Putting Performance Pay to the Test: Effects of Denver ProComp on the Teacher Workforce and Student Outcomes

Allison Atteberry
allison.atteberry@colorado.edu
817-944-0899
University of Colorado-Boulder School of Education,
249 UCB, Boulder, CO 80309

Sarah E. LaCour
sarah.lacour@colorado.edu
University of Colorado School of Education;
249 UCB; Boulder, CO 80309

Updated Citation:
Atteberry, A., LaCour, S. (in press, 2020). "Testing the Denver ProComp Theory of Action: Evidence on Intended Mechanisms for Shaping the Teacher Workforce and Student Outcomes." *Teachers College Record*, Vol. 123(5)

Allison Atteberry, PhD, is an assistant professor of Research and Evaluation Methodology at the University of Colorado-Boulder School of Education. Her research focuses on policies and interventions that are intended to help provide effective teachers to the students who need them most.

Sarah E. LaCour is a Ph.D. candidate in Educational Foundations, Policy, & Practice at the University of Colorado Boulder. Her research primarily uses quasi-experimental designs to explore the impacts of law and larger educational policy on students and communities.

Summary:
In 2005-06, Denver became one of the first U.S. districts to implement a pay-for-performance compensation system, and Denver's ProComp is now the longest-running PFP policy in the country. We use a 16-year panel to analyze the effects of ProComp on both student and teacher outcomes, with a focus on the onset of its second instantiation, ProComp 2.0, which began in 2008-09.

Structured Abstract

Context

In 2005-06, Denver became one of the first U.S. districts to implement a pay-for-performance (PFP) compensation system, and Denver's ProComp is now the longest-running PFP policy in the country. The national proliferation of PFP systems in education has been controversial, with mixed evidence and competing narratives about its impacts. During Denver's 2019 strike, disagreements arose about whether 13 years of ProComp have helped or harmed efforts to retain effective teachers to improve student outcomes. This paper addresses this policy debate.

Research Questions

We use a 16-year panel to analyze the effects of ProComp on both student and teacher outcomes. We focus on the onset of the second version of the policy, ProComp 2.0, in 2008.

Intervention

The ProComp policy is a pay-for-performance teacher compensation system, which includes 10 distinct financial incentives, some of which are awarded schoolwide. Annual payouts represented 12% of base-pay, on average, among full-time teachers.

Research Design

We use comparative interrupted time series (CITS) to examine pre/post ProComp trends in outcomes in DPS relative to similar districts across the same period. When CITS is not possible, we conduct interrupted time series (ITS) analysis in DPS using a panel up to 5 years pre-PC1 and up to 10 years post-onset.

Results

ProComp may have had a positive effect on ELA, math, and writing achievement that was not evident in comparable districts with similar achievement trends prior to 2005. We also find descriptive evidence that more effective teachers were recruited to DPS once ProComp began and that the overall decline in teacher retention across districts in this time period was less precipitous among DPS' highly effective teachers during ProComp.

Conclusions

Our results can help reflect on which of the hypothesized mechanisms undergirding PFP policies find empirical support in the field. The onset of ProComp shifted the composition of the DPS teacher workforce through recruitment and retention of certain kinds of teachers. These results at first appear to contradict teacher perceptions that the program coincided with a dramatic decline in teacher retention and was thus ineffective. However, retention did, in fact, decline throughout the period. Yet DPS retention patterns were not that different from other comparable Colorado districts during this period. Thus, while teachers' perceptions of reduced teacher retention were accurate, it would be very difficult to see from within DPS that retention rates were not necessarily distinct from secular trends outside DPS.

Executive Summary

In recent years, there has been a notable spike in U.S. teacher strikes across a surprisingly diverse landscape of places—from West Virginia to Los Angeles. Among these high-profile events, the January 2019 Denver teacher strike stands out for its focus on differential compensation, which has a particularly long history in Denver Public Schools (DPS). In 2005-06, DPS became one of the first U.S. districts to implement a pay-for-performance (PFP) compensation system, and Denver's ProComp is now the longest-running PFP policy in the country.

The national proliferation of PFP systems in education has been controversial, with mixed evidence and competing narratives about its impacts. In January 2019, the local divide in stakeholders' views of ProComp came to a head, leading to a 3-day teacher strike. After a marathon 20 straight hours of deliberations, DPS and the union agreed on $23 million in teacher pay raises at the cost of primarily cutting about 150 jobs from the central office.

During the January strike, disagreements arose about whether 13 years of ProComp have helped or harmed DPS's efforts to retain effective teachers in service of improving student outcomes. Teachers described experiencing ever-increasing turnover in their schools during the decade after ProComp began, leading many to conclude that it did not improve retention. District leaders viewed rising retention since 2015 as evidence that its current system was having positive effects. Interestingly, both of these descriptive facts are *true*. Yet neither reveal ProComp's causal effect on achievement or retention.

This paper directly addresses this policy debate. We use a 16-year panel to analyze the effects of ProComp on both student and teacher outcomes. We particularly focus on the onset of ProComp 2.0 (PC2) which could have affected outcomes as early as 2009/2010 (data panels in prior ProComp studies end in 2010). In many ways, the full onset of the policy began with PC2: The average annual ProComp payout increased nearly six-fold from about $960 to over $5700.

We examine whether ProComp had an effect on student achievement or the recruitment/retention of (effective) teachers. When possible, we rely on comparative interrupted time series (CITS) to examine pre/post ProComp trends in outcomes in DPS relative to similar districts across the same period. When CITS is not possible, we conduct interrupted time series (ITS) analysis in DPS using a panel up to 5 years pre-PC1 and up to 10 years post-onset. To enhance the teacher retention ITS, we compare these pre/post outcome patterns across teacher effectiveness levels. We explore whether the relationship between effectiveness and retention has changed from pre- to post- ProComp in ways that are consistent with PFP theory and ProComp's design. We supplement the ITS with descriptive analyses of retention as a function of ProComp payout amounts.

The CITS analysis of student achievement suggests that ProComp may have had a positive effect on ELA, math, and writing achievement that was not evident in comparable districts with similar achievement trends prior to 2005. We also find descriptive evidence from ITS analyses that more effective teachers were recruited to DPS once ProComp began and that the overall decline in teacher retention across districts in this time period was less precipitous among DPS' highly effective teachers during ProComp.

In recent years, there has been a notable spike in U.S. teacher strikes across a surprisingly diverse landscape of places—from West Virginia to Los Angeles (Van Dam, 2019). Among these high-profile events, the January 2019 Denver teacher strike stands out for its focus on differential compensation, which has a particularly long history in Denver Public Schools (DPS). In 2005-06, DPS became one of the first U.S. districts to implement a pay-for-performance (PFP) compensation system, and Denver's ProComp is now the longest-running PFP policy in the country.

The national proliferation of PFP systems in education has been controversial, with mixed evidence and competing narratives about its impacts. In January 2019, the local divide in stakeholders' views of ProComp came to a head, leading to a 3-day teacher strike. After a marathon 20 straight hours of deliberations, DPS and the union agreed on $23 million in teacher pay raises at the cost of primarily cutting about 150 jobs from the central office (Ragan, 2019; Will, 2019).

During the January strike, disagreements arose about whether 13 years of ProComp have helped or harmed DPS's efforts to retain effective teachers to improve student outcomes. Teachers described experiencing ever-increasing turnover in their schools during the decade after ProComp began, leading many to conclude that it did not improve retention (Brundin, 2019). District leaders viewed rising retention since 2015 as evidence that its current system was having positive effects (Denver Public Schools, 2019). Interestingly, both of these descriptive facts are *true*. Yet neither reveal ProComp's causal effect on achievement or retention. This unresolved disconnect around ProComp contributed to the strike and strained relations among the district, teachers, and parents.

This paper directly addresses this policy debate. We use a 16-year panel to analyze the effects of ProComp on both student and teacher outcomes. We particularly focus on the onset of ProComp 2.0 (PC2) which could have affected outcomes as early as 2009/2010 (data panels in prior ProComp studies end in 2010). In many ways, the full onset of the policy began with PC2: The average annual ProComp payout increased nearly six-fold from about $960 to over $5700—from

an average of 2% of a teacher's base salary to 12% in PC2. The average annual participation rate among full-time teachers went from 33% in ProComp 1.0 (PC1) to 81% in PC2 (ProComp participation was always mandatory for all new hires). Several of the 10 incentives only began in earnest in PC2, and for all incentives, the percent of participants that received them increased in PC2. The current study extends up to 8 years after PC2 began.

Conditional on the setting, we implement quasi-experimental approaches with the strongest possible causal warrants to explore if ProComp had an effect on student achievement or the recruitment/retention of (effective[1]) teachers. When possible, we rely on comparative interrupted time series (CITS) to examine pre/post ProComp trends in outcomes in DPS relative to similar districts across the same period. When CITS is not possible, we conduct interrupted time series (ITS) analysis in DPS using a panel up to 5 years pre-PC1 and up to 10 years post-onset. To enhance the teacher retention ITS, we compare these pre/post outcome patterns across teacher effectiveness levels. We explore whether the relationship between effectiveness and retention has changed from pre- to post- ProComp in ways that are consistent with PFP theory and ProComp's design. We supplement the ITS with descriptive analyses of retention as a function of ProComp payout amounts.

The current study also illustrates how policy researchers can engage in rigorous social science research in situations that do not lend themselves to straightforward causal inference. When the assumptions are laid out and findings are well-contextualized, policy research has the potential to directly inform the efforts of public agencies as they work to address societal needs. The current research is enhanced by the fact that it was conducted in ongoing communication with, and buy-in from, both DPS and the teachers' union, the Denver Classroom Teachers Association (DCTA). The study is a prime example of a scenario in which policy effects are deeply contested, the answer is far from straightforward, and research can help.

Current Study

We address the following primary research questions about the effects of DPS' ProComp on student outcomes and teacher recruitment/ retention:

(1) Is ProComp associated with improvements in student achievement (ELA, math, writing)?

(2) Is ProComp associated with the recruitment of more effective teachers to DPS?

(3) Did ProComp affect teacher retention overall (3A) or differentially by effectiveness (3B)?

In addition to these key questions, we also explored a set of secondary questions that are not the focus of the current paper and thus only mention them briefly: We did not find evidence that, once ProComp began, teachers became more likely to transfer to schools ever categorized as Hard to Serve, Top Performing, or High Growth—that is, the kinds of schools that were eligible for large schoolwide ProComp bonuses (results available upon request). We also investigated whether voluntary participation in ProComp was associated with improved teacher effectiveness but found the available data could not support this analysis.[2]

As a preview, the CITS analysis of student achievement suggests that ProComp may have had a positive effect on ELA, math, and writing achievement that was not evident in comparable districts with similar achievement trends prior to 2005. We also find descriptive evidence from ITS analyses that more effective teachers were recruited to DPS once ProComp began and that the overall decline in teacher retention across districts in this time period was less precipitous among DPS' highly effective teachers during ProComp.

The paper proceeds as follows: We begin with an introduction to ProComp's bonus structure, participation, and timeline. We then review literature about PFP systems, focusing on prior research on ProComp. This is followed by a description of the data, methods, and modeling approach. After presenting results, we conclude with a discussion of our findings, limitations, and implications.

DPS and ProComp

DPS is the 13th largest urban school district in the U.S., serving about 90,000 students in 2016 (U.S. Department of Education, 2018). Table 1 shows that over two thirds of DPS students come from households with low income (proxied by federal free/reduced-price lunch program (FRPL) eligibility), and the majority are Hispanic (56%), with an additional 23% White, 13% Black, and 3% Asian (U.S. Department of Education, 2018). Rapid enrollment growth has been a defining feature of the district for over a decade, increasing by 25% between 2006 and 2016. The percentage of Black students has decreased between 2001-02 and 2015-16 from 20% to 13%, while the percent of students who are identified as White, Hispanic, Asian, Limited English Proficient (LEP), and FRPL-eligible has changed little over the same period.

[Insert Table 1 about here]

In the early 2000s, members of the DPS community sought to replace the traditional teacher compensation system of uniform step-and-lane increases with targeted teacher pay incentives to "couple teacher compensation more directly with the mission and goals of DPS and DCTA" (Gonring, Teske, & Jupp, 2007; Slotnik, Smith, Glass, Helms, & Ingwerson, 2004). Denver's ProComp was the product of negotiations between DPS and the teachers' union (DCTA). The plan to initiate ProComp was approved by DCTA members in 2004, with funding provided by a $25 million per year property tax increase (adjusts with inflation) to be spent exclusively to fund incentive pay that Denver voters approved in 2005. ProComp began in the 2005-06 school year.

All new hires as of January 1, 2006 were automatically enrolled. Teachers hired prior to 2006 could opt in to ProComp during certain windows between 2005 and 2011. Those who did not opt-in maintained the original compensation model. Table 2 shows numbers and rates of ProComp teacher participation from 2005-06 through 2016-17. Participation increased due to both opt-in and

the automatic enrollment of new hires. In 2006, 14% of DPS teachers participated. In 2011, 73% of DPS teachers participated. By 2017, 90% of all DPS teachers were ProComp participants.

[Insert Table 2 about here]

The ProComp system includes 10 distinct financial incentives that began at different times, vary in terms of size and likelihood of receipt, and target different behaviors. See Table 3 for a high-level overview. The 10 incentives are divided into 4 categories: (1) Knowledge and Skills (i.e., education and professional development completion), (2) Student Growth (at class or school level), (3) Market Incentives (working in certain positions or schools), and (4) Comprehensive Professional Evaluations. Some incentives are paid as a one-time bonus, while others are paid as a permanent, "base-building" salary increase. While most of the incentives are earned individually, 3 of the 10 bonuses (working in Hard to Serve, High Growth, and Top Performing schools) are awarded schoolwide. See Online Appendix A for detailed information on each incentive.

[Insert Table 3 about here]

DCTA voted in August 2008 to approve significant modifications to ProComp, giving rise to the significant changes we refer to as PC2. At the time, both DPS leadership and DCTA agreed that the initial design required revision in order to reduce expenditures on administering the policy and to increase the odds of influencing teacher retention (Proctor, Walters, Reichardt, Goldhaber, & Walch, 2011). While there was some disagreement about how to revise the policy, and DCTA representatives expressed concerns about privileging new teachers at the expense of veteran teachers (Meyer, 2008), an agreement was ultimately reached that more than tripled annual ProComp incentive expenditures from $6.7 million to over 23.8 million in the year PC2 took effect (Proctor et al., 2011). Table 3 summarizes changes from PC1 to PC2 in the average amount of and receipt rate for each incentive. Under PC2, annual ProComp payments increased substantially; for example,

the incentives for Hard to Staff positions and Hard to Serve schools more than doubled (Gonring et al., 2007). The magnitude of this increase is readily apparent in Figure 1, which shows that the average PC1 payout represented only about 2% of a teacher's base salary. However, in PC2, annual payouts represented 12% of base-pay, on average, among full-time participating teachers, with some as high as 35 to 45% of base-pay. ProComp annual payouts vary widely both across teachers in a given year (e.g., across teachers in 2013-14, a standard deviation (SD) of $3,505) and across years for a given teacher (an SD of $2,300 across years within teacher).

[Insert Figure 1 about here]

Prior Research on PFP Policies

Critiques of the uniform teacher compensation schedule have a long history in the U.S. (for an overview, see Podgursky & Springer, 2007). However, in the early 2000s, there was a resurgence of interest in replacing step-and-lane systems with performance pay. This movement is sometimes attributed to Lazear and colleagues' findings of increased productivity under performance pay in other industries (Hall, Lazear, & Madigan, 2000; Lazear, 1996, 2000). Lazear subsequently theorized that teachers, too, could be compensated under a merit-based system. He posited that the merit pay could improve student outcomes via two mechanisms—improvement and selection. Under the former, teachers respond to a monetary incentive to align their performance to a common goal of raising student test scores (if used as the performance metric). Under the latter, teachers who are best able to raise student test scores are attracted the district offering PFP. These mechanisms can also include "de-selection" (i.e., discouraging retention among low-performing teachers), reallocation (incentivizing moves of highly effective teachers to positions of need), and strategic retention (signaling to stronger teachers—via higher pay— that their efforts are valued).

As teacher PFP systems began to appear in the mid-2000s, Podgursky and Springer (2007)

published a comprehensive theoretical review of support for and challenges to performance pay (e.g., difficulty in measuring performance, the team-based nature of educating students, the potential to ignore unmeasured dimensions of the job, misalignment with intrinsic motivation of teachers). At that time, little empirical evidence existed to adjudicate between these competing theories.

Since then, a growing body of empirical work has begun to emerge. Some of these studies focus on estimating policy impacts on *teacher* outcomes. For instance, Glazerman and Seifullah (2012) found short-term impacts of the Chicago Teacher Advancement Program on in-school retention, though not district retention. Springer, Swain, and Rodriguez (2016) found that $5000 retention bonuses for highly-rated teachers in low-performing Tennessee schools increased retention for tested-subject teachers. Feng and Sass (2018) found that loan forgiveness incentives for Florida teachers in hard-to-staff positions reduced attrition among these teachers. Glazerman, Protik, Teh, Bruch, and Max (2013) found in the Talent Transfer Initiative (TTI) that offers of $20,000 incentives to transfer to low-achieving schools only induced 5% of TTI candidates to actually transfer. And at least one study indicated that PFPs do not cause short-term changes in teacher practices (Yuan et al., 2013), while another showed that teachers change their practice in response to a threatened loss of incentives (Fryer, Levitt, List, & Sadoff, 2012).

Other studies have focused on effects of PFP on *student* outcomes. Using nationally representative survey data, Figlio and Kenny (2007) found higher test scores in districts with incentive pay. However, studies of individual policies have not found consistent positive achievement effects (Glazerman & Seifullah, 2012; Springer et al., 2011). While many studies based in the U.S. have found little to no impact of PFP on student achievement (see also Fryer, 2011; Marsh, Springer, McCaffrey, Yuan, & Epstein, 2011), some international studies have found significant effects of incentive pay on student outcomes (see, e.g., Atkinson et al., 2009; Duflo,

Hanna, & Rya, 2012; Kremer, Ilias, & Glewwe, 2003; Muralidharan & Sundararaman, 2008).

In sum, the evidence base on teacher incentive pay is mixed for both student achievement and overall teacher retention. The findings have been mixed not only because the policies and settings being studied differ from one another in important ways, but also because those settings lend themselves to different methodological approaches. Findings have been particularly tepid with respect to stimulating cross-school transfers or raising overall district retention. However a recent study by Dee and Wyckoff (2015) found that Washington D.C.'s IMPACT teacher evaluation and incentive system increased retention among high performing teachers, improved teacher performance, and raised student achievement (Adnot, Dee, Katz, & Wyckoff, 2017). This suggests that the *strategic* retention of effective teachers may be a promising avenue for further study.

Literature on ProComp

ProComp has been studied in three prior publications (see Figure 2 for an overview). In Fulbeck's (2014) descriptive analysis of DPS teacher attrition through 2009-10, receiving a bonus in excess of $5,000 was associated with a 35% decrease in a teacher's odds of exiting DPS (but again no association with cross-school transfers). Goldhaber and Walch (2012) found modest, positive student achievement effects as of 2010 by comparing both teachers who opted in with those who did not, as well as pre- and post-ProComp comparisons among teachers who opted in. Goldhaber, Bignell, Farley, Walch, and Cowan (2016) do not focus on estimating ProComp effects, but instead examine the non-random decisions of teachers who worked in DPS prior to the onset of ProComp to opt-in or not. They find, for instance, that more effective teachers and teachers in high needs assignments or schools were more likely to opt in, while African American teachers were less likely to opt-in. These findings both shed light on teachers' preferences for pay-for-performance policies like ProComp but also raise concerns about any approach leveraging variation in opt-in/out.

[Insert Figure 2 about here]

All three prior studies rely on data only through the 2009-2010 school year. This is an important limitation, because DPS implemented their substantial redesign of PC2 in 2008-2009, at which point annual ProComp payouts increased nearly six-fold. The prior work analyzes 3 years of PC1 but only up to 1 year of PC2. The current data extends up to 8 years after PC2 began, which also covers the time during and after the Great Recession. We also add to this literature by exploring whether ProComp's possible effect on teacher retention might differ by teacher effectiveness. Finally, we extend upon existing research by incorporating data from outside DPS where possible to implement a CITS approach (with data up to 6 years pre- and 10 years post-ProComp onset from both within and outside DPS), to improve upon the limited ITS causal warrant.

Data

The Assessment, Research, and Evaluation Department in DPS provided student-level data including demographic information, student test scores, grade and year indices, and links between student identifiers (IDs), teacher IDs in each subject, and school IDs. We use teacher IDs to merge in data from human resources files that include teachers' years of experience, age, gender, race, and highest level of education. We supplement these files with monthly teacher paycheck data, including total compensation as well as how much was earned for each ProComp incentive and whether it was paid as a base-builder or onetime bonus. The current analyses are limited to employees categorized as classroom teachers.[3] The DPS teacher analytic sample includes a total of 16,628 unique teachers, with an average of approximately 4,500 teachers observed in each of the 16 years.

