Teacher Pay Reform and Productivity: Panel Data Evidence from Adoptions of Q-Comp in Minnesota

Aaron J. Sojourner* (asojourn@umn.edu)

Elton Mykerezi* (myker001@umn.edu)

Kristine L. West** (klwest@stkate.edu)

September 26, 2013

Abstract

This paper studies the impacts of teacher pay-for-performance (P4P) reforms adopted with complementary human-resource management (HRM) practices on student achievement and workforce flows. Since 2005, dozens of Minnesota school districts in cooperation with teachers' unions implemented P4P as part of the state's Quality Compensation program. Exploiting district variation in participation status and timing, we find evidence that P4P-centered HRM reform raises students' achievement by 0.03 standard deviations. Falsification tests suggest that gains are causal. They appear to be driven especially by productivity increases among less-experienced teachers. JEL: M5, I21, J45

Authors' Affilations: * University of Minnesota and ** St. Catherine University. Thanks to Avner Ben-Ner, Jen Brown, John Budd, Brian Cadena, 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 George Henley and Kristie Anderson of the Minnesota Department of Education and the Kingsbury Center for help with restricted-data acquisition and to the Human Capital Research Collaborative and the U. Minnesota Center for Urban and Regional Affairs for funding support. All errors are ours.

1 Introduction

The potential to improve U.S. education through human-resource management (HRM) reforms centered on pay-for-performance (P4P) remains an open, active question for economists and policymakers. After decades of paying teachers almost solely on their education and experience, hundreds of U.S. school districts have begun measuring teacher performance in various ways and incorporating these measures into pay determination. State governments including Colorado, Florida, Minnesota, South Carolina, and Texas now encourage or require local school districts to shift their teacher management systems to include performance evaluation and pay. The federal government also encourages districts to shift towards performance-based HRM systems through its $1.6 billion Teacher Incentive Fund and $4.4 billion Race to the Top program. These reform efforts are driven by theories that P4P-centered HRM reform will raise teacher productivity and student learning.

Despite strong evidence of P4P effects in production environments where output is relatively easy to measure (Lazear 2000; Bandiera et al. 2007), evidence of P4P and broader HRM reforms' impact in more complex environments is limited. According to Bloom and Van Reenen (2011), in organizations in general, "There is certainly a robust positive cross sectional association between bundles of 'modern' HRM practices and productivity, but with some exceptions (e.g. Ichniowski et al., 1997) these are not robust in the time series dimension." In education, evidence on the effects of P4P in U.S. schools is mixed (Podgursky and Springer 2007; Prentice et al. 2007; Neal 2011). P4P field experiments generated null or even negative effects (Springer et al. 2010; Fryer 2011).[1] Aside from P4P, evidence is emerging that another kind of HRM reform, performance measurement and feedback, can generate positive effects through improvement of incumbent teachers' practice (Taylor and Tyler 2012) and increasing low-performers' separation rates (Rockoff et al. 2012).

This paper reports new evidence on how teacher HRM reforms centering on P4P con-

[1] Strong evidence of positive P4P effects has emerged in other countries (Muralidharan and Sundararaman 2011; Lavy 2002, 2009; Glewwe et al. 2010).

tracts affect student achievement and workforce flows using panel data from Minnesota. A shock to local school districts' HRM practices was induced by the state's introduction of its Quality Compensation program (Q-Comp) in 2005. The state began offering local districts additional funding in return for districts' use P4P contracts along with complementary HRM practices. Different districts applied for and joined Q-Comp each year. Dozens of districts have since implemented reform and hundreds of thousands of student-years have been taught in participating districts. As one of the nation's largest and longest-standing programs encouraging P4P-centered HRM reform, Q-Comp attracts significant policy and political attention, yet little is known about the effects of the reforms it spurred. This is the first study to use the Q-Comp program to study effects of reform on student achievement.[2]

This study estimates the effects of P4P-centered HRM reform on achievement for students in grades 3 to 8 and the mechanisms by which the effects operate by integrating data from multiple sources. We use student-level panels from two different standardized achievement tests (one state-mandated and the other optional but widely used), the population of teachers linked to their district each year, Q-Comp program data coded from archives of official documents, the U.S. Schools and Staffing Survey, as well as data on district characteristics and finance from the Minnesota Department of Education (M.D.E.). This enables the study to make five main contributions.

First, this statewide field study based on one of the nation's largest teacher pay programs is uniquely informative about longer-run, general-equilibrium effects. Previous efforts in this direction have relied on cross-national variation (Hanushek and Woessmann 2011). Theory suggests that P4P will work by some mixture of (1) increasing incumbent employees' productivity and (2) attracting more able employees to the organization (Lazear 2000). In order to operate fully, both mechanisms require belief that the reform is here-to-stay. Incum-

[2]A legislative auditor's report (Nobels 2009) and state-commissioned external reports (Hezel Associates 2009; Wahlstrom et al. 2006) provide evidence about Q-Comp's implementation but little about resulting student achievement. Neither dealt with selection or covariates. Nadler and Wiswall (2011) study Q-Comp participation but not impacts. Schwartz (2012) provides qualitative description based on interviews with numerous stakeholders.

bent teachers may require a few years of trial and error to increase their own productivity in response to P4P incentives. They will make these investments only if they believe the reform will endure. Regarding sorting, people will not alter their choice of employer based on a pay system they expect to disappear. Only a permanent policy change gives teachers, administrators, and families incentives and time to adjust to reform.

Second, despite the non-experimental context, the study has several properties that make identification credible. Student achievement is measured in individual panels covering the academic years starting 2003 to 2009 and different districts applied for and adopted Q-Comp in different years. A generalized difference-in-difference framework identifies the effect of reform on districts' productivity net of time effects and fixed differences between individual students. Further, we conduct a series of falsification tests to rule out various alternative explanations. One placebo test uses multiple years of pre-adoption data to estimate Q-Comp "effects" prior to adoption, which tests for the presence of differential trends in unobservables among adopting districts. A second placebo test estimates failed-application "effects" to assesses whether districts' desire to reform is sufficient to generate effects, absent actual reform. Q-Comp's teacher-pay reforms produced an average three percent of a standard deviation increase in reading achievement, with some evidence of a similar effect on math achievement in the full sample. Evidence from the appliers-only sample is weaker.

Third, we assess whether achievement effects reflect real skill gains or simply teaching-to-the-test by assessing Q-Comp's impact using two distinct standardized tests. Gaming is a central concern in the P4P literature generally (Holmstrom and Milgrom 1991) and in the educational P4P literature particularly. According to Neal (2011), "clean evidence on the generalizability of assessment gains is rare, and the existing literature does not speak with one voice." The Minnesota Comprehensive Achievement test (MCA) is state-mandated and administered in all districts. Some districts also voluntarily contract with the Northwest Evaluation Association (NWEA) to use its Measures of Academic Progress achievement

test.[3] We use all existing data on Minnesota students in grades 3 to 8 from the MCA and NWEA in both reading and math. In each subject, we assess estimates' robustness across the tests. Further, we check whether students make larger gains on the higher-stakes test than on the other test. We do this in a unique way. Commonly, higher-stakes are attached to one test for all students and a second test is lower-stakes for everyone, meaning that stakes are confounded with test. In this study, many students also take two tests but, in contrast, which test is higher stakes with respect to school-wide performance bonuses differ by district and sometimes even by school-grade within district. Test and stakes are not confounded.

Fourth, we use data on the universe of Minnesota teachers tied to their employing district each year to examine if changes to districts' teacher HRM policies affect teacher movements, and the extent to which this accounts for achievement gains. While we cannot link particular teachers to particular students, we develop measures of district-year teacher workforce flows through various channels (novice teachers entering, inter-district transfers in and out, retention, retirements out) and school-year teacher experience profiles (share with ≤ 5, 6-15, and ≥ 16 years experience). We study how these relate to changes in student achievement. There is no evidence of significant effects of HRM reform on teacher workforce flows or experience profiles and observed changes in flows and profiles do not explain coincident changes in student achievement. Similarly, we test and reject the possibilities that changes in districts' student populations or total expenditures drive achievement gains. In contrast, we find evidence that the effect of reform interacts significantly with school-level measures of teacher experience; the effect is strongest among schools with more inexperienced teachers.

Finally, Q-Comp provides an opportunity to examine the effects of a "grantor-grantee" structure for program design that mirrors recent U.S. Department of Eduction efforts such as Race to the Top and the Teacher Incentive Fund. In these programs, the funder sets out guidelines and asks local entities to propose tailored plans within them. The grantor delegates some design decisions in order to harness local knowledge about what will work

[3]Though the panels cover many of the same students in the same years, we cannot link a student across tests nor to a particular teacher.

and what is politically feasible. However, this comes with risk that local grantees do not deliver improvements. Q-Comp is an opportunity to examine if decentralized design with centralized approval generated gains or produced mostly rent-seeking behavior.

These features of the study enable contributions to the education, personnel, and labor economics literatures. Having long-term outcomes coupled with a real policy shift permits study of how impacts unfold over a multi-year horizon, which is unusual and valuable. Further, having P4P-centered HRM reform sweeping across a large fraction of an industry while observing output and worker flows for all firms in the industry gives an unprecedented view into the mechanisms of change. This paper reports some of the strongest evidence available that teacher HRM reform can increase teacher productivity and student learning in the U.S. Though the magnitudes of the achievement impacts are not large, a rough estimate suggests the social benefits of the gains exceed the costs of producing them.

2 Background and Process

Description of Q-Comp program. Q-Comp was designed to encourage Minnesota school districts to adopt specific forms of P4P. It also aimed to ensure that districts had complementary HRM practices in place. Minnesota's leading newspaper described Q-Comp at launch this way, "[The enabling legislation] puts $86 million into a teacher merit pay plan called 'Q-comp,' which allows teachers to get raises based on merit, additional duties and student achievement, not only for years on the job and college credits" (deFiebre et al. 2005). More recently, former Gov. Tim Pawlenty described it, "Q Comp is the nation's most comprehensive performance pay for teachers program" (Victory New Hampshire 2012). The Minnesota Department Education (M.D.E.) defines the range of district HRM practices — types of P4P contracts, classroom observation protocols, individual professional development plans, etc. — acceptable under Q-Comp and invites districts to propose a specific set of HRM practices within this range. If the proposal is approved and adopted, the district becomes eligible for

up to $260 in additional annual funding per pupil added to the district's general budget.

M.D.E. encouraged districts to design their P4P so that all teachers could conceivably collect the full bonus amount, rather than using optimal tournaments (Barlevy and Neal 2012), and to use multiple performance measures including a mix of objective and subjective measurements and individual- and team-based goals. In their applications, districts specify the P4P bonuses each teacher is eligible to earn for the following three types of criteria: (a) the formal classroom-observation process; (b) school-wide or district-wide goals for student achievement usually on standardized tests; and (c) quantifiable goals negotiated within the school (between administration and teachers) for student achievement defined at the teacher, team, or grade level but not usually based on standardized tests. Across Q-Comp districts, the average bonus available to teachers tied to each of these criteria was $1,107, $243, and $850 per year, respectively.[4] Classroom observations had the most stakes tied to them. School-wide goals based on standardized achievement tests had the least.

P4P goals were set and measured in the context of a complementary management practice, which the M.D.E. refers to as "job embedded professional development" and which shares many features with high-performance work practices (Ichniowski et al. 1997; Cappelli and Neumark 2001). Specifically, with the support of their administration, teachers form Professional Learning Communities (PLCs). Here, they help select performance targets which form the basis for Q-Comp's individual or small-group, P4P bonuses and help each other achieve those targets. They meet regularly to analyze classroom practice, to learn new instructional strategies and tactics, to field-test them in the classroom and to report the results to each other (Hord and Hirsch 2008; Wei et al. 2009). For classroom observations, the state encouraged districts to use the Danielson evaluation framework (Danielson and McGreal 2000), the most widely used evaluation rubric nationally, and to conduct at least three observations per year using a trained evaluator and with pre- and post-observation conferences. Teachers are rated on measures of planning and preparation, classroom envi-

[4]These averages are computed by weighting each district's bonuses by its number of students tested in MCA reading after reform.

ronment, professional responsibility, and instructional practice. Depending on the district, the evaluator is the principal or other administrator, a peer, or a hired consultant.

