Reforming Teacher Contracts: A Look at the Impact of Q Comp on Student Achievement in Minnesota

By Elton Mykerezi, Aaron Sojourner, and Kristine West

Photo © Steve Schneider, 2007.

Adoption of Minnesota's pay-for-performance teaching bonus program Q Comp leads to an additional month's worth of learning, on average, in reading scores.

Abstract: *Until recently, teacher employment and pay across the United States was strictly determined by education and experience, even though research shows that teachers with similar credentials and experience vary widely in their ability to influence student outcomes. As a result, there has been a surge in interest nationally in "pay-for-performance" contracts that tie pay to various measures of performance. Critics of pay-for-performance, often including teacher unions, raise concerns about incentives to narrow curricula (aka "teaching to the test") and adverse effects on teacher collaboration, among others.*

Minnesota's Q Comp is one of the nation's largest and longest-lasting pay-for-performance programs; it delivers reform via a voluntary "grantor-grantee" format that ensures political feasibility. Districts design alternative contracts with their unions, then propose them to the state for extra funding. We studied the effect that adoption of Q Comp in Minnesota districts had on student achievement growth and found program adoption leads to an additional month's worth of learning, on average, in reading scores. Similar gains were observed in math scores, but with less statistical precision. We found no notable evidence of student or teacher movements in response to the program, nor any significant evidence of "teaching to the test." We concluded that the program increased test scores by providing incumbent teachers with the incentives and tools to adjust their practices in ways that increase test scores. Gains were obtained at relatively low cost, so the program is cost effective. However, the size of the gain was not sufficiently large for such voluntary pay-for-performance to be relied on as the only tool for improving education quality. A well-designed "grantor-grantee" type pay-for-performance program can be a valuable tool in policymaker's arsenals, for improving education quality.

In the United States, teacher contracts have long mandated that compensation and other personnel decisions be either strictly or heavily based on teacher education and experience. This "steps and lanes" approach references the two-dimensional salary schedule at the heart of most teacher contracts. Recent research highlights that teacher effectiveness (typically measured through student outcomes) varies substantially even within cohorts of teachers of similar education and experience. This finding has led to a surge in the design and implementation of policies that tie teacher compensation and other personnel decisions to factors other than education and experience.

Minnesota's Q Comp is one of the nation's largest and longest-lasting programs to, in part, implement departures from traditional teacher contracts. We studied the performance of the program with the hope of understanding whether or not the reforms implemented through Q Comp helped increase student test scores. (Note: We studied the effects on test scores because they are readily available, quantitative, and have unambiguously desirable outcomes. We do not mean to suggest that standardized test scores are the *only* relevant output of an educational system.)

Background and Significance

Minnesota has a rich history of educational reform, and efforts to introduce changes to teacher contracts predate Q Comp. Significant state-wide changes started to appear in the early 2000s. Legislative change in 2002 made it possible for districts, in collaboration with teacher unions, to design pilot compensation programs that augmented the steps and lanes approach and made use of alternative measures of performance. Importantly, these reforms were implemented through a "grantor-grantee" mechanism. The grantor (in this case, the state) set parameters for what is considered acceptable reform, and the grantees (the districts) designed plans that fit within the scope of the program in exchange for extra funds. Five districts applied to be part of this early reform effort, and in 2003 they started participating in the program, which offered an annual increase of $150 per student in funding. 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).

Q Comp became law in 2005 with bipartisan support. It maintains the grantor-grantee structure of the 2003 pilot and some of the definitions of "performance," but it also draws heavily on TAP for defining what reforms are acceptable and how to measure performance. Under Q Comp, districts are invited to design reforms that satisfy certain criteria in exchange for $260 in additional funding per student each year. Because the criteria that Q Comp–eligible contracts need to satisfy were, to a large extent, inspired by those used in TAP, sites that were participating in TAP were automatically enrolled into Q Comp, while the five districts participating in the 2003 pilot were given the option to modify their contracts to meet Q Comp requirements or to continue with their existing programs and receive $150 per year, until 2009 (instead of $260).

