students	1339042	1339042	578414	578414	446951	446951
Adjusted R^2	0.914	0.914	0.947	0.947	0.932	0.932

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

Variables are year-district-grade averages. Lags are prior year, prior grade $(t-1)d(g-1)b$. Data are pooled across grades 3 to 8 and academic years 2005-06 to 2009-10. An indicator, 1(Dropped Q-Comp), as well as district-grade demographic controls are included as in Table VII but not reported.

Table IX: Program design effects on student achievement *levels* - math

Sample	Full		Interested Only		Adopters Only	
Specification	(A)	(B)	(A)	(B)	(A)	(B)
Teacher P4P$	-.031	-.028	-.040	-.033	-.049	-.039
	(0.031)	(0.031)	(0.032)	(0.032)	(0.031)	(0.032)
School P4P$	0.182*	0.157	0.195*	0.169	0.187*	0.166
	(0.107)	(0.11)	(0.107)	(0.11)	(0.108)	(0.109)
Evaluation P4P$	-.005	-.011	-.006	-.009	-.015	-.015
	(0.022)	(0.02)	(0.026)	(0.024)	(0.025)	(0.024)
2+ pre-adoption		-.065*		-.070*		-.061
		(0.039)		(0.038)		(0.041)
1(Dropped Q-Comp)	0.046	0.033	0.052	0.052	0.016	0.027
	(0.111)	(0.111)	(0.125)	(0.122)	(0.115)	(0.113)
Enrollment, 1,000s	-.952**	-.964**	-.958*	-1.000*	-1.060*	-1.106*
	(0.386)	(0.385)	(0.573)	(0.571)	(0.607)	(0.605)
Share free lunch	-1.077***	-1.079***	-1.166***	-1.168***	-1.262***	-1.263***
	(0.135)	(0.135)	(0.207)	(0.208)	(0.249)	(0.249)
Share red. price	-.547***	-.549***	-.675**	-.683**	-1.039***	-1.044***
	(0.139)	(0.139)	(0.317)	(0.316)	(0.365)	(0.367)
Share special Ed.	-1.907***	-1.909***	-2.041***	-2.046***	-1.840***	-1.848***
	(0.122)	(0.122)	(0.222)	(0.221)	(0.235)	(0.233)
Share Male	-.008	-.007	0.146	0.15	0.118	0.122
	(0.078)	(0.078)	(0.146)	(0.145)	(0.175)	(0.174)
Share Afr.-American	-1.653***	-1.643***	-2.051***	-2.021***	-1.940***	-1.892***
	(0.346)	(0.347)	(0.206)	(0.212)	(0.299)	(0.308)
Share Hispanic	-.892***	-.887***	-.715***	-.686***	-.556*	-.530*
	(0.17)	(0.17)	(0.248)	(0.251)	(0.285)	(0.288)
Share Asian-American	0.16	0.179	0.265	0.313	0.16	0.224
	(0.226)	(0.228)	(0.303)	(0.304)	(0.285)	(0.284)
Share Native American	-.691***	-.690***	-1.198***	-1.198***	-.613	-.625
	(0.263)	(0.263)	(0.399)	(0.397)	(0.383)	(0.386)
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	0.86	0.86	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 X: 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$	0.019	0.02	0.004	0.007	-.006	-.002
	(0.039)	(0.04)	(0.038)	(0.039)	(0.039)	(0.04)
School P4P$	0.042	0.028	0.059	0.046	0.066	0.057
	(0.125)	(0.129)	(0.122)	(0.126)	(0.123)	(0.126)
Evaluation P4P$	-.032*	-.036**	-.041**	-.044**	-.049**	-.049**
	(0.016)	(0.016)	(0.02)	(0.019)	(0.02)	(0.02)
Lagged reading	0.164***	0.164***	0.167***	0.166***	0.163***	0.162***
	(0.017)	(0.017)	(0.035)	(0.035)	(0.042)	(0.041)
Lagged math	0.344***	0.345***	0.36***	0.361***	0.368***	0.369***
	(0.016)	(0.016)	(0.03)	(0.03)	(0.035)	(0.035)
2+ pre-adoption		-.041		-.040		-.026
		(0.032)		(0.034)		(0.034)
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	445	445	132	132	98	98
N school-years	1985	1985	584	584	442	442
N students	1295202	1295202	556746	556746	433988	433988
Adjusted R^2	0.9	0.9	0.936	0.936	0.924	0.924

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

Variables are year-district-grade averages. Lags are prior year, prior grade $(t-1)d(g-1)b$. Data are pooled across grades 3 to 8 and academic years 2005-06 to 2009-10. An indicator, 1(Dropped Q-Comp), as well as district-grade demographic controls are included as in Table IX but not reported.