We also use publicly-available data from the Colorado Department of Education (CDE) for every Colorado district on mean student achievement in English language arts (ELA), math, and writing (since 2004) and teacher retention rates (since 2001). Online Appendix A provides a

summary of all data availability timing (Table A7), as well as an overview of the key variables used in the three research questions (Table A8) including the outcome of interest's definition and coding, the analytic method(s) used, unit of analysis, and any analytic sample limitations. Next, we describe these key variables in greater detail.

Measures of Teaching Effectiveness

Teacher median growth percentile (MGP) scores are used as a proxy for teaching effectiveness both in RQ2 as the outcomes of interest for new-to-DPS teachers and in RQ3 as a categorical variable for interactions (refer again to Online Appendix A (Table A8) for an overview). MGP scores are generated from student growth percentiles[4] (SGPs), which are used in Colorado to measure a student's academic progress from year to year, compared with that of their academic peers. DPS characterizes a *teacher's* annual effectiveness by taking the median SGP across the teacher's eligible students in that year.

An SGP approach is used across Colorado, as opposed to a value-added measures (VAM) approach with additional controls (often preferred in research contexts). A practical advantage of MGPs is that they are easier to understand and interpret than VAMs. Because MGPs are the salient policy measure in DPS, we use them as proxies for teacher effectiveness in our main analyses.[5] It is important to note that research has been decidedly mixed on the use of both value-added and MGP measures of teacher effectiveness. Some studies have pointed out the limitations of value-added measures (e.g., McCaffrey, Sass, Lockwood, & Mihaly, 2009; Papay, 2010; Rothstein, 2009) and also specifically of MGP scores (Pivovarova & Amrein-Beardsley, 2018). At the same time, other research suggests bias is less of a concern (Chetty, Friedman, & Rockoff, 2014a; Kane, McCaffrey, Miller, & Staiger, 2013; Koedel & Betts, 2011), that these measures correspond with principal perceptions of effectiveness (Briggs & Dadey, 2017; Jacob & Lefgren, 2008), and they predict

important student post-schooling outcomes (Chetty, Friedman, & Rockoff, 2014b). From a policy perspective, DPS uses MGP scores to make inferences about teaching effectiveness. In addition to examining teacher outcomes by MGP scores, we conduct sensitivity analyses when using MGPs on the right-hand side of the equation, substituting other proxies for teacher effectiveness, including VAMs,[6] experience, or degree attainment.

Key Outcomes for Research Questions

(RQ 1) For the CITS analysis of student achievement effects, the outcome of interest is mean student achievement on statewide assessments in ELA, math, or writing at the district-year level for each Colorado district (unit of analysis is district-year). CDE data is available 2004 - 2014. Scores are expressed in standard deviations, relative to the state mean in the same subject, grade, and year.[7]

(RQ 2). For the ITS analysis of the recruitment of more effective teachers to DPS, the primary outcome of interest is MGP_{py}—that is, teacher p's annual MGP score. Originally, MGP scores range from 0 to 100, though in practice that full scale is not realized and therefore not straightforward to interpret. When these MGP scores are used as an outcome, we standardize them at the teacher level. We leverage MGP scores in three different ways (see Online Appendix B) to characterize the quality of arriving teachers and conduct robustness checks to see if results are consistent across these definitions. The three approaches have tradeoffs between including as many teachers as possible versus making MGP measures as comparable across teachers as possible.

(RQ 3). For the CITS analysis of overall teacher retention effects (3A), the outcome of interest is a continuous variable capturing the percent of teachers who were retained in district d from year y-1 to year y (unit of analysis is district-year). For RQ 3B—the ITS analyses of differential retention in DPS only—the outcome of interest is $Retained_{py}$, a dummy variable coded 1 if teacher p was present in the y-1 and is present in the district in year y, and 0 otherwise (unit of analysis is

teacher-year).We leverage teachers' MGP scores to explore whether any ITS-based ProComp effects vary by teaching effectiveness. See Online Appendix C for a discussion of our two different approaches for using MGP scores to characterize teacher performance for RQ 3B.

Methods

Because all ten of ProComp's incentives were enacted simultaneously and all participants were equally exposed to them, it is not possible to disentangle the causal effect of particular incentives. However, the longevity of ProComp provides the opportunity to leverage a long panel of data (up to 5 years before and 11 years after onset) to explore the policy's overall effects. The essence of the analytic approach is to compare mean outcomes and trends in outcomes in the pre-ProComp period to the PC1 and PC2 eras, net of demographic shift. However, for student achievement (RQ1) and overall teacher retention (RQ 3A), we use district-year aggregated outcome data for all Colorado districts. For these outcomes, we conduct a comparative interrupted time series (CITS) by differencing out secular trends in outcomes based on suitable comparison groups of non-treated districts observed throughout the same time frame (more on this below).

The model below presents the time specification used throughout all analyses to follow. The model separates the panel into three linear segments—the pre-ProComp, PC1, and PC2 eras[8]—each with a distinct mid-point intercept and slope. Note that, for the district-year level CITS analyses, all of the linear segments described below are also interacted with a DPS dummy variable to compare pre/post outcomes in DPS to other, similar districts. However, for simplicity, we explicate the time function using the within-district ITS analysis of student math achievement:

$$Math_{iy} = \beta_0 + \beta_1(PC1_{iy}) + \beta_2(PC2_{iy}) + \beta_3(Time_{iy}) + \beta_4(Time_{iy} \times PC1_{iy}) + \beta_5(Time_{iy} \times PC2_{iy}) + \boldsymbol{X_{i(y)}\beta} + \boldsymbol{W_{sy}\beta} + \varepsilon_{iy} \quad \textbf{(Eq. 1)}$$

In Equation (1), $Math_{iy}$ is modeled as a function of two dummy variables ($PC1_{iy}$ and $PC2_{iy}$), the variable $Time_{iy}$ (captures linear time measured in years, centered at the midpoint of each of the three time segments[9]), and the interactions of these variables. See Figure 3 for a full illustration of this variable coding and how it leads to the following interpretations of coefficients: β_0 captures mean math achievement in the pre-ProComp period, while β_1 represents the change in mean math achievement between the pre-ProComp and PC1 period. Likewise, β_2 captures the estimated change in mean math achievement between the pre-ProComp and PC2 period. For parsimony, we refer to these as "level effects." Because $Time_{iy}$=0 represents the midpoint of each period, these are also akin to difference estimators (or difference-in-differences in the case of CITS). If β_1 and β_2 are positive and statistically significant, mean math scores (relative to the Colorado mean) rose during PC1 and PC2, compared with the pre-period. Given that ProComp was not fully implemented until 2008/2009, we are particularly interested PC2.

[Insert Figure 3 about here]

The coefficient on $Time_{iy}$, β_3, represents the linear trend in math achievement during the pre-period (annual rate of change in scores). β_4 and β_5 capture the *difference* in achievement trends between the pre-period and the PC1 and PC2 periods, respectively. We refer to these as "trend effects." If estimates for β_4 and β_5 are positive and statistically significant, it suggests that math achievement scores are trending more positively (or less negatively) in the ProComp periods than they were in the pre-period—that is, a change in the trajectory of these outcomes coincident with the onset of ProComp. To address the possibility that fluctuations in student composition are driving

observed changes, we include a vector[10] of grand-mean centered, time-invariant and time-varying student covariates, $\boldsymbol{X_{i(y)}}$, as well as a vector of school demographic covariates, $\boldsymbol{W_{sy}}$ (i.e., student variables aggregated to the school-year level, plus annual school enrollment).[11] We use this same model specification when either teacher MGP scores or retention are the outcome of interest.

Like any ITS, the causal warrant rests on several assumptions:[12] The basic logic is that, conditional on controls, the trend in outcomes prior to the onset of the policy serves as a valid counterfactual for what trends in those outcomes would have been in the post-onset years in the absence of the policy. This also implies that nothing else that affects our outcomes changes concurrently with PC1 and then PC2. These are strong assumptions. Where possible, we preference CITS-based estimates, which rely on weaker assumptions (more on this below). When limited to DPS data, we do conduct some robustness checks that attempt to separate the onset of ProComp from other DPS policy changes that occur later in the panel.

For ITS-based teacher retention analyses, we push on any findings with descriptive analyses that explore whether PC1 or PC2 retention is correlated with teacher effectiveness in a way not existent before ProComp began. We also explore whether this strategic retention is stronger in schools eligible for the large ProComp schoolwide bonuses. Though we cannot rule out the possibility that the assumptions above are violated, we look for suggestive evidence that any effects are distributed in a manner that reflects ProComp's design.

Comparative ITS

We use three strategies for identifying sets of districts that are "similar" to DPS in the period leading up to the onset of ProComp. We present the CITS results of two matching methods—(1) matching on a propensity score, and (2) matching to districts with similar trends (but not levels) in both outcomes and demographic characteristics.[13] We also conduct the CITS with (3) a synthetic

control method (SCM) approach (Abadie, Diamond, & Hainmueller, 2012; Abadie & Gardeazabal, 2003). The SCM approach is similar in spirit to the matching techniques except that we estimate a set of district-level weights to construct a synthetic comparison group to DPS. See Online Appendix D for additional detail about the matching and SCM approaches. Because CITS analyses use data pre-aggregated to the district-year level, it is not possible to reasonably estimate the precision of coefficients (there are only a handful of district-year observations in each model; at the same time, standard errors do not reflect the fact that datapoints in each year are based off hundreds or thousands of individuals). For CITS analyses, we therefore focus on the pattern of results, rather than statistical significance. Though we therefore must interpret results with caution, we prefer this approach to presenting and interpreting questionable standard errors.

The CITS assumptions for causal inference are less demanding than for the ITS. We assume that nothing other than ProComp that could affect the *trends* in outcomes changes concurrently with the policy in the treated district but not the comparison districts. In other words, the comparison group captures secular trends in outcomes throughout the period, and we assume these secular trends would have also been observed in DPS in the absence of ProComp. The comparison group trends thus serve to contextualize the outcome patterns observed in DPS throughout the panel.

RQ 1: ProComp and Student Achievement

Student achievement patterns from before to after ProComp, relative to comparable districts, are consistent with the hypothesis that ProComp had a positive effect on student achievement. In Figure 4, we first visually inspect the CITS results (estimates presented in Table 4): In the upper panel (ELA), we can see that—as intended—all three comparison groups exhibit achievement *trends* in the pre-period that are approximately parallel to the positive pre-period trend in DPS, suggesting these comparisons seem appropriate (the levels are not always the same, but the assumptions for

CITS do not require this). For DPS, achievement levels and trends then increase slightly throughout both PC1 and PC2. While DPS looked similar to comparison districts in the pre-period, this is not the case once ProComp begins. Indeed, none of the comparison groups sustain their achievement trends past the year ProComp began. We observe a similar pattern in math and writing, as well. This suggests that the uptick in DPS achievement scores that coincided with the onset of ProComp is not simply reflecting secular trends in achievement.

We report the estimated coefficients that correspond to Figure 4 in Table 4. For the sake of parsimony, we present only estimates of the intercept, the Denver dummy main effects, and the interaction terms between Denver and the time function variables. Positive coefficients on those interactions indicate that Denver had larger gains/more positive trends changes than the given comparison group. Though the point estimates vary in magnitude, both level and trend effects are larger in DPS than in the comparison groups in virtually all cases.

[Insert Table 4 and Figure 4 about here]

At first, there appears to be a few exceptions to these overall positive effects in the SCM comparison (column 3 of Table 4). Results from the SCM comparison indicate that, achievement scores rose more in the *comparison* group than in DPS during PC1 (e.g., in ELA negative level effects of -0.183 for PC1 and $-.005$ for PC2). However, the SCM group's trend in the PC2 period was strongly negative, whereas DPS exhibited increasingly positive trends during that time. The visualization of these SCM patterns in Figure 4 demonstrate that the SCM, even more than the other comparisons, produce a clear picture that something unique happened in DPS during PC1 and PC2. This case also highlights the need to examine both level effects (difference-in-differences) and trend effects alongside one another. We therefore provide visual illustrations of results where possible.

In sum, DPS students' test score increases generally accelerated once ProComp began in all three subjects. The CITS estimates indicate that increased test scores in DPS following ProComp's initiation are unique to DPS; similar Colorado districts showed flat or declining achievement during the ProComp time period. These results suggest that ProComp may have had a positive effect on student achievement. Next, we explore potential ProComp mechanisms—e.g., recruitment and retention of effective teachers—that may have led to these potential achievement effects.

RQ 2: ProComp and the Recruitment of Effective Teachers to DPS

Here, we use the ITS approach to examine whether the effectiveness (MGP scores) of new-to-DPS teachers increased after the onset of ProComp, controlling for school-level demographic changes in the district. We rerun these models using MGP scores in three different ways to characterize the quality of these arriving teachers (refer back to Online Appendix B). In the end, the pattern of results is substantively consistent across the different approaches.

Results presented in Table 5 (and illustrated in Figure 5) suggest that DPS has been able to recruit more effective teachers (higher MGPs) to the district since ProComp began. To facilitate interpretation, the MGP score outcomes have been standardized at the teacher level. All level effects and trend effects in Table 5 are positive across all three columns. We generally find substantively large and often statistically significant increases in the MGP level of arrivers from pre- to post-ProComp. We estimate larger increases during PC2 than PC1 (compare PC1 and PC2 level effects). In PC2, the mean MGP of arrivers increased by between 9 to 17 percent of a standard deviation of teacher MGPs. Five of the six level effects are statistically significant. While all the trend effects are positive, most are not statistically significant. This is indicative of an upwards shift in the MGP of arrives from pre- to post-ProComp, but a weaker change in the overall trends (see Figure 5). The

positive level effects—particularly in PC2—are consistent with the hypothesis that ProComp is associated with modest to moderate positive changes in the effectiveness of arriving teachers.

[Insert Table 5 and Figure 5 about here]

RQ 3(A): ProComp and Overall Retention

It is not entirely clear that ProComp—even if working as intended—would have an impact on *overall* teacher retention, because differential compensation systems target incentives toward *specific* teachers. The theory behind these systems posits that some teachers (ostensibly those contributing to the district's stated objectives) will be more inclined to stay because their contributions are valued, while other teachers become less inclined to do so. If ProComp was working in this way, the retention increases among some could be offset by (intended) decreases among others, and overall retention could appear unaffected. On the other hand, it is possible that the majority of teachers prefer to work (or not) under a system like ProComp—regardless of whether they personally benefit from it—in which case overall retention could shift. Because the CITS analyses provide important context for DPS patterns of overall retention throughout the panel, we begin by examining overall retention effects (3A). However, we anticipate that the strongest evidence of a ProComp effect should more likely manifest itself by inducing a positive relationship between retention and, say, MGP scores that was not present before ProComp began (RQ 3B).

The CITS estimates presented in Figure 6 suggest that the pattern of declining overall teacher retention in DPS throughout the study panel was not unique to DPS. While teacher retention rates appear parallel to DPS in the pre-period for two of the three comparison groups (the PSM comparison does not perform well in the pre-period), retention rates in DPS and comparison districts *also* appear relatively similar during ProComp—particularly in the PC2 era. The corresponding CITS estimates in Table 6 reflect this interpretation: While the level and trend effect estimates vary

somewhat across the columns, they are consistently small—close to zero—indicating that DPS retention does not look very different from comparison districts.

[Insert Table 6 and Figure 6 about here]

These results provide two important takeaways about DPS teacher retention: First, results corroborate teachers' perspectives voiced during the January 2019 strike that retention is declining precipitously in DPS throughout the ProComp period. That said, we also see that overall teacher retention may have been on the decline before ProComp began, and a similar degree of teacher retention declines occurred in comparable Colorado districts. While one should be concerned about the drop in teacher retention across the panel, it does not clearly correspond to the onset of ProComp.

RQ 3(B): ProComp and Differential Teacher Retention

Differential Retention by Bonus Amount

Despite the null results in terms of overall teacher retention (3A), we see some descriptive evidence of a link between ProComp and retention once ProComp began. Figure 7 illustrates that, among ProComp participants, there is a strong correlation between the size of a teacher's total incentives in one year and the likelihood of returning to the district the next year. For teachers receiving less than $2,000 in total bonuses in a given year, 83-84% return the next year. However, among teachers who receive over $14,000 in ProComp payouts in a given year, around 92-95% return. This is not direct evidence of a ProComp effect—by design, the kinds of teachers who receive greater incentive pay should be systematically different than those who receive less incentive pay. However, we do see in Table 7 that a $1,000 increase in a teacher's ProComp payment in y-1 is associated with a positive and statistically significant increase in retention rates, even among

teachers with the same MGP scores (M3), who are also in the same school (M4), and in the same year (M5). This prompted a closer look at ProComp and retention.

[Insert Table 7 and Figure 7 about here]

Differential Retention by MGP Terciles

Four of the 10 ProComp bonuses are explicitly tied to teachers' contributions to student growth, and 4 others seek to reward, perhaps indirectly, more effective teachers (to the extent that teachers with higher evaluation scores, who pursue additional formal education, or participate in professional development are indeed stronger). This raises the possibility that ProComp effects should manifest themselves—not in terms of overall retention—but instead as differentially higher retention among more effective teachers.

We therefore rerun the ITS analyses separately for teachers in the top, middle, and bottom third of the teacher effectiveness distribution (we also re-conduct this analysis below using MGP scores as a continuous interaction, more on this below). We already anticipate, based on the CITS analyses above (3A), estimates will show overall declines in retention over time. However, if ProComp differentially retained stronger teachers, estimated level and trend effects would be most positive (or least negative) among high-MGP teachers, followed by teachers in the middle-MGP tercile, followed by the largest negative estimates concentrated among the bottom third of the MGP distribution. We focus particularly on the PC2 period, once all ProComp incentives were in place and incentives were paid out at their full amounts.

We find that ProComp may have induced *strategic* retention effects concentrated among more effective (higher MGP) teachers (Table 8, Figure 8). When using either definition of MGP scores to create terciles of teacher effectiveness, we see the pattern described above in PC2: For both level and trend effects, estimates are indeed closest to zero for high-MGP teachers, and most

negative for low MGP teachers. These results are quite consistent with a ProComp effect, but this is more readily apparent when plotted in Figure 8: Before ProComp began, there was no association between teachers' MGP scores and retention. However, once ProComp begins we see a decided fanning out of retention by MGP groups. Retention falls less precipitously among highly effective teachers, particularly during PC2. In sum, it appears that teachers' decisions to return to DPS were unrelated to their effectiveness before ProComp, however once the policy began, highly effective teachers were less likely to leave than their less-effective counterparts.

[Insert Table 8 and Figure 8 about here]

We also conduct this analysis using a continuous version of MGP scores, interacted with the time function (results are presented in Online Appendix C, Table C2 and Figure C1). The advantage in this approach is that we need not make choices about category cut points for high, middle, and low-achieving teacher groups that could drive our results. Another advantage is the ability to formally test the hypothesis that level or trend effects depend on MGP (i.e., significant coefficients on MGP interactions). A disadvantage is the difficulty in interpreting the magnitude of coefficients on three-way interactions. Yet the takeaways from the continuous interaction are consistent: The interactions between MGPs and level effects are always positive and statistically significant in Table C2 (the interactions between MGPs and trend effects are positive but very small due to the three-way interactions and thus not significant). Table C2 (see Figure C1) indicates that retention rates decrease between 2 and 3.6 times faster for teachers with MGP=20, relative to an MGP=80 teacher.

Retaining Effective Teachers in Schools with Large ProComp Bonuses

Given the retrospective nature of the current study and how ProComp was rolled out, it is difficult to identify other ways to isolate ProComp's effects. For instance, one may be concerned that some other policy, coincident with ProComp, could have forged the effectiveness-retention

relationship that was not present pre-ProComp. We therefore attempted to think of ways that a ProComp-*specific* differential retention effect might manifest itself that another policy's differential retention effect would not.

We therefore consider the differential retention of effective teachers in schools that might "feel" ProComp the most—that is, schools that became eligible for large, schoolwide incentives once ProComp began. The weakness of this descriptive analysis is that schools eligible for these bonuses are, by design, not random: High growth, top performing, and hard to serve schools were all eligible, and these schools likely employ different kinds of teachers than schools that were never eligible for any ProComp bonuses. At the same time, it seems sensible that a ProComp-specific effect would be concentrated among highly effective teachers working in large-bonus schools.