Once Q Comp was taken to scale, traditional public school districts and charter schools state-wide began applying to the program. In its first official year, 2005, nine districts applied to Q Comp; all adopted. In 2006, 38 more districts applied, of which 29 adopted and 9 failed to adopt. Of all 361 districts in the state 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.

Q Comp has several attributes that make it an interesting policy experiment. First, the program is voluntary (districts chose to apply) and somewhat selective (the state could choose not to fund proposals). Q Comp was adopted by a sizable number of districts, over multiple years, while some districts were rejected. This structure allows us to separate the impact of the program itself from other possible explanations (generally a difficult thing to achieve without an experimental design). For instance, consider a case where all districts adopt in the same year. It would be impossible to determine if any change in test scores was because of the program or because of something else that happened in the same year to encourage applications to Q Comp. Also, because some districts that wanted to join the program were unable to do so, we could rule out the possibility that it is the intent of an ambitious administration to "shake things up" that matters and not the reforms that come with Q Comp.

Second, the program is large and was implemented as a permanent (or at least long-lasting) regime. This is important, because it may take time for teachers to adjust to the new incentives and take advantage of the professional development opportunities. Experiments are generally considered superior to secondary data analysis for dealing with confounding factors in program evaluation, but because they are short-lived, they trail large policy experiments such as Q Comp in terms of external validity and generalizability.

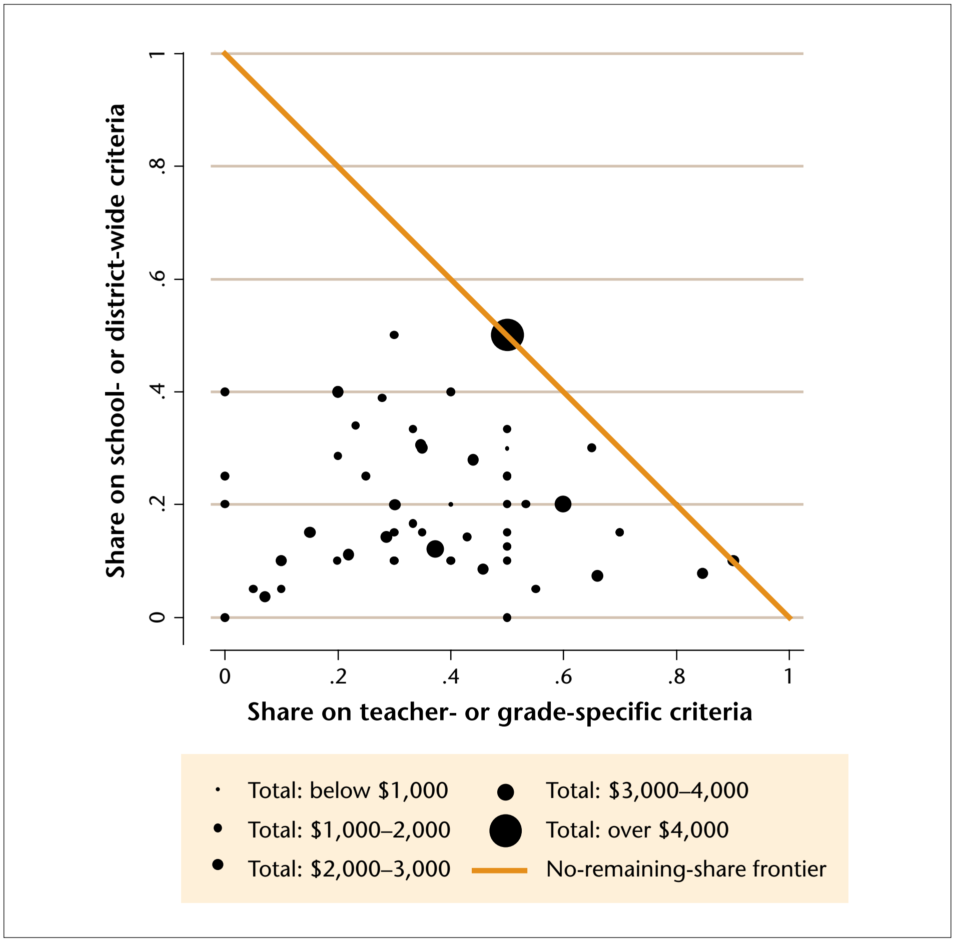
Also, as noted, the program follows a grantor-grantee structure, where the district (grantee) first comes up with a proposal that has the support of the administration and teachers (represented by the union) and then proposes it to the Minnesota Department of Education (grantor), which, in turn, ensures that the proposal is within required reform parameters. This structure ensures that reform is both feasible and that it adheres to the intent of the legislation that made it possible in the first place. It is perhaps due to these desirable features that this grantor-grantee format is now used by large federal programs, such as the $1.6-billion Teacher Incentive Fund and the $4.4-billion Race to the Top program. Q Comp presents the first opportunity to evaluate an instance of such an approach to teacher labor market reform.