Our research design is premised on district's participation in Q-Comp triggering change in HRM practice. To probe this premise, we look to survey data on Minnesota districts' pay practices. To see whether Q-Comp districts actually were more likely to use P4P, we conducted an independent phone survey of Minnesota school district human-resource professionals about district pay practices without making any mention of Q-Comp.[5] Q-Comp districts report starkly different ways of compensating teachers than other districts. Among Q-Comp participants, 86% report paying for student performance and 90% report paying for subjective evaluations. In contrast, none of the non-participating districts report paying on either of these dimensions. P4P in Q-Comp districts is a supplement to, rather than a replacement of, traditional compensation criteria. Basing pay on years of experience and educational credentials is reported by 95% of participating and 100% of nonparticipating districts. Though this evidence of cross-sectional difference is strong, our design requires pay-system *changes*.

Q-Comp adoption did change how districts compensate teachers according to Schools and Staffing Survey (SASS) data on whether districts 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. Fifty-five Minnesota districts were sampled in both the 2003-04 and 2007-08 SASS waves. Among districts *not* participating in Q-Comp during the second wave, 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 and 58% report paying for excellence in the post-adoption wave. Q-Comp adoption triggered a large rise in the likelihood of reporting pay-for-excellence.[6]

[5]This was part of a national survey of districts' policies (West 2012). The Minnesota sub-sample consists of 92 districts (38% response), 21 of which we know participate in Q-Comp from administrative sources.

[6]What about the 43% of Q-Comp districts that told SASS the district does not pay for excellence? West (2012) discusses how vagueness in SASS's pay-for-excellence question induces more measurement error than survey questions that ask whether pay is tied to more specific criteria, such as student performance.

While P4P was certainly new, it likely came with changes in other HRM practices. M.D.E. required districts have or put complementary HRM practices in place. The U.S. Department of Education's Race to the Top and Teacher Incentive Funds works similarly. Because we do not observe districts' professional development or classroom observation schedules prior to Q-Comp adoption, we cannot measure how much of a change in these complementary HRM practices Q-Comp represents. With respect to classroom-observation based evaluations, prior to Q-Comp, untenured teachers were commonly observed once or twice a year and tenured teachers were rarely observed. Pay was not tied to these. Some PLCs existed prior to Q-Comp but they were not as widespread and were not used to set P4P criteria.

Given this, the current paper aims to identify the effect of adopting a locally-designed P4P reform in the context of complementary HRM practices. As in any firm, P4P's impact on labor productivity will depend on employee screening, production, and monitoring technologies (Milgrom and Roberts 1990; Ichniowski et al. 1997; Prendergast 2002; Bloom and Van Reenen 2011). Therefore, school districts and other firms may have incentives to adopt P4P along with bundles of complementary organizational practices in any kind of observational or experimental context. For instance, performance measurement systems must be operating. The organization may also reorganize production and redesign jobs to allow teams to operate effectively and offer new opportunities for workers to improve their skills. Coincident management reforms — more classroom observations (Taylor and Tyler 2012), support for teachers meeting regularly in professional learning communities (Jackson and Bruegmann 2009), job-embedded professional development, improved information systems — are potentially important complements to P4P incentives.

In many studies, including this one, these bundles of practice will tend to vary together and the effect of each element is difficult to separate. We are confident that Q-Comp adoption implies that, for the first time, teachers in adopting districts could earn extra money based on performance; Q-Comp adoption indicates new adoption of P4P. However, districts changed

other organizational practices too. For this reason, we characterize the "treatment" as P4P-centered HRM reform.

Description of adoption process. Each year, nonparticipating districts decide whether to apply for Q-Comp and what specific proposal to make. M.D.E. decides whether to accept each proposal. The local teachers union organizes a district-wide vote on whether to accept any new contract terms required by the proposal. Districts that clear all these hurdles adopt Q-Comp, implement their proposals, and start getting additional funds. A few participating districts subsequently drop out of Q-Comp. Therefore, in each year, each district has either not applied, been rejected, has adopted, or has dropped Q-Comp.

Our sample includes all regular Minnesota public school districts with any students in grades between 3 and 8.[7] We code whether or not each district applied, adopted, or dropped in each year based on archives of districts' applications, letters from the state granting approval, and district progress reports provided by the M.D.E.[8] The number of districts in our sample by Q-Comp participation status and year and the number of unique districts ever in each category are presented in Table 1's top-left panel.

There was some very limited use of P4P in Minnesota districts prior to Q-Comp, which we measure but which necessitates some judgement in coding the timing of P4P-centered HRM reform adoption. In a precursor to Q-Comp in 2002, the State of Minnesota passed its first version of this grantor-grantee reform structure, inviting districts to apply to serve as P4P-pilot adopters. Five districts applied and, in 2003, they started participating in the pilot, which offered $150 per student-year in funding, guaranteed to last through 2009. The funds were offered in return for adopting some elements of what would eventually become Q-Comp. Additionally, in 2004, all schools in the Waseca district and three Minneapolis schools started participating in the Milken Foundation's Teacher Advancement Program (TAP) (Springer et al. 2008; Glazerman and Seifullah 2010). When Q-Comp, which combined elements of

[7] Appendix B describes the design and rationale for the sample in detail.

[8] Two districts first made an initial, failed application and a subsequent, successful one. For these districts, the initial, failed application is ignored.

the 2002 pilot program and TAP, became law in 2005, M.D.E. approved all TAP schools for participation in Q-Comp without requiring any changes to their programs. So, we code the TAP schools as adopting Q-Comp in 2004. In contrast, the five pilot districts' existing P4P policies did not automatically qualify them for Q-Comp. M.D.E. offered them the option to either modify their policies to fit within the range of proposals acceptable under Q-Comp (qualifying them for full Q-Comp funding) or to continue operating under the pilot program until the promised funding expired in 2009. Two pilot districts transitioned into Q-Comp in 2005; they are coded as adopting Q-Comp then despite the fact that the change in their HRM practices is relatively small. The other three pilot districts are coded as never applying to Q-Comp, putting these districts with some elements of P4P in the control group. We use this conservative approach in most analysis and, later, assess robustness by estimating models excluding the five pilot districts from the sample.

In 2003, none of the state's 364 districts had applied to Q-Comp, as the program did not exist. In 2004, TAP schools adopt. In 2005, another nine districts applied; all adopted. In 2006, 38 more districts applied, of which 29 adopted and 9 failed to adopt. Of all 361 districts in 2006, 42 were participating in Q-Comp, 9 were rejected, and 310 had not applied. This second year of the official program had the largest application and adoption cohorts. By 2009, the final year of our achievement data, 56 districts (16%) were participating, 20 (5%) had previously applied but were not participating, and 281 (79%) had never applied.

Table 1's top-right panel describes the analogous number of students in the analytic sample, those who took the state-mandated MCA exam in grades 3 to 8 between 2003 and 2009, from each type of district in each year. For instance, in 2009, participating districts included 98,411 tested students (29%), rejected districts had 26,136 (8%), and not-applied districts had 215,596 (63%). Applying districts are larger on average than districts that did not apply, which is not surprising given that applying for and adopting Q-Comp involved fixed costs but the state subsidy varied linearly per student. The following section studies selection into Q-Comp more carefully and estimates effects of the reforms it spurred.

3 Effects of P4P-centered HRM reform

Identification. 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 this flexible approach perform? What was the average effect of the reforms after six years and over $200 million in state funds allocated? To understand the effect of HRM reform on student achievement, we analyze panels of individual student achievement in reading and math on two different kinds of achievement tests: the MCA and the NWEA. In each panel, we observe an individual student's school-grade each year, though we cannot link to teachers nor any individual student's MCA and NWEA scores across panels.

The MCA, the primary state-mandated achievement test, is used for No Child Left Behind, state-published school report cards, and other accountability programs. Prior to the academic year starting in 2005, Minnesota required all public-school students in grades 3, 5, and 7 to take the MCA tests in spring. Starting in the 2005 academic year, the mandate expanded to include grades 4, 6, and 8. We convert individual MCA scores to z-scores by standardizing against the statewide mean and standard deviation by grade-year-subject. This puts all outcomes in standard deviation units and facilitates pooling across grades.

We use a difference-in-difference framework to explain variation in achievement within-student over time. In year t, student i attends grade g in school s in district d. We assume student achievement (y_{it}) is produced according to the following function:

$$y_{it} = \beta Q_{dt} + \alpha w_{sgt} + \gamma_i + \tau_t + \epsilon_{it}. \tag{1}$$

Observable influences are measures of the district's Q-Comp program in that year (Q_{dt}) and measures of average demographic characteristics of students in the same school-grade-year as student i (w_{sgt}). Unobservables are partitioned into three additive components: a student fixed effect captures stable influences on i's achievement over time (γ_i), a year fixed effect (τ_t) allowing differences across years in average performance, and idiosyncratic influences

(ϵ_{it}). Standard errors allow heteroscedasticity and correlation of ϵ_{it} within district.

To allow β to capture the effect of Q-Comp participation on average, we define Q as a simple post-adoption indicator: 1(post-adoption)$_{dt}$. 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)$_{dt}$, making the reference category the single year immediately prior to adoption (Lovenheim 2009). Additionally, an indicator for district-years where the district once participated in Q-Comp but since dropped out is always included. In a dynamic specification similar to B, we use a full set of leads and lags capturing each year relative to the pre-adoption year separately. Because Q-Comp participation is not randomly assigned, systematic unobserved differences between districts that influence both Q-Comp adoption and outcomes could bias estimates of program effects.

The generalized difference-in-difference analysis relies on differences in the timing of reform across districts and, consequently, students to separate time effects from treatment effects. Student-level achievement panels permit identification of effects by studying changes in achievement within student, over time. We compare how, on average, a student's achievement changed after her district adopted reforms from its average pre-adoption level relative to the average achievement change experienced by students whose districts did not adopt over the same period. In the context of personnel economics, this can be interpreted as a study of how adoption of P4P and associated HRM reforms changes a district's teachers' average productivity, i.e. their ability to produce student learning as proxied by student achievement scores. Year indicators identify counter-factual year effects (τ_t).

Because peers may influence achievement and because variation in peers over time may correlate with changes in district P4P policy, we condition on a vector of school-grade-year average student demographic characteristics (w_{sgt}), including student shares eligible for free lunch, eligible for reduced-price lunch, in special education, male, African American, Hispanic, Asian American, and Native American. w also includes total enrollment measured

in thousands of students. w does not vary across subject, although its effects (α) can. Table 2 describes the average values of these context variables for students tested in reading in grades 3-8 across years 2003 through 2009. The top panel presents summary statistics for the MCA sample, which is analyzed first.[9]

The model is identified by assuming that the timing of the participation decision (Q_{dt}) is conditionally uncorrelated with unobserved influences (ϵ_{it}):

$$Cov[Q_{dt}, \epsilon_{it} | w_{sgt}, 1_i, 1_t] \equiv 0 \tag{2}$$

Within the restrictions of functional form, this model yields unbiased estimates of reform effects if selection into Q-Comp is based on stable differences in students' achievement levels. If, for instance, districts with higher student achievement levels are more likely to adopt or to adopt earlier than districts with lower levels, this is not a problem. The crucial condition is that within-student, time-varying, unobserved influences on scores are not systematically related to whether or when a district adopted Q-Comp.

As a preliminary test of this condition, we estimate a hazard model of Q-Comp adoption. The model predicts whether any district adopts Q-Comp in each year $t = 2005, 2006...2010$. Districts apply in $t-1$ to adopt in t. Changes in each district's average math and reading achievement leading into $t-1$, i.e. the difference between average achievement in $t-2$ and $t-1$, are the main predictors of interest. Additional predictors are districts' $t-1$ levels of average math and reading achievement, student demographics, teacher characteristics, and indicators of the districts' fixed M.D.E. region. Neither achievement changes nor levels significantly predict adoption (Table A-1), evidence consistent with the identifying condition.

Average effects on student achievement. The HRM reforms spurred by Q-Comp adoption are estimated to raise average reading achievement by about 3% of a standard deviation on the state-mandated MCA test. Table 3 presents estimates of equation (1). As reported in the first column under specification A, this analysis estimates $\hat{\beta}$ (SE) of 0.031 (0.015), implying

[9]Peer context variables' means approximately equal student-level means for the underlying variables.

a 3.1 percent of σ increase in students' MCA reading achievement in years after Q-Comp adoption compared to years before, adjusting for the changes that occurred in districts that did not adopt across the same time span. The effect's 95% confidence interval ranges from 0.002 to 0.060 σ. Coefficients on all peer covariates are as expected, except the positive coefficient on the share of students eligible for free lunch.[10]

However, estimates of β would be biased if districts select into participation based on fluctuations in student achievement levels. For example, if a district is more likely to adopt in a year when its students' scores would rise for other reasons than in a year when they would fall, this would violate the identifying condition and bias estimated effects upward.