Table XI: Robustness to dropping any Q-Comp adoption cohort

	Adoption cohort excluded from analysis:					
	2005	2006	2007	2008	2009	2010
			Reading			
Teacher P4P$	0.108***	0.054	0.08***	0.088***	0.087***	0.087***
	(0.03)	(0.046)	(0.024)	(0.026)	(0.025)	(0.025)
School P4P$	-.056	0.085	0.12	-.005	0.035	0.039
	(0.09)	(0.083)	(0.078)	(0.088)	(0.08)	(0.08)
Evaluation P4P$	-.053**	-.038	-.074***	-.041*	-.052**	-.049**
	(0.023)	(0.031)	(0.024)	(0.023)	(0.023)	(0.024)
2+ pre-adoption	-.010	0.0003	-.042	0.002	-.006	0.007
	(0.047)	(0.056)	(0.046)	(0.052)	(0.047)	(0.054)
N districts	464	432	462	461	465	443
N district grades	4509	4041	4475	4591	4637	4474
N tested students	1680075	1428053	1638260	1700066	1743176	1702211
Adj. R^2	0.886	0.881	0.883	0.884	0.886	0.887
			Math			
Teacher P4P$	0.002	-.080	-.037	-.026	-.029	-.028
	(0.032)	(0.063)	(0.03)	(0.032)	(0.031)	(0.031)
School P4P$	0.048	0.217*	0.259*	0.137	0.16	0.16
	(0.099)	(0.131)	(0.139)	(0.12)	(0.111)	(0.111)
Evaluation P4P$	-.007	-.010	-.029	-.003	-.011	-.010
	(0.02)	(0.021)	(0.025)	(0.021)	(0.02)	(0.02)
2+ pre-adoption	-.071*	-.067*	-.068	-.064	-.065	-.061
	(0.04)	(0.038)	(0.05)	(0.046)	(0.041)	(0.044)
N districts	462	430	460	459	463	441
N district grades	4498	4034	4466	4580	4626	4463
N tested students	1631582	1386350	1591116	1650008	1691809	1652454
Adj. R^2	0.86	0.853	0.856	0.857	0.858	0.86

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.
Reading (math) analogous to column 2 of Table VII (IX), except for exclusion of adoption cohorts. Data are pooled across grades 3 to 8 and academic years 2005-06 to 2009-10.

Table XII: Robustness to alternative conditioning sets

	(1)	(2)	(3)	(4)
		Reading		
Teacher P4P$	0.087***	0.087***	0.087***	0.092***
	(0.03)	(0.025)	(0.024)	(0.027)
School P4P$	0.03	0.034	0.035	0.001
	(0.074)	(0.08)	(0.078)	(0.104)
Evaluation P4P$	-.067***	-.051**	-.051**	-.059**
	(0.023)	(0.023)	(0.023)	(0.028)
2+ pre-adoption	-.015	-.007	-.006	0.003
	(0.048)	(0.046)	(0.046)	(0.055)
N districts	471	471	471	436
N district grades	4677	4677	4670	4439
N tested students	1749818	1749818	1749080	1384099
Adj. R^2	0.873	0.886	0.886	0.893
		Math		
Teacher P4P$	-.025	-.028	-.028	-.013
	(0.038)	(0.031)	(0.031)	(0.032)
School P4P$	0.131	0.157	0.159	0.122
	(0.106)	(0.11)	(0.109)	(0.119)
Evaluation P4P$	-.017	-.011	-.011	-.002
	(0.02)	(0.02)	(0.02)	(0.02)
2+ pre-adoption	-.063	-.065*	-.062	-.056
	(0.041)	(0.039)	(0.04)	(0.046)
N districts	469	469	469	434
N district grades	4666	4666	4659	4420
N tested students	1698331	1698331	1697597	1347064
Adj. R^2	0.848	0.86	0.859	0.873
Student observables	No	Yes	Yes	Yes
Teacher observables	No	No	Yes	Yes
District observable	No	No	No	Yes
District-grade FE	Yes	Yes	Yes	Yes
Year-grade FE	Yes	Yes	Yes	Yes

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.
Reading (math) analogous to column 2 of Table VII (IX), except for changes in covariate sets. Data are pooled across grades 3 to 8 and academic years 2005-06 to 2009-10.