The state mandated that, in their applications, districts specify the bonuses each teacher is eligible to earn for three types of criteria:

1. the formal classroom-observation process;
2. school- or district-wide goals for student achievement usually on standardized tests;
3. 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.

For classroom observations, the state encouraged districts to use the Danielson Framework for Teaching, the most widely used evaluation rubric nationally, and to conduct at least three

Figure 1. Variety of Pay-for-performance Bonuses Offered by Q Comp Districts



Elissa Schloesser

observations per year using a trained evaluator and with pre- and post-observation conferences. Teachers are rated on measures of planning and preparation, classroom environment, professional responsibility, and instructional practice. Depending on the district, the evaluator is the principal or another administrator, a peer, or a hired consultant (e.g., senior teacher from other districts or retired teacher).

Bonuses for school- or district-wide goals tied to standardized tests would be based on specific objectives laid out in the initial Q Comp application, and/or updated in annual reports to the Minnesota Department of Education (MDE) prior to the start of each school year. These are awarded to all teachers in the school/district if the goals are met, and to no teachers otherwise.

Individual teacher or small team bonuses were typically linked to goals that were set and measured in the context of a complementary management practice, which MDE refers to as "job embedded professional development." Specifically, with the support of their administration, teachers form professional learning communities. Here, teachers collaborate on selecting performance targets that form the basis for Q Comp's individual or small-group 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.

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. Classroom observations had the most stakes tied to them. School-wide goals based on standardized achievement tests had the least.

There was a lot of variation around these averages. Figure 1 shows the wide variety of programs adopted; the size of the dot indicates the total amount each teacher was eligible to earn, the share tied to teacher or grade-level goals is along the *x* axis, and the share tied to school- or district-wide goals is along the *y* axis. The distance from the frontier measures the amount tied to the formal observation cycle. For example, the small dot at the origin represents a district that offered a bonus of between $1,000 and $2,000, none of which was tied to teacher or grade-level goals, or to school- or district-wide goals, thus all of it was tied to formal observations. Alternatively, the large dot in the middle of the frontier represents a district that offered a bonus of more than $4,000, half of which was tied to teacher or grade-level goals, half of which was tied to school- or district-wide goals, and none of which was tied to formal observations. The key point is that districts that adopted Q Comp came up with a large variety of contract designs.

Questions

We attempted to answer two questions:

1. **Did Q Comp adoption cause growth on student test scores that would not have been realized in its absence?** Increasing student achievement is the primary purpose of most recent reforms, so Q Comp's impact on the path of student test scores is a question of primary interest.

