Evaluation of the Teacher Incentive Fund: Final Report on Implementation and Impacts of Pay-for-Performance Across Four Years

December 2017

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

U.S. Department of Education

THIS PAGE IS INTENTIONALLY BLANK

Evaluation of the Teacher Incentive Fund: Final Report on Implementation and Impacts of Pay-for-Performance Across Four Years

December 2017

Hanley Chiang
Cecilia Speroni
Mariesa Herrmann
Kristin Hallgren
Paul Burkander
Alison Wellington
Mathematica Policy Research

Elizabeth Warner
Project Officer
Institute of Education Sciences

NCEE 2018-4004
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

U.S. Department of Education
Betsy DeVos
Secretary

Institute of Education Sciences
Thomas W. Brock
Commissioner of the National Center for Education Research
Delegated Duties of the Director

National Center for Education Evaluation and Regional Assistance
Ricky Takai
Acting Commissioner

December 2017

The report was prepared for the Institute of Education Sciences under Contract No. ED-IES-14-C-0115. The project officer is Elizabeth Warner in the National Center for Education Evaluation and Regional Assistance.

Chiang, Hanley, Cecilia Speroni, Mariesa Herrmann, Kristin Hallgren, Paul Burkander, Alison Wellington. (2017). *Evaluation of the Teacher Incentive Fund: Final Report on Implementation and Impacts of Pay-for-Performance Across Four Years* (NCEE 2017-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at http://ies.ed.gov/ncee.

ACKNOWLEDGMENTS

This study would not have been possible without the contributions of many individuals. We are grateful for the cooperation of many TIF administrators, teachers, principals, district leaders, and central office staff who assisted with the study's data collection and provided important information that shaped the study. A dedicated technical assistance team helped TIF districts implement the programs examined in this study. This team was led by Duncan Chaplin and Jeffrey Max and included Lauren Akers, Kevin Booker, Julie Bruch, Albert Liu, Allison McKie, Debbie Reed, Alex Resch, Christine Ross, and Margaret Sullivan from Mathematica and Patrick Schuermann and Eric Hilgendorf from the Peabody College of Education at Vanderbilt University.

Several individuals made enormous efforts to collect data successfully for this study. Sheila Heaviside and Annette Luyegu provided excellent leadership over our administration of teacher, principal, and district surveys, and Kathy Shepperson oversaw the design of key systems for collecting this survey data. Lauren Akers, Nickie Fung, Chris Jones, Margaret Sullivan, Sarah Wissel, and Claire Smither Wulsin patiently conducted and summarized numerous interviews with TIF administrators. Acquiring and processing administrative data required a large effort led by Jacqueline Agufa and Mary Grider with assistance from Michael Brannan, Kai Filipczak, Chris Jones, William Leith, Mickey McCauley, Sarah Osborn, and Juha Sohlberg.

Many people contributed to the analysis and interpretation of the study's data and the production of this report. The study received useful advice from our technical working group, consisting of David Heistad, James Kemple, Daniel McCaffrey, Anthony Milanowski, Richard Murnane, Jeffrey Smith, and Jacob Vigdor. At Mathematica, Jill Constantine and Steven Glazerman helped shape the evaluation and provided expert advice. Sarah Osborn helped with a variety of critical tasks including facilitating the management of the project. The analysis was made possible by an excellent team of programmers, consisting of Raúl Torres Aragon, Michael Brannan, Molly Crofton, Kai Filipczak, and John Hotchkiss. John Kennedy edited the report, and Jill Miller carefully and patiently prepared the report for publication.

THIS PAGE IS INTENTIONALLY BLANK

CONTENTS

THIS PAGE IS INTENTIONALLY BLANK

TABLES

FIGURES

EXECUTIVE SUMMARY

Research indicates that effective teachers are critical to raising student achievement. However, there is little evidence about the best ways to improve teacher effectiveness, or how schools that serve the students most in need can attract and retain effective teachers. Traditional salary schedules, which pay teachers based on their years of teaching experience and degree attainment, may not reward effective teaching or provide incentives for the most effective teachers to teach in high-need schools.

In 2006, Congress established the Teacher Incentive Fund (TIF), which provides grants to support performance-based compensation systems for teachers and principals in high-need schools. This congressionally mandated study, conducted by the U.S. Department of Education's Institute of Education Sciences, examined the implementation of TIF programs in over 130 districts that received grants awarded in 2010. Ten of these districts, the evaluation districts, agreed to participate in a random assignment study of the pay-for-performance component of TIF. Within those districts, this evaluation provided a more in-depth examination of TIF implementation and measured the impacts of pay-for-performance bonuses on educator effectiveness and student achievement.

This report, the final report from the evaluation, covers all four years of program implementation (2011–2012 through 2014–2015) under the 2010 TIF grants. The main findings include:

- **Within the ten evaluation districts, pay-for-performance led to slightly higher student achievement in reading and math by the second year of implementation.** Student reading achievement was higher by 2 percentile points at the end of the first year in schools that offered pay-for-performance bonuses than in schools that did not. The total difference remained at 1 to 2 percentile points across the subsequent three years and was statistically significant in most years. From the second year onward, the total difference in math achievement was similar in magnitude, but was statistically significant in only one year. In both subjects, these differences were equivalent to about three to four weeks of learning. Impacts on student achievement also differed across districts and schools, but these differences were mostly unrelated to either district or school characteristics.
- **Most 2010 TIF districts implemented each individual component of the comprehensive, performance-based compensation system required under their grant, and about half implemented all four components for teachers.** Districts were required to (1) use measures of both student achievement growth and observations of classroom or school practices to evaluate teachers' and principals' effectiveness, (2) offer educators bonuses based on their performance, (3) offer educators opportunities to earn additional pay for taking on extra roles or responsibilities, and (4) provide professional development to help educators understand the measures on which they were evaluated and improve their performance on those measures. Starting from the first year and continuing through all subsequent years, nearly all the districts reported offering pay-for-performance bonuses (over 90 percent) and additional pay opportunities (over 85 percent). Fewer districts, but still over half, implemented the required measures of teacher effectiveness (about 80 percent) and principal effectiveness (about 60 to 70 percent) and offered the required professional development to their teachers (59 to 74 percent). The percentage of districts that implemented all four components for teachers was similar across the four years (45 to 52 percent).

- **Many 2010 TIF districts reported that sustainability of their program was a major challenge, and slightly fewer than half planned to offer pay-for-performance bonuses after their grant ended.** In each year, about half or more of the districts reported that sustainability of the TIF program was a major challenge (63 percent in the second year, 48 percent in the third year, and 58 percent in the fourth year). Consistent with these concerns, slightly fewer than half (47 percent) of the districts planned to offer bonuses to teachers based on their performance in the 2015–2016 school year, the year after their grant ended. However, most districts planned to continue other key components of their program, including measures of teacher effectiveness similar to those used in TIF (at least 80 percent), additional pay for taking on extra roles or responsibilities (74 percent), and professional development based on teachers' actual performance ratings to help improve their instructional practices (90 percent).

Background on TIF and This Study

TIF Grants and Requirements

Across four rounds of TIF grants (in 2006, 2007, 2010, and 2012), the U.S. Department of Education awarded about $1.8 billion to help states and districts create comprehensive, performance-based compensation systems for teachers and principals.[1]

Box ES.1. Required Components of 2010 TIF Programs

Measures of educator effectiveness based on student achievement growth and at least two observations of classroom or school practices

Pay-for-performance bonuses that were substantial in size, differentiated, challenging to earn, and based solely on educators' effectiveness

Additional pay opportunities for educators to take on extra roles or responsibilities, such as becoming a master or mentor teacher who directly counsels other teachers

Professional development to help educators understand how they were evaluated and to provide feedback based on educators' actual performance ratings to improve their instructional practices

The 2010 TIF grant notice differed from the other rounds in that it included a main and an evaluation competition. Applicants for evaluation grants had to meet the same requirements for the performance-based compensation system as non-evaluation grantees (Box ES.1), but were subject to some additional requirements and guidance. One important requirement was that evaluation grant applicants had to agree to participate in a random assignment evaluation of pay-for-performance bonuses in which only half of their participating schools would offer those bonuses. In addition, evaluation grantees received more specific guidance about the structure of their pay-for-performance bonuses. They received examples of pay-for-performance bonuses that were *substantial* (with an average bonus worth 5 percent of the average educator's salary), *differentiated* (with at least some educators expecting to receive a bonus worth three times the average bonus), and *challenging* to earn (with only those performing significantly better than average receiving bonuses).