Exploiting the fact that pre-adoption data are available, we perform a falsification test by testing directly for the presence of differential trends among non-adopters and future adopters *prior to adoption*. In specification B, the coefficient on $1(2+ \text{ pre-adoption})_{dt}$ tests for unexplained pre-adoption differences in outcomes between adopters and non-adopters and, thereby, provides a falsification test of the parallel trends condition necessary for identification. If the coefficient on the pre-adoption indicator differs from zero, it would appear that Q-Comp has an effect on achievement *before* it is implemented. As this is not possible, it would suggest that adopters were experiencing systematically different trends than non-adopters. The estimated coefficient -0.010 (0.018) is evidence that there were *not* systematic differences in reading achievement trends in adopting districts in the years leading up to adoption, supporting the validity of the identifying condition.

We next estimate the same models using NWEA, rather than MCA, scores as outcomes. The MCA and NWEA both have many strengths. Both are student panels covering hundreds of thousands of Minnesota students in math and reading. Both cover multiple years prior to and subsequent to program adoption and both cover many districts that adopted Q-Comp, many that applied but failed to adopt, and many that never applied for Q-Comp, allowing

[10]Net of peers' race/ethnicity, more low-income peers appears to increase achievement, perhaps because some school funding increases with low-income share. The effect of share eligible for free lunch changes to negative if racial/ethnic shares are excluded from the regression.

falsification tests. They complement each other well. Where one is weak, the other is strong. While the MCA was revised from version I to version II in 2005, the NWEA was stable throughout.[11] While the NWEA covers only students whose districts chose to test them, the MCA covers all students. Table 1's panel B reports by year how many districts used the NWEA and how many students took it.[12]

Comparison of estimated effects and pre-adoption trends between the MCA and NWEA is useful for two main reasons. First, NWEA provides a second falsification test for the identifying condition that adoption is exogeneous to achievement trends and one which is not subject to the possibility of a spurious null due to the 2005 MCA-revision. Second, comparison of effects in the MCA and NWEA offers evidence on the extent to which any apparent effect reflects real gains to reading skill rather than simply efforts to "teach to the test." The comparison yields evidence on the generalizability of learning (Neal 2011). To make the NWEA analysis as parallel as possible to the MCA analysis, we use NWEA spring achievement scores for grades 3-8 for academic years starting 2003 to 2009 standardized within grade-subject using national student norms (Hauser and Kingsbury 2008).[13]

Q-Comp participation is estimated to raise NWEA reading achievement by 0.032 (0.016) σ (Table 3: Column 3). The falsification test in specification B shows that the timing of

[11] We use three strategies to deal with the MCA revision in the year of Q-Comp's start. First, converting to z-scores within grade-year-subject helps ensure the test's units are comparable across versions. Second, we explore whether year-to-year correlation in MCA achievement within student is different across the version-transition year than across other years. If the correlation between versions I and II were much weaker than the year-to-year correlation within version, this could weaken our ability to test for differential trends in unobservable influences on achievement between adopters and non-adopters. We find that the within-student, year-to-year correlation between the MCA-I and MCA-II is substantially equivalent to the correlation across two years of MCA-II, suggesting the revision is not an issue in testing for differential pre-adoption trends. To test for differential year-to-year correlation, we regress each student MCA score on its lag, an indicator that the lag score is MCA-I (rather than MCA-II), and the interaction of the lag score and the MCA-I indicator. For reading, the estimated coefficients are 0.787 (0.001) on the lag score and -0.006 (0.002) on the interaction term. The MCA-I to MCA-II correlation is not meaningfully different than the MCA-II to MCA-II correlation, though the estimate is very precise and so statistically significant. For math, the corresponding estimates are 0.819 (0.001) and -0.023 (0.002), also evidence of very high correlation across versions. Third, we bring in evidence from the NWEA, a single version of which spans the period.

[12] Table 2 allows comparison of the NWEA sample's (school-grade-year peer) demographics to those for the statewide MCA sample. They are broadly similar, with the NWEA sample slightly higher-income and more white. Our models include student fixed effects and these measures of peer characteristics.

[13] NWEA test dates are not state-mandated and vary somewhat across districts. We include elapsed days between NWEA spring and fall tests in all NWEA models.

Q-Comp adoption is uncorrelated with pre-adoption NWEA reading trends (Column 4). Though the NWEA sample contains only about 32 percent of student-years from the MCA sample and is a nonrandom subsample selected by the districts, results on reading achievement are remarkably consistent across the MCA and NWEA.

Analogous results for math achievement are presented in the right-side of Table 3. The estimated effect of Q-Comp participation on MCA math achievement is 0.004 (0.021) σ and NWEA math is 0.038 (0.026) σ. The 95% confidence interval for the MCA ranges between -0.04 and 0.04 σ and, for the NWEA, between -0.013 and 0.088 σ. These estimates are consistent with a null effect or a true effect similar to that estimated for reading. Importantly for the credibility of the design, the falsification tests for math come back clean using both the MCA and NWEA.

In sum, the estimates in Table 3 suggest three main points. First, the fact and timing of adoption appear uncorrelated with pre-adoption achievement trends. Second, reform is estimated to raise achievement by about 3% of a standard deviation in reading, with 95% confidence intervals from just above zero up to just above 6% of a standard deviation. Third, the estimated effect on math achievement is more ambiguous and less precise.

Dynamics of effects. Next, we estimate models that allow the effect of Q-Comp participation to differ depending on the number of years elapsed relative to Q-Comp adoption. The post-adoption indicator is replaced with a set of indicators for the first year post-adoption, the second year post-adoption, etc. By replacing the single indicator for any year two or more pre-adoption with indicators for each pre-adoption lead, more refined falsification testing is also possible. As in specification B, the year immediately prior to adoption is omitted.

Table 4 shows the effect of Q-Comp participation on student achievement by years elapsed relative to Q-Comp adoption. The four columns of results correspond to effects on reading as measured by the MCA and NWEA and math on the same two tests. Point estimates and 95% confidence intervals are graphed in the four panels of Figure 1.[14] The post-adoption

[14]In the Table 4 results, the effect of the final lag is identified off of a single cohort so that lag is confounded with a cohort. As coefficients are similar across the final two lags, the figure presents estimates where the

effect estimates suggest that, in reading, the effect started small and grew. Districts that were in their first year of Q-Comp adoption saw their MCA reading scores increase by an average of 0.023 (0.011) σ above the prior year more than never-adopting districts did over the same two years. By the fourth year after adoption, the estimated effect is 0.076 (0.025) σ. Results are similar for NWEA reading achievement where the first-year effect estimate is 0.020 (0.010) σ and the fourth-year effect estimate is 0.062 (0.027) σ. Effect estimates for math are mixed across tests as before. Regarding falsification, soon-to-adopt districts are not experiencing systematically different achievement trends compared to districts that are not adopting over the same interval, as all coefficients for leads are small and insignificant (Table 4; Figure 1).

Teaching-to-the-test or generalizable skills? Within the range of state-permitted policies and in agreement with the teachers' local union, Q-Comp districts have flexibility to choose the performance standards that trigger payment of Q-Comp bonuses to teachers. The apparent congruence of Q-Comp's reading effects across the MCA and NWEA could result from teaching to different high-stakes tests in different districts rather than real increases in students' reading skill.[15]

To assess this possibility, we use a pair of indicators for whether the MCA and/or NWEA is named as a high-stakes test in determining award of school-wide student achievement bonuses, the only kind of performance bonus consistently tied to student standardized achievement tests. These are also intended as proxies for the relative emphasis each test receives within the school. Whether these bonuses are tied to MCA and/or NWEA outcomes

final two post-adoption years are pooled, which avoids confounding and improves precision.

[15]Consider this extreme example. Suppose half of Q-Comp districts tie stakes to the MCA reading test and the other half tie to the NWEA reading test. Each district produces 0.06 σ gains on the high-stakes test purely by teaching to that test, while producing no skill gain and, consequently, no effect on the other test. This would show up as a 0.03 σ effect on both tests, as in Table 3, but without any real learning. (Neal 2011) explains the idea this way, " 'Suppose that in every period, the fifth grade students in this district had also taken a second math assessment, B, and teachers were not rewarded or punished as a result of student outcomes on assessment B. Would one have observed gains on assessment B following the introduction of incentive pay that were comparable to the gains observed on assessment A?' In sum, do gains measured on the assessments used to determine incentive payments reflect increases in skill that create general improvements in math assessment results or only improvements specific to one assessment format or a particular set of questions?"

varies across districts and, in some districts, varies across schools and grades. In contrast, previous literature on the generalizability of achievement gains produced by P4P-centered HRM reform use variation in outcomes and stakes at the state-year level (Koretz et al. 1996; Vigdor 2009). Here, within state-year, we can hold the achievement test fixed and use variation in stakes across district-grades.[16]

There is no evidence that larger gains appear on the higher-stakes test (Table A-2). For reference, the first column reproduces the specification A result reported earlier (Table 3: Column 1). The second column adds an indicator of the MCA being a high-stakes test for school-wide bonuses in a student's school-grade-year. The estimated coefficient on this indicator for reading is -0.004 (0.02) σ and including it leaves the estimated Q-Comp effect, now referring to the Q-Comp effect among districts that do not tie school-wide bonuses to the MCA, basically unchanged. The NWEA results are very similar (Table A-2: Column 4). The effect of Q-Comp participation on NWEA reading achievement among school-grades that do not tie school-wide bonuses to the NWEA is 0.031 (0.021) σ and no different among school-grades that tie to the NWEA. Taken together, this provides further evidence that the estimated effect represents generalized gains in reading skill.[17]

The bottom panel presents analogous results for math. Again, there are no significant differences in achievement changes between students in Q-Comp school-grades taking tests with school-wide bonuses attached compared to students in Q-Comp school-grades taking tests without school-wide bonuses attached. The difference in estimated math effects between the MCA and NWEA does not appear to be driven by teaching to a specific test. The fact that most teacher bonuses are tied to classroom observations or individually-negotiated performance standards, rather than to standardized test scores, may explain this null result and does give assurance that any achievement effects are not driven by teaching-to-the-test.

[16]Among districts ever adopting Q-Comp, 54 (23) percent of the post-adoption MCA sample comes from school-grades where school-wide bonuses are tied to MCA (NWEA) performance. Some tie to both or neither.

[17]This is evidence against teaching to a *specific* test within subject. It does not rule out the possibility that the curriculum narrowed to focus on tested skills to the detriment of non-tested skills and subjects.

Robustness and additional falsification. Most Minnesota districts never adopted or applied to Q-Comp. The previous analysis included students in never-adopting districts to help identify the counter-factual time trends and peer effects. This generated evidence that estimated reform "effects" prior to adoption were zero, meaning that adopting and non-adopting districts were not experiencing differential achievement trends leading up to adoption. This section reports additional robustness tests.

We first restrict the sample to observations only in districts that adopt and then to those only in districts that apply. We estimate the reform effects and the pre-adoption placebo "effects" in both the adopters-only and the appliers-only subsamples using the same specification B as in Table 3. The MCA sample includes students from the 81 adopting and the 102 applying districts (Table 5: Columns 1 & 2). The estimates of Q-Comp adoption effects on MCA reading are 0.012 (0.015) σ among adopters-only and 0.018 (0.013) σ among appliers-only, which are both smaller in magnitude and less precise than the 0.028 (0.012) σ estimate obtained from the corresponding full sample analysis (Table 3). These estimates are not statistically distinguishable either from zero or from the full-sample estimate. The estimates for NWEA reading are 0.019 (0.015) σ among adopters-only and 0.023 (0.014) σ among appliers-only. Evidence of a reading effect is weaker in the adopters-only and appliers-only samples than in the full sample.

In addition, we estimate placebo failed-application "effects" along with effect of adoption in the full sample. This is similar to dropping never-appliers from the sample but is superior in that it exploits information on the timing of failed applications. If the crucial selection process creating bias divides never-appliers from appliers and what produces gains were district management's desire for reform (proxied by application) rather than policy reform itself, then we would see a positive failed-application "effect" similar to the effect of adoption. On both reading tests, failed applications appear to have no effect (Table 5: Columns 3 & 6), evidence that the effects are not driven solely by district motivation but that reform itself matters. Also, there is evidence that failed appliers experience no differential pre-application

trends compared to never-appliers.