2. **What were the mechanisms that produced the gains?** Q Comp is a large change and could operate through several channels. We examined if changes in test scores were explained by:

   (a) adopting districts attracting better students,
   (b) adopting districts simply getting extra money,
   (c) adopting districts attracting different kinds of teachers,
   (d) incumbent teachers in adopting districts changing practices, or
   (e) assessment-specific gains (often indicative of what is referred to as "teaching to the test").

The answer to the second question is important because the different channels have very different implications. If adopters experienced better outcomes because the program attracted better students (option *a*), then there is no average gain across all students, just a reshuffling, with the potential for widening achievement gaps and little/no overall gain. Next, adopting districts got extra money, and that fact alone could bring about positive change (option *b*). If the additional dollars explained the effect, the implication would be that it is not necessary to design elaborate labor-market reform to achieve gains, just increase budgets.

CC/Flickr/Melanie Holtsman

Gains that districts had in standardized test scores after implementing Q Comp do not seem to be connected with teachers "teaching to the test."

Clearly, explanation *d* whereby the reform provides incumbent teachers with the tools and incentives needed to change their practices in a way that achieves higher test scores, is the state's best-case scenario. Explanation *c* might be equally desirable to the state, but only under certain circumstances. Explanation *c* works through differential selection of different types of workers into teaching positions. Presumably, individuals who are able to thrive in a high-stakes environment will be more likely to be drawn to teaching positions that have some performance stakes attached. Explanations *c* and *d* may be equally desirable from an overall policy perspective, depending on whether the "differential selection" is into the State's teaching force, or a mere reshuffling of existing teachers across the Q Comp and non Q Comp schools. To the central authority (MDE in this case), it may not matter if scores increased because of differential selection into the teaching force or if the reform gave incumbent teachers the tools and incentives to approach their jobs differently.

If explanation *e* is at play, in practice, one would expect the program to initially reshuffle existing teachers. However, with growing coverage and tenure of Q Comp, all differential selection would eventually have to occur into the teaching force. In other words, pay-for-performance regimes, such as Q Comp, would have to become an expected, non-trivial, and widespread component of teacher pay in order to be reflected in the occupational decisions of young college students.