[1] The 2015 reauthorization of the Elementary and Secondary Education Act renamed TIF the Teacher and School Leader Incentive Grants program.

The TIF Study

This study addressed four research questions:

1. What were the characteristics of all TIF districts and their performance-based compensation systems? What implementation experiences and challenges did TIF districts encounter?
2. How did teachers and principals in schools that did and did not offer pay-for-performance bonuses compare on key dimensions, including their understanding of TIF program features, exposure to TIF activities, allocation of time, and attitudes toward teaching and the TIF program?
3. How did pay-for-performance bonuses affect educator effectiveness and the retention and recruitment of high-performing educators?
4. What was the impact of pay-for-performance bonuses on students' achievement on state assessments in math and reading?

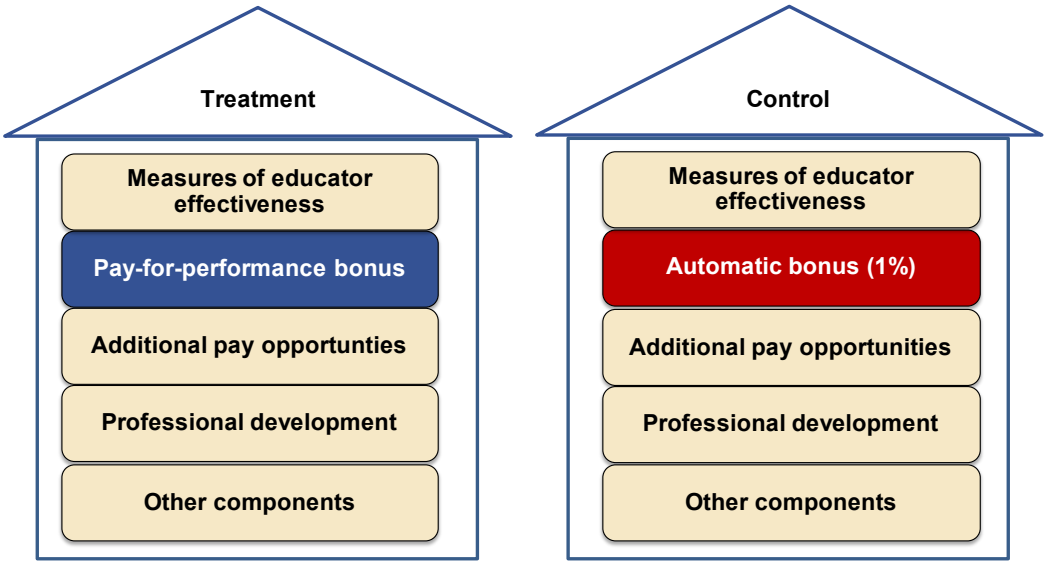
This report addresses each of the research questions with information from all four years of TIF implementation (2011–2012 to 2014–2015). Previous reports from this study (Max et al. 2014; Chiang et al. 2015; Wellington et al. 2016) covered prior years within the grant period.

Districts and schools in the study. The study provides a broad overview of program implementation by all 2010 TIF districts based on districts that had a TIF program in place in each year—153 districts in the first year, 155 in the second year, 144 in the third year, and 139 in the fourth year. This report's in-depth analyses of TIF implementation and the effects of pay-for-performance were based primarily on the ten evaluation districts and 131 schools in grades 3 to 8 that completed all four years of TIF implementation.

Compared with the average U.S. district, TIF districts were larger, were more likely to be urban and located in the South, and had a higher proportion of students who were racial or ethnic minorities and eligible for free or reduced-price lunches (Max et al. 2014). On average, evaluation districts were larger than non-evaluation TIF districts and had smaller percentages of students who were white (38 versus 50 percent). Evaluation districts were also more likely than non-evaluation districts to be in urban areas (69 versus 29 percent) and less likely to be in states with right-to-work laws (54 versus 67 percent).

Random assignment study design. To assess the impacts of pay-for-performance on educator and student outcomes, the study team assigned schools randomly—that is, completely by chance—to offer pay-for-performance bonuses (treatment group, 65 schools) or not (control group, 66 schools). As shown in Figure ES.1, treatment and control schools were expected to implement all other components of TIF. Because the two groups were expected to differ only in the opportunity for educators to receive pay-for-performance bonuses, differences in outcomes between the groups could be attributed to the impact of pay-for-performance. Specifically, the study measured the impact of pay-for-performance bonuses implemented within the context of broader performance-based compensation systems.

Figure ES.1. Random Assignment Evaluation Design



Data sources. Data for this report came from multiple sources (Table ES.1), all with response rates greater than 85 percent. The sources enabled us to examine implementation broadly in all TIF districts and, within evaluation districts, to report on more detailed aspects of implementation and the impacts of pay-for-performance on educator and student outcomes.

Table ES.1. Main Data Sources for This Report

Data Source	How the Study Used this Data Source
Data Collected from Evaluation and Non-Evaluation Districts	
District survey	Describe broadly the districts' programs and implementation challenges
Data Collected from Evaluation Districts Only	
District interview	Obtain more detail on districts' programs and implementation challenges
Principal and teacher surveys	Describe educators' understanding of TIF and measure impacts of pay-for-performance on educators' attitudes
Student administrative data	Measure impacts of pay-for-performance on student achievement
Educator administrative data	Describe teachers' and principals' performance ratings and bonuses; measure impact of pay-for-performance on educator performance ratings

Analysis approach. To describe districts' implementation of TIF, we calculated averages (such as the average of the largest bonuses districts awarded) and percentages (such as the percentage of districts that used specific effectiveness measures). To measure the impacts of pay-for-performance on educator and student outcomes, we compared outcomes in treatment and control schools.

Additional Findings from the Evaluation Districts

To provide context for interpreting the student achievement impacts that we found within the evaluation districts, we examined the districts' implementation of TIF in detail and measured impacts on a variety of other outcomes that could shape the student achievement impacts. In particular, we examined (1) the districts' overall implementation of the TIF required components, (2) the measures used to evaluate educators' effectiveness, (3) the structure of pay-for-performance bonuses, (4)

educators' understanding of and experiences with key components, (5) impacts on teachers' attitudes, and (6) impacts on educator effectiveness.

Overall Implementation of TIF Required Components

Most evaluation districts reported implementing all required components for teachers at the start of their programs. In each year, the only component for teachers that was not implemented by all of the districts was the professional development required by the grant. Starting from the first year, all evaluation districts reported using student achievement growth and at least two observations by trained observers to evaluate teachers, and by the second year, all districts used such measures to evaluate principals as well. In all years, every district also offered bonuses based on how educators performed on the effectiveness measures and offered additional pay to take on extra roles or responsibilities. However, only 6 to 8 of the 10 evaluation districts reported providing the required professional development for teachers, with no clear pattern of increasing or decreasing implementation over the course of the grant.

Measures of Educator Effectiveness

When evaluating educators, districts were required to use student achievement growth and observation measures, but they were given discretion to design the details of these measures. For example, when evaluating teachers based on student achievement growth, districts could measure the achievement growth of the teachers' own students (classroom achievement growth); all students in the same grade, team, or subject area (achievement growth of student subgroups); all students in the school (school achievement growth); or some combination of these measures. When considering their performance ratings from multiple measures, teachers may be more likely to believe the ratings are meaningful and accurate if the ratings are based on their individual performance, are more similar across measures, and are consistent across years.