In Table 9, we therefore limit the sample to above-average MGP teachers and compare ITS retention patterns between schools that were ever eligible for *any* ProComp schoolwide bonus (column 1) and schools never eligible for these bonuses (column 2). Here it is important to focus on PC2 since some of the schoolwide bonuses were not enacted in PC1. In column 1, we see that the retention of effective teachers fell by 11.4 percentage points from pre-ProComp to PC2. However, the drop was even larger—14.6 percentage points—among schools that never experienced schoolwide bonuses (PC2 level effect in column 2).

[Insert Table 9 about here]

The trend effects are even more striking: Among bonus-eligible schools, retention rates increased by +1 percentage point each year in the pre-period, but during PC2, retention rates were decreasing by 2 percentages points annually—a 3 percentage point drop (the PC2 trend effect in column 1). However, for never-eligible schools, the pre-period retention trend of +2.7 percentage points fell to a retention trend of -5.6 percentage points in PC2. This 8 percentage point drop (PC2

trend effect in column 2) is more than twice as large the one observed in bonus-eligible schools. In columns 2 through 5, we categorize the ever-eligible schools by which of the three schoolwide bonuses the school was eligible for. In PC2, level and trend effects are modestly more negative in schools eligible for the Hard to Serve incentive, but the differences across bonus types are not large or systematic. It does not appear that one of these three is driving the pattern evident in the column 1 and 2 comparison. Taken together, this suggests that—though retention generally fell throughout the panel, it fell less precipitously among effective teachers in schools that became eligible for ProComp's large schoolwide incentives.

Conclusion

We use a 16-year panel to estimate the effects of ProComp, one of the longest-standing and largest PFP systems in the U.S. We rely on CITS (whenever possible) and ITS approaches to estimate both level and trend effects of ProComp across a range of outcomes. Results are suggestive of a positive ProComp student achievement effect: We find increases in achievement outcomes across all subjects in DPS, controlling for demographic shifts, and—importantly—these increases are not observed in comparable Colorado districts (Figure 4). While this suggestive evidence is intriguing, an important question remains: If ProComp caused the observed achievement gains, through what change to the teacher workforce did this occur?

A key finding in this paper is evidence of both strategic teacher recruitment and retention during ProComp. Teachers recruited to DPS appear to be somewhat more effective—as measured by MGP scores—in the ProComp era (Figure 5). For retention, descriptive evidence suggests strategic ProComp effects that reflect the design of the policy: Even among teachers with the same MGPs working in the same school in the same year, the size of one's ProComp payout in the preceding year positively predicts retention (Figure 7). While teacher effectiveness bore no relation

to retention in the pre-ProComp period, MGP scores became predictive of retention after ProComp began. While high-MGP teachers *did* exhibit a decline in retention throughout the period, their decline was significantly less precipitous than that of their low-MGP counterparts (Figure 8). Finally, the ITS retention effects among effective teachers were stronger in the set of schools that became eligible for large schoolwide bonuses, particularly in PC2 (Table 9).

We can also reflect on which of the hypothesized mechanisms undergirding PFP policies find empirical support in the current study (e.g., via overall teacher improvement, recruitment, retention, reallocation, or strategic recruitment/retention of more effective teachers). Our findings lend support to the possibility that the onset of ProComp shifted the composition of the DPS teacher workforce through recruitment and retention of certain kinds of teachers. To capture this pattern descriptively, Figure 9 presents the mean MGP of teachers *arriving* in DPS in any given year, relative to the MGP of those who left the district. During ProComp, this relationship has generally flipped: The effectiveness of arriving teachers exceeds the effectiveness of those who depart.

Limitations

The noted weakness of the current analysis lies in the strong assumptions for ITS methods to produce unbiased effect estimates. One potential threat to this assumption could be the Great Recession (2008-2009), which began in the midst of ProComp and likely exerted an independent effect on the teacher labor market. If the Great Recession increased teacher retention, the ITS approach may over-estimates true effects. However, the Great Recession resulted in reductions in public K–12 education resource allocations (Engel, Claessens, Watts, & Stone, 2016; Leachman, Albares, Masterson, & Wallace, 2016), which may have included layoffs and staffing reductions. In this case, the ITS approach may underestimate retention effects. CITS analyses that incorporate secular trends from comparable districts could also partially address this concern.

Two other statewide policies are worth discussing: Senate Bill (SB)-191 and Race to the Top (RTTT). SB-191 is a statewide statute that addresses a teacher's professional trajectory from the probationary to non-probationary period.[14] By July 2013, all Colorado districts had to have an evaluation system in place that aligns with SB-191 (ratings did not have consequences until 2013-14). Colorado was also awarded RTTT funding in 2011-12, and those systems were piloted in 27 districts starting in 2012-13 (though DPS was not a pilot district). In theory, all Colorado districts were exposed to these statewide changes and thus their presence does not represent a violation of the CITS assumptions if their effect can be differenced out using the secular trends in comparison districts. However, we cannot rule out the possibility that DPS experienced these policies uniquely.

In addition, the assumptions of both ITS and CITS approaches could be violated by the confounding effect of other concurrent, DPS-specific policies. DPS's *Leading Effective Academic Practice* (LEAP) program is a multiple measure teacher evaluation and support system, which was piloted DPS-wide in 2011-12, with formal evaluations beginning in 2013-14 (Jerald, 2013). LEAP, SB-191, and RTTT all took effect late in the panel. When we restrict the panel to end in 2012-13—to avoid any potential confounding from SB-191, RTTT, or LEAP, our main findings are substantively similar (available upon request).

Another challenge is the fact that MGP scores only become available in 2004-05. The lack of such measures in the pre-period complicates our analyses, though we present a number of different solutions to this challenge (see Online Appendix B and Online Appendix C). Nonetheless, it would be preferable that MGP scores were available consistently throughout the panel.

Reflections on ProComp

This study provides a comprehensive assessment of ProComp, the pay-for-performance teacher compensation model at the heart of the January 2019 Denver teacher strike. A primary

contribution to the existing ProComp literature is that we are able to reflect on the ProComp 2.0 era—arguably the real start of full ProComp implementation. The evidence suggests that ProComp may have improved student achievement and led to the strategic recruitment and retention of effective teachers. We are limited in terms of the causal warrant of the approaches we can adopt, however we supplement CITS and ITS results with supporting descriptive analyses.

These results at first appear to contradict teacher perceptions about ProComp—that the program coincided with a dramatic decline in teacher retention and was thus ineffective—a central assertion during the strike (Far Northeast Teachers, 2019; Turkewitz & Goldstein, 2019). However, some of the evidence presented here is useful for reconciling the results with this important viewpoint. First, the CITS retention patterns illustrated in Figure 6 provide a key insight: Consistent with teachers' reports of increasingly problematic teacher turnover during ProComp, retention did, in fact, decline throughout the period. However, DPS retention patterns were not that different from other comparable Colorado districts during this period. Thus, while teachers' perceptions of reduced teacher retention were accurate, it would be very difficult to see from within DPS that retention rates were not necessarily distinct from secular trends outside DPS.

Second, while the ITS approach yields positive differential retention effect estimates, that increased retention could perhaps more accurately described as "mitigated attrition" from DPS. In other words, results suggest that the rate of departures among highly effective teachers might have been even worse in the absence of ProComp. This kind of an impact—while real and important—may also be a difficult one to perceive on the ground. Third, our results suggest that retention effects were not distributed evenly across all teachers (i.e., by MGP); since Denver is a quite large and segregated city,[15] retention effects may have been concentrated in some parts of the district and not equally visible to all. Overall, the nuanced results of the current study, relative to the public

discourse, highlight the importance of conducting careful investigations with an eye on issues of false causality.

There are reasons that teachers could be dissatisfied with ProComp that have little to do with its causal effects. For instance, with so many moving parts to ProComp's 10 different incentives, teachers may have difficulty predicting how their actions will translate into incentive pay. A teacher's ProComp pay clearly fluctuates significantly from year to year, and teachers report difficulty in anticipating their annual income (Brundin, 2019; Hess, 2019). These fluctuations may be felt acutely by DPS teachers, who are living in one of the most expensive U.S. cities (Kiersz, 2018), which is in turn located within in a state that has been ranked 50th in teacher wage competitiveness (Baker, Sciarra, & Farrie, 2014). Moreover, the cost of living in Denver has risen sharply—the cost of a median price home in Denver has increased by 85% in the last 10 years (Black, 2018). Those conditions—along with the Great Recession—may put many DPS teachers in a position where every penny counts, and fluctuations in pay at the margin are likely readily felt.

There may also be opportunities for DPS to improve its communication with teachers about the ProComp incentives they receive. Paycheck data indicates that ProComp payments are usually divided across a teacher's 9 or 12 monthly paychecks and appear on direct deposit forms (if one even looks at them) as one of many line items.[16] This is a lost opportunity given that most teachers do, in fact, receive higher compensation than they would in the absence of ProComp. This is especially true given that the legislation that funds ProComp specifies that it can only be used for merit pay—that is, it cannot be reallocated toward base pay or uniform increases.

Finally, there are likely ways that DPS could facilitate teachers' ability to compare their pay to what they would have earned in the absence of ProComp under the previous uniform pay schedule. Though the district's perceptions of teacher pay are framed in comparison to a pre-

ProComp traditional salary schedule, most teachers likely do not reflect on what their earnings would have been like under another system that has not been in place for 13 years now. Goldhaber et al. (2016) execute a well-considered approach to estimating which of these pay systems DPS teachers would have benefited from most, which could likely be adopted and adapted by DPS.

Reflections on Real-World Policy Research

Unlike policy research that can rely on a clean experiment or regression discontinuity, estimating ProComp effects is not straightforward and requires more context, robustness checks, and caveats. Nonetheless, empirical evidence can be most useful in the very moments when policy makers and practitioners are actively debating important policy choices, like the future of DPS' ProComp. Fortunately, Denver is a district that has long engaged in policy innovation, and both DCTA and DPS have a history of valuing research evidence in determining their policy positions.

This is reflected in the content of the Tentative Agreement that ended the January 2019 strike. In the final stages of the bargaining deliberations, a separate teacher workforce policy (not a part of ProComp) became a lightning rod issue. Though DPS and DCTA disagreed about that particular policy, they resolved the issue in the Tentative Agreement by requiring a joint research project by an external party to determine if the program should be continued (Denver Classroom Teachers Association, 2019). Working alongside both the central office and DCTA on this ProComp study has highlighted the potential role for policy research in DPS. Collaborative work on ProComp has established a certain level of trust in external research relationships (Oliver, Innvar, Lorenc, Woodman, & Thomas, 2014; Penuel & Gallagher, 2017). Providing high-quality, actionable research when it is needed most may give DPS a reason to further integrate empirical research into the organization's policy pipeline. This, in turn, could lead to future opportunities to implement and evaluate policies in ways that can be best deployed for practical and scholarly value.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2012). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*.

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, 113-132.

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis, 39*(1), 54-76.

Atkinson, A., Burgess, S., Croxson, B., Gregg, P., Propper, C., Slater, H., & Wilson, D. (2009). Evaluating the impact of performance-related pay for teachers in England. *Labour Economics, 16*(3), 251-261.

Atteberry, A., Loeb, S., & Wyckoff, J. (2016). Teacher Churning Reassignment Rates and Implications for Student Achievement. *Educational Evaluation and Policy Analysis*, 0162373716659929.

Baker, B. D., Sciarra, D. G., & Farrie, D. (2014). Is school funding fair? A national report card. *Education Law Center*.

Black, J. (2018, 12/13/2018). A 10 Year Look at the Denver Real Estate Market. *USAJ Realty*.

Briggs, D. C., & Dadey, N. (2017). Principal holistic judgments and high-stakes evaluations of teachers. *Educational Assessment, Evaluation and Accountability, 29*(2), 155-178.

Brundin, J. (2019, 01/08/2019). With A Strike On The Horizon, Denver Teachers Ready To Fight For Wages, Bonuses. *Colorado Public Radio*. Retrieved from https://www.cpr.org/news/story/with-a-strike-on-the-horizon-denver-teachers-ready-to-fight-for-wages-and-benefits

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review, 104*(9), 2593-2632. doi:doi: 10.1257/aer.104.9.2593

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review, 104*(9), 2633-2679. doi:doi: 10.1257/aer.104.9.2633

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34*(2), 267-297.

Denver Classroom Teachers Association. (2019). Tentative Agreement Between School District #1 Denver Public Schools and Denver Classroom Teachers Association *Counter Proposal: 02/14/2019 5:30AM*.

Denver Public Schools. (2019). 2018-19 Teacher Retention Update. https://www.dpsk12.org/wp-content/uploads/TeacherRetention_UPDATEDwithMYdata_19129.pdf

Duflo, E., Hanna, R., & Rya, S. P. (2012). Incentives work: Getting teachers to come to school. *The American Economic Review, 102*(4), 1241-1278.

Engel, M., Claessens, A., Watts, T., & Stone, S. (2016). Socioeconomic inequality at school entry: A cross-cohort comparison of families and schools. *Children and Youth Services Review, 71*, 227-232.

Fact Sheet: What is Senate Bill 10-191? Sopporting Great Teachers and Leaders. (2014). *Fact Sheets and FAQs.* Retrieved from https://www.cde.state.co.us/educatoreffectiveness/sb191factsheet

Far Northeast Teachers. (2019, 01/22/2019). We work at Denver's Title I schools, too. Here's why we're ready to strike. *Chalkbeat*. Retrieved from

https://www.chalkbeat.org/posts/co/2019/01/22/we-work-at-denvers-title-i-schools-too-heres-why-were-ready-to-strike/

Feng, L., & Sass, T. R. (2018). The Impact of Incentives to Recruit and Retain Teachers in "Hard-to-Staff" Subjects. *Journal of Policy Analysis and Management, 37*(1), 112-135.

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

Fryer, R. G. (2011). *Teacher incentives and student achievement: Evidence from New York City public schools*. Retrieved from

Fryer, R. G., Levitt, S. D., List, J., & Sadoff, S. (2012). *Enhancing the efficacy of teacher incentives through loss aversion: A field experiment*. Retrieved from

Fulbeck, E. S. (2014). Teacher Mobility and Financial Incentives A Descriptive Analysis of Denver's ProComp. *Educational Evaluation and Policy Analysis, 36*(1), 67-82.

Glazerman, S., Protik, A., Teh, B.-r., Bruch, J., & Max, J. (2013). Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment. NCEE 2014-4004. *National Center for Education Evaluation and Regional Assistance*.

Glazerman, S., & Seifullah, A. (2012). An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after Four Years. Final Report. *Mathematica Policy Research, Inc*.

Goldhaber, D., Bignell, W., Farley, A., Walch, J., & Cowan, J. (2016). Who chooses incentivized pay structures? Exploring the link between performance and preferences for compensation reform in the teacher labor market. *Educational Evaluation and Policy Analysis, 38*(2), 245-271.

Goldhaber, D., & Walch, J. (2012). Strategic pay reform: A student outcomes-based evaluation of Denver's ProComp teacher pay initiative. *Economics of Education Review, 31*(6), 1067-1083.

Gonring, P., Teske, P., & Jupp, B. (2007). *Pay-for-performance teacher compensation: An inside view of Denver's ProComp plan*: ERIC.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2013). *Strategic Staffing: Examining the Class Assignments of Teachers and Students in Tested and Untested Grades and Subjects*. CEPA Working Paper.

Hall, B. J., Lazear, E., & Madigan, C. (2000). Performance pay at Safelite Auto Glass (A).

Hess, F. (2019, 01/28/2019). Denver's Teacher Strike Puts Pay-For-Performance In The Spotlight. *Forbes Magazine*. Retrieved from https://www.forbes.com/sites/frederickhess/2019/01/28/denvers-teacher-strike-puts-pay-for-performance-in-the-spotlight/#2760b3bf2caf

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136.

Jerald, C. D. (2013). *Beyond Buy-In: Partnering with Practitioners to Build A Professional Growth and Accountability System for Denver's Educators*. Retrieved from http://careers.dpsk12.org/wp-content/uploads/2016/10/Stakeholder-Engagement-white-paper.pdf

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kiersz, A. (2018, 12/28/2018). The most and least expensive places to live in America. *Business Insider*.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy, 6*(1), 18-42.

Kremer, M., Ilias, N., & Glewwe, P. (2003). *Teacher incentives*. Retrieved from

Lazear, E. P. (1996). *Performance pay and productivity*. Retrieved from

Lazear, E. P. (2000). The future of personnel economics. *The Economic Journal, 110*(467), 611-639.

Leachman, M., Albares, N., Masterson, K., & Wallace, M. (2016). Most states have cut school funding, and some continue cutting. *Center on Budget and Policy Priorities, 4*.

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., & Epstein, S. (2011). *A big apple for educators: New York City's experiment with schoolwide performance bonuses: Final evaluation report*: Rand Corporation.

McCaffrey, D. F., Sass, T. R., Lockwood, J., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572-606.

Meyer, J. P. (2008, August 10, 2008). Merit pay splits DPS, union. *The Denver Post*. Retrieved from https://www.denverpost.com/2008/08/10/merit-pay-splits-dps-union/

Muralidharan, K., & Sundararaman, V. (2008). The Impact of School Block Grants on Student Learning Outcomes," Harvard.

Oliver, K., Innvar, S., Lorenc, T., Woodman, J., & Thomas, J. (2014). A systematic review of barriers to and facilitators of the use of evidence by policymakers. *BMC health services research, 14*(1), 2.

Papay, J. P. (2010). Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Educational Research Journal*.

Penn, K. (2015). *ProComp Teacher Perspectives on Design, Usability and Impact. Focus Group Findings, Implications & Recommendations* Retrieved from https://www.issuelab.org/resources/26768/26768.pdf

Penuel, W. R., & Gallagher, D. J. (2017). *Creating Research Practice Partnerships in Education*: ERIC.

Pivovarova, M., & Amrein-Beardsley, A. (2018). Median Growth Percentiles (MGPs): Assessment of Intertemporal Stability and Correlations with Observational Scores. *Educational Assessment, 23*(2), 139-155.

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

Proctor, D., Walters, B., Reichardt, R., Goldhaber, D., & Walch, J. (2011). Making a difference in education reform: ProComp external evaluation report 2006-2010. *Prepared for the Denver Public Schools*.

Ragan, K. (2019, 02/14/2019). Denver Public Schools teacher strike ends after three days. *The Coloradan*. Retrieved from https://www.coloradoan.com/story/news/education/2019/02/14/denver-public-schools-teacher-strike-ends/2842477002/

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537-571.

Slotnik, W. J., Smith, M. D., Glass, R., Helms, B., & Ingwerson, D. (2004). Catalyst for Change: Pay for Performance in Denver Final Report. *Boston: Community Training and Assistance Center*.

Springer, M. G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J., McCaffrey, D. F., . . . Stecher, B. M. (2011). Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT). *Society for Research on Educational Effectiveness*.

Springer, M. G., Swain, W. A., & Rodriguez, L. A. (2016). Effective teacher retention bonuses: Evidence from Tennessee. *Educational Evaluation and Policy Analysis, 38*(2), 199-221.

Turkewitz, J., & Goldstein, D. (2019, 02/11/2019). Denver Teachers' Strike Puts Performance-Based Pay to the Test. *New York Times*. Retrieved from https://www.nytimes.com/2019/02/11/us/denver-teacher-strike.html

U.S. Department of Education. (2018). National Center for Education Statistics, Elementary/Secondary Information System. Retrieved from https://nces.ed.gov/ccd/elsi/

Van Dam, A. (2019, 02/14/2019). Teacher strikes made 2018 the biggest year for worker protest in a generation. *Washington Post*. Retrieved from https://www.washingtonpost.com/us-policy/2019/02/14/with-teachers-lead-more-workers-went-strike-than-any-year-since/

Will, M. (2019, 02/14/2019). Denver Teachers' Union and District Reach Deal to End Strike. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/teacherbeat/2019/02/denver_strike_ends.html

Yuan, K., Le, V.-N., McCaffrey, D. F., Marsh, J. A., Hamilton, L. S., Stecher, B. M., & Springer, M. G. (2013). Incentive pay programs do not affect teacher motivation or reported practices: Results from three randomized studies. *Educational Evaluation and Policy Analysis, 35*(1), 3-22.

Endnotes

[1] It is important to complicate the term "effective", which appears in several of the research questions above. In this study, we measure teacher effectiveness using the measure used by DPS, teacher median growth percentile (MGP) scores—one approach to estimating teacher value-added. However, the research base is mixed on such measures. In the Methods section when we introduce the MGP scores, we highlight the important research that has debated the appropriateness of using value-added and/or MGP scores to characterize teacher effectiveness for policy decisions. In addition to examining teacher outcomes by MGP scores, we conduct sensitivity analyses using other proxies for teaching effectiveness including value-added scores, teacher age, years of experience, and highest degree earned.