Table XIII: Robustness to the inclusion of district-specific time trends

Specification	1	2	3	4	5
			Reading		
Teacher P4P$	0.087***	0.031	0.081**	0.094***	0.320***
	(0.025)	(0.026)	(0.035)	(0.013)	(0.140)
School P4P$	0.034	0.114	-.148	0.184**	-0.290
	(0.08)	(0.09)	(0.145)	(0.089)	(0.339)
Evaluation P4P$	-.051**	-.043	0.086**	-.075***	-0.135
	(0.023)	(0.04)	(0.038)	(0.026)	(0.105)
N districts	471	471	471	463	343
N district-grades	4677	4677	4677	2530	3237
N tested students	1749818	1749818	1749818	1230784	689109
Adj. R^2	0.886	0.896	0.902	0.881	0.715
			Math		
Teacher P4P$	-.028	0.01	0.039	-.005	.047
	(0.031)	(0.027)	(0.052)	(0.025)	(0.109)
School P4P$	0.157	-.053	0.005	0.068	0.122
	(0.11)	(0.116)	(0.153)	(0.106)	(0.289)
Evaluation P4P$	-.011	0.022	0.031	0.044	0.062
	(0.02)	(0.02)	(0.049)	(0.035)	(0.078)
N districts	469	469	469	461	344
N district-grades	4666	4666	4666	2528	3228
N tested students	1698331	1698331	1698331	1204226	690676
Adj. R^2	0.86	0.873	0.88	0.852	0.746
Includes:					
1(2+ pre-adoption)	Y	N	N	N	N
District-specific trend	N	Linear	Quadratic	Linear	Linear
Sample includes:					
Test	MCA-II	MCA-II	MCA-II	MCA/MCA-II	NWEA
Years	'05-'09	'05-'09	'05-'09	'03-'09	'02-'09
Grades	3-8	3-8	3-8	3, 5, & 7	3-8

Coefficient (within-district-correlation corrected SE). Significance: $*$: 10% $**$: 5% $***$: 1%. Student observables, year-grade and school-grade fixed effects always included. Specification 1 reproduces Table VII (IX) Full-B result for reading (math). While 2005-2009 outcomes are normalized MCA-II scores, the 2003 and 2004 outcomes are normalized MCA.

Table XIV: Program design effects by test and alternate test availability

	District-years with NWEA scores		District-years without NWEA scores	
Test:	MCA	NWEA	MCA	NWEA
			Reading	
Teacher P4P$	0.053	0.266	0.099***	Scores
	(0.058)	(0.163)	(0.028)	do
School P4P$	0.077	-.251	-.079	not
	(0.14)	(0.296)	(0.11)	exist
Evaluation P4P$	-.033	-.021	-.041	
	(0.029)	(0.123)	(0.035)	
2+ pre-adoption	0.007	0.020	-.034	
	(0.071)	(0.072)	(0.054)	
N districts	334	334	447	
N school-grades	2,990	2,990	3,487	
N student-years	951,452	497,265	798,366	
Adj. R^2	0.876	0.698	0.915	
			Math	
Teacher P4P$	-.047	-0.070	-.008	
	(0.057)	(0.160)	(0.028)	
School P4P$	0.297**	0.331	-.028	
	(0.133)	(0.300)	(0.121)	
Evaluation P4P$	-.023	0.085**	0.023	
	(0.028)	(0.041)	(0.025)	
2+ pre-adoption	-.051	0.061	-.048	
	(0.041)	(0.064)	(0.042)	
N districts	333	333	444	
N school-grades	2,968	2,968	3,481	
N student-years	928,817	496,742	769,514	
Adj. R^2	0.857	0.74	0.893	

Table XV: Program design effects on alternative outcomes

	Log (Enrollment)	Inter-district net flow
Teacher P4P$	-.0009	-30.62
	(0.032)	(45.83)
School P4P$	0.149	133.70*
	(0.094)	(69.62)
Evaluation P4P$	-.060***	-15.59
	(0.021)	(21.49)
2+ pre-adoption	-.096***	-51.99***
	(0.028)	(14.83)
N districts	558	516
N district-years	3974	3244
Adj R^2	0.986	0.934

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.
Data is district-year level. Year effects and district effects included.
Includes academic years from 2003-2004 to 2009-2010 and all grades (K-12)
Student characteristics included as in Table VII and Table IX but not reported.