All evaluation districts reported using observations and achievement growth to evaluate teachers, as required, and most used classroom achievement growth. More than half (60 to 70 percent) of the districts reported evaluating teachers based on classroom achievement growth. Within these districts, about 40 to 60 percent of teachers received classroom achievement growth ratings, typically because they taught grades and subjects tested by state assessments. All districts also used school achievement growth.

Most teachers received similar performance ratings from one year to the next, with many teachers receiving higher ratings on classroom observations than on achievement growth. For example, more than half of teachers received similar ratings, based on a 1-to-4 rating scale, in Year 4 as they did in Year 3 on classroom observations (56 percent), school achievement growth (51 percent), and classroom achievement growth (51 percent). However, in each year, teachers often earned higher ratings on observations than on achievement growth. For example, in Year 4, more than half (56 percent) of teachers received a higher rating on observations than on school achievement growth, and about half (49 percent) received a higher rating on observations than on classroom achievement growth.

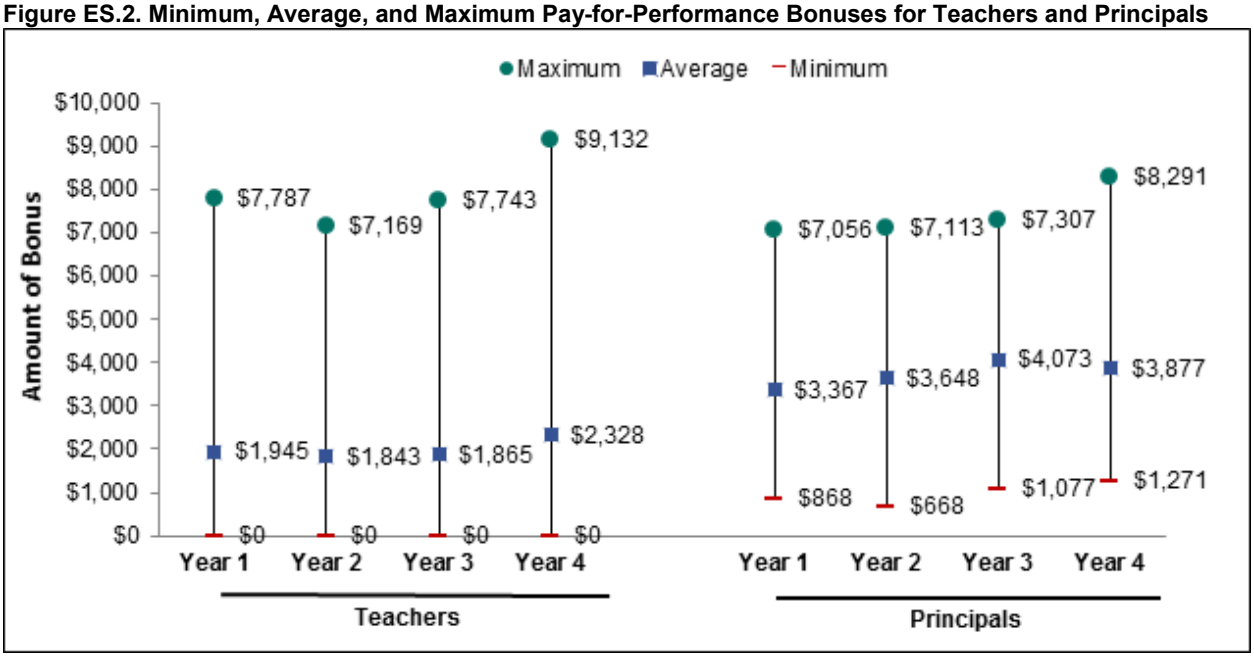
Structure of Pay-for-Performance Bonuses

To enhance the potential for performance bonuses to motivate teachers and principals to improve, the TIF notice required that the bonuses had to be challenging to earn, substantial in size, and differentiated.

Most teachers and principals received a bonus, a finding inconsistent with making bonuses challenging to earn. The average bonus fell somewhat short of the guidance to make bonuses substantial. The structure of the bonuses was similar across all four years of TIF implementation. In each year, within schools that offered performance bonuses, most teachers (about 70 percent) received a bonus, and the average bonus was about $2,000 (Figure ES.2), representing about 4 to 5 percent of the average teacher salary. Similar to teachers, most principals (more than 70 percent) received a bonus, and the average bonus (about $4,000) was about 3 to 5 percent of their average salary.

Bonuses were differentiated, with the highest-performing teachers earning bonuses significantly larger than the average bonus. In each year, the maximum bonus for teachers (about $7,000 to $9,000) was roughly four times the average bonus (Figure ES.2). For principals, bonuses were less differentiated, with the maximum bonus consistently about twice the average bonus.

Figure ES.2. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers and Principals



Source: Educator administrative data (N = 2,151 teachers in Year 1, N = 2,160 teachers in Year 2, N = 2,236 teachers in Year 3, and N = 2,083 in Year 4; N = 65 principals in Year 1, N = 68 principals in Year 2, N = 65 principals in Year 3, and N = 64 principals in Year 4).

Figure reads: In Year 1, on average across the evaluation districts, the minimum pay-for-performance bonus for teachers was $0, the average pay-for-performance bonus was $1,945, and the maximum pay-for-performance bonus was $7,787.

Educators' Understanding of and Experiences with Key Components

For the components of TIF to lead to improvements in educators' practices, districts had to effectively communicate information about those components to educators, and educators needed to know how to improve their performance.

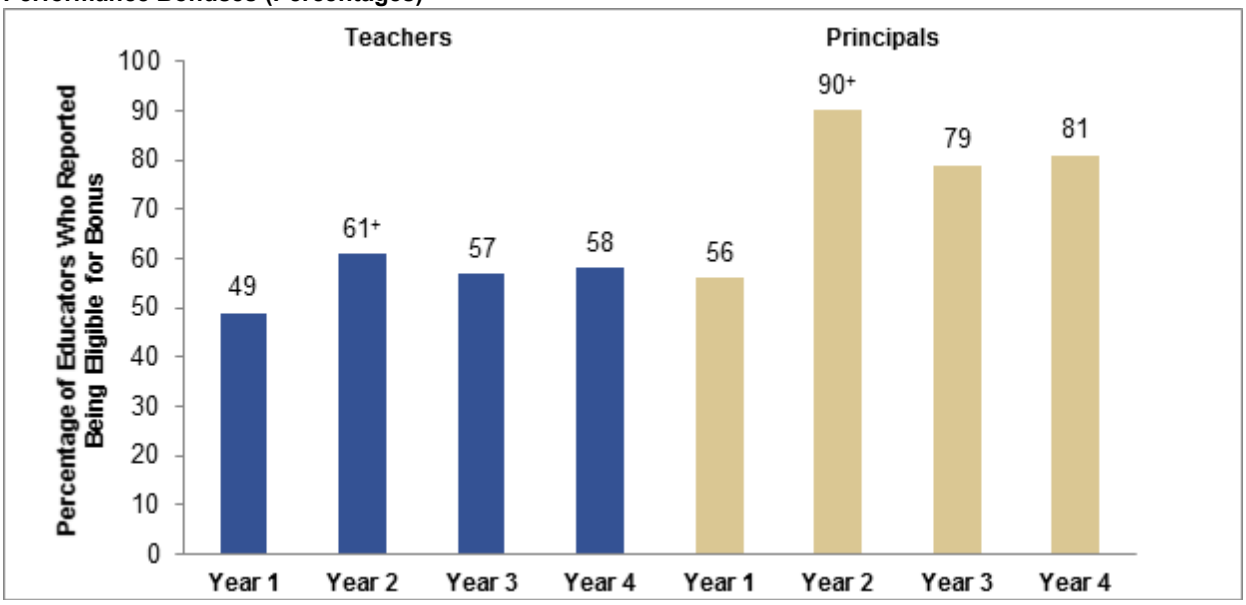
Most teachers were aware of being evaluated based on student achievement growth and classroom observations early in TIF implementation, and their awareness of these performance measures improved over time. At least 70 percent of teachers in Year 1 reported being evaluated on student achievement growth, and more than 70 percent reported being evaluated on at least two classroom observations. Furthermore, the percentage of teachers who reported being evaluated on these measures increased over time, with significant improvements in some years for both treatment and control teachers. By Year 4, 89 percent of treatment teachers and 78 percent of control teachers reported being evaluated on student achievement growth, and more than 85 percent of teachers reported being evaluated on at least two classroom observations. Similar to teachers, about 85 percent of principals in Year 4 reported being evaluated on student achievement growth; however, a smaller percentage of principals (less than 60 percent) reported being evaluated on at least two observations of their school practices.

