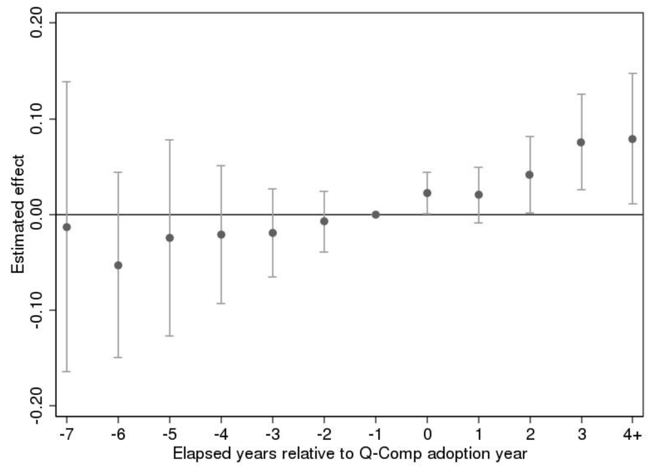
Online appendix for

Teacher Pay Reform and Productivity

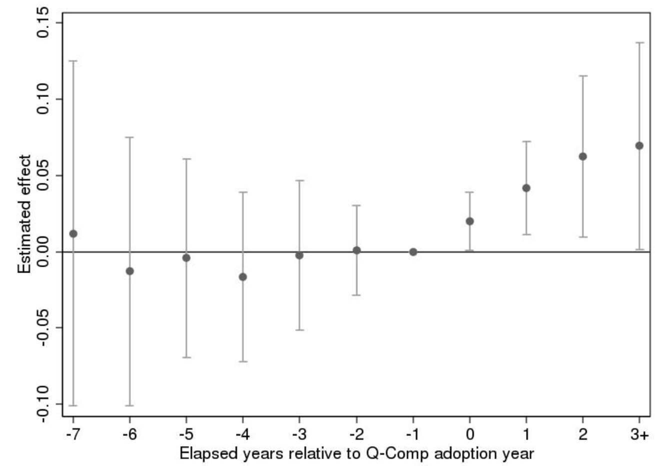
by Sojourner, Mykerezi & West

September 26, 2013

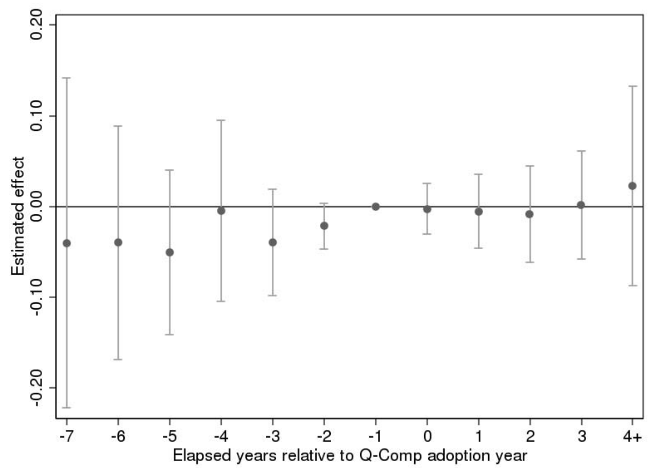
Appendix A Additional results



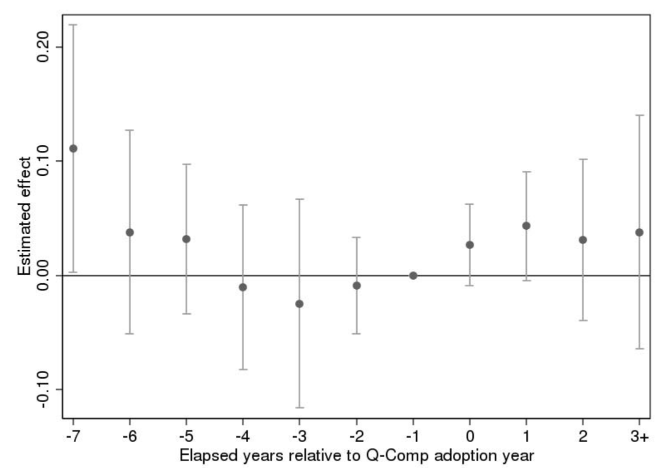
(a) MCA reading



(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.