Table XVI: Allowing for differential effects by whether School P4P$ are tied to student achievement exclusively in a single subject

Dep. Variable:	Reading	Math
Teacher P4P $	0.077***	-.029
	(0.029)	(0.033)
1(Goal for this subject only) · Teacher P4P$	0.026	-.046
	(0.063)	(0.073)
School P4P$	0.01	0.158
	(0.08)	(0.112)
1(Goal for this subject only) · School P4P$	0.101	0.326
	(0.248)	(0.506)
Evaluation P4P$	-.041	-.018
	(0.025)	(0.022)
1(Goal for this subject only) · Evaluation P4P$	-.047	0.074
	(0.031)	(0.084)
2+ pre-adoption	-.003	-.066*
	(0.047)	(0.04)
N districts	471	469
N school-grades	4,677	4,666
N tested students	1749818	1698331
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 VII and IX.

Table XVII: Participation effects on achievement levels

Sample	Full		Interested Only		Adopters Only	
Specification	(A)	(B)	(A)	(B)	(A)	(B)
			Reading			
Post-adoption	0.016	0.001	0.014	0.006	-.004	-.004
	(0.024)	(0.024)	(0.033)	(0.032)	(0.033)	(0.035)
2+ yrs. pre-adoption		0.013		0.011		0.004
		(0.054)		(0.057)		(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	0.915	0.915	0.909	0.909
			Math			
Post-adoption	0.016	0.001	0.014	0.006	-.004	-.004
	(0.024)	(0.024)	(0.033)	(0.032)	(0.033)	(0.035)
2+ yrs. pre-adoption		-.074**		-.081**		-.073*
		(0.038)		(0.037)		(0.04)
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 VII and Table IX, except 1(Post-adoption)$_{tsg}$ substituted for the three P4P$ variables. The single year immediately prior to adoption is always omitted.

8 Figures

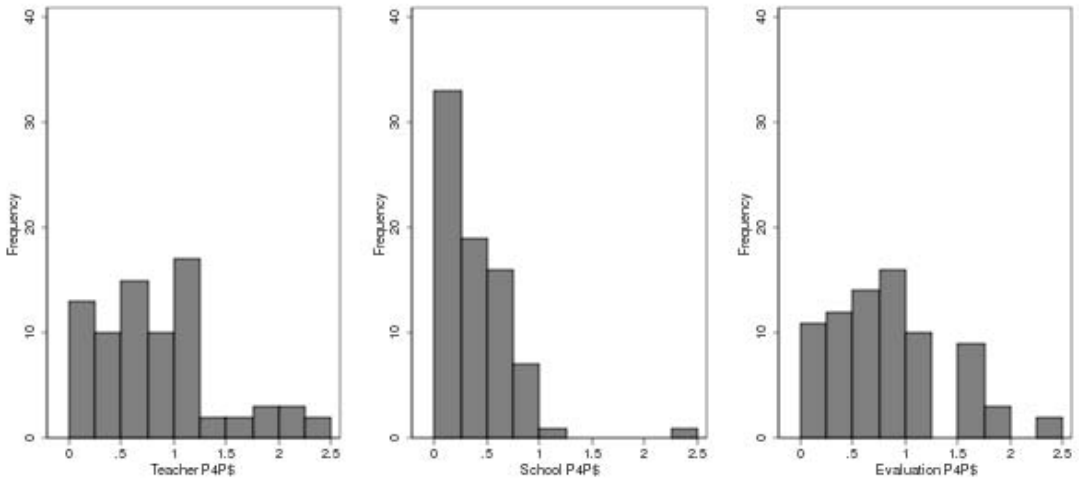

Figure I: Marginal frequencies of P4P design variables across Q-Comp districts, in $1,000