Many teachers in treatment schools did not understand that they were eligible to earn a performance bonus, and their understanding did not improve after the second year of implementation. Although teachers' and principals' understanding of their bonus eligibility improved significantly from Year 1 to Year 2, there was no further improvement beyond Year 2. In Years 2 through 4, about 60 percent of treatment teachers (for example, 58 percent in Year 4) were aware that they could potentially earn a performance bonus, implying that about 40 percent were still unaware (Figure ES.3). Although understanding of eligibility was better among principals than teachers, about 20 percent of principals in Year 4 still did not know they were eligible to earn a bonus based on their performance.

Teachers underestimated how much they could earn from a performance bonus. In each year, teachers in treatment schools believed that the maximum bonus they could earn was no more than 40 percent of the actual maximum bonus districts awarded (Figure ES.4). Principals also underestimated the potential amount of performance bonuses they could receive, but their beliefs better aligned with actual bonus payouts than did teachers' beliefs. Across years, the maximum pay-for-performance bonus that principals reported they could receive ranged from 67 to 91 percent of the actual maximum bonus that districts awarded.

Most teachers reported receiving professional development required under the TIF grant but indicated they received only a few hours of it over the school year. In each year, more than half of teachers reported that they received or expected to receive professional development focused on understanding performance measures used in TIF (ranging from 77 percent of treatment teachers in Year 1 to 53 percent in Year 4). Most teachers (50 to 60 percent) also reported receiving or expecting to receive feedback based on their performance ratings. Of those who expected to receive any professional development on these two topics, the expected amount of time on each topic per year ranged from two to six hours.

Figure ES.3. Teachers and Principals in Treatment Schools Who Reported Being Eligible for Pay-for-Performance Bonuses (Percentages)

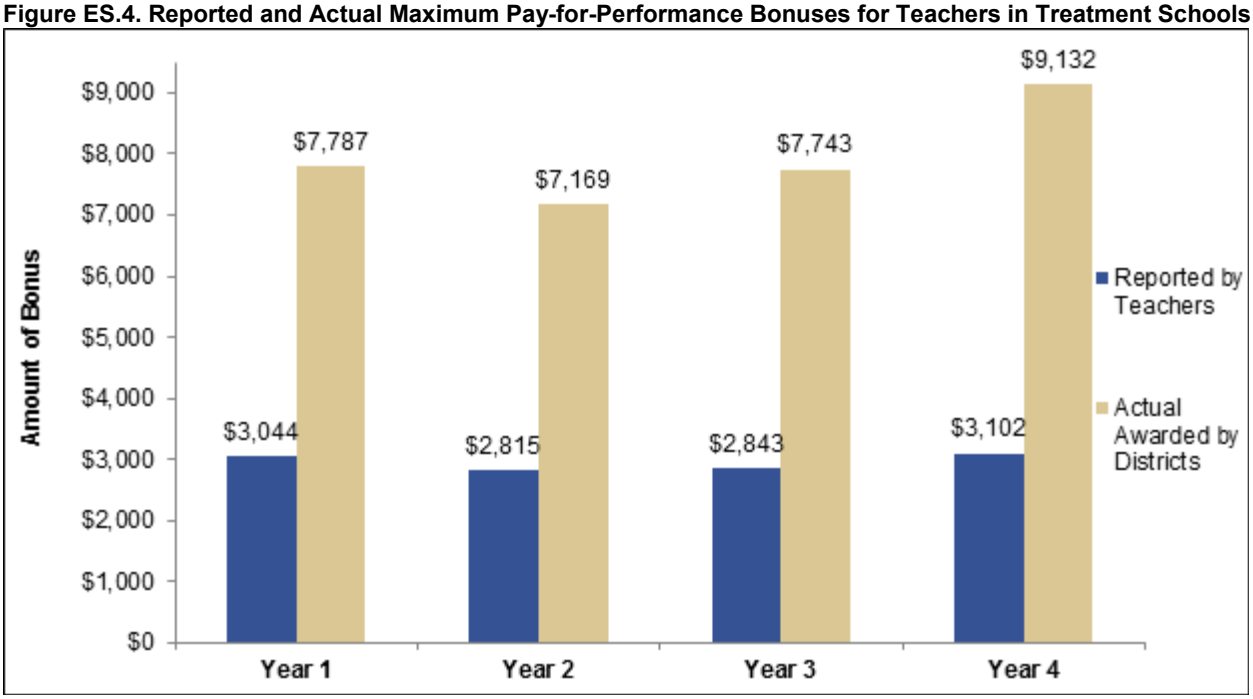


Sources: Teacher and principal surveys, 2012, 2013, 2014, and 2015 (N = 368 teachers in Year 1; N = 439 teachers in Year 2; N = 420 teachers in Year 3; N = 391 teachers in Year 4; N = 63 principals in Year 1; N = 62 principals in Year 2; N = 57 principals in Year 3; and N = 60 principals in Year 4).

Figure reads: In Year 1, 49 percent of teachers in treatment schools reported being eligible for a pay-for-performance bonus.

+Difference with prior year is statistically significant at the .05 level, two-tailed test.

Figure ES.4. Reported and Actual Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools



Sources: Teacher surveys, 2012, 2013, 2014, and 2015 (N = 218 teachers in Year 1; N = 229 teachers in Year 2; N = 229 teachers in Year 3; and N = 210 teachers in Year 4) and educator administrative data.

Figure reads: In Year 1, on average, the actual maximum pay-for-performance bonus that evaluation districts awarded to teachers was $7,787, and the maximum pay-for-performance bonus teachers reported they could earn was $3,044.

Impacts of Pay-for-Performance on Teachers' Attitudes

To the extent that pay-for-performance enhances or worsens teachers' job satisfaction, their motivation to improve could increase or decrease as a result.

Most teachers were satisfied with their jobs and the TIF program. In each year of TIF implementation, at least two-thirds of teachers in both treatment and control schools reported being satisfied with their job overall and were glad to be participating in TIF. On each specific aspect of their professional opportunities, evaluation system, and school environment, at least half of teachers reported being satisfied.

Although initially less satisfied with their jobs and TIF, teachers in treatment schools were as satisfied as, and sometimes more satisfied than, teachers in control schools by the third year of implementation. In the first two years of TIF implementation, teachers in treatment schools tended to report being less satisfied than teachers in control schools. For example, in Year 2, treatment teachers reported being less satisfied than control teachers with the use of student achievement scores to assess their performance (61 versus 69 percent) and recognition of their accomplishments (61 versus 66 percent). They were also more likely to report that TIF reduced their freedom to teach (39 versus 30 percent) and harmed collaboration (29 versus 21 percent). However, in Years 3 and 4, treatment teachers were no longer less satisfied than control teachers on any of the measured dimensions and, in fact, reported being more satisfied on some dimensions. For example, in Year 4, more treatment teachers than control teachers were satisfied with feedback on their performance (87 versus 77 percent) and believed TIF caused teachers to work more effectively (59 versus 51 percent).