[2] The PFP theory of action posits that such systems could stimulate teachers to improve their performance. We explored this by examining changes in teachers' annual MGP scores from before- to after- opting in to ProComp (in essence running an ITS model within-teacher by adding teacher fixed effects). However, because all new hires after the onset of ProComp were automatically enrolled in the program (and not observed prior), we can only conduct this analysis for teachers who were hired before the onset of ProComp, opted in to the system, and have at least one MGP score both before and after opting in. These analytic restrictions raise immediate concerns over selection bias, generalizability, and sample size. In fact, only 6.4% of teachers meet the data requirements for inclusion. Estimated level and trend effects were not significant, did not follow a clear pattern, and were not practically meaningful in magnitude. We therefore do not present these results (available upon request). While this was the best approach available, this context cannot provide insight into this question.

[3] Some ProComp polices also apply to other job categories. We focus on teachers only.

[4] A student's current performance is compared with that of students with matching test score performance in the previous two to three years (one year for fourth graders) and then translated to percentiles, which become each student's SGP score.

[5] Using MGP scores as an outcome has many complications: DPS did not produce MGP scores prior to 2004-05, nor in 2015 (due to the switch to the PARCC exam). Further, not all teachers *have* MGP scores, because they do not teach in tested subjects or grades. Analyses that rely on MGP scores are systematically limited to the kinds of teachers who have MGP scores. Other research suggests that which teachers do and do not have value-added scores is likely not random (Atteberry, Loeb, & Wyckoff, 2016; Grissom, Kalogrides, & Loeb, 2013). To test the robustness of any findings with respect to MGP of arrivers, we use MGP scores in several different ways to characterize teacher effectiveness. See Online Appendix B and C for a full description of these approaches, and an analysis of how the analytic sample is affected by these choices.

[6] To estimate teacher value added measures (VAM), we first standardize test scores within subject-grade-year. Models are run separately by subject. Our VAM model specifies a student's current achievement score as a function of a prior year achievement score, teacher-by-year fixed effects (which become the VAM scores), grade and year fixed effects, and of typically-available student and school covariates from the district administrative records—the same that are used in controls in the main analyses herein). We find that teachers' MGP and VAM score (aggregated across subjects and years) are correlated at 0.61.

[7] We use CDE-reported state-level test score means and standard deviations (for each year by subject and grade) to standardize both the student-level test scores and district-level mean test scores within subject-grade-year to facilitate achievement analyses that span Colorado districts.

[8] The timing of these periods depends on the outcome of interest. The first full school year of PC1 was 2006-07. PC2 was approved in August 2008, and we hypothesize that PC2 could first affect student outcomes in 2008-09, but teacher recruitment, retention, and transfers at the start of 2009-10. The final year of available data varies across outcomes, between 2014 and 2017.

[9] $Time_{iy}$ captures linear time measured in years, centered at the midpoint of each of the three time periods. For instance, when the outcome is student achievement, $Time_{iy}$ is coded as follows: Pre-ProComp (5 yrs.): 2002= -2.0, 2003= -1.0, 2004= 0.0, 2005= 1.0, 2006= 2.0; PC1 (2 yrs.): 2007= -0.5, Midpoint of PC1= 0,

2008= 0.5; PC2 (8 yrs.): 2009= -3.5, 2010= -2.5, 2011= -1.5, 2012=0.5 Midpoint of PC2= 0, 2013= 0.5, 2014= 1.5, 2015= 2.5, 2016=3.5. See Figure 3 for a recap.

[10] The vector of student control variables includes dummy variables for student sex, race/ethnicity categories, enrollment in Gifted and Talented programs, English Language Learner status, special education designation, grade, and eligibility for free- or reduced-price lunch).

[11] We selected control variables both based on prior literature and based on the data available to us that improves the model's ability to separate possible causal effects of the policy on outcome trends from potential confounding factors, such as demographic shift in the student population over time. For teacher-level analyses, only the school-level covariates, $\boldsymbol{W_{sy}}$,can be included.

[12] Quasi-experimental methods based on the counterfactual model of causality also require the stable unit treatment value assumption—each unit has only one potential outcome per treatment status.

[13] We also considered—but ultimately chose not to use— several other definitions of "comparable districts" from which to construct the comparison group: We examined whether nearby districts exhibited similar mean trends in retention and achievement in the pre-period but found that both the level and trends in outcomes were quite dissimilar to DPS. When we compared DPS to *all* other Colorado districts, pre-period trends and levels were not comparable.

[14] Previously, teachers automatically exited the probationary period upon tenure, the statute requires that teacher effectiveness be evaluated (with at least 50% of the evaluation coming from measures of student growth) prior to conferring non-probationary status.

[15] In terms of racial/ethnic segregation statistics, 50% of White individuals in Denver would need to change census tracts to equalize racial distribution. In terms of income inequality, the top 90th percentile of Denver's income distribution earns 11.7 times more than bottom 10th percentile—a greater disparity that San Francisco or Los Angeles, and on par with New York City or Chicago.

[16] In an external report on ProComp from 2016 involving interviews, teachers said, "I have no idea what my paycheck means…What is in the miscellaneous line on my pay check?" (Penn, 2015).

Tables

Table 1. DPS Average Student Demographics, Every Other Year 2001-02 through 2015-16

	SY 2001-02	SY 2003-04	SY 2005-06	SY 2007-08	SY 2009-10	SY 2011-12	SY 2013-14	SY 2015-16
% Asian	3%	3%	3%	3%	4%	3%	3%	3%
% Black	20%	19%	18%	18%	16%	14%	14%	13%
% Hispanic	55%	57%	58%	56%	54%	58%	58%	56%
% White	21%	20%	20%	21%	25%	20%	21%	23%
% FRPL Eligible	62%	62%	65%	66%	71%	72%	72%	68%
% Designated LEP	25%	30%	36%	26%	29%	31%	31%	26%
Total Enrollment	72,361	72,100	72,312	73,053	77,267	80,863	86,046	90,235

FN: Source: U.S. Department of Education, National Center for Education Statistics, Common Core of Data (CCD), "Local Education Agency (School District) Universe Survey", 2016-17 v.1a; "Local Education Agency (School District) Universe Survey Directory Data", 2015-16 v.1a; "Local Education Agency (School District) Universe Survey Geographic Data (EDGE)", 2015-16 v.1a; "Public Elementary/Secondary School Universe Survey Membership Data", 2015-16 v.2a.

Table 2. ProComp Participation Rates, by School Year

Year	Total Eligible Teachers	Participating Teachers	Percent Participating	ProComp Status
SY'01-'02	4,223	0	0%	Not enacted
SY'02-'03	4,413	0	0%	Not enacted
SY'03-'04	3,883	0	0%	Not enacted
SY'04-'05	3,848	0	0%	Not enacted
SY'05-'06	3,794	522	14%	ProComp 1.0 enacted Nov '05
SY'06-'07	3,793	1361	36%	ProComp 1.0
SY'07-'08	3,859	1855	48%	ProComp 1.0
SY'08-'09	4,044	2611	65%	ProComp 2.0 enacted Aug '08
SY'09-'10	4,220	2934	70%	ProComp 2.0
SY'10-'11	4,290	3212	75%	ProComp 2.0
SY'11-'12	4,409	3620	82%	ProComp 2.0
SY'12-'13	4,484	3794	85%	ProComp 2.0
SY'13-'14	4,646	4024	87%	ProComp 2.0
SY'14-'15	4,826	4183	87%	ProComp 2.0
SY'15-'16	4,929	4166	85%	ProComp 2.0
SY'16-'17	4,922	4435	90%	ProComp 2.0

Table 3. ProComp Incentive Structure, Receipt Rate, and Average Amount in PC1 and PC2

				% That Received Given Bonus		Average Amount ($) for Given Bonus	
Bonus Name	**Target**	**Category**	**Incentive Type**	PC 1.0	PC 2.0	PC 1.0	PC 2.0
Top Performing School	Schoolwide	Student Growth	One-Time Bonus	11%	37%	$397	$2,358
High Growth School	Schoolwide	Student Growth	One-Time Bonus	--	42%	--	$2,362
Hard to Serve (Hi FRL) School	Schoolwide	Market Incentives	One-Time Bonus	34%	56%	$958	$2,327
Hard to Staff Position	Individual	Market Incentives	One-Time Bonus	23%	33%	$974	$2,292
Exceeds SGP Expectations	Individual	Student Growth	One-Time Bonus	--	12%	--	$2,390
Comprehensive Evaluations	Individual	LEAP Evaluation	Base Building	2%	50%	$356	$547
Met 1 or 2 SGOs	Individual	Student Growth	One-Time Bonus	12%	69%	$361	$375
Advanced Degree	Individual	Knowledge & Skills	Base Building	8%	8%	$1,083	$3,570
Tuition/Stud Loan Reimb.	Individual	Knowledge & Skills	Combination	10%	26%	$763	$950
PDUs	Individual	Knowledge & Skills	Combination	53%	55%	$754	$834

FN: PC 1.0 indicates ProComp 1.0, and PC 2.0 indicates ProComp 2.0. Comprehensive Evaluations and SGOs only began in the final year of ProComp 1.0.

Table 4. CITS Results: Student Achievement Outcomes in ELA, Math, and Writing

	Relative to…			Relative to…			Relative to…		
	(1) …PSM Group	**(2) …Group Matched on Pre-Trends**	**(3) … SCM Group**	**(1) …PSM Group**	**(2) …Group Matched on Pre-Trends**	**(3) … SCM Group**	**(1) …PSM Group**	**(2) …Group Matched on Pre-Trends**	**(3) … SCM Group**
	ELA			**Math**			**Writing**		
Constant: Comparison Pre-Mean	-0.010	0.051	-0.599	-0.024	0.056	-0.663	-0.020	0.034	-0.581
Denver Dummy (1= Denver, 0 = Comparison)	-0.584	-0.645	0.005	-0.640	-0.720	0.000	-0.559	-0.613	0.002
Denver x ProComp 1.0 *(Level Effect)*	0.020	0.031	-0.183	0.069	0.112	-0.024	0.040	0.060	-0.090
Denver x ProComp 2.0 *(Level Effect)*	0.159	0.152	-0.005	0.231	0.282	0.104	0.212	0.226	-0.010
Denver x Time *(Linear, Pre-Period)*	0.010	-0.009	0.000	0.023	0.014	0.000	-0.015	-0.033	-0.001
Denver x Time x ProComp 1.0 *(Trend Effect)*	0.087	0.114	0.117	0.003	0.042	0.090	0.076	0.117	0.083
Denver x Time x ProComp 2.0 *(Trend Effect)*	0.023	0.044	0.104	0.018	0.025	0.107	0.060	0.081	0.087
Adjusted R^2	0.997	0.997	0.888	0.999	0.998	0.699	0.998	0.997	0.921
N	22	22	22	22	22	22	22	22	22

FN: The outcome is mean achievement in a given subject in year y for district d. Analyses are run on a dataset that includes mean outcomes for two group—DPS and the given comparison group—in each of 11 school years (N=22). As discussed in the narrative, we focus on the pattern of results, rather than inferential statistics. Results are shown for three separate models, in which the comparison group is defined differently (see Online Appendix D). For the sake of parsimony, only the intercept, Denver main effect (dummy indicating the group-year observation is for Denver (0= comparison group)), and Denver dummy interactions on the time function are shown. PC1 level and trend effects are highlighted in lighter grey; PC2 in darker grey. Positive interaction coefficients indicate greater level- or trend- differences in Denver than in the given comparison group.

Table 5. ITS Results: MGP of New-to-DPS Teachers

	Outcome MGP Defintion #1	**Outcome MGP Defintion #2**	**Outcome MGP Defintion #3**
Pre-Period Mean	-0.091 **	-0.097 **	-0.057
	(0.032)	(0.033)	(0.033)
ProComp 1.0 *(Level Effect)*	0.115 **	0.128 **	0.048
	(0.042)	(0.043)	(0.042)
ProComp 2.0 *(Level Effect)*	0.146 ***	0.173 ***	0.085 *
	(0.037)	(0.038)	(0.037)
Time (Linear, Pre-Period)	-0.008	0.006	-0.026
	(0.026)	(0.026)	(0.026)
Time x ProComp 1.0 *(Trend Effect)*	0.070	0.050	0.089 *
	(0.042)	(0.044)	(0.042)
Time x ProComp 2.0 *(Trend Effect)*	0.013	0.005	0.036
	(0.027)	(0.028)	(0.027)
ProComp 1.0 Mean	0.025	0.031	-0.009
ProComp 2.0 Mean	0.055	0.076 ***	0.028
ProComp 1.0 Trend	0.062	0.055	0.063
ProComp 2.0 Trend	0.005	0.010	0.010
Adjusted R-Squared	0.003	0.004	0.001
N	5850	5208	5850
School Covs?	Yes	Yes	Yes

FN: Samples are limited to teachers who are new to DPS and have MGP score data available. The outcome is the MGP score of a given arriving teacher, and it has been standardized at the teacher level. Coefficients are therefore reported in teacher standard deviations. All models include time varying, school-level control variables (% male, white, Black, Hispanic, Asian, Native American, other race, GATE, SPED, FRL, RDL, and total enrollment). Results are shown across three different models, in which the way MGP scores are used to characterize the effectiveness of incoming teachers varies. See Online Appendix B for more. Briefly: For Definition 1 we use the MGP a teacher actually received in the year of his/her arrival. For Definition 2 we use HLM to estimate an MGP for teachers in their 3rd year in the district. For Definition 3 we take the mean of all MGPs earned by a teacher over his/her observed tenure. We also control for the available number of MGPs.

.

Table 6. CITS Results: Teacher Retention Outcomes in DPS

	(1) Relative to PSM Group	**(2) Relative to Group Matched on Pre-Trends**	**(3) Relative to SCM Group**
Constant: Comparison Pre-Mean	81.5%	82.9%	81.8%
Denver Dummy (1= Denver, 0 = Comparison)	-0.04%	-1.48%	-0.32%
Denver x ProComp 1.0 *(Level Effect)*	-1.06%	-1.08%	0.57%
Denver x ProComp 2.0 *(Level Effect)*	-1.53%	-1.37%	0.76%
Denver x Time (Linear, Pre-Period)	-0.56%	0.22%	-0.02%
Denver x Time x ProComp 1.0 *(Trend Effect)*	-0.62%	-1.02%	0.35%
Denver x Time x ProComp 2.0 *(Trend Effect)*	-0.31%	-0.94%	-0.37%
Adjusted R^2	0.487	0.460	0.430
N	32	32	32

FN: The outcome of interest in the teacher retention rate in year y for district d. Coefficients are reported in percentage points for ease of interpretation (e.g., the constant for column (1) is 0.818). We do so because presenting changes in proportional trends often places the desired information third or fourth decimal position. Analyses are run on a dataset that includes mean outcomes for two group—DPS and the given comparison group—in each of 16 school years (N=32). As discussed in the narrative, we focus on the pattern of results, rather than inferential statistics. Results are shown for three separate models, in which the comparison group is defined differently (see Online Appendix D). For the sake of parsimony, only the intercept, Denver main effect, and Denver interactions are shown. Positive coefficients on the Denver x PC1 and PC2 (and x Time) variables indicate Denver exhibited higher retention rates (levels and trends) than the comparison group. PC1 level and trend effects are highlighted in lighter grey; PC2 in darker grey.

Table 7. Teacher Retention in the District, as a Function of Total ProComp Incentive Payout from Prior Year

	M1	M2	M3	M4	M5
ProComp Bonus Y-1	0.65% ***	0.34% ***	0.26% ***	0.29% ***	0.78% ***
	(0.05%)	(0.06%)	(0.06%)	(0.07%)	(0.08%)
Teacher MGP Score			2.6% ***	1.9% ***	1.9% ***
			(0.28%)	(0.30%)	(0.30%)
Intercept	83.1% ***	86.8% ***	87.2% ***	87.0% ***	95.9% ***
	(0.28%)	(0.41%)	(0.41%)	(0.42%)	(1.59%)
Adjusted R-Squared	0.004	0.001	0.005	-0.005	0.009
N	47,874	21,348	21,348	21,348	21,348
Limit to MGP Teachers		X	X	X	X
Control for MGP?			X	X	X
School FEs				X	X
Year FEs					X

FN: The outcome is a dummy variable, indicating whether a given teacher returned to DPS in a given year. Coefficients and their standard errors are reported in percentage points for ease of interpretation (e.g., the first coefficient for column (1) is 0.0065). We do so because presenting changes in proportional trends often places the desired information third or fourth decimal position. Total ProComp incentive payouts are reported in $1000 increments so that a one-unit different in the bonus amount equals $1000. MGP scores are scaled such that a one unit difference equals a 10 MGP-point difference on the original 0 to 100 scale, and an MGP of zero equals an MGP of 50. The intercept therefore represents the predicted retention for an MGP=50 teacher who received $0 in ProComp bonuses in the prior year. The M1 sample is limited to ProComp participants. Models M2 through M5 are is limited to ProComp participants eligible for MGP scores (in tested subjects and grades).

Table 8. ITS: Teacher Retention, Separately for High, Middle, and Low MGP Teachers

	All Eligible Participants in MGP Grades	MGP Definition A			MGP Definition B		
		Top Third	Middle Third	Bottom Third	Top Third	Middle Third	Bottom Third
Pre-Period Mean	96.6% ***	96.3% ***	96.3% ***	96.2% ***	98.1% ***	96.1% ***	96.1% ***
	(0.33%)	(1.09%)	(0.44%)	(1.14%)	(0.81%)	(0.54%)	(0.74%)
ProComp 1.0 *(Level Effect)*	-8.8% ***	-8.8% ***	-7.5% ***	-10.4% ***	-7.1% ***	-8.0% ***	-8.9% ***
	(0.43%)	(1.44%)	(0.58%)	(1.49%)	(1.03%)	(0.71%)	(0.99%)
ProComp 2.0 *(Level Effect)*	-14.9% ***	-12.6% ***	-14.1% ***	-21.0% ***	-13.0% ***	-14.5% ***	-18.7% ***
	(0.39%)	(1.26%)	(0.52%)	(1.39%)	(0.91%)	(0.63%)	(0.94%)
Time (Linear, Pre-Period)	0.13%	-0.60%	0.06%	-0.11%	-0.70%	0.27%	-0.05%
	(0.27%)	(0.87%)	(0.35%)	(0.95%)	(0.65%)	(0.44%)	(0.60%)
Time x ProComp 1.0 *(Trend Effect)*	-1.9% ***	-2.4%	-2.2% ***	-5.6% ***	-1.3%	-2.4% **	-4.6% ***
	(0.45%)	(1.49%)	(0.59%)	(1.55%)	(1.04%)	(0.73%)	(1.02%)
Time x ProComp 2.0 *(Trend Effect)*	-2.7% ***	-1.4%	-2.5% ***	-2.6% **	-1.4% *	-2.7% ***	-3.3% ***
	(0.28%)	(0.90%)	(0.37%)	(1.00%)	(0.66%)	(0.45%)	(0.63%)
ProComp 1.0 Mean	90.8% ***	90.5% ***	91.8% ***	88.8% ***	94.0% ***	91.1% ***	90.3% ***
ProComp 2.0 Mean	84.7% ***	86.7% ***	85.2% ***	78.2% ***	88.1% ***	84.6% ***	80.5% ***
ProComp 1.0 Trend	-1.81% ***	-3.01% *	-2.15% ***	-5.73% ***	-1.99% *	-2.12% ***	-4.62% ***
ProComp 2.0 Trend	-2.52% ***	-1.96% ***	-2.38% ***	-2.72% ***	-2.08% ***	-2.45% ***	-3.38% ***
Adjusted R-Squared	0.038	0.034	0.036	0.065	0.034	0.036	0.065
N	57672	10638	14880	10572	16568	18820	15592
School Covs?	Yes	Yes	Yes	Yes	Yes	Yes	Yes

FN: The outcome is a dummy variable (1= teacher returned to DPS in a given year). For readability, coefficients and standard errors are reported in percentage points. The analysis is limited to teachers-years with observed MGP scores (i.e., in tested subjects and grades). We separate teachers into the top, middle, and bottom third of the MGP score distribution and conduct the analysis twice—using MGP Definition A, and then using MGP Definition B. See Online Appendix C for full details. Briefly: Def. A: Mean of any available MGPs (controlling for the # available). Def. B: An HLM-based estimate of MGP. Models include time varying, school-level control variables.