Another source of potential bias is the pilot districts. We have used a conservative coding rule in the primary analysis.[18] Because there is room for judgement in how to treat these districts, we repeat the robustness and falsification tests excluding pilot districts from the sample (Table 5: bottom panel). Excluding pilots, the positive effect of Q-Comp on MCA reading scores in the appliers-only sample moves from insignificant to marginally significant. We repeat the robustness analysis for math achievement and find the results very stable. For math, estimated effects in the adopters-only and applier-only samples (Table 6: Columns 1, 2, 4 & 5) look similar to those from the full sample (Table 3). Coefficients on the pre-adoption indicator in MCA math in the adopters- and appliers-only samples are similar in magnitude to the full-sample results but become marginally significant, giving some weak evidence that adopting districts' were more likely to adopt when their MCA math scores were improving. In math, as in reading, there is no evidence of a placebo "failed-application" effect (Table 6: Columns 3 & 6). Also, excluding pilots does not change results (bottom panel).

3.1 Mechanisms: students, money, teachers, or efforts

Lazear (2000) posits that P4P adoption might have a positive effect through either changing the composition of the workforce towards higher ability individuals or increasing effort of the incumbent workforce. In education, a third possibility is that parents and students sort themselves differently across districts once P4P is introduced. Here, a fourth possibility is simply that the additional $6,500 per 25 student class in annual funding allowed productive increases in district expenditures, quite apart from teacher pay reform. This section develops evidence on the extent to which changes in districts' students, teachers, or expenditures matter. Otherwise, effects would appear attributable to changes in the internal operations

[18] As discussed in Section 2, the five pilot districts had elements of P4P-centered HRM policies in place throughout the entire 2003-2009 study period. Three of them applied and qualified for Q-Comp in 2005. We code them as adopting that year, though this represents a smaller change in policy than non-pilot adopters experienced. The other districts stuck with the pilot program and ran a light version of Q-Comp throughout. We code them as never appliers though they adopted some P4P-centered HRM reform in 2003.

of the district and efforts from incumbent teachers.

New students. Given mounting evidence that peers matter for student achievement (Hoxby 2000; Whitmore 2005; Graham 2008; Sojourner 2013), it is important to ask whether reform had its effect by triggering changes in student sorting across districts. If so, then effects may be zero-sum, with losses in non-Q-Comp districts offsetting gains in Q-Comp districts. To assess this possibility, we remove peer observables from the conditioning set. As can be seen by comparing the first and second columns of Table 7, this barely changes the estimated effect of Q-Comp participation on MCA reading. Similarly, this change to the conditioning set barely affects the estimated Q-Comp effect on NWEA reading or on either test of math achievement (Table 8). This suggests that changes in student sorting (on observables) across districts are not responsible for the observed effect.

New money. The gains could simply be due to the increased funding that districts receive through Q-Comp participation, rather than due to the HRM reforms themselves. Achievement gains may come simply because Q-Comp participation brings additional funding to the district and so the district increases expenditures in a variety of useful ways. Districts' budgets increase, not their productivity. If this were true, then effects should be interpreted primarily as the effect of an increase in expenditures rather than of P4P-centered HRM reform. To assess this, we include the log of district-year expenditures as a covariate and see how much of the effect on student achievement this accounts for. While the effect of additional expenditures appears positive, this does not account for the effect on MCA reading (Table 7). The coefficient on 1(post-adoption) remains unchanged. Similar results are obtained for NWEA reading (column 7) and MCA and NWEA math (Table 8).

New teachers or new efforts? This leaves the two standard mechanisms by which P4P might raise productivity: shifting the composition of the workforce (sorting) or increasing the productivity of the incumbent workforce (incentives/effort). Any change in district-level productivity, ΔY, can be decomposed between effort and sorting channels: $\Delta Y =$

$\Delta Y^R r + (Y^H - Y^E)(1-r)$.[19] The effect of a reform treatment (T) is then,

$$\frac{\partial Y}{\partial T} = \underbrace{\frac{\partial Y^R}{\partial T} r + Y^R \frac{\partial r}{\partial T}}_{effort} + \underbrace{\frac{\partial (Y^H - Y^E)}{\partial T}(1-r) - (Y^H - Y^E)\frac{\partial r}{\partial T}}_{sorting} \tag{3}$$

Available data on retention rates allow us to measure $\frac{\partial r}{\partial T}$ directly. We do not observe individual-teacher productivity measures, so we cannot separate $\frac{\partial Y^R}{\partial T}$ from $\frac{\partial (Y^H - Y^E)}{\partial T}$ as cleanly as Lazear (2000). We present evidence on the extent to which (changes in) various, observable teacher characteristics can or cannot account for the observed effects of reform on student achievement.[20] We observe two kinds of changes. First, we observe changes in the flow rates of workers into, out of, and between districts. Appendix C describes the workforce flow measures in detail. Second, we observe changes in teacher experience profiles and education levels. Specifically, we observe the school's share of teachers with ≤ 5, 6-15, and ≥ 16 years of experience and the school-grade share with at least a masters degree.

We assess the degree to which changes in teacher flows and characteristics account for gains in student achievement in two steps. First, we estimate the dynamic effects of HRM reform on workforce flows and experience profiles.[21] Second, we add the vector of teacher workforce flows, experience profiles, and education as predictors to the student achievement models and assess the extent to which this set of teacher workforce measures account for the reform's achievement effects.

Reform had little effect on districts' novice teacher rate, new hire rate, rate of flow to ever-

[19] The effort channel is the change in average retained incumbent productivity, ΔY^R, times the retained-worker share, r. Sorting is the difference between the average productivity of new hires and exiters, $(Y^H - Y^E)$, weighted by turnover.

[20] Though we cannot observe these sub-populations separately, suppose $Y^j = X^j \pi + \varepsilon^j$ for $j = R, H, E$ where X^j are mean observable determinants and ε^j are mean unobservable determinants, i.e. effective effort representing changes in effort levels and complementary changes in organizational practice that raise the productivity of effort. If incumbents' observables are fixed, changes in Y^R must be due to changes in effort $\frac{\partial Y^R}{\partial T} = \frac{\partial \varepsilon^R}{\partial T}$. In contrast, effects due to (Arriver-Exiter) contrasts could be in either observables or effort, $\frac{\partial (Y^H - Y^E)}{\partial T} = \frac{\partial (X^H - X^E)}{\partial T}\pi + \frac{\partial (\varepsilon^H - \varepsilon^E)}{\partial T}$.

[21] Depending on if the unit of observation (u) is the district or school, we study effects of reform on workforce observables (X_{ut}) using a model that parallels equations (1) and (2) but omits w_{sgt} and replaces γ_i with a unit effect.

adopting districts, or retention rate (Table A-3: Columns 1-4) or on teacher experience or education (Columns 5-7).[22] Responses to adoption look generally flat. There is little evidence that reform increased the hiring of novice teachers, increased new hires more generally, increased the flow from never-adopting to ever-adopting districts, decreased retention of incumbent teachers, or affected experience or education.

To assess more formally whether these might be driving the positive effect of reform adoption on student achievement, we add this vector of workforce measures to the achievement model. The estimated effect of adoption on MCA reading does not change (Table 7: Column 4). Similar stability is obtained for NWEA reading (Table 7: Column 8) and both tests of math achievement (Table 8: Columns 4 & 8). Both $\frac{\partial r}{\partial T}$ and $\frac{\partial(X^H - X^E)}{\partial T}$ appear to be approximately zero. Changes in flow rates or shares with a given experience or education level do not appear to generate the effects.

Therefore, changes appear due to the combination of changes in effective effort among incumbents ($\frac{\partial \varepsilon^R}{\partial T}$) and from improvements in effective effort gains from turnover ($\frac{\partial(\varepsilon^H - \varepsilon^E)}{\partial T}$). Under plausible conditions, the effect of reform through both channels would be most positive among the least experienced teachers and present itself as a positive interaction of reform with share of teachers with low experience.[23] To investigate this further, we allow for heterogeneity in the effect of reform depending on the share of a school's teachers in each experience category while also conditioning on experience-share levels. Estimates are presented in Table 9. For MCA reading (column 1), the effect of reform is estimated at 0.116 (0.054) for teachers with ≤ 5 years of experience, 0.058 (0.032) for teachers with 5-16 years

[22] In two districts (Minneapolis and Roseville), some schools adopted Q-Comp in different years than other schools in the same district. In the analysis up to now, we treated each within-district adoption cohort as a separate synthetic district and exploited the within-district differences in adoption timing to help identify Q-Comp effects. However, our workforce flow measures derive from data that connect teachers to districts, not to schools. Consequently, in some years, we cannot tell if Minneapolis and Roseville teachers are in a Q-Comp or non-Q-Comp school. Therefore, we exclude Minneapolis and Roseville schools from analyses involving workforce flow measures. Student achievement effects look similar in this restricted sample.

[23] If P4P-centered HRM reform attracts new hires with higher productivity than exiters and if less experienced teachers are more likely to churn, this sorting story would show up as a positive interaction of reform with share of teachers with low experience. If P4P-centered reform inspires increased effort and this operates most strongly among less experienced teachers who are less secure in their jobs, less set in their ways, and for whom potential bonuses are a bigger share of compensation.

of experience, and -0.043 (0.027) for teachers with 16+ years of experience. It is also worth noting that, outside of Q-Comp, achievement increases with the share of very experienced teachers, as implied by the significant 0.129 (0.025) effect of share of teachers with 16+ years of experience compared to share with 6-15 years experience. For NWEA reading and both math tests (columns 2-4), reform effects are not statistically significantly different across experience levels but a similar qualitative pattern appears. This evidence seems to corroborate the theoretical prediction of larger positive effects among less experienced teachers but cannot separate the sorting and effort channels among less experienced teachers.

Taken together, results suggest that reform produced achievement gains primarily through changes in district policies and effects among less experienced teachers rather than through changes students demographics, additional expenditures, workforce flow rates, experience profiles, or increased effort among the most senior teachers.

4 Conclusion

This study exploits large shifts in dozens of Minnesota school districts' HRM policies towards P4P that occurred since the 2005 introduction of the state's Q-Comp program. We find that, on average, P4P-centered HRM reform is associated with a 3% of a standard deviation increase in student achievement. This shows up on two different reading assessments and on one of two math assessments. Gains were not driven by districts that put higher stakes on a particular subject-test nor are they simply the effect of being motivated enough to apply.

This is some of the strongest evidence yet that teacher pay reform can raise U.S. teacher productivity and student learning. Prior U.S. field studies that found any positive effects of P4P on achievement used cross-sectional comparisons (Figlio and Kenny 2007), pre-post designs (Vigdor 2009), focus on a single district (Ladd 1999), or found that gains appeared only on the high-stakes test and not on a low-stakes one (Koretz 2002). The two largest P4P experiments, each conducted within a single district over 3 years or less, found null or

negative effects on high-stakes achievement tests (Springer et al. 2010; Fryer 2011).

Because the effect of a given P4P incentive may vary depending on its context, unmeasured heterogeneity in pre- and post-adoption contexts might also explain some discrepant results in the literature. Like the present study, most P4P studies in education really measure the effects of P4P-centered HRM reform as proxied by P4P. Complementary HRM practices are correlated with P4P but not well measured. The fact that Q-Comp changed professional development and evaluation procedures as well as compensation criteria may help explain why it was effective, in contrast to reforms focused exclusively on compensation reform (Springer et al. 2010; Fryer 2011). Despite a limited ability to unbundle which pieces of the reform matter, this study contributes new evidence that adoption of HRM reforms centered on P4P can raise productivity. Alternatively, the difference in results compared to Springer et al. (2010); Fryer (2011) may be attributable to Q-Comp's relative permanence and the consequent larger impact on workers' expectations and willingness to adjust, as well as a longer time horizon for outcome measurement. Our results are similar to Lavy (2009), where the introduction of a P4P tournament among Israeli teachers raised student achievement by changing practice among incumbent teachers without evidence of gaming.