Impacts of Pay-for-Performance on Educator Effectiveness

The rationale behind pay-for-performance is that it can improve student achievement by enhancing educator effectiveness. In light of the positive impacts on student achievement, we examined whether pay-for-performance led to improved classroom practices in ways that were detected by trained observers and were related to higher student achievement.

Pay-for-performance led teachers to earn slightly higher classroom observation ratings by the third year of implementation. Differences between the classroom observation ratings of teachers in treatment schools and those in control schools grew over the four years of implementation and became statistically significant by Year 3. In Years 3 and 4, treatment teachers earned observation ratings that were 0.05 and 0.09 points higher on a 1-to-4 scale than those of control teachers. Pay-for-performance had no impact on the observation ratings of principals.

The impacts of pay-for-performance on classroom observation ratings did not appear to explain the impacts on student achievement. Schools that experienced larger impacts of pay-for-performance on observation ratings did not experience larger impacts on student math or reading achievement. Therefore, although classroom observation ratings detected some improvements in practices due to pay-for-performance, they did not identify the improvements that were actually associated with student achievement.

Concluding Thoughts

Overall, the 2010 TIF districts were able to implement most components of a comprehensive, performance-based compensation system, typically starting early in their grant period and lasting throughout all four years of their programs. However, many districts anticipated that sustaining the TIF program beyond the life of their grant would be difficult. In particular, fewer than half of the 2010 TIF districts planned to continue offering performance bonuses.

A primary objective of TIF grants is to raise student achievement in high-need schools. Within the evaluation districts, this study found that the pay-for-performance component of TIF had small, positive impacts on student achievement by the second year of implementation. From that year onward, reading and math achievement was higher by 1 to 2 percentile points in schools that offered performance bonuses than in schools that did not.

To draw lessons from these impact findings for future policies on performance-based compensation, it is useful to consider possible explanations for why performance bonuses within the TIF program had any positive impacts on student achievement and why those impacts were small. According to the rationale behind pay-for-performance bonuses, they can improve student achievement only if educators (1) face a bonus structure that provides meaningful incentives for improvement, (2) understand key components of the program, (3) feel motivated to adjust their practices or their choice of where to work to earn these bonuses, and (4) know how to change their practices in ways that improve student achievement. As we summarize next, some, but not all, of these factors were in place within the evaluation districts.

First, the structure of the bonuses provided a mix of stronger and weaker incentives for teachers to improve. The highest-performing teachers could earn a performance bonus worth about four times the average bonus, which provided an incentive for teachers to demonstrate high performance. However, in each year, the criteria for earning any bonus resulted in about 70 percent of teachers earning bonuses within schools that offered them. Therefore, even some teachers who performed worse than the typical teacher earned a bonus. If failing to receive a bonus represented a clear signal about having room for improvement, the bonus structure gave a minority of teachers this type of encouragement to improve.

Second, educators' misunderstanding of these bonuses may have hampered the degree to which this policy could influence educators' behavior. Across all four years of implementation, many teachers were unaware they were eligible for a performance bonus or underestimated the amount they could earn. These teachers perceived limited or no monetary incentives to improve their performance even though their districts had actually structured the bonuses to provide stronger incentives.

Third, some teachers perceived schools that offered performance bonuses to be a more appealing place to work, potentially enhancing their motivation to remain at these schools and improve their practices. By the third year of TIF implementation, teachers' satisfaction with key aspects of their jobs was either unchanged or improved as a result of pay-for-performance. However, we have no evidence that these favorable impacts on teachers' job satisfaction contributed to improvements in student achievement. In fact, positive impacts on student achievement emerged in the first two years—a period when pay-for-performance was actually *lowering* satisfaction.

Fourth, it is unclear whether teachers knew how to change their classroom practices in ways that could improve student achievement. By the third year of implementation, pay-for-performance led to small increases in teachers' classroom observation ratings. However, schools that experienced larger impacts of pay-for-performance on observation ratings did not experience larger impacts on student achievement. Therefore, we found no indication that changes in teachers' measured practices were the source of the improvements in student achievement. The disconnect between changes in measured practices and changes in achievement could have been due to a number of factors, which this study did not have the data to examine. One possibility is that the amount of targeted professional development teachers received—no more than six hours over the school year—was insufficient to promote changes in practices that were substantial enough to improve student achievement. Another possibility is that the observation measures encouraged teachers to focus on aspects of instruction that were not related to student achievement.

Although the impacts of pay-for-performance were small, the costs of the bonuses were also low enough such that this policy was at least as cost-effective as some alternative policies that have been evaluated. Specifically, a cost-effectiveness analysis suggests that pay-for-performance was more cost-effective than class-size reduction (through four years of program implementation) and about as cost-effective as providing transfer incentives for high-performing teachers to move to low-performing schools (at the end of two years).[2] However, the available evidence cannot predict the policies' cost-effectiveness beyond the limited number of years in which these policies were implemented and evaluated.

[2] For example, after four years, raising student achievement by the actual impact of pay-for-performance required spending $499 per student on pay-for-performance, but would have required spending $767 per student on class-size reduction. However, to achieve the same impact as pay-for-performance did after two years, transfer incentives would require $193 per student, nearly identical to the $196 per-student cost of pay-for-performance.

THIS PAGE IS INTENTIONALLY BLANK

I. INTRODUCTION

Research indicates that effective teachers are critical to raising student achievement. However, there is little evidence about the best ways to improve teacher effectiveness, or how schools that serve the students most in need can attract and retain effective teachers. Traditional salary schedules, which pay teachers based on their years of teaching experience and degree attainment, may not reward effective teaching or provide incentives for the most effective teachers to teach in high-need schools.

In 2006, Congress established the Teacher Incentive Fund (TIF), which provides grants to support performance-based compensation systems for teachers and principals in high-need schools.[1] The TIF grants have two goals:

- Reform compensation systems to reward educators for improving student achievement
- Increase the number of high-performing teachers in high-need schools and hard-to-staff subject areas

The incentives and support offered through TIF grants aim to improve student achievement by improving educator effectiveness and the quality of the teacher workforce.

The U.S. Department of Education (ED) commissioned a multiyear study of the TIF grants awarded in 2010.[2] The study had two main goals. First, it sought to inform program development and improvement by describing how grantees implemented their performance-based compensation systems and the implementation challenges they faced. Second, it sought to test whether pay-for-performance bonuses as part of a comprehensive reform system could lead to increases in educator effectiveness and student achievement.

This report is the fourth and final report from the evaluation of the 2010 TIF grants. The first report (Max et al. 2014) examined grantees' implementation experiences and educators' perspectives on the program near the end of the first year of program implementation, before the first pay-for-performance bonuses were awarded to teachers and principals. The second and third reports (Chiang et al. 2015; Wellington et al. 2016) examined grantees' implementation experiences and educators' understanding of, and attitudes toward, the program near the end of the second and third year of program implementation, as well as changes in educators' understanding and attitudes. They also examined the impacts of pay-for-performance bonuses on educator effectiveness and student achievement after one, two, and three years of TIF implementation. This final report provides findings on grantees' implementation of TIF, educators' understanding and attitudes toward the program, and the impacts of pay-for-performance bonuses on educator effectiveness and student achievement over all four years of the program.

[1] The 2015 reauthorization of the Elementary and Secondary Education Act renamed TIF the Teacher and School Leader Incentive Grants program. This program provides grants to eligible entities to develop, implement, improve, or expand performance-based compensation systems or human capital management systems in schools.

[2] The U.S. Department of Education has awarded four rounds of TIF grants – in 2006, 2007, 2010, and 2012. It also awarded a round of Teacher and School Leader Incentive Grants in 2016. For this report, all references to TIF are for the 2010 awardees.