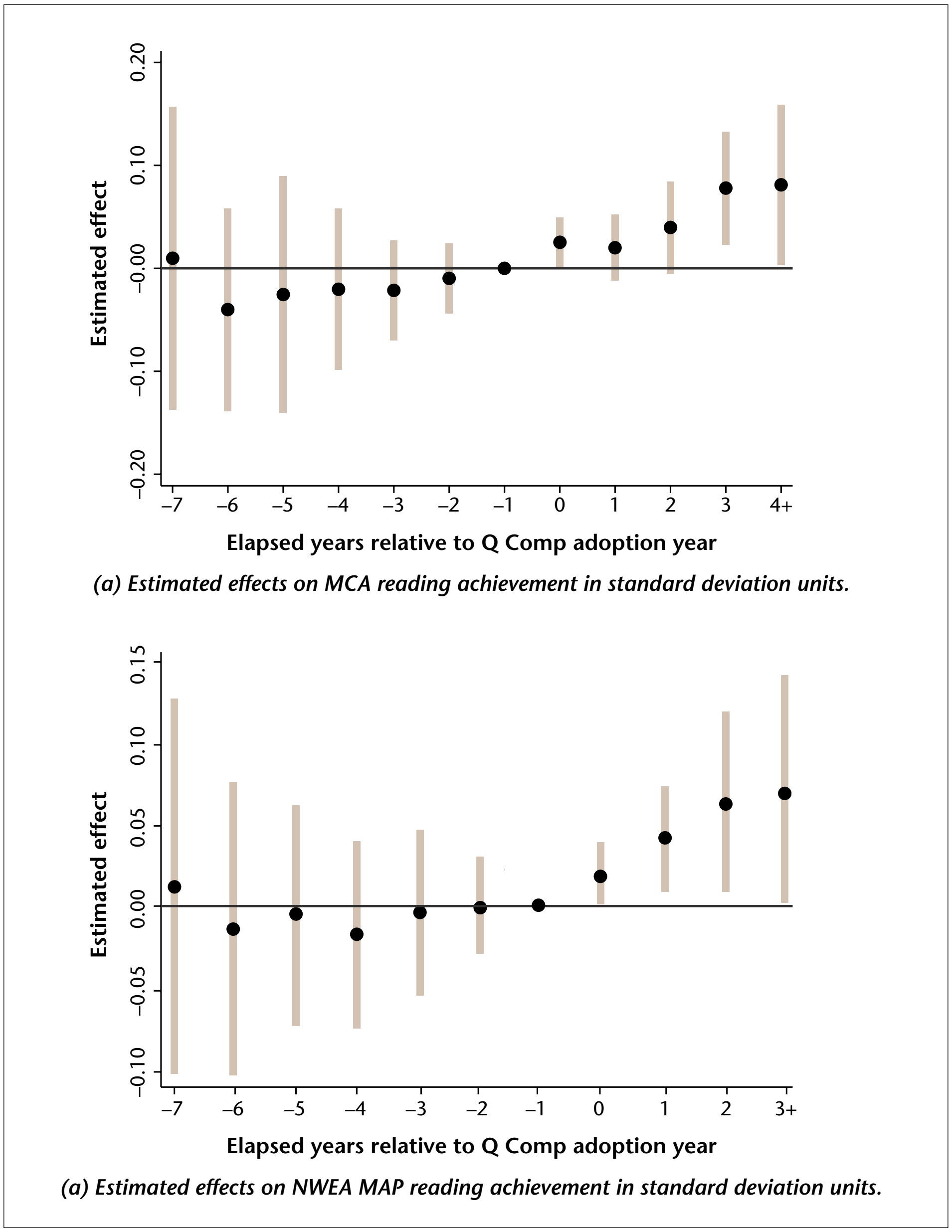
Explanation *e* underscores one of the most frequent objections to teacher labor reform that ties personnel decisions to student test scores. The idea is that, in order to increase test scores, teachers may either narrow the curriculum to focus heavily or explicitly on topics and skills that are directly tested and/or spend time teaching test-taking skills/techniques rather than substance. If the curriculum were narrowed, observed gains in test scores would represent an actual increase in knowledge, but at an unknown cost to other skill/knowledge that is not directly tested. This can, to some extent, be mediated by designing better tests. Effort put into merely improving test-taking skill rather than subject-matter knowledge, on the other hand, generates relatively unproductive test score gains.

Methods and Results

To answer each question, we used the test score history of all Minnesota students in grades 3 through 8 at any point between 2003 and 2010. This includes full student histories of Minnesota Comprehensive Assessments (MCA) reading and math scores, as well as full student histories of Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) when available.

The use of two separate assessments is important because they complement each other well. The MCA is used by all districts, but it is also high stakes for teachers and administrators in ways that are not related to Q Comp. Since this is the metric that is generally used for accountability to the state and under the No Child Left Behind Act, schools already may have strong incentives to raise MCA scores and it could be subject to "teaching to the test" and other less-than-productive practices (explanation *e*). MAP is available only in districts that purchase it, so it does not have the widespread coverage that the MCA does, but it is used only as a diagnostic tool, so it is inherently a low-stakes assessment and more reliable in that sense.

Figure 2. Differences Between Districts Adopting Q Comp and the State as a Whole by Year Relative to Adoption (Reading Scores)



(a) Estimated effects on MCA reading achievement in standard deviation units.

(a) Estimated effects on NWEA MAP reading achievement in standard deviation units.

Elissa Schloesser

To answer the first question of "Did Q Comp adoption cause growth on student test scores that would not have been realized in its absence?" we then compared the test score histories of students whose districts adopted Q Comp at some point after the student took his or her first test, to the histories of similar students but whose districts never adopted Q Comp, and then to the test score histories of students whose districts attempted to adopt Q Comp but were unable to do so.