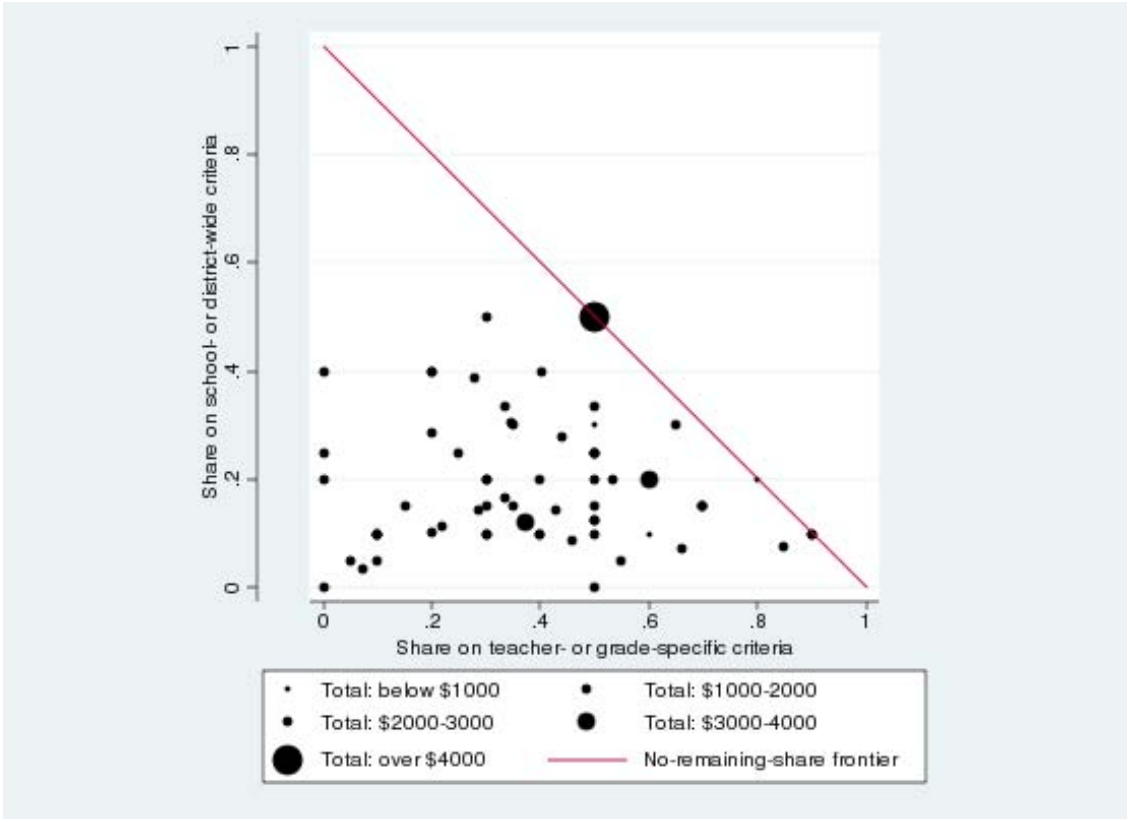

Figure II: Joint distribution of P4P designs across Q-Comp districts

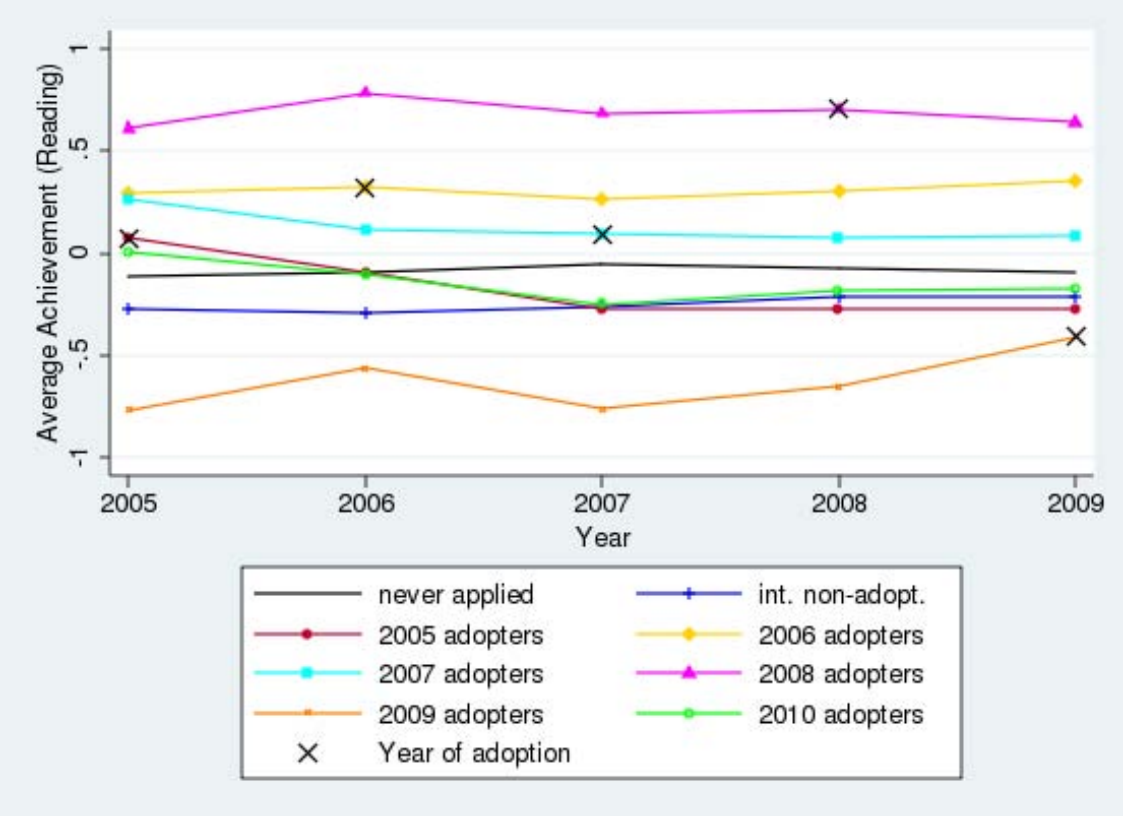

Figure III: Trend in average reading achievement by Q-Comp adoption cohort

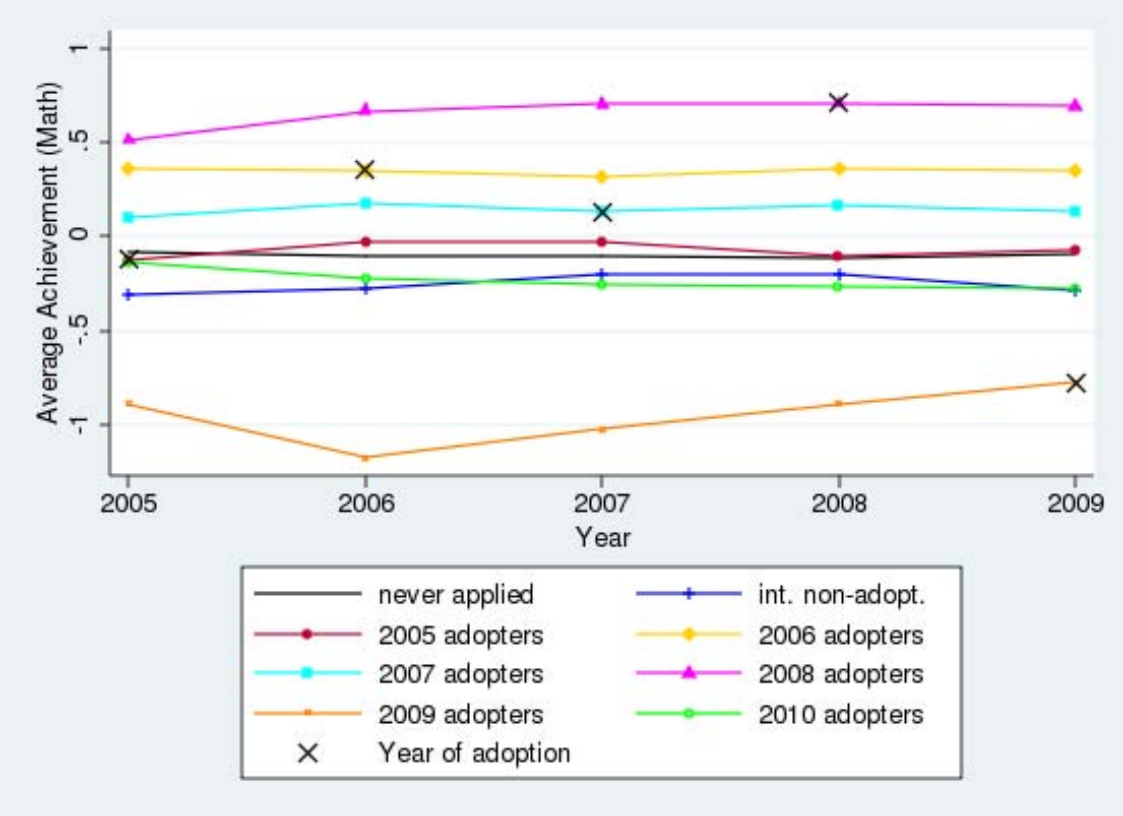

Figure IV: Trend in average math achievement by Q-Comp adoption cohort

A Web Appendix

Table 1: Program design effects on average teacher salary

	Log (Average teacher salary)
Teacher P4P$	0.029***
	(0.010)
School P4P$	-.036
	(0.039)
Evaluation P4P$	0.027**
	(0.012)
2+ pre-adoption	0.019*
	(0.011)
N districts	498
N district-years	3120
Adj R^2	0.804