While the magnitudes of the reforms' achievement impacts are not large, they appear to be socially productive given their modest costs. The social value of a 20% of a standard deviation achievement gain for a kindergarten class has been estimated conservatively at $200,000 in net present value (Krueger 2003; Hanushek 2010; Chetty et al. 2011).[24] This implies that the point-estimated 3% of a standard deviation achievement gain induced by Q-Comp pay reforms would be worth approximately $30,000 in present value, though given the statistical uncertainty of the estimate, benefits are possibly much lower or higher. In terms of costs, Q-Comp adoption involved an increase in funding of only about $6,500 per year per 25 student class, suggesting a 5 to 1 benefit to cost ratio. The true costs of the pay reforms may be greater if districts additionally allocate their own funds to finance bonuses or

[24]Updating Krueger's most pessimistic estimate of a $5,718 per child benefit to 2012 dollars and multiplying times 25 children per teacher yields an estimate of just over $200,000.

associated reforms, such as time for teachers to meet in professional learning communities, creation of data and assessment teams, or cadres of trained classroom observers. However, the $4,000 per classroom gap between the new $6,500 in funding and the $2,500 average maximum total bonus leaves Q-Comp districts with significant cushion to cover the cost of complementary administrative reforms without adding additional resources. This suggests that the set of P4P programs funded by Q-Comp are yielding positive social returns.

By studying a long-standing education reform in the field, valuable new insights are gained. First, the reforms studied here have proven politically and operationally feasible. This is not a fragile model program. This is happening at scale, long-term, and with the support of teachers and their unions through a process of labor-management cooperation.[25] Second, we provide unique new evidence on workforce flows suggesting that unmeasured changes among incumbent teachers and administration drove the achievement gains, rather than teacher turnover. At least with respect to changes to the pay schedule of this magnitude and this time span, P4P did not induce important changes in the observable composition of the teacher workforce. But the fact that effects appeared larger among less experienced teachers suggests that unobserved productivity changed in this group. As teachers cannot transfer tenure across districts, the lack of mobility across districts may not be surprising. Reforms that facilitate inter-district mobility of teachers and ease entry into the teacher workforce could be useful complements.

Within the bounds of what the state allowed, we can see what kinds of programs districts chose to adopt and which they did not. Despite theoretical reasons for preferring tournaments where teachers compete for a limited pool of money and where higher-performing teachers are guaranteed to earn more than others (Barlevy and Neal 2012) and evidence that tournaments have worked in some settings in Israel (Lavy 2009), no district adopted a tournament. A teacher's bonus is usually not tied directly to the standardized test scores of his or her own

[25]Dillon (2007) wrote, "A major reason [Q Comp] is prospering, Gov. Tim Pawlenty said... is that union leaders helped develop and sell it to teachers. "As a Republican governor, I could say, 'Thou shalt do this,' and the unions would say, 'Thou shalt go jump in the lake,' " Mr. Pawlenty said. "But here they partnered with us.... This is a complex process of changing a culture, and it will fail if teachers don't support it.""

students. The bonuses tied to standardized achievement targets tend to be small, tied to school-wide performance, and paid according to generally very simple schedules.[26] We do not observe any districts that link rewards to "value-added" measures of teacher effectiveness. Thus we cannot comment on the debate over the use of these measures in personnel decisions, except to say that they have not been popular in Minnesota thus far.

Although bonuses are largely tied to local goals and subjective evaluations that appear corruptible, some theories and evidence support this approach. By tying stakes to criteria other than standardized tests, coaching, gaming and other waste through unproductive hidden action may be avoided. Our findings also echo Marsden and Belfield (2007) and Marsden (2010) who argue from U.K. evidence that P4P can succeed in education when it promotes a process of locally-negotiated goal setting. This process of setting goals, when taken seriously by all sides, may harness teachers' local information about the most productive strategies for success better than a centrally-defined standard. Also, evidence is emerging from other states that attaching stakes to a process with regular feedback from classroom observers can produce achievement gains (Taylor and Tyler 2012). 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. Q-Comp districts also devote resources and attach incentives to teachers meeting regularly in professional learning communities and to meeting goals set negotiated within those groups, which may constructively harness peer effects among teachers (Jackson and Bruegmann 2009).

Finally, the paper increases understanding of locally-designed education reform. Here, the grantor-grantee relationship between education authorities and districts allowed use of local information and experimentation in finding appropriate, feasible P4P designs and complementary HRM practices. Grantees may have captured some information or agency rents, but this evidence suggests that the process can generate meaningful social returns.

[26]Herweg et al. (2010) shows that simple payment schedules may be optimal for loss-averse teachers.

References

Bandiera, O., I. Barankay, and I. Rasul, "Incentives for managers and inequality among workers," *Quarterly Journal of Economics*, 2007, *122* (2), 729–773.

Barlevy, G. and D. Neal, "Pay for Percentile," *American Economic Review*, 2012, *102* (5), 1805–1831.

Bloom, N. and J. Van Reenen, "Human Resource Managament," *Handbook of Labor Economics*, 2011, *4b*, 1697–1763.

Cappelli, P. and D. Neumark, "Do "High-Performance" Work Practices Improve Establishment-Level Outcomes?," *Industrial & Labor Relations Review*, 2001, *54*, 737–882.

Chetty, R., J. Friedman, N. Hilger, E. Saez, D. Schanzenbach, and D. Yagan, "How does your kindergarten classroom affect your earnings?," *Quarterly Journal of Economics*, 2011, *126* (4), 1593.

Danielson, C. and T.L. McGreal, *Teacher evaluation to enhance professional practice*, Alexandria, Virginia: Association for Supervision and Curriculum Development, 2000.

deFiebre, C., P. Lopez, and J. Hopfensperger, "Legislature 2005 Wrapping Up," *Star Tribune*, July 14 2005.

Dillon, S., "Long Reviled, Merit Pay Gains Among Teachers," *New York Times*, June 18 2007.

Editorial, "Teachers' salaries: Q-Comp has potential," *Star Tribune*, August 14 2005.

Figlio, D.N. and L.W. Kenny, "Individual teacher incentives and student performance," *Journal of Public Economics*, 2007, *91* (5), 901–914.

Fryer, R.G., "Financial incentives and student achievement: Evidence from randomized trials," *The Quarterly Journal of Economics*, 2011, *126* (4), 1755–1798.

Glazerman, S. and A. Seifullah, "An Evaluation of the Teacher Advancement Program (TAP) in Chicago," 2010.

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

Graham, B.S., "Identifying Social Interactions Through Conditional Variance Restrictions," *Econometrica*, 2008, *76* (3), 643–660.

Hanushek, E.A., "The economic value of higher teacher quality," *Economics of Education Review*, 2010, *30* (3), 466–479.

_ **and L. Woessmann**, "Overview of the symposium on performance pay for teachers," *Economics of Education Review*, 2011, *30* (3), 391–393.

Hauser, C. and G.G. Kingsbury, "RIT Scale Norms: For Use with Measures of Academic Progress," Technical Report, Northwest Evaluation Association 2008.

Herweg, F., D. Muller, and P. Weinschenk, "Binary payment schemes," *American Economic Review*, 2010, *100* (5), 2451–2477.

Hezel Associates, "Quality Compensation for Teachers Summative Evaluation," Technical Report, Syracuse, NY 2009. archive.leg.state.mn.us/docs/2009/other/090321.pdf.

Holmstrom, B. and P. Milgrom, "Multitask principal–agent analyses," *Journal of Law, Economics, & Organization*, 1991, *7*.

Hord, S.M. and S.A. Hirsch, "Making the Promise a Reality," in A.M. Blankenstein, P.D. Houston, and R. W. Cole, eds., *Sustaining Professional Learning Communities*, Thousand Oaks, California: Corwin Press, 2008.

Hoxby, C.M., "Peer Effects in the Classroom," 2000. NBER Working Paper 7867.

Ichniowski, C., K. Shaw, and G. Prennushi, "The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines," *The American Economic Review*, 1997.

Jackson, C.K. and E. Bruegmann, "Teaching Students and Teaching Each Other," *American Economic Journal: Applied Economics*, 2009, *1* (4), 85–108.

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

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

_ , S. Barron, K.J. Mitchell, and B.M. Stecher, *Perceived effects of the Kentucky instructional results information system*, RAND: Santa Monica, CA, 1996.

Krueger, A.B., "Economic considerations and class size," *Economic Journal*, 2003, *113* (485), F34–F63.

Ladd, H.F., "The Dallas school accountability and incentive program," *Economics of Education Review*, 1999, *18* (1), 1–16.

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

_ , "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics," *American Economic Review*, 2009, *99* (5), 1979–2011.

Lazear, E.P., "Performance Pay and Productivity," *American Economic Review*, 2000, *90* (5), 1346–1361.

_ , "Teacher incentives," *Swedish Economic Policy Review*, 2003, *10* (2), 179–214.

Lovenheim, M., "The Effect of Teachers' Unions on Education Production," *Journal of Labor Economics*, 2009.

Marsden, D., "The Paradox of Performance-Related Pay Systems," *Paradoxes of Modernization*, 2010, *1* (9), 185–203.

_ **and R. Belfield**, "Pay for Performance Where Output Is Hard to Measure," *Advances in Industrial and Labor Relations*, 2007, *15*, 1–34.

Milgrom, P. and J. Roberts, "The economics of modern manufacturing," *American Economic Review*, 1990, pp. 511–528.

Muralidharan, K. and V. Sundararaman, "Teacher Performance Pay: Experimental Evidence from India," *Journal of Political Economy*, 2011, *119* (1), 39–77.

Nadler, C. and M. Wiswall, "Risk Aversion and Support for Merit Pay," *Education Finance & Policy*, 2011, pp. 1–31.

Neal, D., "The Design of Performance Pay in Education," Technical Report, National Bureau of Economic Research Working Paper No. 16710 2011.

Nobels, J., "Evaluation Report: Q Comp Quality Compensation," Technical Report, Minnesota Office of the Legislative Auditor 2009.

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

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

Prentice, G., S. Burgess, and C. Propper, "Performance pay in the public sector," Technical Report, Office of Manpower Economics 2007.

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

_ **, D.O. Staiger, T.J. Kane, and E.S. Taylor**, "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools," *American Economic Review*, 2012, *102* (7), 3184–3213.

Schwartz, N., "Aligning Teacher Improvement Strategies: A Mixed Method Study of Teaching Reform in Minnesota." PhD dissertation, University of Michigan 2012.

Sojourner, A.J., "Identification of peer effects with missing data," *Economic Journal*, 2013, *123* (569), 574–605.

Springer, M.G., D. Ballou, and A. Peng, "Impact of the Teacher Advancement Program on student test score gains," Technical Report, National Center on Performance Incentives, Vanderbilt University 2008.

_ , **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," Technical Report, National Center on Performance Incentives, Vanderbilt University 2010.

Taylor, E.S. and J.H. Tyler, "The Effect of Evaluation on Teacher Performance," *The American Economic Review*, 2012, *102* (7), 3628–3651.

The INFOHRM Group, *The Metrics Standard*, 1 ed., Washington, D.C.: INFOHRM Group Inc., 2006.

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

Victory New Hampshire, "Gov. Tim Pawlenty The 60 Second Education Update," 2012. http://60secondupdates.com/pawlenty/pawlenty-2.

Vigdor, J., "Teacher salary bonuses in North Carolina," in M. Springer, ed., *Performance Incentives: Their Growing Impact on American K-12 Education*, Brookings, 2009.

Wahlstrom, K., T. Sheldon, and K. Peterson, "Implementation of the Quality Compensation Program (Q Comp)," Technical Report, University of Minnesota's Center for Applied Research and Educational Improvement 2006.

Wei, R.C., L. Darling-Hammond, A. Andree, N. Richardson, and S. Orphanos, "Professional learning in the learning profession," Technical Report, National Staff Development Council 2009.

West, K.L., "Teachers Unions, Compensation, and Tenure," Technical Report, University of Minnesota 2012.

Whitmore, D., "Resource and peer impacts on girls' academic achievement," *American Economic Review*, 2005, *95* (2), 199–203.