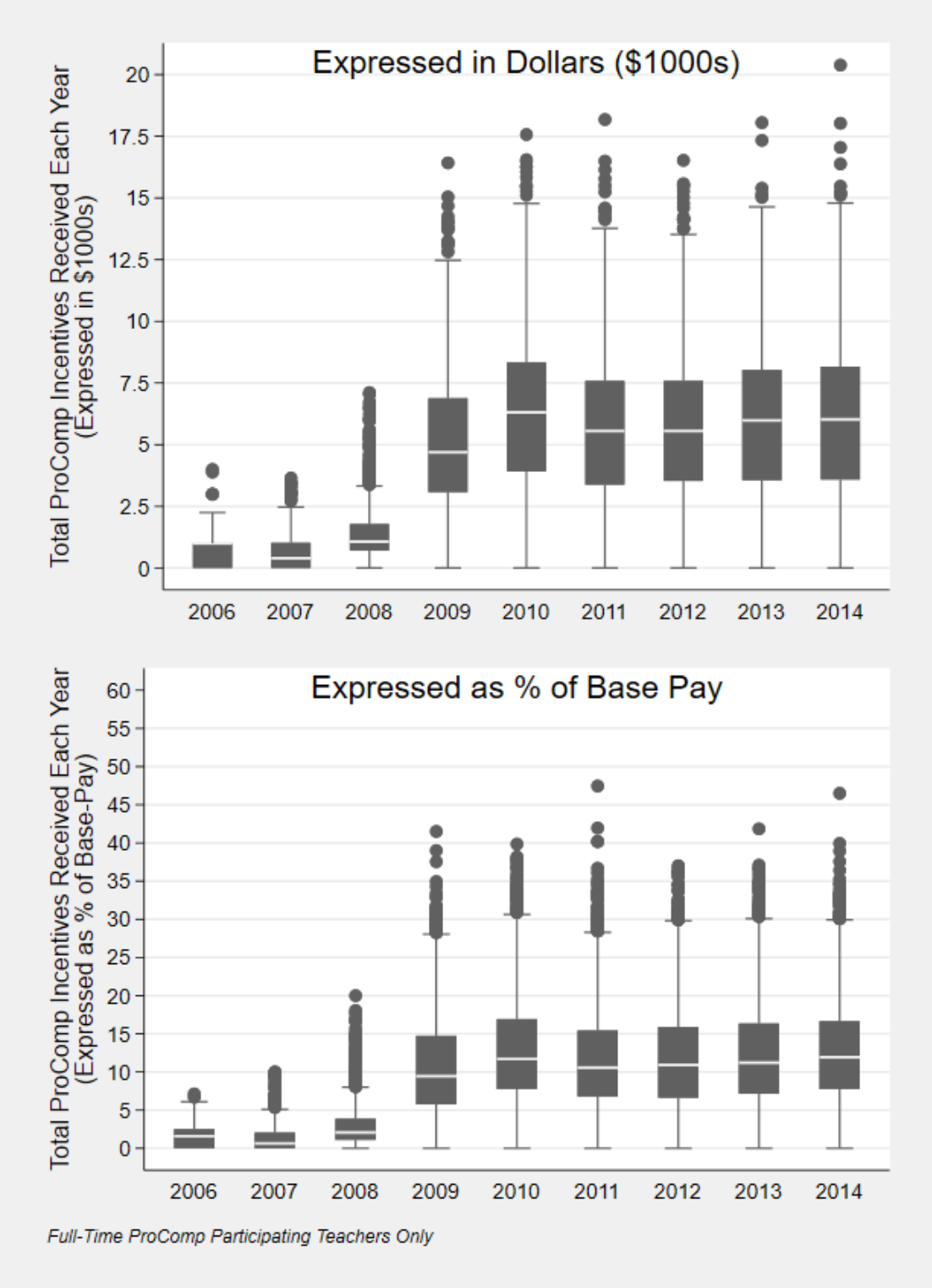
Table 9. ITS: Retention of High-MGP Teachers, Separately for Groups of Schools Eligible for ProComp Schoolwide Incentives

	Eligible for Any School Bonus	Inegible for Any School Bonuses	High Growth Schools	Top Performing Schools	Hard-To-Serve Schools
Pre-Period Mean	95.5% ***	97.9% ***	96.1% ***	96.1% ***	94.3% ***
	(0.74%)	(2.18%)	(0.79%)	(0.80%)	(1.07%)
ProComp 1.0 *(Level Effect)*	-6.9% ***	-6.3% *	-7.2% ***	-7.9% ***	-5.4% ***
	(0.93%)	(2.87%)	(0.98%)	(0.99%)	(1.35%)
ProComp 2.0 *(Level Effect)*	-11.4% ***	-14.6% ***	-11.1% ***	-11.0% ***	-12.1% ***
	(0.84%)	(2.74%)	(0.88%)	(0.89%)	(1.20%)
Time (Linear, Pre-Period)	1.0%	2.7%	1.1%	0.6%	1.1%
	(0.82%)	(2.58%)	(0.87%)	(0.88%)	(1.19%)
Time x ProComp 1.0 *(Trend Effect)*	-1.2%	-5.7%	-2.0%	-1.2%	0.3%
	(1.10%)	(3.64%)	(1.15%)	(1.16%)	(1.60%)
Time x ProComp 2.0 *(Trend Effect)*	-3.0% ***	-8.3% **	-3.0% ***	-2.4% **	-3.4% **
	(0.84%)	(2.66%)	(0.88%)	(0.89%)	(1.21%)
ProComp 1.0 Mean	91.6% ***	94.6% ***	91.8% ***	91.2% ***	91.9% ***
ProComp 2.0 Mean	87.1% ***	86.4%	88.0% ***	88.1% ***	85.2% ***
ProComp 1.0 Trend	-0.2%	-3.0%	-0.9%	-0.6%	1.4%
ProComp 2.0 Trend	-2.0% ***	-5.6% ***	-1.9% ***	-1.8% ***	-2.3% ***
Adjusted R-Squared	0.027	0.085	0.027	0.025	0.029
N	15036	3828	13084	12932	8230
School Covs?	Yes	Yes	Yes	Yes	Yes

FN: *The outcome is a dummy variable, indicating whether a given teacher returned to DPS in a given year. The sample is limited to teachers with MGP scores in the top third of the MGP score distribution (using MGP definition A from Table 8 and described in Online Appendix C). For readability, coefficients and standard errors are reported in percentage points. We do so because presenting changes in proportional trends often places the desired information third or fourth decimal position. Column 1 includes teacher-year observations from schools that ever became eligible for one of the three ProComp schoolwide incentives. Column 2 includes teacher-year observations from schools that were never eligible for any of the three ProComp schoolwide incentives. Columns 3 – 5 separate the schools by the particular schoolwide bonus for which they were ever eligible: High Growth schools, Top Performing schools, or Hard to Serve schools. These groups are not necessarily mutually exclusive. PC1 level and trend effects are highlighted in lighter grey; PC2 in darker grey.*

Figures

Figure 1. Box Plot of Total ProComp Incentives (in Dollars, % of Base Pay), by School Year



FN: Sample is limited to teachers participating in ProComp that are employed full-time.

Figure 2. Overview of ProComp Publications

Author (Year)	Fulbeck (2013)	Goldhaber & Walch (2012)	Goldhaber, Bignell, Farley, Walch, and Cowan (2016)	Current Study	
Outcomes	Teacher retention/ attrition (in schools & in district)	Student achievement (math and reading); 0/1 teacher receives a given ProComp reward in yr y or not	0/1 teacher decides to opt-in to ProComp or not.	Student Outcomes: Student achievement (ELA, math, writing)	Teacher Outcomes: MGP of recruited teachers; 0/1 teacher returns to DPS in yr y or not
Y1 of Outcome Data	2000-01 DPS teacher attrition	2002-03 DPS student test scores	2005-06 ProComp opt-in decisions	2001-02 DPS student test scores; 2003-04 district mean scores across CO.	2000-01 district turnover data across CO; 2002-03 DPS quality recruitment and retention data
Last Yr of Outcome Data	2009-10 DPS teacher attrition	2009-10 DPS student test scores	2009-10 ProComp opt-in decisions	2015-16 DPS student test scores; 2013-14 district mean scores across CO	2015-16 district turnover data across CO; 2015-16 quality recruitment; 2016-17 retention
Years Pre-PC	6	4	n/a	5 (CITS); 3 (in DPS)	6 (CITS); 5 (in DPS)
Years of PC 1	3 (all)	2 (all)	3 (all)	2 (all)	3 (all)
Years of PC 2	1	1*	1	5*	7 (CITS, recruitment); 8 (retention in DPS)
Research Questions:	(1) To what extent is ProComp associated with changes in teacher turnover from all schools in the district and from high-poverty schools specifically? (2) How do the effects of ProComp vary for teachers' within-district moves and teachers' exits from the district?	(1) How does student achievement change with onset of PC? (2) How effective are teachers who opt in to Pro-Comp, compared to those who choose not to? (3) How well does the allocation of rewards in the ProComp system correspond to VAMs of teacher effectiveness?	(a) What type of incentive structure is more appealing to teachers? (b) To what extent do teacher characteristics and reward eligibility predict decisions to enroll in ProComp? (c) Are more effective teachers, defined as those who are more likely to receive a PFP reward, more likely to opt in?	(1) Is ProComp associated with improvements in student achievement (ELA, math, writing)?	(2) Is ProComp associated with the recruitment of more effective teachers to DPS? (3) Did ProComp affect teacher retention overall (3A) or differentially by effectiveness (3B)?
Method	Hazard models to conduct a descriptive analysis of teacher mobility across teachers who did and did not receive a ProComp incentive, with special attention to teacher mobility in high-poverty schools.	OLS models (sometimes with teacher fixed effects framework) to predict student test scores as a function of ProComp onset, and teachers' decisions to opt in to ProComp.	Logit model to predict opt in decision as a function of teacher characteristics and reward eligibility and likelihood to receive PFP rewards.	CITS-based OLS model to predict student achievement using achievement patterns within DPS relative to comparable, non-DPS districts	CITS-based OLS model to predict retention using achievement patterns within DPS relative to comparable, non-DPS districts; ITS of retention by teacher MGP scores
Use Teacher VA/ MGP Effective-ness Measures ?	No	N/A for student achievement analysis. Yes, used to predict likelihood of ProComp opt-in and specific reward receipt.	Yes, used to predict likelihood of ProComp opt-in	N/A for student achievement analysis.	Yes, both as an outcome and as a predictor of retention

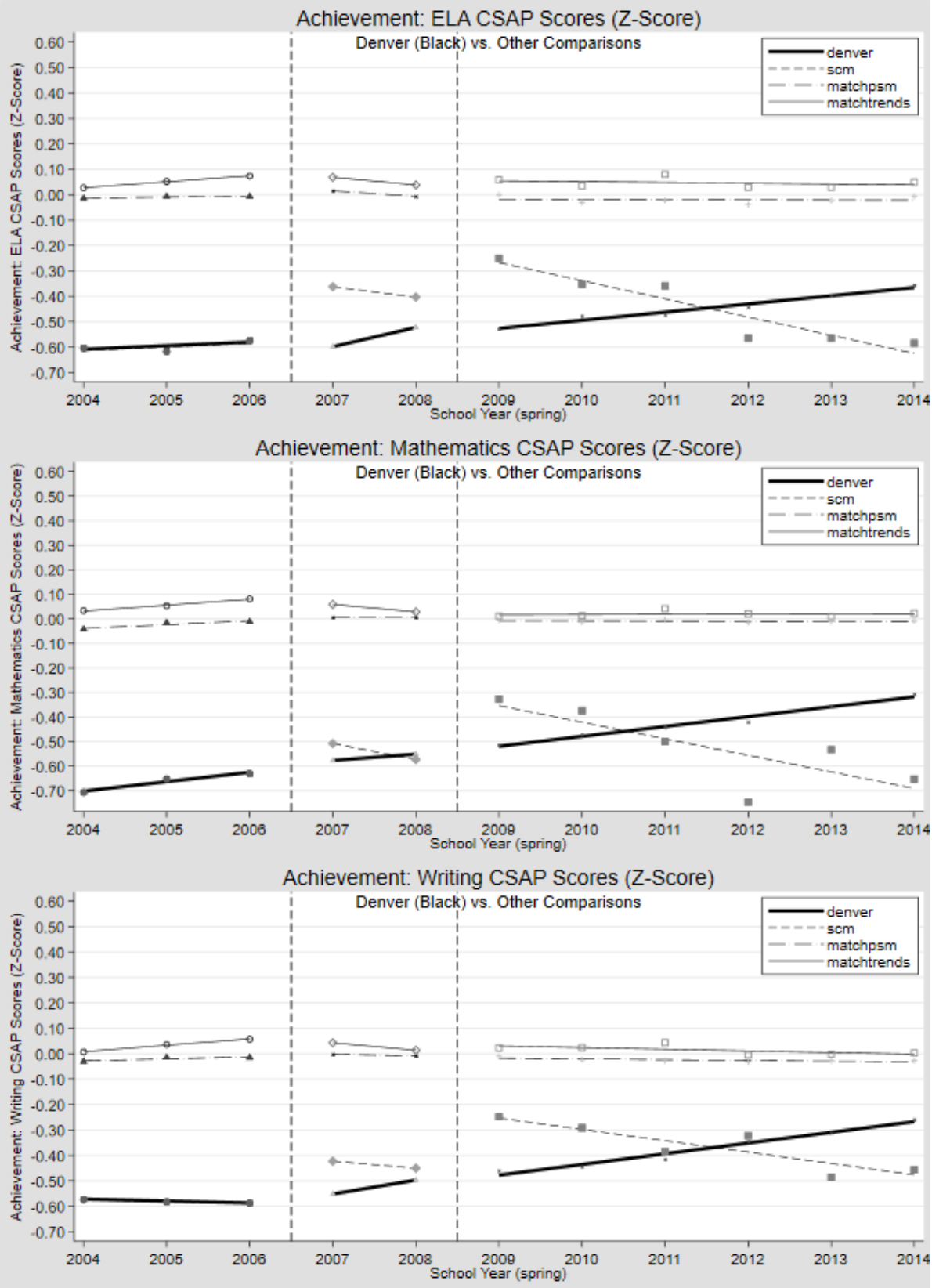
*FN: Description of research questions reflect those stated in these publications. *It is not clear whether PC2--which as approved in August 2008-- would have yet had a chance to affect student outcomes in spring of 2008-09.*

Figure 3. Model Variable Coding, Period-Specific Equations, and Predicted Outcomes

Year	PC1	PC2	Time	B0	B1	B2	B3	B4	B5	Equation for Period	Predicted Outcome
2002	0	0	-2	[B0		] +	[B3*(-2)		]		B0+B3(-2)
2003	0	0	-1	[B0		] +	[B3*(-1)		]		B0+B3(-1)
Midpoint of Pre-: 2004	**0**	**0**	**0**	[**B0**		]				B0+B3(Time)	B0
2005	0	0	1	[B0		] +	[B3*(1)		]		B0+B3(1)
2006	0	0	2	[B0		] +	[B3*(2)		]		B0+B3(2)
2007	1	0	-0.5	[B0	+ B1	] +	[B3*(-0.5)	+ B4*(-0.5)	]		[B0_B1]+[B3+B4](-0.5)
Midpoint of ProComp 1.0: 1	**0**	**0**	**0**	[**B0**	+ **B1**	]				[B0+B1]+[B3+B4](Time)	[B0_B1]
2008	1	0	0.5	[B0	+ B1	] +	[B3*(0.5)	+ B4*(0.5)	]		[B0_B1]+[B3+B4](0.5)
2009	0	1	-3.5	[B0	+ B2] +	[B3*(-3.5)	+	B5*(-3.5)]			[B0_B2]+[B3+B4](-3.5)
2010	0	1	-2.5	[B0	+ B2] +	[B3*(-2.5)	+	B5*(-2.5)]			[B0_B2]+[B3+B4](-2.5)
2011	0	1	-1.5	[B0	+ B2] +	[B3*(-1.5)	+	B5*(-1.5)]			[B0_B2]+[B3+B4](-1.5)
2012	0	1	-0.5	[B0	+ B2] +	[B3*(-0.5)	+	B5*(-0.5)]			[B0_B2]+[B3+B4](-0.5)
Midpoint of ProComp 2.0: 0	**1**	**0**	**0**	[**B0**	+ **B2**]					[B0+B2]+[B3+B5](Time)	[B0_B2]
2013	0	1	0.5	[B0	+ B2] +	[B3*(0.5)	+	B5*(0.5)]			[B0_B2]+[B3+B4](0.5)
2014	0	1	1.5	[B0	+ B2] +	[B3*(1.5)	+	B5*(1.5)]			[B0_B2]+[B3+B4](1.5)
2015	0	1	2.5	[B0	+ B2] +	[B3*(2.5)	+	B5*(2.5)]			[B0_B2]+[B3+B4](2.5)
2016	0	1	3.5	[B0	+ B2] +	[B3*(3.5)		B5*(3.5)]			[B0_B2]+[B3+B4](3.5)

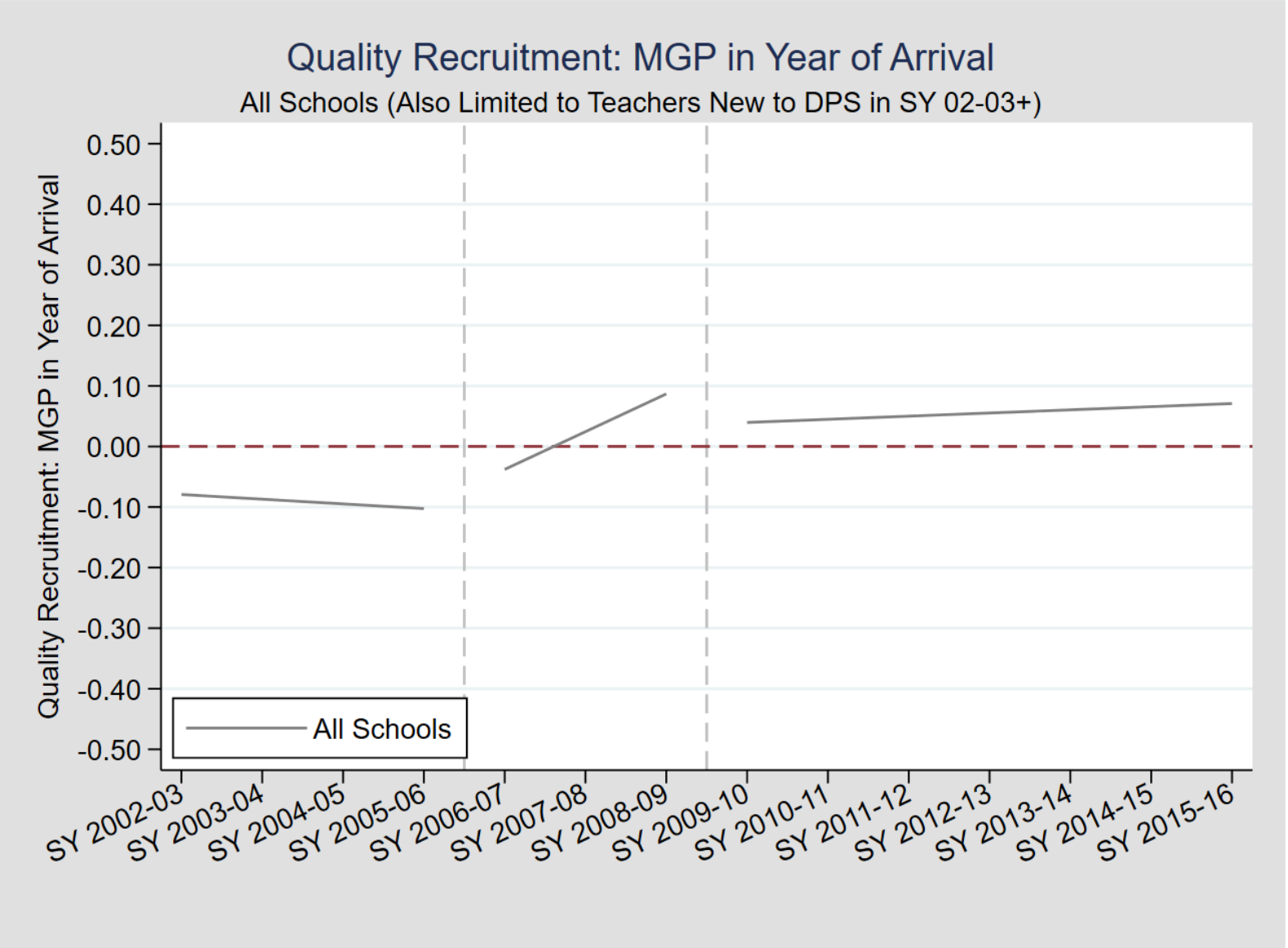
FN: For reference, Equation (1) is as follows $Y_{iy} = \beta_0 + \beta_1(PC1_{iy}) + \beta_2(PC2_{iy}) + \beta_3(Time_{iy}) + \beta_4(Time_{iy} \times PC1_{iy}) + \beta_5(Time_{iy} \times PC2_{iy}) + X_{i(y)}\beta + W_{sy}\beta + \varepsilon_{iy}$. *Columns PC1, PC2, and Time correspond to the variables entered into the model. B0, B1, and B2 clarify which coefficients are multiplied by 1 or 0 and thus contribute to the intercept for a given time period. B3, B4, and B5 clarify which coefficients contribute to the estimated slope for a given time period. We also write out the predicted outcome for each value of time.*