Previous Research on Pay-for-Performance Programs for Educators

Existing research has produced inconsistent findings on the effectiveness of pay-for-performance programs in U.S. public schools. Although some studies have found positive impacts of these programs on student achievement, most have not.[3]

However, the existing studies have one or more key limitations (see Max et al. [2014] and Chiang et al. [2015] for a more detailed discussion of the literature and its limitations). First, one limitation for many studies was their research design. For example, many studies did not have a random assignment design and, therefore, left open the possibility that observed outcomes were due to unobserved school, educator, or student characteristics, rather than the offer of pay-for-performance programs. All of the random assignment studies included schools from only one district, making it difficult for policymakers to determine whether the study findings can be generalized more broadly. Second, several of the pay-for-performance programs examined by previous studies provided bonuses that were small, similar for all teachers regardless of their effectiveness, easy to earn, or not well-explained to teachers. Third, the performance bonuses were not always part of a more comprehensive reform package that would help teachers change their teaching practices. Overall, there is still little high-quality evidence on comprehensive, well-implemented pay-for-performance programs.

Previous research on the design, implementation, and effects of pay-for-performance has informed the design and evaluation of the TIF grants. In addition, targeted technical assistance supported program implementation to help ensure programs were well designed. This series of reports is the first to present findings from a large, multisite random assignment study of the impact of pay-for-performance, as part of a comprehensive reform system, on educator effectiveness and student achievement.

In the following sections, we provide a framework for the evaluation by describing key components of TIF grants and presenting a logic model of how pay-for-performance could influence student outcomes.

TIF Grant Competition

From 2006 to 2012, ED awarded about $1.8 billion to support 131 TIF grants. ED awarded 16 grants in 2006, 18 in 2007, 62 in 2010, and 35 in 2012. The TIF grants awarded in 2010 ranged from $607,211 to $62,325,746 over a five-year period.[4] Among the 62 TIF grantees in 2010, more than two-thirds were states or school districts (69 percent), 16 percent were nonprofits, 13 percent were charter schools or charter management organizations, and 2 percent were universities. Grantees that were not states or school districts had to partner with a state or local education agency. The 2010

[3] Studies of how pay-for-performance programs affect, or are associated with, student achievement or teacher retention in U.S. public schools include Balch and Springer (2015); Bayonas (2010); Chiang et al. (2015); Dee and Wyckoff (2015); Fryer (2013); Fryer et al. (2012); Fulbeck (2014); Glazerman et al. (2009); Glazerman and Seifullah (2010, 2012); Goldhaber and Walch (2012); Goodman and Turner (2011); Imberman and Lovenheim (2015); Marsh et al. (2011); Shifrer et al. (2013); Slotnick et al. (2013); Sojourner et al. (2014); Springer et al. (2009a, 2009b); Springer et al. (2011); Springer et al. (2012); Springer et al. (2014); Springer et al. (2016); Springer and Taylor (2016); and Wellington et al. (2016).

[4] A full list of the 2010 TIF grantees can be found at http://www2.ed.gov/programs/teacherincentive/awards.html.

grants were supported, in part, by the American Recovery and Reinvestment Act of 2009 (ARRA). As part of this funding, Congress required a rigorous evaluation of the 2010 grantees, which are the focus of this report.

The 2010 TIF grants were designed to create comprehensive, performance-based compensation systems that could provide (1) incentives for educators to become more effective in improving student achievement in high-need schools, and (2) support for educators to improve their performance. Consistent with the grants' focus on high-need schools, every school that received TIF funds needed to have at least half of its students receiving free or reduced-price lunch. The 2010 TIF grants differed from prior TIF grants by providing more detailed guidance on the measures used to evaluate educators and on the design of the pay-for-performance bonuses. The 2010 grants required four components in performance-based compensation systems implemented in districts, as well as five core elements needed to support the initial and ongoing implementation of the compensation systems. Next, we summarize these four required components.

Required Components of the Performance-Based Compensation Systems

1. **Measures of educator effectiveness.** Grantees were required to use a comprehensive, multiple-component measure of effectiveness for teachers and principals. The measures had to include student achievement growth and at least two observations of classroom or school practices. In addition, the evaluation had to give significant weight to student achievement growth—defined as the change in student achievement for an individual student between two or more points in time. Only trained observers using objective, evidence-based rubrics could conduct the observations. Grantees had discretion to include additional measures.
2. **Pay-for-performance bonuses.** Grantees were required to offer bonuses to educators based on how they performed on the effectiveness measures. The bonuses were designed to incentivize educators and to reward them for being effective in their classroom and schools. There were no additional requirements for earning the bonuses beyond performing well on the effectiveness measures. To provide a strong incentive for the most effective educators, bonuses were to be differentiated and substantial enough to lead to changes in the behavior of teachers and principals to improve student outcomes.
3. **Additional pay opportunities.** The performance-based compensation systems had to include pay opportunities for educators to take on additional roles or responsibilities. These roles might include becoming a master or mentor teacher who directly counsels other teachers or develops or leads professional development sessions for teachers. Limiting these additional pay opportunities to educators identified as effective could also provide an incentive for educators to improve their effectiveness. However, those educators would need to agree to take on leadership roles and perhaps work additional hours.
4. **Professional development.** TIF grantees were required to support teachers and principals in their performance improvement efforts. Support included providing information about measures on which educators would be evaluated and more targeted professional development based on an educator's actual performance on the effectiveness measures. Specifically, districts were required to provide educators with

feedback and professional development on how to alter their pedagogy or practices to improve along the measures.

These four components of a performance-based compensation system were required of all grantees. In addition, ED encouraged the use of other components that would provide additional pay by awarding points to applicants that included these features in their performance-based compensation systems. For example, districts could offer additional pay to effective educators who agreed to work in hard-to-staff subjects, such as secondary math and science in high-need schools.

Core Elements Designed to Support Implementation of the Performance-Based Compensation System

TIF grantees also were required to have the proper supports to implement and maintain the performance-based compensation system. The five core elements were (1) the involvement and support of teachers, principals, unions (if applicable), and other personnel needed to carry out the TIF grant; (2) a rigorous, transparent, and fair evaluation system for teachers and principals; (3) a plan to effectively communicate the components of the grantee's performance-based compensation system; (4) a plan for ensuring educators understood the measures of educator effectiveness; and (5) a data management system that could link student achievement data to educator payroll and human service systems (see Max et al. [2014] for more details on the core elements).

The required components of the performance-based compensation system are comprehensive and designed to work together, so grantees had to have the core elements in place before implementing their compensation systems. Grantees that did not have all the core elements in place when they were awarded their grants in 2010 were required to spend the 2010–2011 school year planning and developing the support for implementation, and most grantees used the 2010–2011 school year as a planning year (Max et al. 2014). All grantees were required to begin implementation of their performance-based compensation systems by the 2011–2012 school year.

Areas of Discretion in Performance-Based Compensation System Designs

Although the TIF grant required grantees to include specific components in the performance-based compensation system, it gave them substantial discretion in designing and implementing these components. For example, grantees could assess a teacher's measured effectiveness based on the achievement growth of that teacher's students, all students in the same grade, the entire school, or some combination of these measures. Grantees could measure student achievement growth using a value-added model or by calculating the change in students' achievement on a standardized test from one year to the next. They could use models developed by the district, a vendor, or the state. Grantees could decide which rubrics they wanted to use to observe teachers and principals, the number of observations in a year (as long as there were at least two), and which staff members to train as observers. The criteria for earning a bonus based on the effectiveness measures also could vary (for example, criteria might require scoring above a predetermined threshold or in the top percentiles on individual measures or a combination of measures). Grantees could choose bonus amounts based on educator performance. Finally, grantees could choose whether to offer retention and recruitment incentives (such as stipends) to educators to teach in high-need schools or to teach hard-to-staff subjects in those schools.

Additional Requirements for Evaluation Grantees

The 2010 TIF grant notice differed from the other rounds of the TIF grants in that it included a main competition and an evaluation competition (Max et al. 2014). By holding two separate competitions, ED created a sample of grantees that, by virtue of having applied for an evaluation grant, had indicated their interest and willingness to participate in a more in-depth evaluation of their TIF grants.