5 Tables

Table 1: District and student Q-Comp participation in full sample and in NWEA sub-sample by year

Academic Year Starting	Number of Districts					Number of Tested Students in grades 3-8				
	Q-Comp Status					Q-Comp Status				
	Not applied	Rejected	Adopted	Exited	Total	Not applied	Rejected	Adopted	Exited	Total
Sample: all districts										
2003	364	0	0	0	364	178506	0	0	0	178506
2004	359	0	4	0	363	173119	0	1053	0	174172
2005	351	0	13	0	364	319194	0	13748	0	332942
2006	310	9	42	2	361	260976	5469	77613	1221	345279
2007	298	10	53	3	361	236092	5845	97137	2917	341991
2008	291	13	57	5	361	219919	11703	104375	3610	339607
2009	281	20	57	8	357	204507	26134	98327	10872	339840
Unit-years	2254	52	225	18	2531	1592313	49151	392253	18620	2052337
Unique units	369	20	65	8	369	545998	18475	125375	7122	696970
Sub-sample: district-years with NWEA test										
2003	74	0	0	0	74	63529	0	0	0	63529
2004	151	0	0	0	151	102072	0	0	0	102072
2005	186	0	6	0	192	110746	0	5288	0	116034
2006	187	4	26	1	218	100879	775	26601	402	128657
2007	192	8	34	2	236	69477	928	30258	1708	102371
2008	192	8	41	2	243	53670	1149	24762	1348	80929
2009	185	15	50	7	257	37291	2401	15846	2761	58299
Unit-years	1167	35	157	12	1371	537664	5253	102755	6219	651891
Unique Units	273	8	57	7	273	206311	1789	36686	2240	247026

Though outcomes for 2010 are not available, we observe another 14 districts adopt Q-Comp in 2010 and 1 applier that did not adopt. A total of 22 schools in the Minneapolis and Roseville school districts adopted Q-Comp at the school level; these are considered separate districts for this analysis.

Table 2: Descriptive statistics

Variable	Mean	Std. Dev.	Min.	Max.
Panel A: MCA Sample				
Peer context variables (2,052,337 student-years)				
Share male	0.513	0.060	0	1
Share free lunch	0.242	0.188	0	1
Share spec. ed.	0.136	0.07	0	1
Share Afr.-Am.	0.082	0.126	0	1
Share Hispanic	0.059	0.084	0	1
Share Asian-Am.	0.057	0.084	0	1
Share Native Am.	0.021	0.072	0	1
Enrollment (school-grade-year)	170.7	139.0	1	826
Expenditures and teacher characteristics (district-year)				
District-year				
Log(expenditures)	3.98	1.33	0.54	6.25
Percent novice teacher	3.68	2.01	0	27.27
Percent external hires	5.17	2.55	0	42.86
Percent retained	91.95	2.60	59.09	100
Percent flowed to ever-Q-Comp	0.19	1.97	-15.38	14.89
School-year				
Share with experience ≤ 5 years	0.168	0.105	0	1
Share with experience $6-15$ years	0.409	0.124	0	1
Share with experience ≥ 16 years	0.417	0.155	0	1
School-grade-year				
Percent with MA+	21.92	14.12	0	92.2
Panel B: NWEA sample				
Peer context variables (651,891 student-years)				
Share male	0.512	0.052	0	1
Share free lunch	0.197	0.119	0	1
Share spec. ed.	0.133	0.052	0	1
Share Afr.-Am.	0.037	0.051	0	1
Share Hispanic	0.049	0.062	0	1
Share Asian-Am.	0.035	0.041	0	1
Share Native Am.	0.023	0.079	0	1
Enrollment (school-grade-year)	193.2	152	1	826

Table 3: Effect of P4P-centered HRM reform on student achievement z-score by subject and test

Outcome Subject:	Reading				Math			
Outcome Test:	MCA		NWEA		MCA		NWEA	
Specification:	A	B	A	B	A	B	A	B
1(post-adoption)	0.031**	0.028**	0.032**	0.032**	0.004	-0.004	0.038	0.035*
	(0.015)	(0.012)	(0.016)	(0.012)	(0.021)	(0.017)	(0.026)	(0.020)
1(2+ yrs. pre-QComp)		-0.010		0.001		-0.026		-0.009
		(0.018)		(0.017)		(0.019)		(0.029)
Share male	-0.083***	-0.083***	-0.057	-0.057	-0.134***	-0.134***	-0.121**	-0.121**
	(0.029)	(0.029)	(0.052)	(0.052)	(0.038)	(0.038)	(0.061)	(0.061)
Share free lunch	0.104***	0.103***	0.029	0.029	0.104***	0.100***	0.045	0.044
	(0.036)	(0.036)	(0.047)	(0.047)	(0.034)	(0.033)	(0.066)	(0.065)
Share spec. ed.	-0.046	-0.046	0.132*	0.133**	-0.209***	-0.211***	0.097	0.095
	(0.040)	(0.040)	(0.067)	(0.067)	(0.071)	(0.072)	(0.080)	(0.080)
Share Afr.-Am.	-0.143***	-0.142***	-0.001	-0.001	-0.095***	-0.092***	-0.154	-0.151
	(0.030)	(0.030)	(0.088)	(0.088)	(0.030)	(0.030)	(0.113)	(0.111)
Share Hispanic	-0.075*	-0.076*	-0.102	-0.102	-0.007	-0.008	-0.084	-0.087
	(0.043)	(0.043)	(0.108)	(0.108)	(0.045)	(0.046)	(0.109)	(0.111)
Share Asian-Am.	0.021	0.021	0.086	0.086	0.068	0.068	0.009	0.010
	(0.046)	(0.045)	(0.149)	(0.149)	(0.043)	(0.042)	(0.171)	(0.172)
Share Native Am.	-0.131***	-0.131***	0.028	0.029	-0.050	-0.052	0.016	0.016
	(0.050)	(0.050)	(0.063)	(0.064)	(0.052)	(0.053)	(0.093)	(0.094)
Enrollment	-0.004	-0.005	0.005	0.005	-0.011	-0.012	-0.010	-0.010
	(0.006)	(0.006)	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)
Districts	369	369	273	273	369	369	273	273
Students	696,970	696,970	247,026	247,026	686,483	686,483	247,767	247,767
Student-years	2,052,337	2,052,337	651,891	651,891	2,007,029	2,007,029	655,341	655,341
Adj. R^2	0.774	0.774	0.793	0.793	0.793	0.793	0.838	0.838

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%. All specifications include year indicators, student fixed effects, and indicator for having dropped Q-Comp. Covariates are measured in student's school-grade-year.

Table 4: Effect of P4P-centered HRM reform on achievement by years elapsed from adoption

Outcome Subject:	Reading		Math	
Outcome Test:	MCA	NWEA	MCA	NWEA
7th year prior	-0.013	0.012	-0.040	0.111**
	(0.077)	(0.058)	(0.093)	(0.055)
6th year prior	-0.053	-0.013	-0.040	0.038
	(0.049)	(0.045)	(0.065)	(0.045)
5th year prior	-0.024	-0.004	-0.050	0.032
	(0.052)	(0.033)	(0.046)	(0.033)
4th year prior	-0.021	-0.017	-0.005	-0.010
	(0.036)	(0.028)	(0.051)	(0.036)
3rd year prior	-0.019	-0.003	-0.039	-0.025
	(0.023)	(0.025)	(0.030)	(0.046)
2nd year prior	-0.007	0.001	-0.022*	-0.009
	(0.016)	(0.015)	(0.013)	(0.021)
1st year prior omitted				
1st year post-adoption	0.023**	0.020**	-0.002	0.027
	(0.011)	(0.010)	(0.014)	(0.018)
2nd year post-adoption	0.021	0.042***	-0.006	0.043*
	(0.015)	(0.015)	(0.021)	(0.024)
3rd year post-adoption	0.041**	0.063**	-0.009	0.031
	(0.020)	(0.027)	(0.027)	(0.036)
4th year post-adoption	0.076***	0.069**	0.002	0.038
	(0.025)	(0.034)	(0.030)	(0.052)
5th year post-adoption	0.079**		0.023	
	(0.035)		(0.056)	
Districts	369	273	369	273
Students	696,970	247,026	686,483	247,767
Student-years	2,052,337	651,891	2,007,029	655,341
Adj. R^2	0.774	0.793	0.793	0.838

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.
Specification includes year indicators, student fixed effects, peer covariates, and indicator for having dropped Q-Comp as in Table 3. Outcomes are in standard deviation units. There is no fifth-year post-adoption effect for NWEA achievement because none of the 2004 TAP adoption cohort opted into NWEA testing.

Table 5: Robustness and further falsification of effect on reading achievement

Outcome Test:	MCA			NWEA		
Sample:	Adopters Only	Appliers Only	All	Adopters Only	Appliers Only	All
Specification:	B	B	B	B	B	B
1(post-adoption)	0.012	0.019	0.028**	0.019	0.024*	0.032***
	(0.015)	(0.013)	(0.012)	(0.015)	(0.013)	(0.012)
1(2+ yrs. pre-adoption)	0.009	0.006	-0.011	0.015	0.012	0.001
	(0.017)	(0.019)	(0.018)	(0.016)	(0.016)	(0.018)
1(post failed application)			0.007			-0.032
			(0.014)			(0.033)
1(2+ yrs. pre-failed app.)			-0.025			-0.001
			(0.028)			(0.015)
Districts	81	102	369	66	83	273
Students	257,039	316,250	696,970	88,016	96,790	247,026
Student-years	707,958	880,412	2,052,337	225,397	249,446	651,891
Adj. R^2	0.857	0.781	0.774	0.874	0.792	0.793
Excluding pilot districts						
1(post-adoption)	0.014	0.024*	0.029**	0.018	0.023*	0.032**
	(0.016)	(0.014)	(0.013)	(0.015)	(0.014)	(0.013)
1(2+ yrs. pre-adoption)	0.009	-.002	-.011	0.015	0.011	0.001
	(0.017)	(0.018)	(0.018)	(0.016)	(0.016)	(0.017)
1(post failed application)			0.007			-.026
			(0.017)			(0.031)
1(2+ yrs. pre-failed app.)			0.014			-.002
			(0.016)			(0.015)
Districts	60	80	345	55	71	257
Students	240,651	281,890	664,330	86,884	95,530	244,537
Student-years	672,869	790,612	1,956,707	222,193	246,083	644,725
Adj. R^2	0.849	0.766	0.768	0.874	0.792	0.793

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%. Specification includes year indicators, student fixed effects, peer covariates, and indicator for having dropped Q-Comp as in Table 3. Outcomes are in standard deviation units.

Table 6: Robustness and further falsification of effect on math achievement

Outcome Test:	MCA			NWEA		
Sample:	Adopters Only	Appliers Only	All	Adopters Only	Appliers Only	All
Specification:	B	B	B	B	B	B
1(post-adoption)	0.006	0.004	-0.004	0.041	0.042*	0.037*
	(0.018)	(0.017)	(0.017)	(0.027)	(0.024)	(0.019)
1(2+ yrs. pre-adoption)	-0.025*	-0.030**	-0.027	-0.019	-0.014	-0.012
	(0.013)	(0.014)	(0.019)	(0.029)	(0.028)	(0.029)
1(post failed application)			0.002			-0.006
			(0.017)			(0.041)
1(2+ yrs. pre-failed app.)			-0.016			-0.004
			(0.019)			(0.015)
Districts	81	102	369	66	83	273
Students	251,600	310,023	686,483	88,387	97,140	247,767
Student-years	690,991	858,302	2,007,029	226,732	250,294	655,341
Adj. R^2	0.872	0.803	0.793	0.902	0.901	0.899
Excluding pilot districts						
1(post-adoption)	0.007	0.007	0.0008	0.041	0.04*	0.035*
	(0.019)	(0.019)	(0.018)	(0.027)	(0.024)	(0.02)
1(2+ yrs. pre-adoption)	-0.024*	-.027*	-.025	-0.017	-.011	-.006
	(0.014)	(0.015)	(0.019)	(0.029)	(0.027)	(0.029)
1(post failed application)			0.003			0.016
			(0.019)			(0.031)
1(2+ yrs. pre-failed app.)			-.024			-.006
			(0.033)			(0.017)
Districts	60	80	345	55	71	257
Students	236399	277,199	655,211	87,254	95,864	245,270
Student-years	659,824	775,176	1,918,192	223,533	246,918	648,187
Adj. R^2	0.867	0.793	0.789	0.902	0.839	0.838

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%. Specification includes year indicators, student fixed effects, peer covariates, and indicator for having dropped Q-Comp as in Table 3. Outcomes are in standard deviation units.

Table 7: Is effect on reading explained by changes in peers, expenditures, or teacher turnover?