Figure 4. CITS for DPS vs. Comparison Districts: Student Achievement (ELA, Math, Writing)



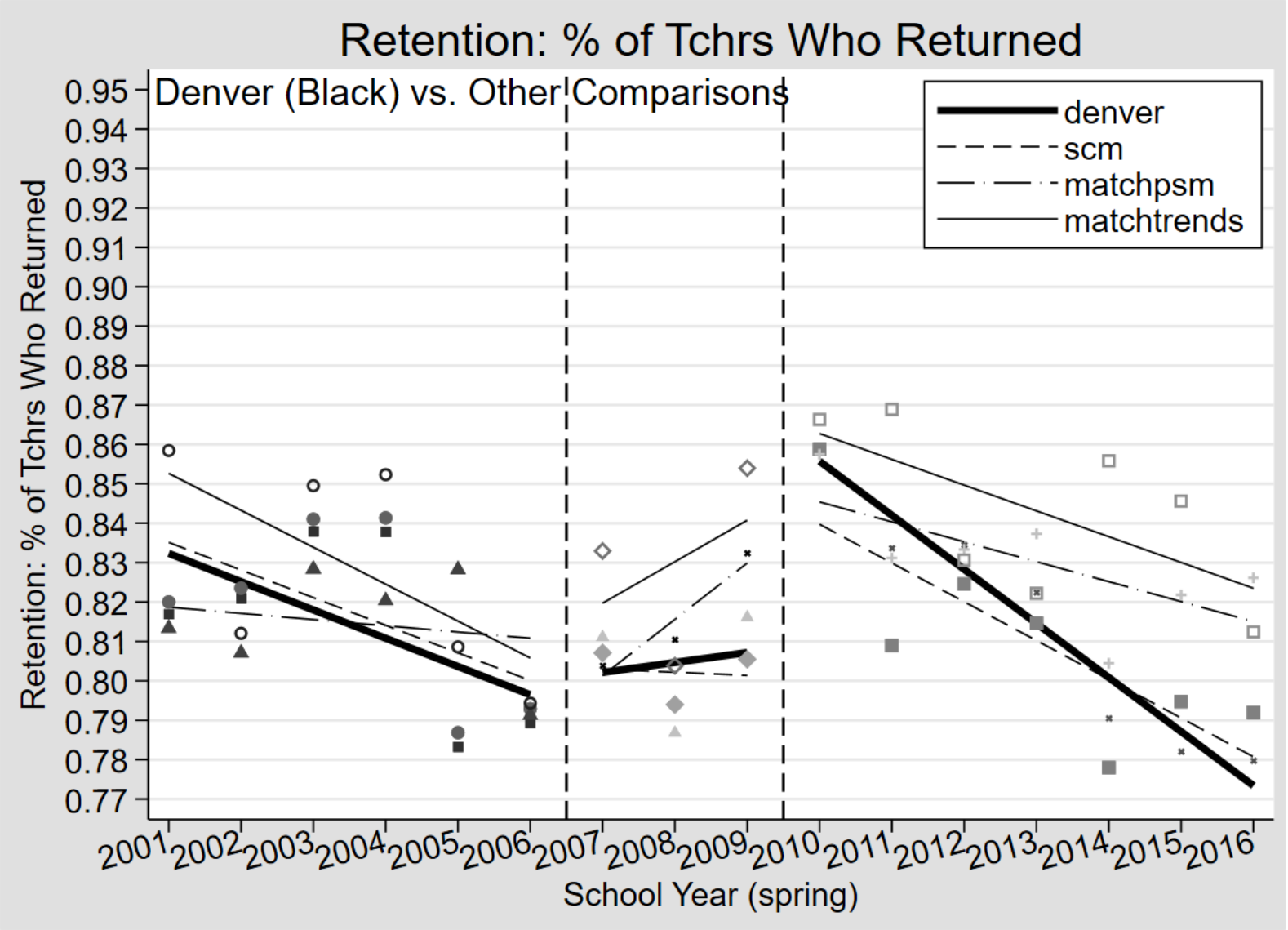
FN: This figure corresponds to results presented in Table 4.

Figure 5. ITS: Effectiveness (MGP) of New-to-DPS Teachers



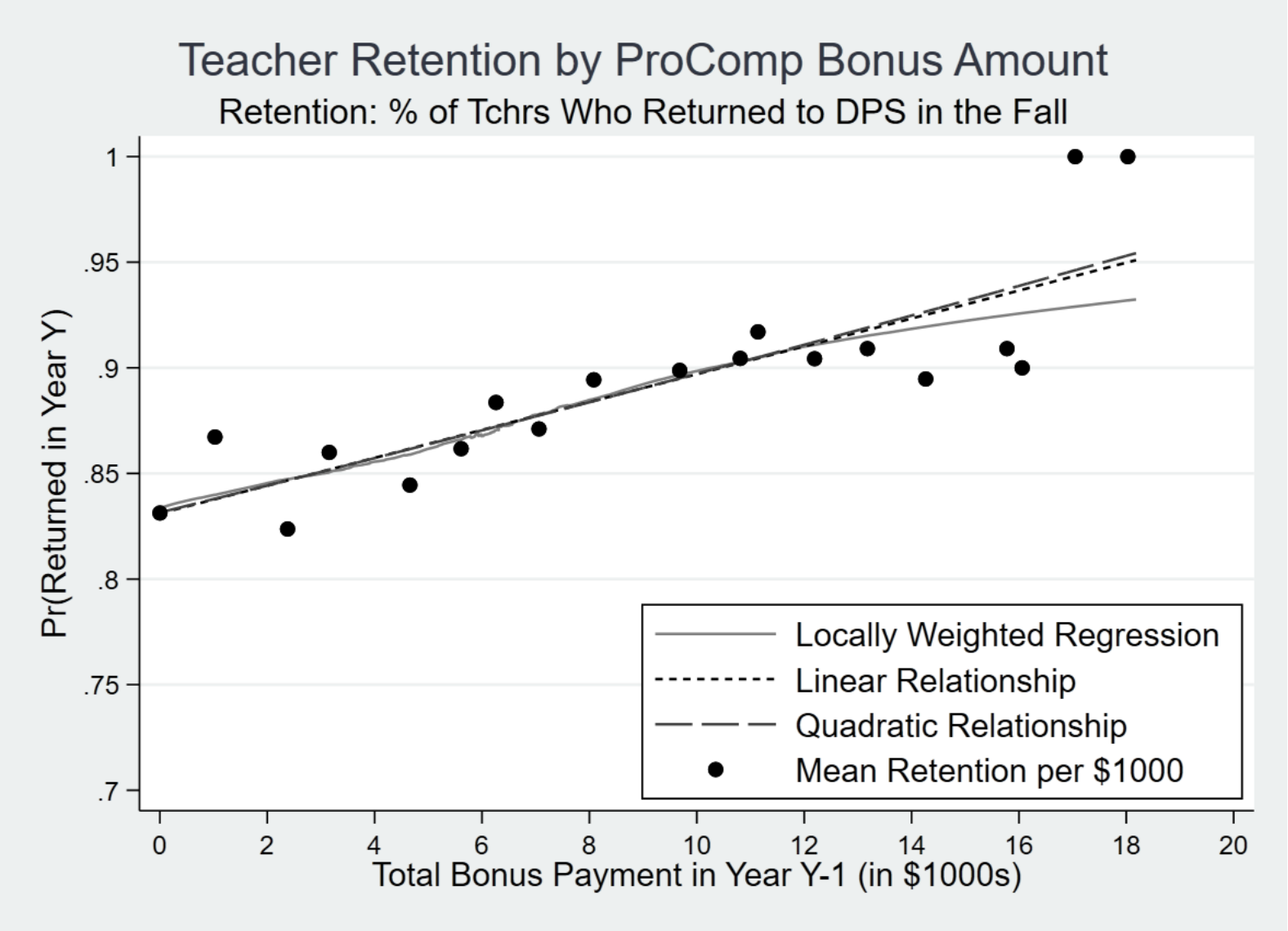
FN: This figure corresponds to results presented in Table 5.

Figure 6. CITS for DPS vs. Comparison Districts: Overall Teacher Retention



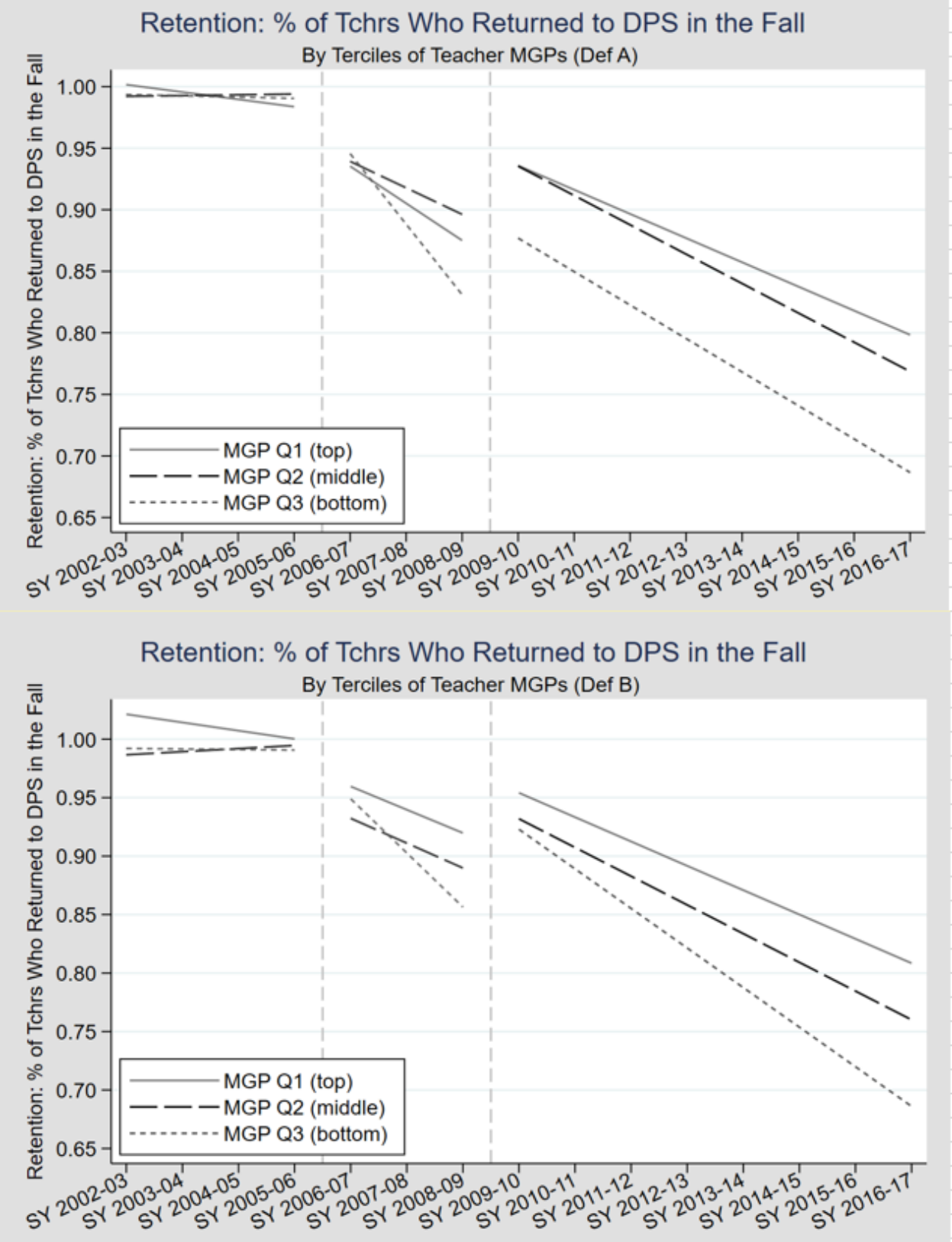
FN: This figure corresponds to results presented in Table 6.

Figure 7. Teacher Retention in the District, as a Function of Total ProComp Incentive Payout from Prior Year



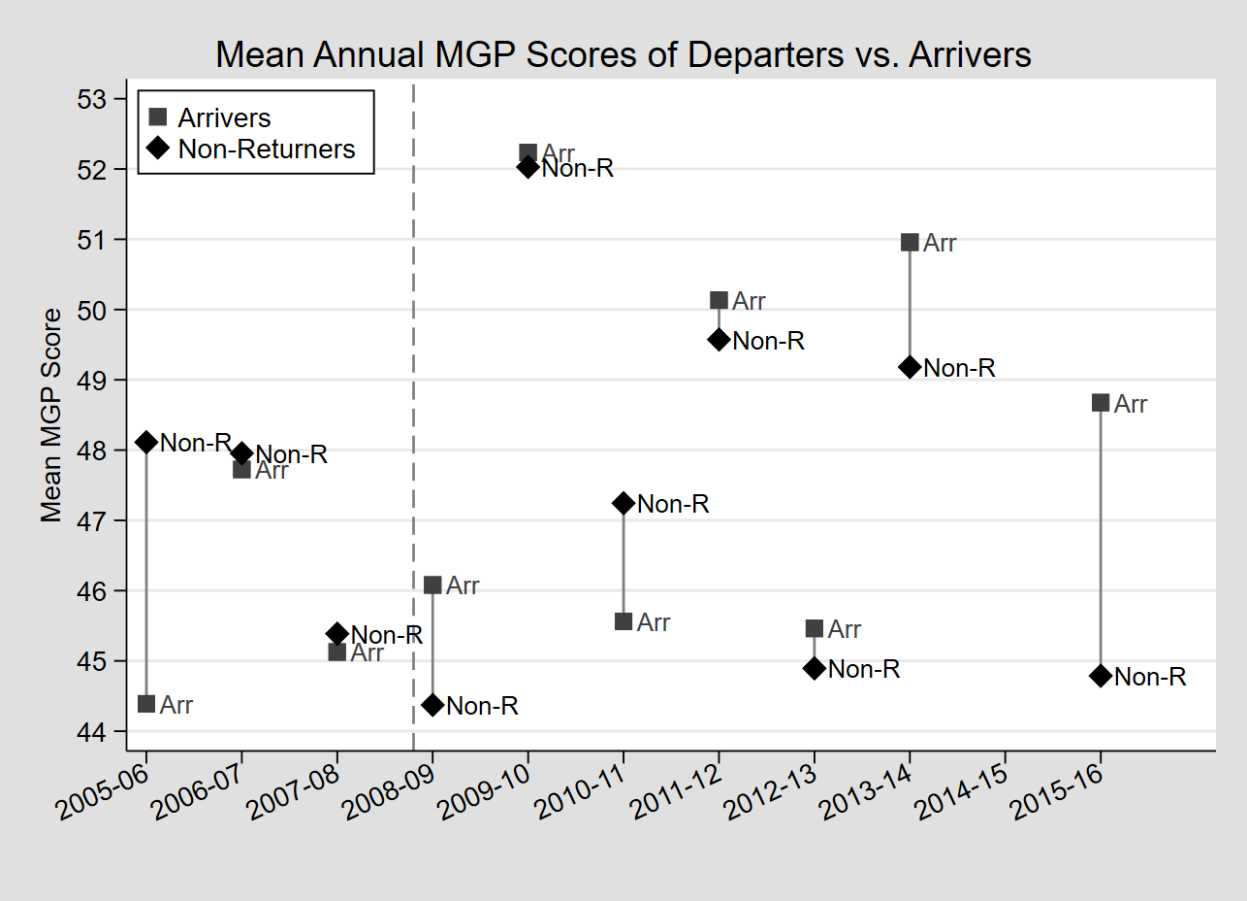
FN: This figure corresponds to results presented in Table 7.

Figure 8. ITS: Teacher Retention, Separately for High, Middle, and Low MGP Teachers



FN: This figure corresponds to results presented in Table 8.

Figure 9. Comparison of Mean MGP Scores of Arriving and Departing Teachers, by Year



FN: The mean annual MPG score for each teach is the mean of all the teacher's available MGP scores across grades in a given year. MGP scores are, by definition, only available for teachers in tested subjects and grades.

Online Appendix A: Additional Detail on Denver ProComp Incentives, Data Availability, and Research Questions

Additional Detail on Denver ProComp Incentives.

The ProComp system includes 10 distinct financial incentives, divided into four broad categories: Knowledge and Skills, Student Growth, Market Incentives, and Comprehensive Professional Evaluations. As Appendix A, Table A1 below shows, Student Growth Objectives (SGOs) were the most commonly earned incentive in the last year for which we have data, followed by Professional Development Units (PDUs). Only a small minority of teachers earned the Exceeds Expectations (because this requires that a teacher be in a tested grade and subject, which is true of a minority of teachers) or Advanced Degree incentives, while most of the other ProComp incentives were achieved by 30-60% of teachers in each year for the period beginning with the 2008-2009 school year, when ProComp's current iteration took effect.

[Insert App. A, Table A1 about here]

Student Growth Incentives.

The student growth component of ProComp consists of four separate incentives, two of which are awarded for school-wide achievements, and two of which are individual awards. At the school level, teachers are eligible to earn Top Performing and High Growth incentives; as individuals, teachers can earn Student Growth Objective incentives (SGOs) or an Exceeds Expectations bonus, both of which are based on measures of student achievement. Because the Exceeds Expectations bonus is tied to student performance on state tests, only those teachers in tested grades and subject areas are eligible to receive it. Therefore, a very small percentage of the overall participant population is eligible to receive Exceeds Expectations.

Teacher-Level Incentives: Student Growth Objectives (SGOs) and Exceeds Expectations. At the beginning of each academic year, each DPS teacher confers with school

leaders to establish two goals for their students' progress over the course of the year. These goals, called Student Growth Objectives (SGOs), may incorporate a wide variety of quantitative or non-quantitative measures, including nationally standardized tests, subject area exams created by DPS, or teacher-created tests, among other data sources.

Teachers who achieve both of their SGOs in a given year earn a salary increase equivalent to 1% of the ProComp index, which was $381.18 for the 2013-2014 school year. Teachers who achieve only one SGO earn a one-time bonus. In the period beginning with the 2008-2009 school year, the second year in which this incentive was awarded, more than half of participants received at least one SGO incentive each year, and in most years, it was much more common for participants to meet both SGO's.

[Insert App. A, Table A2 about here]

The Exceeds Expectations (ExEx) award is offered to teachers whose students achieved substantial growth on the applicable state assessment (e.g. Colorado State Assessment Program (CSAP) or Transitional Colorado Assessment Program (TCAP) exams). Because students take these exams only in certain grades and subjects, the ExEx incentive is available only to teachers in grades 4 through 10 who teach mathematics or language arts. The incentive is a one-time bonus of 6.4% of the index, which was $2,439.55 in the 2013-2014 school year. Appendix Table A3 shows the percentage of eligible (those with MGPs in at least one subject for a given year) teachers as well as total participants receiving ExEx for each year of the study.

[Insert App. A, Table A3 about here]

School-wide Bonuses: High Growth and Top Performing. Both the Top Performing and High Growth bonuses are based on the School Performance Framework (SPF), which rates schools in seven performance categories: Academic Growth, Academic Proficiency, College & Career Readiness, Improvement in College & Career Readiness Over Time, Student

Engagement, Enrollment Rates, and Parent Satisfaction. These seven categories encompass dozens of variables, including CSAP and TCAP scores, parent satisfaction surveys, dropout rates, and attendance rates, among many others.

The Top Performing incentive is awarded to schools ranked in the top half of the annual SPF ratings. The High Growth incentive, however, is based exclusively on schools' Academic Growth score, which is determined by median student growth percentiles (MGP's) for each tested subject. Teachers earned a one-time bonus of 6.4% of the index ($2439.55 for the 2013-2014 school year) for each incentive.

As Appendix Table A4 illustrates, the number of teachers receiving the Top Performing incentive increased considerably in the 2009-2010 school year. Moreover, an overwhelming majority of teachers who earned the Top Performing bonus also earned the High Growth incentive. For example, during the 2013-2014 school year, 43.5% of all ProComp teachers received the Top Performing bonus, while 39.4% of teachers earned both the High Growth and Top Performing incentives; in other words, only 4% of teachers earned the Top Performing incentive but not the High Growth award.

[Insert App. A, Table A4 about here]

Teachers in elementary schools were much more likely to receive the Top Performing and High Growth awards than their counterparts in middle and high schools. As shown in Appendix Table A5, 56% of teachers in elementary schools received the Top Performing bonus during the 2013-14 school year, compared with only 33 and 38% of teachers in middle and high schools. Receipt of the High Growth Incentive is similarly varied across school levels. This suggests that teachers in elementary schools disproportionately benefit from certain ProComp incentives. While the High Growth incentive is based on MGPs, it is a school-wide

award, so the relative percentage of teachers teaching in tested grades/subjects should not be the sole reason for the disparity.