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

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

Table 2: Heterogeneous effects by grade-subject

Grade:	3	4	5	6	7	8
			Reading			
Teacher P4P$	0.093**	-.020	0.038*	0.061	0.165***	0.186***
	(0.047)	(0.03)	(0.023)	(0.072)	(0.039)	(0.06)
School P4P$	0.108	0.217**	0.015	-.083	0.113	-.153
	(0.138)	(0.087)	(0.102)	(0.226)	(0.11)	(0.205)
Evaluation P4P$	-.067*	-.069***	-.049	-.050	-.113***	0.032
	(0.039)	(0.025)	(0.039)	(0.04)	(0.04)	(0.044)
2+ pre-adoption	0.02	0.026	-.042	0.112	-.133*	-.038
	(0.075)	(0.051)	(0.06)	(0.1)	(0.068)	(0.078)
			Math			
Teacher P4P$	-.035	-.045	-.103***	-.024	-.023	0.066
	(0.037)	(0.049)	(0.034)	(0.073)	(0.048)	(0.053)
School P4P$	0.226*	0.2	0.052	-.020	0.288	0.211
	(0.12)	(0.149)	(0.139)	(0.193)	(0.202)	(0.224)
Evaluation P4P$	-.067*	-.069***	-.049	-.050	-.113***	0.032
	(0.039)	(0.025)	(0.039)	(0.04)	(0.04)	(0.044)
2+ pre-adoption	-.123**	-.037	-.034	-.006	-.146**	-.053
	(0.055)	(0.053)	(0.095)	(0.067)	(0.063)	(0.093)

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 VII and IX.

Table 3: Robustness of growth model to alternative conditioning sets

	(1)	(2)	(3)	(4)
		Reading		
Teacher P4P$	0.073**	0.073**	0.073**	0.098**
	(0.036)	(0.035)	(0.035)	(0.042)
School P4P$	-.092	-.121	-.114	-.243*
	(0.093)	(0.103)	(0.102)	(0.132)
Evaluation P4P$	-.032	-.028	-.030	-.033
	(0.032)	(0.037)	(0.037)	(0.037)
2+ pre-adoption	0.013	0.02	0.021	0.03
	(0.054)	(0.051)	(0.051)	(0.07)
N districts	446	446	446	415
N district grades	1989	1989	1987	1890
N tested students	1339042	1339042	1338696	1038698
Adj. R^2	0.91	0.914	0.914	0.92
		Math		
Teacher P4P$	0.021	0.02	0.02	0.07**
	(0.04)	(0.04)	(0.04)	(0.031)
School P4P$	0.024	0.028	0.027	-.113
	(0.125)	(0.129)	(0.13)	(0.115)
Evaluation P4P$	-.037**	-.036**	-.037**	-.037**
	(0.016)	(0.016)	(0.016)	(0.016)
2+ pre-adoption	-.043	-.041	-.042	-.035
	(0.03)	(0.032)	(0.031)	(0.032)
N districts	445	445	445	415
N district grades	1985	1985	1983	1886
N tested students	1295202	1295202	1294863	1005407
Adj. R^2	0.899	0.9	0.9	0.911
Student observables	No	Yes	Yes	Yes
Teacher observables	No	No	Yes	Yes
District observable	No	No	No	Yes
District-grade FE	Yes	Yes	Yes	Yes
Year-grade FE	Yes	Yes	Yes	Yes

Coefficient (within-district SE). Significance: *: 10% **: 5% ***: 1%.
Reading (math) analogous to column 2 of Table VIII (X), except for changes in covariate sets.