Evaluation grantees had to meet three additional grant requirements. First, they had to agree to participate in a random assignment evaluation of pay-for-performance bonuses. Schools within a district were randomly assigned to implement either all four required components of the performance-based compensation system program, including pay-for-performance bonuses (the treatment group), or all components *except* pay-for-performance bonuses (the control group). Second, evaluation grantees were required to include at least eight elementary or middle schools in the evaluation. Third, they were obligated to cooperate with all data collection activities for the evaluation.

Applicants for the evaluation grants were also given more specific guidance about the structure of their pay-for-performance bonuses. They received examples of pay-for-performance bonuses that were *substantial* (with an average bonus worth 5 percent of the average educator salary), *differentiated* (with at least some educators expecting to receive a payout worth three times the average bonus), and *challenging* to earn (with only those performing significantly better than the average receiving bonuses). Although applicants had discretion over the proposed structure of the pay-for-performance bonuses, these examples provided additional guidance to evaluation applicants and may have influenced how they designed their performance-based compensation systems.

In return for meeting the additional grant requirements, evaluation grantees received an extra $125,000 per school that participated in the evaluation. The money could be used to support the implementation of TIF—for example, to cover the cost of academic coaches or release time for professional development activities—as well as costs associated with the evaluation, such as data collection activities. The use of the funds also had to be consistent with the evaluation. For example, they could not be used to offer pay-for-performance in control schools.

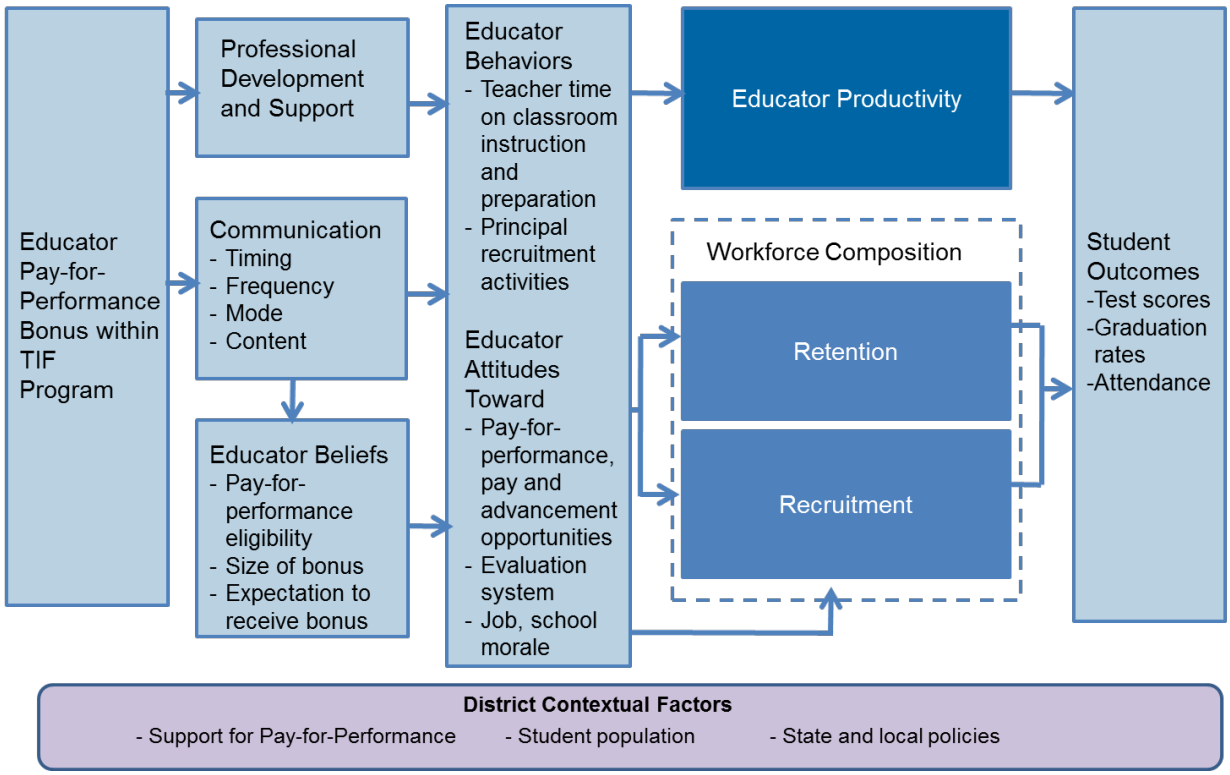
ED monitored all grantees to ensure implementation was consistent with grant requirements. Although ED ensured all grantees received technical assistance, it used two providers—one for the non-evaluation grantees and one for the evaluation grantees. Resources for the evaluation grantee technical assistance team helped ensure that the evaluation grantees received intensive and targeted assistance. The evaluation grantee technical assistance team encouraged and supported evaluation grantees to incorporate criteria for their pay-for-performance bonuses consistent with their specific grant and in keeping with the examples provided in the grant notice. The goal of the technical assistance provided to all grantees was to ensure strong implementation that could bring about change in educational practices to improve student achievement, as specified in the logic model described below.

Logic Model: How Pay-for-Performance Could Influence Student Outcomes

The requirements of the TIF grant, as well as the design of the evaluation of pay-for-performance bonuses, were informed by a theory of change for how pay-for-performance, within a comprehensive TIF performance-based compensation program, might lead to improved student

outcomes. We developed a logic model to show the pathways by which the pay-for-performance component of TIF could influence student outcomes (Figure I.1). These pathways show the type of information needed to determine whether pay-for-performance is having a positive, negative, or neutral effect and thus informed the data collected as part of the evaluation.

Figure I.1. Logic Model



As the starting point for the theory of change, districts adopt a TIF program that includes pay-for-performance bonuses for rewarding educators based on their measured effectiveness. Pay-for-performance bonuses are meant to improve student outcomes (final column of Figure I.1) by enhancing educator effectiveness (fourth column), either through motivating educators to become more productive or attracting and retaining effective educators at the schools that offer those bonuses.

Improvements in educator effectiveness must come from changes in educators' behaviors. Educators might change any of several behaviors in response to the opportunity to earn performance bonuses (third column of Figure I.1). Many of these possible responses involve changing *how much* time educators spend on their jobs or *how they use* that time. For example, to increase their effectiveness ratings and thus their prospects for earning a bonus, teachers could (1) find ways to improve their practices, such as collaborating more frequently with colleagues or soliciting feedback more actively; (2) increase their effort, such as devoting more hours to preparation or instruction outside of the school day; or (3) behave strategically by shifting attention toward activities that can improve their ratings, such as practices aimed at raising test and classroom observation scores, and away from activities that have less bearing on their effectiveness ratings. Principals might also alter their recruitment practices by putting emphasis on the opportunity to earn performance bonuses when attempting to attract more effective teachers to their schools. However, for pay-for-performance bonuses to be an effective recruitment tool, educators must actually find it

appealing to work at schools that offer such bonuses, underscoring the importance of both educators' attitudes and their behaviors in shaping the impacts of pay-for-performance.

Whether and how pay-for-performance bonuses actually lead to changes in educator productivity and the composition of the teaching workforce depend on many other factors as well. For example, educators must be aware they are eligible to earn a bonus. Simply adopting a well-designed pay-for-performance program will not change teaching practices if educators do not know they are eligible. In addition, educators may be incentivized by pay-for-performance bonuses only if they understand how they are being evaluated and how they can change their teaching practices to improve their performance. They also must believe they are being evaluated consistently and fairly and that the bonuses are attainable and large enough to warrant changing their behavior. The critical role communication and professional development play in the logic model (second column of Figure I.1) highlights the emphasis on these activities required by the grant.