One of the most important concerns with evaluation of programs like Q Comp, that are voluntary, is that those districts that chose Q Comp are not randomly chosen. They made a deliberate effort to adopt a novel labor reform system. Such districts that are willing to put effort into reform, and able to successfully collaborate with their teachers (represented by the union) to design and implement this reform have demonstrated inclinations and capabilities that may not be present in other districts. Adopting districts may also be taking other measures to improve their scores, and Q Comp may only be a part of their overall improvement plan. If this were the case, it could be easy to mistakenly attribute higher growth in these districts, that may be due to a variety of factors, to Q Comp alone.

To rule out this possibility, we compared data on Q Comp and non Q Comp districts before and after adoption of Q Comp. For our evidence to be reliably interpreted as a Q Comp effect, we would have to find that Q Comp districts do not start to look any different from all other districts in the state until *after* they adopt the program.

Figure 2 presents highly reliable evidence that Q Comp caused higher test score growth in districts that adopted it. To produce Figure 2, regression techniques have been used to re-align student test score data for all students in the state in such a way that sets the year-to-year growth of all non Q Comp districts to zero, thus representing it on the horizontal axis. Test scores for Q Comp and non Q Comp districts are also fixed to be the same (and set to zero) on the last year prior to adoption. Differences in test score growth patterns between Q Comp districts and the rest of the state are then represented by the dots for each year, before and after adoption of Q Comp. Year 0 is the adoption year, negative numbers indicate years prior to adoption, and positive numbers indicate years into the program.

There are three plausible scenarios related to the effects of Q Comp: (A) districts that chose Q Comp were similar to those that didn't, but Q Comp had no effect; (B) districts that chose Q Comp were already making other efforts to improve, so growth in these districts is really a continuation of pre-existing overall effort, not due to Q Comp per se. This scenario is often mistakenly interpreted as a program effect if one is not careful to look at relative growth in test scores before and after adoption; (C) districts that chose Q Comp were similar to those that didn't, but once they adopt they experience higher growth (most likely due to Q Comp). This scenario can be reliably interpreted as evidence of a Q Comp effect.

Under scenario A, the graph would show dots that overlap the horizontal axis throughout. Under scenario B the dots would form a line that approaches the horizontal axis from below, crosses it at "–1" and continues to grow at about the same rate in post-adoption years. Our actual results (Figure 2) show that before adopting Q Comp, those districts that end up adopting look no different than the rest of the state, on average. The lines around the dots in Figure 2 represent statistical bounds that reflect margins of error. Only in cases when the line does not cross the horizontal axis are we reasonably certain that Q Comp districts are performing differently than the rest of the state. Figure 2 shows that this is nearly always the case after adoption, and never the case before. We thus conclude that the program did make a difference in increasing reading scores in adopting districts. We found the average size of the effect to be about one month's worth of typical student learning.

We also found results that look relatively similar in Math, but the statistical margins of error were higher, so we are not able to interpret those as conclusive evidence of a Q Comp effect.

In terms of mechanisms, the evidence suggests that the program brought about reform that changed the practice of incumbent teachers. We use data on student demographics and expenditures by districts each year to rule out the possibility that the effect that we observe in Figure 2 is driven by movements of students in pursuit of, or away from Q Comp districts, or that the effect is simply explained by the fact that Q Comp comes with an additional $260 per pupil each year. We also include teacher flows between districts in our analysis to test whether Q Comp changed the composition of participating districts' teaching force in any way. We do not find evidence that Q Comp caused differential movements in teachers by experience or education levels, thus we conclude that teacher selection into Q Comp districts was not a factor in explaining the result.