Outcome Test:	MCA				NWEA			
1(post-adoption)	0.03*	0.032**	0.032**	0.031*	0.034**	0.032**	0.032**	0.031**
	(0.016)	(0.016)	(0.016)	(0.016)	(0.015)	(0.016)	(0.016)	(0.016)
Student context								
Total students		-.004	-.004	0.002		0.005	0.005	0.006
		(0.006)	(0.006)	(0.006)		(0.008)	(0.008)	(0.007)
Share male		-.086***	-.083***	-.076**		-.053	-.053	-.044
		(0.032)	(0.032)	(0.031)		(0.052)	(0.052)	(0.051)
Share free lunch		0.124***	0.128***	0.129***		0.03	0.035	0.037
		(0.036)	(0.036)	(0.036)		(0.047)	(0.048)	(0.047)
Share spec. ed.		-.049	-.049	-.060		0.132*	0.129*	0.115*
		(0.047)	(0.047)	(0.046)		(0.068)	(0.067)	(0.066)
Share Afr.-Am.		-.113***	-.128***	-.093**		0.013	-.006	0.017
		(0.041)	(0.041)	(0.04)		(0.093)	(0.094)	(0.093)
Share Hispanic		-.113**	-.122**	-.090*		-.104	-.111	-.088
		(0.05)	(0.051)	(0.05)		(0.109)	(0.109)	(0.103)
Share Asian-Am.		0.045	0.02	0.033		0.072	0.053	0.068
		(0.045)	(0.053)	(0.049)		(0.157)	(0.158)	(0.147)
Share Native Am.		-.102**	-.105**	-.073*		0.033	0.032	0.054
		(0.044)	(0.045)	(0.043)		(0.064)	(0.064)	(0.061)
Finance								
Log(expenditures)			0.01*	0.008			0.01	0.008
			(0.006)	(0.005)			(0.006)	(0.007)
Teachers								
% Novices				-.001				-.0008
				(0.001)				(0.001)
% New hires				-.003***				-.001
				(0.0009)				(0.001)
% to ever-QComp				-.001				0.001
				(0.002)				(0.002)
% Retained				0.0006				0.0006
				(0.0008)				(0.001)
Share $\leq$ 5 yrs. exper.				-.043				-.057
				(0.037)				(0.049)
Share $\geq$ 16 yrs. exper.				0.123***				0.061
				(0.024)				(0.038)
% with MA+				-.0007				-.00003
				(0.0009)				(0.001)
Adj. R^2	0.768	0.768	0.768	0.768	0.793	0.793	0.793	0.793

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%. These are variations on specification (A). All include year indicators, student fixed effects, and indicator for having dropped Q-Comp. Outcomes are in standard deviation units. The samples for each MCA (NWEA) model includes 1,950,163 (650,300) student-years from 662,549 (246,945) students in 345 (255) districts.

Table 8: Is effect on math explained by changes in peers, expenditures, or teacher turnover?

Outcome Test:	MCA				NWEA			
1(post-adoption)	0.002	0.007	0.006	0.005	0.035	0.038	0.038	0.037
	(0.023)	(0.023)	(0.023)	(0.024)	(0.026)	(0.027)	(0.026)	(0.026)
Student context								
Total students		-.011	-.012	-.004		-.010	-.010	-.011
		(0.008)	(0.008)	(0.008)		(0.009)	(0.01)	(0.011)
Share male		-.131***	-.128***	-.121***		-.121**	-.121**	-.114*
		(0.042)	(0.042)	(0.041)		(0.061)	(0.061)	(0.061)
Share free lunch		0.114***	0.12***	0.118***		0.047	0.051	0.057
		(0.039)	(0.038)	(0.038)		(0.066)	(0.067)	(0.067)
Share spec. ed.		-.184**	-.184***	-.194***		0.1	0.097	0.089
		(0.072)	(0.071)	(0.071)		(0.081)	(0.081)	(0.081)
Share Afr.-Am.		-.084*	-.105**	-.076*		-.149	-.163	-.163
		(0.043)	(0.044)	(0.041)		(0.121)	(0.122)	(0.123)
Share Hispanic		-.067	-.080	-.051		-.085	-.091	-.088
		(0.061)	(0.061)	(0.06)		(0.11)	(0.109)	(0.105)
Share Asian-Am.		0.122**	0.088	0.101*		0.013	-.002	0.0002
		(0.052)	(0.06)	(0.056)		(0.179)	(0.181)	(0.173)
Share Native Am.		-.044	-.047	-.020		0.021	0.02	0.034
		(0.055)	(0.055)	(0.055)		(0.094)	(0.094)	(0.092)
Finance								
Log(expenditures)			0.013**	0.014*			0.008	0.006
			(0.007)	(0.007)			(0.007)	(0.008)
Teachers								
% Novices				-.002				-.002
				(0.002)				(0.002)
% New hires				-.0009				-.002
				(0.001)				(0.001)
% to ever-QComp				-.0009				0.0006
				(0.002)				(0.003)
% Retained				-.0003				0.001
				(0.001)				(0.001)
Share $\leq$ 5 yrs. exper.				-.030				0.041
				(0.048)				(0.071)
Share $\geq$ 16 yrs. exper.				0.11***				0.068
				(0.031)				(0.049)
% with MA+				-.001				0.0002
				(0.001)				(0.001)
Adj. R^2	0.788	0.788	0.788	0.788	0.838	0.838	0.838	0.838

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%. These are variations on specification (A). All include year indicators, student fixed effects, and indicator for having dropped Q-Comp. Outcomes are in standard deviation units. The samples for each MCA (NWEA) model includes 1,912,021 (653,730) student-years from 653,475 (246,769) students in 345 (255) districts.

Table 9: Do effects differ by teacher experience levels?

Outcome Subject:	Reading		Math	
Outcome Test:	MCA	NWEA	MCA	NWEA
1(post-adoption)x(Share tchrs. 1-5 yrs. exp.)	0.058	0.004	0.099	0.029
	(0.068)	(0.087)	(0.098)	(0.130)
1(post-adoption)	0.058*	0.057	-0.005	0.041
	(0.032)	(0.039)	(0.046)	(0.053)
1(post-adoption)x(Share tchrs. 16+ yrs. exp.)	-0.101**	-0.069	-0.028	-0.025
	(0.048)	(0.051)	(0.080)	(0.080)
Share teachers 1-5 yrs. exp.	-0.039	-0.061	-0.052	0.026
	(0.033)	(0.048)	(0.044)	(0.069)
Share school's teachers with 6-15 years experience omitted				
Share teachers with 16+ yrs. exp.	0.129***	0.071*	0.108***	0.073
	(0.025)	(0.039)	(0.032)	(0.050)
Districts	369	273	369	273
Students	696,843	247,015	686,336	247,756
Student-years	2,051,570	651,878	2,006,113	655,326
Adj. R^2	0.851	0.872	0.863	0.899
Implied estimate of 1(post-adoption) effect for:				
Teachers with 1-5 yrs. exp.	0.116**	0.060	0.094	0.070
	(0.054)	(0.061)	(0.071)	(0.100)
Teachers with 6-15 yrs. exp.	0.058*	0.057	-0.005	0.041
	(0.032)	(0.039)	(0.046)	(0.053)
Teachers with 16+ yrs. exp.	-0.043	-0.012	-0.033	0.016
	(0.027)	(0.029)	(0.054)	(0.052)

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%. These are variations on spec. (A) in Table 3. All include year indicators, student fixed effects, peer covariates, and indicator for having dropped Q-Comp. Outcomes are in standard deviation units.

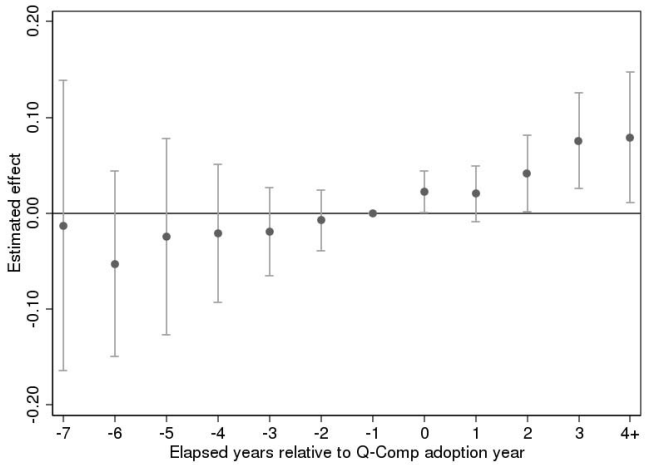
Online appendix for

Teacher Pay Reform and Productivity

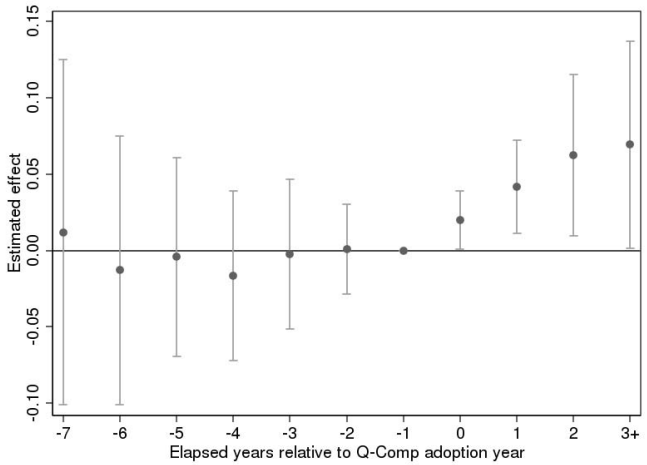
by Sojourner, Mykerezi & West

September 26, 2013

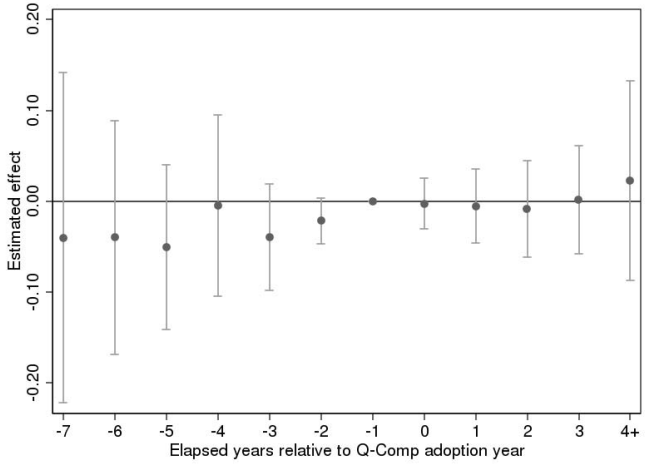
Appendix A Additional results



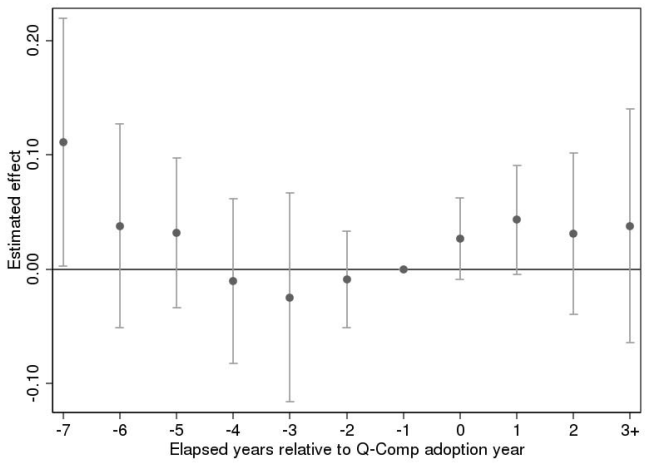
(a) MCA reading



(b) NWEA reading



(c) MCA math



(d) NWEA math

Figure 1: Effects of P4P-centered HRM reform on reading and math achievement as measured by the MCA and NWEA tests by years elapsed to adoption. These are point estimates and 95% confidence intervals from estimates in Table 4. All outcomes are in standard deviation units.

Table A-1: Hazard model of district Q-Comp adoption

DV: District adopts Q-Comp next year		
	Estimates	
Predictors	Hazard Ratio(SE)	
Change in average MCA math	1.045	(0.275)
Change in average MCA reading	1.079	(0.327)
Average MCA math level	1.067	(0.247)
Average MCA reading level	1.035	(0.256)
Students enrolled, thousands	1.001	(0.004)
Pct. free lunch	0.928	(0.584)
Pct. reduced-price lunch	0.698	(0.734)
Pct. special education	0.089*	(0.114)
Pct. male	6.640	(12.429)
Pct. African-American	2.086	(1.162)
Pct. Hispanic	1.144	(0.617)
Pct. Asian-American	1.201	(0.742)
Pct. Native American and other	0.782	(0.913)
Average teacher experience	1.009	(0.014)
Pct. teachers with MA+	1.007	(0.005)
Nine region dummies	Yes	
Districts	369	
District years	1108	
Log-L	-188.58	

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%. Estimates from Cox proportional hazard model predicting Q-Comp adoption in year $t+1$ based on district's characteristics in t for each t from 2003 to 2009.