[Insert App. A, Table A5 about here]

Market Incentives.

Hard-to-Serve (School-Wide) and Hard-to-Staff (Teacher-Level). The Hard-to-Serve bonus is designed to encourage DPS teachers to accept positions in high-needs schools. The Hard-to-Serve incentive is offered to teachers as a one-time bonus of 6.4% of the index ($2,439.55 for the 2013-14 school year) for working in schools with a high percentage of students living in poverty.

Like the Hard-to-Serve incentive, the Hard-to-Staff incentive was also paid as a one-time bonus. Each year, DPS designates positions as Hard-to-Staff if the supply of licensed professionals is low and the rate of turnover is high. In recent years, the list of eligible Hard-to-Staff positions has included special education, ELA-S, and mathematics assignments, among others. [17]

During the last year of the study, more than half of all ProComp participants received the Hard-to-Serve bonus, while slightly less than one-third of participants served in Hard-to-Staff positions (see Appendix Table A6). Teachers who received the Hard-to-Staff incentive were somewhat more likely than other DPS teachers to receive the Hard-to-Serve bonus. In 2013-14, approximately two-thirds of teachers serving in Hard-to-Staff positions also received the Hard-to-Serve Bonus.

[Insert App. A, Table A2 about here]

Knowledge and Skills.

Higher Education Incentives: Advanced Degrees. ProComp participants receive a base-building salary increase of 9% of the index for earning each degree or approved professional

license beyond a bachelor's degree. During the 2013-14 school year, earning an advanced degree incentive added $3,430.62 to a teacher's base salary. Teachers are eligible to earn this incentive only once in any three-year period. Since the 2007-08 school year, less than 10% of teachers have received the incentive in any given year (see Appendix Table A1). For the 2013-14 school year, 426 teachers, or 9.3% of participants, received this incentive.

Higher Education Incentives: Tuition Reimbursement. In addition to the salary increase for completing an advanced degree, DPS also offers teachers a Tuition Reimbursement incentive, which can be used to pay for preexisting student loans, fees for conferences and professional development workshops, or tuition for an advanced degree program. Teachers can earn up to $1,000 per year in Tuition Reimbursement, with a lifetime cap of $4,000. During the 2013-14 school year, 946 teachers, 20.8% of participants, received reimbursement for student loans; 271 teachers, 5.9% of participants, received reimbursement for current tuition expenses (see Appendix Table A1).

Professional Development Units. In an effort to encourage ongoing training, ProComp participants receive an incentive of 2% of the index ($762.36 for the 2013-14 school year) for completing approved Professional Development Units (PDUs). Teachers who have 14 or fewer years of credited tenure with DPS earn a base-building salary increase for successfully earning the PDU incentive; teachers with more than 14 years of service earn a one-time bonus.

Teachers may complete multiple PDUs, but they may only earn a single award in any given school year. However, PDU credits are "bankable", meaning that a teacher who completes two PDUs in a given school year can earn the award for the second completed PDU in a subsequent school year. This has been a relatively highly attained incentive in most years, with 61% of teachers receiving the award in the 2013-14 school year.

App. A, Table A1. Percentage of ProComp Participants Earning each Incentive, by School Year

		ProComp 1.0			*ProComp 2.0*					
Category	Incentive Name	2006	2007	2008	2009	2010	2011	2012	2013	2014
Student Growth	Met 1 or 2 SGOs	0.0%	0.0%	10.4%	54.1%	69.6%	74.8%	65.3%	69.5%	63.9%
	Exceeds Expectations	0.0%	0.0%	0.0%	8.1%	12.0%	12.8%	11.9%	13.8%	12.4%
	High Growth School	0.0%	0.0%	0.0%	28.4%	47.3%	40.8%	40.4%	44.0%	46.2%
	Top Performing School	0.0%	8.1%	13.7%	28.2%	41.4%	32.1%	37.4%	40.1%	43.4%
Market Incentives	Hard to Serve School	38.6%	32.7%	31.2%	54.5%	50.6%	56.5%	56.7%	55.3%	56.5%
	Hard to Staff Position	16.1%	19.9%	32.0%	34.7%	34.8%	33.7%	33.2%	28.3%	27.5%
Knowledge and Skills	Advanced Degree	11.9%	7.2%	6.6%	7.6%	7.2%	7.8%	6.8%	8.5%	8.9%
	Tuition/School Loan Reimbursement	17.8%	5.1%	4.5%	17.7%	24.0%	27.5%	28.7%	26.3%	24.9%
	Complete PDUs	0.0%	0.0%	48.7%	61.2%	58.9%	26.4%	42.0%	64.7%	58.9%
Evaluation	Comprehensive Evaluation, CPEs	0.0%	0.0%	2.3%	53.0%	62.0%	52.0%	46.7%	37.4%	42.5%
	Total ProComp Participants	522	1,361	1,855	2,611	2,934	3,212	3,621	3,795	4,025

FN: Each year represents the spring of a given school year (e.g., 2006 represents 2005-2006 school year). Cells contain percentages of ProComp participants who received each incentive, by year of ProComp implementation. SGOs are student growth objectives. CPE is comprehensive performance evaluation. The total number of ProComp participants is shown in the final row. Recall that all new hires in the district after January 1, 2006 are automatically enrolled in ProComp.

App. A, Table A2. Percentage of Participants Earning 0, 1, or 2 SGO Incentives, by School Year

	ProComp 1.0			*ProComp 2.0*					
	2006	2007	2008	2009	2010	2011	2012	2013	2014
Did Not Meet SGOs	N/A	N/A	88.8%	43.7%	27.8%	22.0%	30.6%	27.6%	34.3%
Met 1 SGO	N/A	N/A	8.9%	8.5%	10.6%	12.1%	69.4%	11.8%	9.9%
Met 2 SGOs	N/A	N/A	2.3%	47.8%	61.6%	66.0%	0.0%	61.0%	55.9%
Total ProComp Participants	522	1,361	1,855	2,611	2,934	3,212	3,621	3,795	4,025

FN: Each year represents the spring of a given school year (e.g., 2006 represents 2005-2006 school year). Cells contain percentages of ProComp participants who received each incentive, by year of ProComp implementation. SGOs are student growth objectives. The total number of ProComp participants is shown in the final row. Recall that all new hires in the district after January 1, 2006 are automatically enrolled in ProComp.

App. A, Table A3. Percent Earning "Exceeds Expectations" (ExEx) Incentive, by School Year

	ProComp 1.0			*ProComp 2.0*					
As a Percent of…	2006	2007	2008	2009	2010	2011	2012	2013	2014
… Teachers in Eligible Subjects/Grades	0.0%	0.0%	0.0%	24.7%	33.2%	36.1%	28.9%	33.4%	31.4%
… All ProComp Participants	0.0%	0.0%	0.0%	2.5%	4.8%	4.4%	4.3%	5.1%	3.9%
Total ProComp Participants	522	1,361	1,855	2,611	2,934	3,212	3,621	3,795	4,025

FN: Each year represents the spring of a given school year (e.g., 2006 represents 2005-2006 school year). Cells contain percentages of ProComp participants who received each incentive, by year of ProComp implementation. The total number of ProComp participants is shown in the final row. Recall that all new hires in the district after January 1, 2006 are automatically enrolled in ProComp.

App. A, Table A4. Percentage Earning High Growth/ Top Performing Incentives, by Year

	ProComp 1.0			*ProComp 2.0*					
	2006	2007	2008	2009	2010	2011	2012	2013	2014
Top Performing School	N/A	8.0%	13.6%	27.9%	41.0%	32.0%	37.1%	40.2%	43.8%
High Growth School	N/A	N/A	N/A	28.1%	46.8%	40.8%	40.3%	44.1%	46.8%
Both	N/A	N/A	N/A	23.1%	39.4%	29.1%	33.4%	36.2%	39.7%
Neither	N/A	92.0%	86.4%	67.2%	51.6%	56.3%	55.9%	51.9%	49.1%
Total ProComp Participants	522	1,361	1,855	2,611	2,934	3,212	3,621	3,795	4,025

FN: Each year represents the spring of a given school year (e.g., 2006 represents 2005-2006 school year). Cells contain percentages of ProComp participants who received each incentive, by year of ProComp implementation. The total number of ProComp participants is shown in the final row. Recall that all new hires in the district after January 1, 2006 are automatically enrolled in ProComp

App. A, Table A5. Schoolwide Student Growth Incentives: Percent Earning Incentives for 2013-14, by School Level

	Elementary Schools	Middle Schools	High Schools	Alternative/ Other
Top Performing School	53%	31%	36%	21%
High Growth School	55%	33%	45%	18%
Total ProComp Participants	2,435	364	559	558

FN: Cells contain percentages of ProComp participants who received each school-wide incentive in 2013-04 in the given school level.

App. A, Table A6. Percentage Earning Hard-to-Serve and Hard-to-Staff Incentives, by Year

	ProComp 1.0			*ProComp 2.0*					
	2006	2007	2008	2009	2010	2011	2012	2013	2014
Hard-to-Serve School	38.6%	32.7%	31.2%	54.5%	50.6%	56.5%	56.7%	55.3%	56.5%
Hard-to-Staff Position	16.1%	19.9%	32.0%	34.7%	34.8%	33.7%	33.2%	28.3%	27.5%
Earned Both	6.6%	9.7%	11.4%	21.6%	20.2%	21.7%	22.1%	19.0%	18.6%
Earned Either	48.1%	42.9%	51.8%	67.6%	65.3%	68.4%	67.7%	64.6%	65.4%
Total ProComp Participants	522	1,361	1,855	2,611	2,934	3,212	3,621	3,795	4,025

FN: Each year represents the spring of a given school year (e.g., 2006 represents 2005-2006 school year). Cells contain percentages of ProComp participants who received each incentive, by year of ProComp implementation. The total number of ProComp participants is shown in the final row. Recall that all new hires in the district after January 1, 2006 are automatically enrolled in ProComp.

App. A, Table A7. Timeline of ProComp Policy Implementation, with Data Availability

1998-99	(A) Begin statewide, district-year level NCES demographics
1999-00	--
2000-01	(B) Begin statewide, district-year level teacher turnover rates
2001-02	(C) Begin (in spr) DPS, student level achievement data (but no links to schools)
2002-03	(D) Begin (in fall) DPS, teacher-year "new to DPS in fall" outcome ("recruitment")
	(E) Begin (in fall) DPS, teacher-year "returned in fall" teacher outcome ("retention")
2003-04	(F) Begin (in spr) statewide, district-year level achievement mean's & SD's
	(G) Begin (in spr) DPS, student level achievement data now has links to schools
2004-05	(G) Begin (in spr) DPS, teacher-year MGP's
2005-06	Policy: Nov. 2005, **PC1** approved by voters, & tchrs immediately opt in/out *(16% opt in)*
	Policy: Jan 01, 2006, all teachers hired after this date are automatically enrolled in PC
	Policy: Summer 2006, teachers making first retention decisions post-ProComp
2006-07	First school year PC1 could affect student achievement and teachers' decisions to return
2007-08	--
2008-09	Policy: August 08: **PC2** finalized *(likely too late to affect fall 2008 retention decisions)*
	Policy: First full year of **PC2** implementation
	First school year PC2 could affect student spring achievement outcomes
2009-10	First school year PC2 could affect teachers' decisions to return or not
2010-11	Policy: Last year teachers (hired pre 01/2006) can opt in (for 2011-12)
2011-12	--
2012-13	(C) End DPS, student level achievement data in science
2013-14	(C) End DPS, student level achievement data in writing
	(A) End statewide, district-year level NCES demographics
	(F) End statewide, district-year level achievement mean's & SD's (mth, rdg, wrt)
2014-15	<<no MGP's in 2014-15 due to transition to PARCC>>
2015-16	(G) End DPS, teacher-year MGPs
	(C) End DPS, student level achievement data math and reading
	(B) End statewide, district-year level teacher turnover rates
2016-17	(D) End DPS, teacher-year "new to DPS in fall" outcome ("recruitment")
	(E) End DPS, teacher-year "returned in fall" teacher outcome ("retention")

App. A, Table A8. Summary of Research Questions, Respective Outcomes, and Analytic Samples

Research Question	Method	Unit of Analysis	Analytic Sample	Outcome Variable(s)	Outcome Variable Definition/ Coding	Time Function Interacted with MGP^*_{py} Scores?
(1) Has ProComp improved student achievement (ELA, math, writing)?	CITS	District-Year	DPS and various sets of "comparable" CO districts	$Math_{dy,}$ $Read_{dy}$ $Write_{dy}$	e.g., $Math_{dy}$ represents district d's mean standardized math achievement score in year y. All achievement outcomes are standardized within subject-grade-year at the state level.	N/A
(2) Has ProComp attracted more effective teachers to DPS?	ITS	Teacher-year	Subset of teachers who are new to DPS each school year.	MGP_{py}	MGP_{py} is the median growth percentile score of teacher p in year y, ranging from 0 to 100.	No. *(MGP_{py} scores are the outcome in RQ 2)*
(3A) Has ProComp improved teacher retention in DPS?	CITS	District-Year	DPS and various sets of "comparable" CO districts	$Retained_{dy}$	$Retained_{dy}$ is a continuous variable between 0 and 1 that represents the percentage of the teacher in district d that were retained in year y.	N/A
(3B) Did ProComp have a more positive impact on retention among more effective teachers?	ITS	Teacher-year	All teachers with MGP scores.	$Retained_{py}$	(same as above).	Yes.

FN: MGP^*_{py} : For the sake of interpreting the coefficients more meaningfully, we re-scale teachers' MGP scores (originally on a scale of 0 to 100) when they appear on the right-hand side of the equation as interaction variables. To make MGP^*_{py}, we first subtract 50 from MGP_{py} and then divide by 10, so that MGP is on a scale of -5 to 5, and a one-unit difference in MGP^*_{py} represents a 10-point difference in original MGP scores.*

Online Appendix B

Using MGP Scores as Outcomes in Recruitment Analysis

In the paper, we follow the practices of DPS to use MGP scores to characterize teaching performance. Recall that our time function breaks the 2003 to 2014 data panel into three periods: Pre-ProComp (2003-2006), PC1 (2007- 2009), and PC2 (2010- 2016). We would like to explore whether the district is able to attract stronger teachers to DPS in the ProComp era. For this recruitment analysis, MGP scores of new-to-DPS teachers become the outcome of interest. When we seek to use MGP scores as an outcome, we are confronted with the complication that DPS did not produce student-teacher links (and therefore MGP scores) until 2004-05. Therefore, should we want to characterize a teacher's performance only by their observed MGP score in the exact year of arrival, we would only be able to include two years of Pre-ProComp data. We therefore explore other ways to use MGP scores to characterize each teacher's performance once they have joined DPS. In the paper, we refer to these as Definitions 1, 2, and 3, each of which is described below (note that these are distinct from the MGP *Interaction* Definitions A and B that are described in Appendix C and used on the right-hand side of the model in the differential retention analysis).

MGP Outcome Definition 1: HLM-Based Estimate of MGP in Year of Arrival.

We use a two-level model (repeated observations of teacher MGP scores, nested within teacher) to impute teacher-year MGP scores in a teacher's first year. For each teacher, we model the repeated observations of available MGP scores in a given subject (across years) as a simple, linear function of years of experience, with the experience predictor centered at the first year. Because the intercept and slope of that experience function are allowed to vary randomly across teachers, we estimate a unique, Bayes-estimated (shrunk) intercept and slope for each teacher. The estimated intercepts themselves represent the predicted MGP score for each teacher in their year

of arrival. Quality of recruitment analyses that use MGP Outcome Definition 1 are presented in column 1 of Table 5.

We assess the quality of these HLM-based, time-varying MGP scores by correlating observed first year MGP scores with predicted first year MGP scores when both are available: The correlation is 0.871. The RMSE from a model predicting observed MGP scores as a function of HLM-estimated MGP scores is 7.54. Taken together, this suggests that the HLM-based MGP scores closely mirror the observed MGP scores in the first year when they are both available, which lends some support to the assertion that they may also estimate what the teacher's MGP scores would have been in the year of arrival for teachers who are missing that MGP score.

MGP Outcome Definition 1 has three primary advantages: First, we can use time-varying information on teacher performance so that we can characterize how a teacher is performing at the time of their arrival in the district. Second, we can create measures of teacher performance even in the Pre-Period. Third, we can include any arriving teacher with at least two MGP scores, allowing us to maximize the population of study. However, MGP Outcome Definition 1 rests on the performance of the multi-level model to estimate missing MGP scores in the first year.

MGP Outcome Definition 2: HLM-Based Estimate of MGP in Third Year.

We again use estimates from the use a two-level model described above, but instead of characterizing teacher quality based on the year in which they arrive, for MGP Outcome Definition 2, we instead use the estimated MGP score in each teacher's *third* year. The idea here is that teachers often do not exhibit their strongest performance or longer-term potential in their first year either new to the district or new to teaching. The third year MGP score may capture a sense of longer-term potential for a teacher who has arrived in DPS either before or after the onset of ProComp. Quality of recruitment analyses that use MGP Outcome Definition 2 are presented in column 2 of Table 5.

This approach has both advantages and disadvantages. On the one hand, more teachers who arrived in the Pre-Period have an *actual* MGP score in their third year (as opposed to an HLM-estimated MGP score), whereas no teachers who arrived in the Pre-Period have an actual MGP score in their first year (because MGP scores are not available in DPS until 2004-05). Another advantage is that each teacher's future performance is assessed on the same number of MGP scores, so that the number of scores available does not bias the analysis. On the other hand, MGP scores in any single year may be imprecisely estimated, and thus we may wish to combine MGP scores from many years to characterize the future performance of arriving teachers. Furthermore, using this measure limits the analysis to the subset of arriving teachers who remain in DPS for at least three years. These disadvantages motivate the third definition, presented next.

MGP Outcome Definition 3: Mean Across Any Available Year and Subject.

In order to include as many teachers as possible in the analysis, we calculate an average MGP for every teacher across all subject and school years in which she is present, and we call this MGP Outcome Definition 3 (see column 3 of Table 5). Of the 16,628 unique teachers for which we observe their arrival to the district in our dataset, 3,971 (approximately 23.8%) are observed in their year of arrival and have at least one MGP score (and thus have a mean MGP score). This approach has the advantage of being as inclusive as possible, allowing any arriving teacher who has ever had at least one MGP score to be part of the recruitment analysis. In addition, it allows us to characterize the performance of teachers who arrive even in the Pre-Period (despite the absence of MGP scores in those years). On the other hand, one disadvantage is that these averaged MGP scores may be more or less imprecise across teachers, depending on how many MGP scores are available (we do control for number of MGP scores available in our models). Another disadvantage is that the use of a time-invariant MGP measure assumes that teacher effectiveness is essentially stable across years but imprecisely estimated. If one believes that teacher effectiveness changes

systematically over the course of several years and that the change can be captured by changes in MGP scores, then using an overall MGP score may not be desirable.

Implications of MGP Outcome Definitions on Sample.

Appendix Table B1 compares the samples of teachers (using each of the definitions) to one another and to the overall analytic sample we access at various points in this paper.

We consider only those teachers we observe in their year of arrival (i.e. teachers whose first year in the district is 2003 or later). As indicated by the first two columns ("All Teachers" and "All MGP Eligible Teachers"), restricting our sample to only those teachers who were ever eligible for an MGP (tested grades, ELA and Math, SY2004-05 and later), reduces our sample size from 16,628 to 3894. Definitionally, these teachers, on average, taught in higher grade levels than the overall teacher sample, though only slightly higher. Nonetheless, not all teachers who were ever eligible for an MGP actually received one. Thus, the sample is further restricted for each of the subsequent definitions.

App. B, Table B1. Summary Characteristics of Restricted Samples for Arrival Quality Analysis

	All DPS Teachers	Teachers Ever Eligible for MGPs	MGP Outcome Definition #1	MGP Outcome Definition #2	MGP Outcome Definition #3
N	16,626	3,893	1,721	3,646	3,970
First Year	2008	2008	2011	2008	2008
Observed	*2002 - 2017*	*2002 - 2017*	*2005 - 2016*	*2002 - 2016*	*2002 - 2017*
Last Year	2012	2014	2014	2014	2014
Observed	*2002 - 2017*	*2005 - 2017*	*2005 - 2017*	*2005 - 2017*	*2005 - 2017*
# Years	4.63	6.55	3.15	5.92	5.58
Observed	*1 - 13*	*1 - 13*	*1 - 13*	*1 - 13*	*1 - 13*
Years ProComp	4.01	5.19	3.92	5.05	4.76
Participation	*0 - 12*	*0 - 12*	*0 - 12*	*0 - 12*	*0 - 12*
Number of MGPs	0.72	2.25	2.55	2.81	2.67
	0 - 10	*0 - 10*	*1 - 10*	*1 - 10*	*1 - 10*
Grade Taught (Modal)	5	5	6	6	6
Percent Male	25%	23%	31%	26%	27%
Percent White	76%	78%	82%	77%	77%
Average Year of Birth	1971	1971	1978	1972	1972

FN: This table presents sample sizes and typical characteristics for the groups of teachers who are included in the analyses when we use each of the four definitions of MGP's as outcomes, described in this appendix. For relevant variables, we also include the minimum and maximum values of the characteristic, below in grey. Here is a brief summary of the MGP-as-Outcome Definitions:

For Definition 1 we use the MGP a teacher actually received in the year of his/her arrival.