Additionally, districts that use both assessments MCA and NWEA-MAP were able to choose to tie their incentives to either assessment in either subject (Math or Reading). We performed additional tests to see if districts experienced disproportionally high gains only on the incentivized test, rather than in the incentivized subject (across both assessments) and found no evidence of such. This implies that there is no evidence of less/unproductive practices, such as "teaching to the test." We concluded that Q Comp had its impact primarily by changing incumbent teachers' practices, resources and incentives.

How does Q Comp change teachers' actions? Q Comp introduced incentives to pay closer attention to measureable outcomes but also the capacity to do so in practice. Teachers were given more time to consult with peers within professional learning communities and increased feedback by both administrators (in high-stakes observations) and colleagues (in low-stakes observations). While it is apparent that the program changed teaching practices, it is less obvious which particular part of Q Comp was most important, or even if the multiple components of the overall reform package can work separately. Because the program encouraged the introduction of reform as a bundle, it was impossible to evaluate the relative merits of the different kinds of incentives and the various reform elements.

Q Comp's Five Components

1. ***Pay for performance:*** *The state encouraged districts to tie bonuses to individual/small-team goals, school- or district-wide goals, and each teacher's formal classroom observations. Districts varied in the amount allocated to each of these.*

2. ***Classroom observations:*** *The state encouraged districts to model their formal observations on Charlotte Danielson's Framework for Teaching. Observation cycles included pre- and postobservation conferences. All observations were done by trained evaluators, often a school administrator.*

3. ***Job-embedded professional development:*** *The state encouraged districts to reform professional development using the model of professional learning communities, which provide time to devote analysis of student data, peer observation, and coaching.*

4. ***Career ladders: The state encouraged districts to provide teachers with opportunities for taking different career roles and accelerating their advancement, through mentoring, involvement in program design and leadership roles.***

5. ***Alternative compensation:*** *The state asked each district to make a commitment to move away from basing pay increases on longevity and coursework. This component was often the most vague as details were left to future labor-management negotiations.*

We attempted to disentangle these components (and even the impact of the various bonuses within the pay-for-performance component). In the end, there was no clear story about which component, or combination of components, was most important. The results reported here treat Q Comp as a single reform that encompasses different combinations of the five components in each of the adopting districts.

Conclusions

Our study showed with high reliability that the implementation of Q Comp had a positive effect on student test scores, at least in reading. Two points about Q Comp are noteworthy:

The effect is relatively small, so one cannot rely on Q Comp as the sole tool for making substantial improvements in test scores or substantial reductions (or eliminations) in achievement gaps across student populations.

The program was highly cost-effective. To put this in perspective, the cost of educating one student in Minnesota over nine months is around $10,000. So, to obtain the growth caused by Q Comp by simply adding resources to schools (even if we were to ignore the practical problems of adding one month to the school year), the cost would still be more than $1,000. Q Comp gets the same result with $260. Thus, efforts that would reallocate Q Comp's revenue to general funds and do away with the program's structure are likely ineffective.

Overall, we believe that Q Comp is an important tool in education policy in Minnesota. One of the shortcomings of the study was that it could not pinpoint the relative importance of the different types of incentives and the various reform elements. However, because Q Comp's framework is sufficiently flexible and able to incorporate diverse reforms into contracts, in politically feasible ways, we suggest that it can be used to deliberately pilot candidate interventions for future study.

This study also sheds new light on various matters of importance to education policy in general. Designing labor market reform for teachers is a very complex task, from both a theoretical perspective and in terms of practical application and political feasibility. We believe that the "grantor-grantee" format used by Q Comp overcomes many difficulties, but does impose some limits. For instance, one of the most important objections to pay for performance based on student test scores is the possibility that teachers will teach to the test or, as has sometimes been the case (e.g., in documented incidents in Chicago and Atlanta), intentionally manipulate test scores. One possible method to mediate this would be to attach bonuses not only to test scores, but also to peer and/or principal observations of teachers' in-classroom practice. Observations, on the other hand, come with added subjectivity; the need for observers raises questions of reliability across observers. The optimal mix of objective and subjective components of evaluation is not obvious, and can vary based on setting.