Table A-2: Performance by presence of high-stakes goal by school-grade's test

Outcome test:	MCA		NWEA	
Specification:	A	C	A	C
	Outcome subject: Reading			
1(post-adoption)	0.031**	0.033**	0.032**	0.031*
	(0.015)	(0.016)	(0.017)	(0.021)
1(test is high-stakes)*1(post)		-.004		-.004
		(0.02)		(0.021)
Districts	369	369	273	273
Students	696,969	696,969	247,026	247,026
Student-years	2,052,337	2,052,337	651,891	651,891
Adj. R^2	0.774	0.774	0.793	0.793
	Outcome subject: Math			
1(post-adoption)	0.004	0.025	0.038	0.038
	(0.021)	(0.02)	(0.026)	(0.034)
1(test is high-stakes)*1(post)		-.038		-.008
		(0.028)		(0.056)
Districts	369	369	273	273
Students	686,483	686,483	247,767	247,767
Student-years	2,007,029	2,007,029	655,341	655,341
Adj. R^2	0.792	0.792	0.838	0.838

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.
Specification includes year indicators, student fixed effects, peer covariates, and indicator for having dropped Q-Comp as in Table 3. Outcomes are in standard deviation units.

Table A-3: Effect of Q-comp participation on districts' teacher workforce flows and experience shares by years elapsed from Q-Comp adoption

	Novice %	New Hire %	Flow to Ever %	Retention %	Share $\geq$ 16 yrs.	Share $\leq$ 5 yrs.	% MA+
6+ years prior	0.6	0.637	-1.190*	-.303	-0.012	-0.001	0.023
	(0.806)	(0.952)	(0.651)	(0.942)	(0.022)	(0.023)	(0.031)
5th year prior	0.1	-.038	-.867	0.466	-0.010	0.010	-0.002
	(0.625)	(0.73)	(0.847)	(0.75)	(0.017)	(0.017)	(0.029)
4th year prior	-.193	-.294	-.874	0.435	0.003	0.004	0.013
	(0.564)	(0.672)	(0.61)	(0.678)	(0.017)	(0.020)	(0.018)
3rd year prior	-.396	-.660	-.543	0.63	0.016**	-0.007	0.010
	(0.485)	(0.581)	(0.432)	(0.582)	(0.008)	(0.009)	(0.011)
2nd year prior	-.440	-.605	-.652*	0.53	0.010	-0.005	0.009
	(0.327)	(0.4)	(0.367)	(0.541)	(0.006)	(0.007)	(0.008)
1st year prior omitted							
1st year post-adoption	-.221	-.448	0.44	1.194***	-0.001	0.002	0.006
	(0.486)	(0.584)	(0.357)	(0.452)	(0.006)	(0.007)	(0.009)
2nd year post-adoption	-.061	-.173	0.178	0.912*	-0.003	0.012	0.003
	(0.518)	(0.646)	(0.4)	(0.534)	(0.010)	(0.010)	(0.010)
3rd year post-adoption	-.719	-.938	0.393	1.314*	0.007	-0.013	0.012
	(0.46)	(0.594)	(0.489)	(0.707)	(0.014)	(0.012)	(0.015)
4+ years post-adoption	-.493	-.653	-.096	0.797	0.015	-0.002	0.014
	(0.518)	(0.578)	(0.415)	(0.679)	(0.016)	(0.016)	(0.015)
Districts	346	346	346	346	369	369	369
Unit-years†	2,379	2,379	2,379	2,379	9,573	9,573	9,573
Adj. R^2	0.008	0.018	0.015	0.014	0.782	0.589	0.848

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%. †Unit is district for teacher flow measures (Columns 1-4) and school for teacher experience shares and education (Columns 5-7). Specification also includes indicators for dropped-Q-Comp, year, and unit.

Appendix B Sample Design

The sample includes public operating elementary and secondary independent districts, special school districts, and intermediate school districts. We exclude charter schools because they may have had performance pay in place prior to Q-Comp adoption, in which case our Q-Comp adoption variable would mismeasure changes in P4P regime. No other type of district (e.g. special education, vocational, integration, or telecommunication) ever applied for Q-Comp, so we exclude these other types. We focus on grades between 3 and 8 primarily because the state never mandated testing in lower grades and it mandates much less extensive testing in higher grades, making student achievement outcomes far less available outside grades 3 to 8.

We focus on districts that adopt through 2009 since outcomes data for 2010 are not available. Summary statistics in Tables 1 include adopters through the 2009 cohort. Another 14 districts adopt Q-Comp in 2010 and there was one other applier that did not adopt. For the 2010 adopters, the years 2003-2008 are indicated as years 2 or more pre-adoption.

Generally, each district decides whether or not its schools as a group will participate in Q-Comp and the district submits a single state application offering a common, district-wide P4P teacher contract. Therefore, we consider Q-Comp adoption and P4P design as a district-level variable. However, in a handful of cases, individual schools within districts adopted Q-Comp using school-specific P4P applications and designs. Because they are exercising "district"-like authority and because the variation in the timing of adoption within district can help identify Q-Comp effects, each of these sub-district adoption cohorts is coded as a separate synthetic district.

Appendix C Construction of teacher workforce flows

To measure teacher workforce flows, we draw on M.D.E. data describing the population of Minnesota public school teachers between 2002 and 2009. Each teacher-district-year match is a record.[27] This panel allows study of changes in the flows of teachers into each district's workforce, across districts, and into retirement. To develop evidence on mechanisms driving potential changes in productivity, we build and analyze four personnel flow rates defined for each district-year for the years 2003 to 2009. These are chosen to illuminate whether any effects of Q-Comp operate through triggering new sorting patterns of teachers to districts.

Constructing flow rates requires defining types of teacher movements. Let T_{dt} be the number of teachers in district-d in year-t. For each district-year, the number of new hires ($HIRES_{dt}$) is the number of teachers working in district-d in year-t who are either novice teachers entering the Minnesota teacher workforce panel for the first time ($NOVICES_{dt}$) or transferring teachers who have their most recent experience in a different Minnesota district. Incumbent teachers can exit by either transferring to another district or retiring from the profession (disappearing from the panel). The relevant law of motion for each district is that the number of teachers in the current year must equal the number of teachers from the previous year minus exits plus new hires, $T_{dt} = T_{dt-1} - EXITS_{dt} + HIRES_{dt}$.

From this, the retention rate for each district-year is measured. In personnel management, retention rates describe the percentage of the total number of workers that a firm employed over a given period of time who are retained in the firm's employment at the end of the period (The INFOHRM Group 2006). In the current setting, the retention rate is the district's number of teachers in the current year over the total number of unique teachers who worked for the district in either the previous or current year ($100 \cdot \frac{T_{dt}}{T_{dt-1}+HIRES_{dt}}$). For example, consider a district with 5 teachers in year $t-1$ and 6 in year t. If the district had two teachers exit and three new hires, its retention rate would be $75 = 100 \cdot \frac{6}{5+3}$. If instead the change from 5 to 6 occurred simply by adding an external hire, the retention rate would be $100 = 100 \cdot \frac{6}{5+1}$. This retention rate reflects personnel decisions made primarily over the summer leading up to the academic year starting in t.

According to economic theory and as illustrated vividly by Lazear (2000), adoption of P4P can increase firm productivity by causing low-productivity incumbents to exit at higher rates than before and by increasing the flow of high-productivity hires into the firm. Either of these kinds of changes would lower a district's retention rate. On the other hand, it may also boost retention of high productivity incumbents, which would offset this. To unpack these issues further, we study changes in the percentage of a district's teachers who are external hires ($100 \cdot \frac{HIRES_{dt}}{T_{dt}}$) and directly measure a component of this, the percentage who are novices in the Minnesota teacher workforce ($100 \cdot \frac{NOVICES_{dt}}{T_{dt}}$). If P4P attracts higher-ability candidates to the profession, this rate might increase.

A subtle issue arises given that we are analyzing flow rates in a difference-in-difference framework using the whole population of teachers. P4P adoption might operate by increasing the rate of swaps, where high-productivity teachers move from non-P4P districts to newly-

[27] A few full-time teachers work in multiple districts in a given year. We use the single record with the highest base salary to create a panel where each teacher is matched to exactly one district in a given year.

P4P districts and low-productivity teachers move the opposite way, from newly-P4P districts to non-P4P districts. This would increase productivity in adopting districts and decrease it in non-adopting districts, yielding a positive difference-in-difference estimated effect of adoption on student achievement. However, difference-in-difference analysis of retention rates and new-hire rates would miss this channel completely. Retention rates would fall in both kinds of districts, so the second difference would erase the effect. New hire rates would not change in either type of district.

To address this possibility, we define a novel kind of transfer rate measuring the flow towards ever-adopting districts so that we can look for changes from the baseline rate of swapping between districts. For each district that ever adopts Q-Comp and in each year, we measure the number of teachers who came in that year as transfers from districts that never adopt Q-Comp (F_{dt}^N) as a percentage of the district's teachers: $\%F_{dt}^N = 100 \cdot \frac{F_{dt}^N}{T_{dt}}$. Analogously, for each district that never adopts Q-Comp and in each year, we measure the number of transfers to that district from districts that ever adopt Q-Comp (F_{dt}^E) as a percentage of the never-adopting district's teachers: $\%F_{dt}^E = 100 \cdot \frac{F_{dt}^E}{T_{dt}}$. To avoid the second differences wiping these out, we reverse the sign on F_{dt}^E. Together these define a new rate, $\%F_{dt}$, which equals $\%F_{dt}^N$ for ever-adopting districts and $-\%F_{dt}^E$ for never adopters. $\%F_{dt}$ measures the percentage of a district's current workforce that flowed from never-adopters to ever-adopters, where those flowing the opposite direction have a negative sign. If P4P adoption triggers an increase (decrease) in the rate of personnel swaps between ever- and never-adopting districts, difference-in-difference analysis will identify a positive (negative) effect of Q-Comp adoption on $\%F_{dt}$.

Appendix D Preliminary versions of paper

This section aims to clarify the evolution of the project, what has changed in the design and the results, and why. There were 2 prior versions, which were similar to one another but quite different from the current version. The earlier versions used publicly-available data on average student MCA achievement and characteristics measured at the school-grade-year level. The current version instead uses two new, separate individual student-level data sets and it also incorporates data on teacher turnover for the first time.

Version 1 used school-grade-year average achievement data for the academic years 2005-2009. Without the ability to follow individual students over time, it was based on models specifying school-grade fixed effects. For each subject-grade-year, achievement outcomes were divided by the standard deviation of mean achievement across schools, as the standard deviation of student-level achievement was not available. This version was presented at the National Bureau of Economic Research in May 2011 and submitted to the Quarterly Journal of Economics. Comments suggested (A) adding prior years of data since the program started in 2005 and this would allow for more robust testing for differential pre-adoption trends and (B) use of student-level data that would deal with unobserved individual student differences and rule out many sorting stories.

Version 2 implemented suggestion (A). Prior to 2005, only grades 3, 5, and 7 were required to implement MCA tests. Achievement outcomes for grades 4, 6, and 8 are not available. In this version, we used two samples: the all-grades sample (all grades 3-8 but only in years 2005-2009, same as version 1) and the all-years sample (all years 2003-2009 but only for grades 3, 5, and 7). Again, all analysis used school-grade fixed effects. This version was presented at the Association for Public Policy Analysis and Management and at the Minnesota Economics Association meetings in late 2011.

The current version represents a large improvement in data and design over earlier versions. For the first time, we secured student-level panel data (suggestion B) and the data cover 2003-2009 (suggestion A). In fact, we secured two separate student-achievement panels (MCA and NWEA) as well as a panel tracking teachers' movements into, out of, and between districts each year. Given this new data structure, we shift to using student fixed effects, rather than school-grade fixed effects. That is, we identify treatment effects off of variation in achievement *within-student* over time adjusting for differences in peers, rather than off of variation in achievement *within-school-grade* over time adjusting for school-grade average student demographics as in earlier versions. This increases precision and removes many sources of potential omitted-variable bias.[28] The additional years of pre-achievement data allow stronger falsification tests than previously possible.

The current version also drops charter schools from the sample. While we are confident assuming that regular districts have not offered P4P outside of Q-Comp, charter schools may have. Results including them are very similar and can be provided on request.

[28]The units of MCA achievement measurement change because now we standardize by the standard deviation of student-level achievement in a given subject-grade-year, rather than the standard deviation of school mean achievement in a given subject-grade-year as we did in earlier versions.