For Definition 2 we use HLM to estimate an MGP for teachers in their third year with the district.

For Definition 3 we take the mean of all MGPs earned by a teacher over his/her observed tenure, controlling for the available number of MGPs.

Online Appendix C

Using MGP Scores as Interactions in Retention Analysis

In the paper, we follow the practices of DPS to use MGP scores to characterize teacher performing characterize his or her teaching performance. Recall that our time function breaks the 2003 to 2014 data panel into three periods: Pre-ProComp (2003-2006), PC1 (2007- 2009), and PC2 (2010- 2016). When we seek to interact this time function with MGP scores, we are confronted with the complication that DPS did not produce student-teacher links (and therefore MGP scores) until 2004-05. Therefore, should we want to examine whether ITS results depend on MGP scores *in the current year* (i.e., a time-varying MGP score), we would only have two years of Pre-ProComp data that also includes MGP scores. We therefore explore other ways to use MGP scores to characterize each teacher's overall performance. In the paper, we refer to these as Definition A, and B, each of which is described below. (These definitions of MGP interactions should not be confused with the definition of MGP *outcomes* described in Appendix B, which are identified by number (i.e., Definitions 1, 2, and 3).

MGP Interaction Definition A: Mean Across Any Available Year and Subject.

In order to include as many teachers as possible in the analysis, we calculate an average MGP for every teacher across all subject and school years in which she is present, and we call this MGP Interaction Definition A (see columns 2 - 4 of Table 8). Of the 16,628 unique teachers that appear in our dataset, 3971 (approximately 23.8%) are eligible for and receive at least one MGP score and thus have a mean MGP score. This approach has the advantage of being as inclusive as possible, allowing any teacher who has ever had at least one MGP score to be part of the analysis. In addition, it allows us to characterize the performance of teachers even in the pre-period. On the other hand, one disadvantage is that these averaged MGP scores may be more or less imprecise across teachers, depending on how many MGP scores are available (we do control for number of

MGP scores available in our models). Another disadvantage is that the use of a time-invariant MGP measure assumes that teacher effectiveness is essentially stable across years but imprecisely estimated. If one believes that teacher effectiveness changes systematically over the course of several years and that the change can be captured by changes in MGP scores, then using an overall MGP score may not be desirable. We thus explore other methods to address the particular weaknesses of MGP Interaction Definition A—losing time-varying MGP scores and having different numbers of MGP scores for different teachers.

MGP Interaction Definition B: HLM-Based Estimate of MGP in Each Year.

We use a two-level model (repeated observations of teacher MGP scores, nested within teacher) to impute teacher-year MGP scores in a teacher's first year through their tenth year. For each teacher, we model the repeated observations of available MGP scores in a given subject (across years) as a simple, linear function of years of experience. Because the intercept and slope of that experience function are allowed to vary randomly across teachers, we estimate a unique, Bayes-estimated (shrunk) intercept and slope for each teacher. We use these to predict scores in years the teacher may not possess scores. We lag this variable, because the decision to return or not in the fall could depend on the teacher's performance in the spring immediately preceding this decision. Differential retention analyses that use MGP Interaction Definition B are presented in column 5 - 7 of Table 8.

We assess the quality of these HLM-based, time-varying MGP scores by correlating observed scores with predicted scores when both are available: The correlation is 0.8131. The RMSE from a model predicting observed MGP scores as a function of HLM-estimated MGP scores is 8.328. Taken together, this suggests that the HLM-based MGP scores closely mirror the observed MGP scores in years they are both available, which lends some support to the assertion

that they may also estimate what the teacher's MGP scores would have been in years where the score is missing.

MGP Definition B has three primary advantages: First, we can use time-varying information on teacher performance so that we can characterize how a teacher is performing at the time of their retention decisions. Second, we can create measures of teacher performance even in the Pre-Period. Third, we can include any teacher with at least two MGP scores, allowing us to maximize the population of study. However, MGP definition B rests on the performance of the multi-level model to estimate missing MGP scores.

Implications of MGP Interaction Definitions on Sample.

Appendix Table C1, below, summarizes the characteristics of the samples for each of the definitions as well as all teachers and all MGP-eligible teachers for comparison. Less than one-third of all unique teachers observed were ever eligible for an MGP. These teachers also, on average, had a longer observed tenure in the district, than did their ineligible counterparts.

Using MGP Scores as Continuous Interactions

In Table 8 and Figure 8 of the paper, we use MGP Definitions A and B to create three groups of teachers: Top, middle, and bottom third of the MGP distribution. Here, we also conduct this analysis using a continuous version of MGP scores, interacted with the time function. A major disadvantage of using the continuous MGP score as an interaction is the difficulty in interpreting the magnitude of coefficients on three-way interactions. However, there are important advantages: We need not make choices about category cut points for high, middle, and low-achieving teacher groups that could drive our results. Another advantage is the ability to formally test the hypothesis that level and trend effects depend on MGP (i.e., significant coefficients on MGP interactions).

We return to the ITS approach described in Equation (1). To detect any differential effects, we interact the elements of the ITS function ($PC1_{py}$, $PC2_{py}$, $Time_{py}$, $Time_{py} \times PC1_{py}$, and

$Time_{py} \times PC2_{py}$) with the continuous MGP^{*}_{py} scores for each teacher. We run this model twice times using MGP definitions A and B described above.

To facilitate interpretation of interaction coefficients, we use a re-scaled version of those scores, MGP^{*}_{py}. To re-scale MGP_{py} (teachers' MGP scores originally on a scale of 0 to 100), we first center them at 50 and then divide by 10 so that MGP^{*}_{py}, is on a scale of -5 to 5. As a result, a one-unit difference in MGP^{*}_{py} corresponds to a 10-percentile point difference in MGP scores. The relevant coefficients in Appendix Table C2—particularly the two- and three-way interactions with the continuous MGP variable—are still somewhat difficult to interpret. We therefore present the results from columns 1 and 2 of Appendix Table C2 visually in Appendix Figure C1. To illustrate differential effects, we graph estimated retention rates from these models during the Pre-Period, PC1, and PC2 for hypothetical teachers with an MGP of 20, 40, 60 and 80. Appendix Figure C1 shows that, prior to ProComp, there is no evidence of a relationship between retention and effectiveness (indeed, if anything, higher MGP teachers have slightly *lower* retention rates). However, this changes markedly during PC1 and PC2: There is now a clear pattern that arises only after the onset of ProComp: Higher MGP teachers are more likely to remain in DPS than their lower MGP counterparts.

The differential retention effects shown in Appendix Figure C1 are captured formally by the MGP interaction coefficients presented in Appendix Table C2 (for parsimony, we only present the MGP main effect and the MGP interaction coefficients in this table). When these are positive and statistically significant, they suggest that ITS retention level or trend effects are stronger for high-MGP teachers than for low-MGP teachers. Indeed, all of the interactions between MGP scores and retention means (or trends) in both the PC1 and PC2 periods are positive, and in several cases statistically significant.

The lower panel of Appendix Table C2 shows estimated mean retention rates from these models during the Pre-Period, PC1, and PC2 for hypothetical teachers with an MGP of 20, 40, 60 and 80. In column 1 (using MGP Interaction Definition A), we predict that the expected retention rate for a low-performing teacher (MGP = 20) is 95% prior to ProComp, decreases to 83.5% during PC1, and declines further to 75.2% during PC2—that is, a 19.8 percentage point drop from pre-ProComp to PC2 among low-performing teachers. In contrast, the expected retention rate for a high-performing teacher (MGP = 80) declined by only 9.6 percentage points across the same period. Retention rates therefore decreased *twice as fast* for low-MGP teachers compared to high-MGP teachers (column 1). The pattern is even more striking in column 2 of Table 8, wherein we use MGP Definition B (also depicted in the lower panel of Figure 8). According to this model, rates decrease about *3.6 times faster* for low-MGP teachers compared to high-MGP teachers.

We also explore whether the differential retention results are consistent across various proxies for teacher effectiveness. We rerun the differential retention analysis using VAM scores rather than MGP scores and find that results (available upon request) are substantively similar to our MGP findings. We also replace MGP scores with other possible effectiveness proxies, including teachers' years of experience, age, or highest degree earned. We do *not* find evidence of differential ProComp retention effects when using these proxies (i.e., interactions are not significant or consistent). These results (available upon request) suggest that ProComp did not cause older/ veteran teachers or teachers with advanced degrees to remain in DPS at higher rates.

The analysis of differential retention rates by MGP/VAM scores is consistent with the hypothesis that—while MGP/VAM scores did not predict retention prior to ProComp—*during* ProComp, retention was higher among high-performing teachers than low-performing teachers.

App. C, Table C1. Characteristics of Restricted Samples for MGP Interaction Analysis.

	All DPS Teachers	Teachers Ever Eligible for MGPs	MGP Interaction Definition A	MGP Interaction Definition B
N	16,626	3,893	3,959	3,970
First Year Observed	2008 *2002 - 2017*	2008 *2002 - 2017*	2008 *2002 - 2017*	2008 *2002 - 2017*
Last Year Observed	2012 *2002 - 2017*	2014 *2006 - 2017*	2014 *2006 - 2017*	2014 *2006 - 2017*
# Years Observed	4.63 *1 - 13*	6.55 *1 - 13*	6.16 *1 - 13*	5.58 *1 - 13*
Years ProComp Participation	4.01 *0 - 12*	5.19 *0 - 12*	5.00 *0 - 12*	4.76 *0 - 12*
Number of MGPs	0.72 *0 - 10*	2.25 *0 - 10*	2.98 *1 - 10*	2.67 *1 - 10*
Grade Taught (Modal)	5	5	6	6
Percent Male	25%	23%	27%	27%
Percent White	76%	78%	78%	77%
Average Year of Birth	1971	1971	1972	1972

FN: This table presents sample sizes and typical characteristics for the groups of teachers who are included in the analyses when we use each of the four definitions of MGP's as interactions, described in this appendix. For relevant variables, we also include the minimum and maximum values of the characteristic, below in grey. Here is a brief summary of the MGP-as-Interactions definitions:

For Definition A, we use the mean of any available MGPs, controlling for the number available.

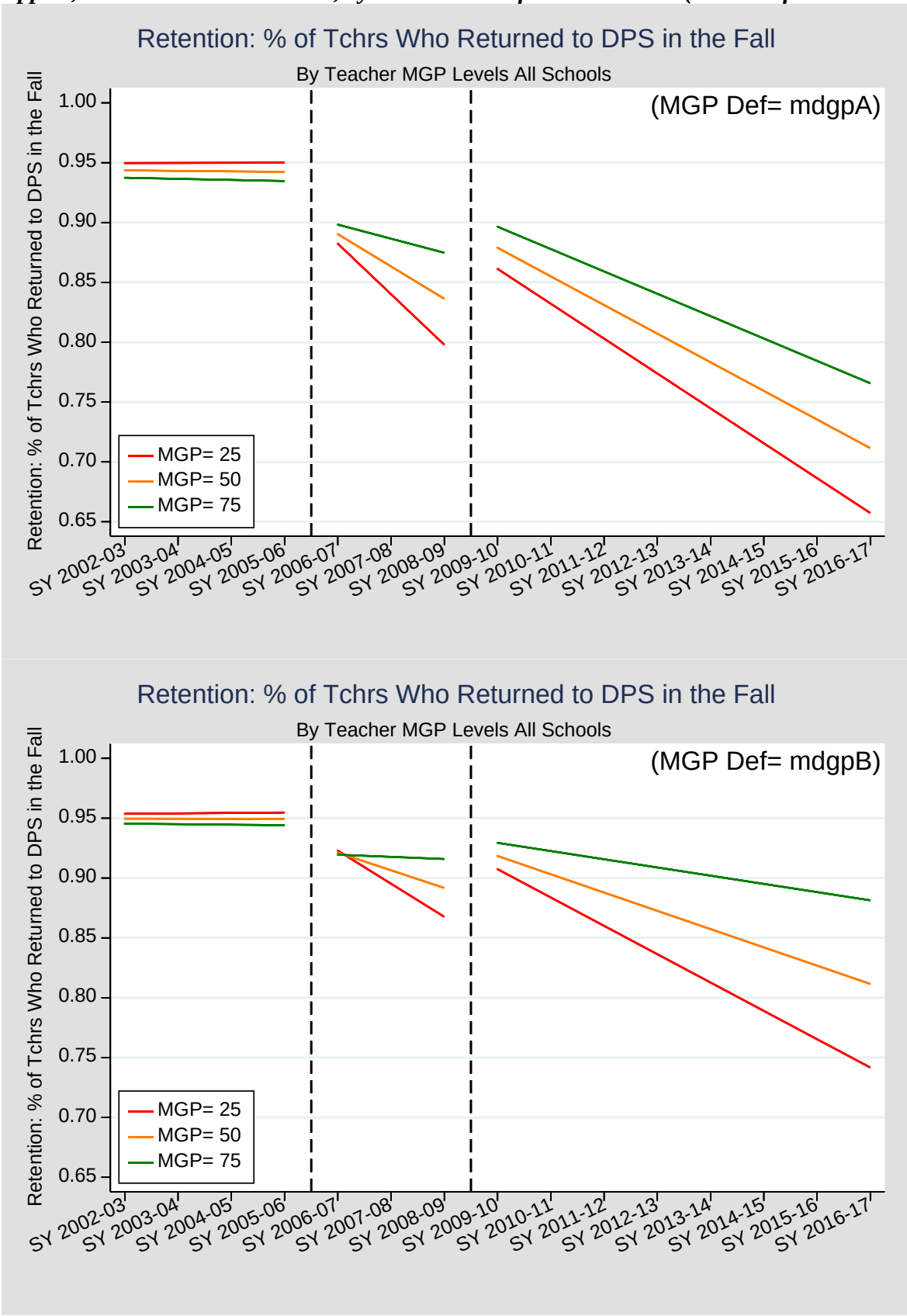
For Definition B, we use an HLM-based estimate of MGP, controlling for the number actually available.

App. C, Table C2. ITS Results: Teacher Retention in DPS, with MGP Score Interactions

`	MGP Interaction Definition A		MGP Interaction Definition B	
MGP* Scores *(rescaled -5 to 5)*	-0.28%		-0.19%	
	(0.25%)		(0.34%)	
MGP* x ProComp 1.0 *(Level Effect)*	1.20%	***	0.64%	*
	(0.34%)		(0.34%)	
MGP* x ProComp 2.0 *(Level Effect)*	1.71%	***	1.80%	***
	(0.28%)		(0.39%)	
MGP* x Time	-0.02%		-0.02%	
	(0.22%)		(0.31%)	
MGP* x Time x ProComp 1.0 *(Trend Effect)*	0.63%		0.53%	
	(0.37%)		(0.51%)	
MGP* x Time x ProComp 2.0 *(Trend Effect)*	0.23%		0.35%	
	(0.23%)		(0.32%)	
MGP of 25 : Pre-Period Mean	95.0%		95.2%	
ProComp 1.0 Mean	84.0%		89.5%	
ProComp 2.0 Mean	75.9%		82.5%	
MGP of 50 : Pre-Period Mean	94.3%		95.0%	
ProComp 1.0 Mean	86.3%		90.6%	
ProComp 2.0 Mean	79.5%		86.5%	
MGP of 75 : Pre-Period Mean	93.6%		94.5%	
ProComp 1.0 Mean	88.6%		91.8%	
ProComp 2.0 Mean	83.1%		90.5%	
Adjusted R-Squared	0.041		0.025	
N	46,090		43,289	
Omitted Period?	Pre vs. PC1 & PC2		Pre vs. PC1 & PC2	
Time Frame	2003-2014		2003- 2014	
School Covariates?	Yes		Yes	

FN: Coefficients and their standard errors are reported in percentage points for ease of interpretation (e.g., the first coefficient on Time in column (1) would be 0.0010). Def. A: Mean of any available MGPs (controlling for the # available). Def. B: An HLM-based estimate of MGP. All models include time varying, school-level control. A one-unit difference in MGP^{*}_{py} corresponds to a 10-percentile difference in actual MGP scores.

App. C, Table C1. ITS Retention, by Four Levels of Teacher MGP (MGP Definition A & B)



Online Appendix D: Matching and SCM Methods

For two outcomes—student achievement and teacher retention—we are able to use district-year level data from the Colorado Department of Education to compare the trajectory of Denver outcomes to those of Colorado school districts that look "similar" to Denver in the pre-period. The idea here is to see whether any observed ProComp effect on Denver's outcomes are isolated to Denver (rather than reflected in other places that did not enact ProComp). We refer to these analysis as Comparative Interrupted Time Series (CITS).

Findings from the CITS approach could naturally depend on how one defines "similarity," and we therefore present results across four reasonable approaches to constructing the comparison groups. In this appendix, we describe the details of those four approaches: (1) a synthetic control method, and two matching methods: (2) matching on a propensity score and (3) matching to districts with similar trends in both outcomes and demographic characteristics during the pre-period.

(1) Synthetic Control Method.

We first use a synthetic control method (SCM) approach (Abadie et al., 2012; Abadie & Gardeazabal, 2003) to estimate a set of district-level weights to construct a synthetic comparison group to DPS. The idea is to execute a comparative case-study using a similar causal inference approach as a difference-in-differences approach. The synthetic control method was specifically designed to address the challenges of using aggregated, time-varying data in which only one unit (here, a district) is exposed to a particular treatment event (here, ProComp) while other units are not. We execute the method using the Stata "synth" program, also developed by Hainmueller, Abadie, and Diamond. Stata's synth program constructs a "…synthetic control group by searching for a weighted combination of control units chosen to approximate the unit affected by the

intervention in terms of the outcome predictors. The evolution of the outcome for the resulting synthetic control group is an estimate of the counterfactual of what would have been observed for the affected unit in the absence of the intervention" (Stata 15 Help file).

When implementing the SCM approach via "synth", the analyst specifies the covariates and pre-treatment years on which units will be matched upon. We include: the district-year mean, standard deviation, and count of math scores (as well as for reading and writing). These are publicly available from the Colorado Department of Education. We also include Common Core of Data district-year level demographics including average pupil-teacher ratio, percentage of students who have an IEP, are free-lunch eligible, are reduced-lunch eligible, are male, and are White, American Indian, Hispanic, Asian, or Black. In total, there are 170 other Colorado school districts that could be used and weighted in the SCM.

When the outcome of interest is teacher retention rate, the pre-period includes 2001- 2006, and the ProComp policy can first affect outcomes in 2007 (the last year of available retention data is 2016). In the pre-period, the difference between Denver's observed teacher retention rate and that of the synthetic control group was less than 0.32 percentage points (i.e., less than 1 percentage point), suggesting a very good comparison.

When the outcome of interest is mean state test scores, the first available year of data is 2004, and the ProComp policy can first affect outcomes in 2007 (the last available year of test score data at the state level is 2014). In the pre-period, the difference between Denver's observed mean test score and that of the synthetic control group was less than 0.001 standard deviations, again suggesting a very good comparison.

(2) Propensity Score Matching.

We use the same data described above to execute a propensity score matching approach. We begin by estimating a mean level and trend in the pre-period for every available district-year level measure (see full list above) and collapse the data to the district level. We then estimate a propensity model (predict the binary outcome of experiencing ProComp or not) as a function of all pre-treatment means and trends. We use that model to estimate a predicted probability of treatment for every district. We then used a nearest neighbor approach to identify matching districts: The 16 districts in the top 10 percent of propensities were selected to comprise the PSM group.

(3) Matching based on all Pre-Policy Variable Trends.

The causal warrant of the CITS model does not require that the pre-period *levels* of variables are comparable, but instead simply the trends. Therefore, for the third matching approach, we match Denver to districts that are trending in the same direction in the pre-period in terms of mean test scores, retention rates, percent free-lunch eligible, percent white, percent Hispanic, and student enrollment. Throughout all of Colorado, there are 13 other districts that exhibit this exact same pattern of pre-treatment trends in both outcomes and demographics, and this group becomes the fourth and final comparison group.