Another issue relates to an inherent tradeoff between strength of incentives and collaborative work. Economists believe that tournament-style competitions for bonuses, where agents compete for set prizes, are able to deliver the strongest incentives. However, if collaboration among teachers is important to productivity, tournament-style incentives or individual bonuses could prove unproductive by limiting incentives to collaborate. A solution to this might be to offer group bonuses to meaningful teams of teachers. Group bonuses, however, introduce a "commons" problem, in that the likelihood of a bonus depends less on a teachers' own efforts, and more on those of others; the larger the group, the weaker the added incentive for each teacher. The optimal mix of individual and group bonuses, size of group, whether collaboration is meaningful within teachers of the same grade, subject or both are also questions that do not have easy answers. Teachers of the same subject can collaborate on techniques or content development, but teachers of the same group of students can synchronize curricula (e.g., to make sure that the math teacher covers content needed to make progress in science class).

The "grantor-grantee" model has a central authority (state or federal funder) that sets some boundaries on the types of reforms that are acceptable, but ultimately delegates the design details on the desired mix of stakes tied to standardized test scores, other student outcomes, or in-class observations of teacher practices to the district. The district also chooses the extent to which individual or group bonuses are awarded, and choses how to determine group sizes and settings. Because districts are likely to have far better information on their own settings than a central designer (at a state or federal agency), they might be able to make choices that work better for them. In Minnesota, for example, some of the bonuses were operationalized through the Professional Learning Community structure, which was part of professional development practice in many districts statewide. Local design also ensures political feasibility, which is not to be taken for granted for a centrally designed reform package.

Local design, however does have its drawbacks. Local authorities may be able to design incentives and set targets that are not as challenging as they would be under central design, making sure that more teachers get the bonus. This would weaken incentives. This is where the grantor can play a role, by setting reform parameters that make it less likely for local designers to propose weaker incentive schemes, and by overseeing implementation and updating of reform packages from the viewpoint of an advocate of a strong incentive system.

Our study showed that Q Comp gave Minnesota districts an opportunity to design reform packages and districts did, in fact, choose a wide variety of designs. With all things considered, the widespread adoption of the program and impact on test scores, shows that the policy design is capable of producing a positive cost-effective net effect on achievement.

Dr. Elton Mykerezi is an economist and associate professor in the Department of Applied Economics. He is also an Extension Economist with the Center for Community Vitality, University of Minnesota, Twin Cities. He completed his PhD in economics at Virginia Tech. Research interests include the study of human capital, causes of poverty, food insecurity and poor nutrition, the role of public assistance in enhancing household well-being, and rural business and rural labor markets. His extension and outreach program aims to improve economic opportunities for vulnerable populations by focusing on access to healthy foods, a quality education, and opportunities for entrepreneurship and employment.

Dr. Aaron Sojourner is an economist and assistant professor in the Department of Work and Organizations in the Carlson School of Management. He completed his PhD in economics at Northwestern University. He also has an MA in public policy analysis from the University of Chicago and a BA in history from Yale University. Research interests include sources of human capital, impacts of unions in the economic and political arenas, and behavioral consumer finance decisions. Complementing his research interest, Sojourner has a wide range of policy experience including service as a fellow with the U.S. Senate's Health, Education, Labor and Pensions Committee.

Dr. Kristine West is an economist and assistant professor in the Department of Economics at St. Catherine University. She completed her PhD in applied economics at the University of Minnesota. She also has a BA in economics from Macalester College. Research interests include the economics of education, quantitative program evaluation, and labor economics. Complementing her research interest, she taught social studies at Washburn High School in Minneapolis for seven years before earning her doctorate.

The research upon which this article is based was supported through a grant from CURA's New Initiatives program. These grants support projects that are initiated by faculty, community organizations, government agencies, or students and that fall outside CURA's existing program areas.