Table 4: Robustness of growth model to dropping any adoption cohort

	Adoption cohort excluded from analysis:					
	2005	2006	2007	2008	2009	2010
			Reading			
Teacher P4P$	0.121***	0.03	0.068**	0.071*	0.074**	0.072**
	(0.043)	(0.049)	(0.033)	(0.036)	(0.035)	(0.035)
School P4P$	-.302**	-.096	-.008	-.125	-.118	-.111
	(0.124)	(0.074)	(0.108)	(0.115)	(0.106)	(0.102)
Evaluation P4P$	-.025	0.019	-.071**	-.024	-.028	-.026
	(0.037)	(0.033)	(0.032)	(0.04)	(0.038)	(0.039)
2+ pre-adoption	0.011	0.04	-.0009	0.003	0.023	0.04
	(0.053)	(0.059)	(0.039)	(0.063)	(0.052)	(0.066)
N districts	439	407	438	436	440	419
N district grades	1954	1808	1951	1942	1962	1877
N tested students	1292480	1094541	1257031	1301639	1335331	1306279
Adj. R^2	0.914	0.907	0.911	0.91	0.914	0.914
			Math			
Teacher P4P$	0.085**	-.055	0.01	0.02	0.019	0.019
	(0.035)	(0.06)	(0.039)	(0.04)	(0.04)	(0.04)
School P4P$	-.194	0.134	0.057	0.072	0.041	0.032
	(0.122)	(0.112)	(0.159)	(0.138)	(0.13)	(0.129)
Evaluation P4P$	-.030*	-.016	-.045*	-.044**	-.036**	-.036**
	(0.016)	(0.019)	(0.023)	(0.017)	(0.016)	(0.017)
2+ pre-adoption	-.054*	-.032	-.031	-.063*	-.039	-.037
	(0.032)	(0.036)	(0.043)	(0.036)	(0.032)	(0.034)
N districts	438	406	437	435	439	418
N district grades	1950	1804	1947	1938	1958	1873
N tested students	1249991	1058062	1215480	1258601	1291574	1263516
Adj. R^2	0.901	0.89	0.897	0.897	0.9	0.901

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

Reading (math) analogous to column 2 of Table VIII (X), except for exclusion of adoption cohorts.

Table 5: Proportional hazard model for Q-Comp adoption

DV: 1(start Q-Comp next year)		
Predictors	Coefficient	(SE)
Time-varying		
Average math score	1.061	(0.048)
Average reading score	0.968	(0.043)
Total enrollment	1.014**	(0.006)
Percent free lunch	0.756	(0.158)
Percent red. price lunch	1.811	(0.849)
Percent special education	0.581	(0.348)
Percent male	1.443	(0.502)
Percent Afr.-American	1.313	(0.240)
Percent Hispanic	0.988	(0.322)
Percent Asian-American	1.399*	(0.255)
Percent Native American	1.402	(0.395)
Teachers average years experience	0.998	(0.009)
Percent of teachers with Masters	1.011***	(0.003)
Log(Mean teacher salary)	0.911	(0.134)
Net interdistrict flow, thousands	1.309***	(0.115)
Reserve Fund/Expenditure	0.997	(0.002)
Time-invariant		
1(Charter)	3.108**	(1.644)

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

Table 6: Test of relationship between program design and changes in district observables leading into the application year

DVs:	Teacher P4P$	School P4P$	Evaluation P4P$
Δ average math achievement	-0.136	-0.111	-0.036
	(0.183)	(0.079)	(0.187)
Δ average reading achievement	0.107	0.015	-0.058
	(0.158)	(0.068)	(0.162)
Δ student enrollment (1,000)	0.977	0.580**	-0.809
	(0.642)	(0.278)	(0.658)
Δ percent free lunch	2.387	-0.598	2.339
	(1.597)	(0.691)	(1.639)
Δ percent reduced price lunch	1.069	-0.374	1.564
	(2.754)	(1.192)	(2.826)
Δ percent special education	3.250	2.370*	0.975
	(3.209)	(1.389)	(3.293)
Δ percent male	-6.583**	1.109	-1.589
	(3.254)	(1.408)	(3.339)
Δ percent Afr.-American	-4.051	-0.021	1.212
	(3.962)	(1.715)	(4.066)
Δ percent Hispanic	-7.933	0.001	-1.373
	(5.565)	(2.409)	(5.711)
Δ percent Asian-American	-5.704	1.451	13.428**
	(6.399)	(2.770)	(6.567)
Δ percent Native American	-3.862	8.902	-36.192***
	(12.997)	(5.626)	(13.337)
Δ General Fund/Expenditures	0.020	0.008	-0.006
	(0.016)	(0.007)	(0.017)
Δ Log(Mean teacher salary)	0.034	0.450	-0.023
	(0.727)	(0.315)	(0.746)
Δ Net interdistrict flow	-0.004***	0.000	0.000
	(0.001)	(0.001)	(0.001)
Constant	0.874***	0.322***	0.865***
	(0.079)	(0.034)	(0.081)
N districts	69	69	69
P-value of joint null test	.16	.21	.23

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

Estimated by SUR. For district-d adopting Q-Comp in year t, $\Delta x_d \equiv x_{d,t-1} - x_{d,t-2}$.