Educators' understanding of their TIF program will depend on districts' communication activities, the timing of this communication, and educators' receiving the information. How often districts communicate information about the program, the type of information they provide, and how they provide the information can affect how well teachers and principals understand their TIF program. Furthermore, educators must be made aware of the program when there is still sufficient time to affect their school choice (for example, request a school transfer) or to alter their teaching practices. Yet even a well-communicated program may be misunderstood if the program is complicated or if educators do not attend informational meetings or read the materials offered.

The ability of pay-for-performance bonuses to affect educator behaviors and attitudes also depends on the district context, such as educators' support for performance bonuses and the presence of other policies. If few educators in a school support pay-for-performance initiatives, adopting such a program may diminish school morale and job satisfaction, thereby decreasing productivity or inducing effective educators to leave the school. District hiring policies, such as hiring freezes, may restrict mobility and negate potential benefits. Other existing policies, such as the requirements for teacher tenure, may already provide strong incentives for educators to improve student outcomes, diminishing the potential impact of performance bonuses. Finally, for schools at risk of closing because they have been designated as needing improvement, the introduction of a pay-for-performance program may not provide additional incentive for change.

Even a well-designed and well-implemented comprehensive compensation reform program may take more than a year before it can have an impact on student achievement. For example, educators may not initially understand the incentives they are eligible to receive, know how to effectively change their teaching practices based on feedback provided through the district evaluation system, or be willing to change their behavior until they experience performance bonus payouts. Districts may need time to (1) design or revise performance measures so they can provide useful and accurate information to educators, (2) effectively explain to educators how they are being evaluated and how bonuses are determined, and (3) understand how to provide professional development that can help educators improve on the performance measures. It also may take time for the policy to cause changes in the overall quality of the educator workforce through the retention and recruitment of high quality teachers and principals. Because these learning and feedback processes may take multiple school years, it could take several years for impacts on student outcomes to be realized.

Research Questions

The purpose of this multiyear study was to describe the program characteristics and implementation experiences of 2010 TIF grantees and measure the impact of pay-for-performance bonuses within a well-implemented, performance-based compensation system. Because educators' understanding of and response to this policy can change over time, the study followed the grantees for all four years of TIF implementation.

The study addressed four research questions:

1. What were the characteristics of all TIF districts and their performance-based compensation systems? What implementation experiences and challenges did TIF districts encounter?
2. How did teachers and principals in schools that did and did not offer pay-for-performance bonuses compare on key dimensions, including their understanding of TIF program features, exposure to TIF activities, allocation of time, and attitudes toward teaching and the TIF program?
3. How did pay-for-performance bonuses affect educator effectiveness and the retention and recruitment of high-performing educators?
4. What was the impact of pay-for-performance bonuses on students' achievement on state assessments in math and reading?

The first report from this study (Max et al. 2014) described implementation of TIF for all 2010 grantees and, for a subset of 10 evaluation districts, provided detailed findings on implementation and the effect of pay-for-performance bonuses on educators' reported satisfaction, attitudes, and behaviors. This report found that fewer than half of all 2010 TIF districts reported implementing all four required components of their TIF program. For the 10 evaluation districts, the report indicated that (1) many educators misunderstood the measures used to evaluate their performance, their eligibility for a pay-for-performance bonus, and the potential amount of the performance bonus they could earn; (2) most educators were satisfied with their professional opportunities, school environment, and the TIF program; and (3) educators in schools that offered pay-for-performance bonuses tended to be less satisfied than those in schools that did not offer performance bonuses.

The second and third reports (Chiang et al. [2015] and Wellington et al. [2016], respectively) focused on ongoing implementation of TIF and the effect of pay-for-performance bonuses in the 10 evaluation districts. The second report covered two years of program implementation, and the third report included a third year of implementation. These reports found that pay-for-performance had small, positive impacts on students' reading and math achievement. The reports also indicated that few evaluation districts structured pay-for-performance bonuses to align well with TIF grant guidance, and that educators' understanding of key program components improved somewhat over the period, but many teachers still misunderstood whether they were eligible for performance bonuses or the amount they could earn.

This fourth and final report also focuses on implementation of TIF and the effect of pay-for-performance in the 10 evaluation districts, but includes a fourth year of program implementation. It captures educators' views and attitudes that, by the end of the TIF program, were shaped by several years of pay-for-performance bonuses. The report also presents impacts of pay-for-performance on educator effectiveness and student achievement, and compares the cost of generating those

achievement impacts to that of other interventions. In addition, it presents findings about educators' perspectives on key TIF components and districts' plans for continuing these components in the year following the end of the grant. These analyses are based on information obtained from educator and district surveys, interviews with TIF district administrators, and student and educator administrative data provided by the evaluation districts. Although the report focuses on the 10 evaluation districts, it also includes information on implementation of TIF for all 2010 grantees.

Road Map for the Remainder of the Report

In the rest of this report, we describe in detail the study's design and findings. In Chapter II, we describe the study sample, design of the experimental evaluation, data used for this report, and analytic approaches. In Chapter III, we describe the programs of all 2010 TIF districts and challenges the districts encountered in implementing TIF. In Chapter IV, we provide more detailed information on implementation experiences in TIF evaluation districts, and, in Chapter V, we examine the impact of eligibility for pay-for-performance bonuses on teachers' and principals' attitudes and behaviors. In Chapter VI, we present findings on the impact of pay-for-performance on educator effectiveness and student achievement, and examine the cost-effectiveness of pay-for-performance bonuses in light of those impact findings. Finally, in Chapter VII, we describe educators' attitudes and districts' plans toward continuing TIF components.

THIS PAGE IS INTENTIONALLY BLANK

II. STUDY SAMPLE, DESIGN, DATA, AND METHODS

In this chapter, we describe the study sample, design, and data used for this report. We also present an overview of the study's analytic approaches.

Study Sample

This study is based on school districts and schools that were part of the Teacher Incentive Fund (TIF) grants awarded in 2010 by the U.S. Department of Education (ED). That year, ED awarded 62 TIF grants that included 183 districts. As explained in Chapter I, the 2010 grants were awarded under two separate competitions: (1) a main competition; and (2) an evaluation competition, for which grantees agreed to participate in a study that involved random assignment of schools to a treatment group or a control group. Most of this report focuses on the TIF districts that were part of the evaluation competition, which we refer to as evaluation districts.[5] We refer to the remaining TIF districts as non-evaluation districts.

Most, but not all, districts that received a grant participated in TIF through the 2014–2015 school year, the final year of the 2010 TIF grant program. A total of 171 districts implemented TIF—that is, they had a performance-based compensation system supported by TIF funds—in 2011–2012; 164 districts implemented TIF in 2012–2013; 158 districts implemented TIF in 2013–2014; and 148 districts implemented TIF in 2014–2015 (Table II.1).[6] Among the districts that implemented TIF in 2014–2015, 13 were evaluation districts.

Table II.1. Number of Districts that Implemented TIF and Responded to the District Survey, by Year

	2011–2012	2012–2013	2013–2014	2014–2015
Implemented TIF				
Non-Evaluation Districts	159	151	145	135
Evaluation Districts	12	13	13	13
Total	171	164	158	148
Responded to District Survey				
Non-Evaluation Districts	141	142	131	126
Evaluation Districts	12	13	13	13
Total	153	155	144	139

Sources: U.S. Department of Education and TIF grantee reports.

Notes: A district is regarded as implementing TIF if it had at least some components of a performance-based compensation system supported by TIF funds. The counts show the total number of districts that had a TIF program in place during the school year.

[5] For this study, one set of charter schools that were part of the same TIF evaluation grant, were in the same state, and belonged to a common charter school association was considered to be a single evaluation district.

[6] From 2011–2012 to 2012–2013, 8 non-evaluation districts withdrew from their grants, and 1 evaluation grantee added a district to its TIF grant. From 2012–2013 to 2013–2014, 6 non-evaluation districts withdrew from their grants. From 2013–2014 to 2014–2015, 10 non-evaluation districts withdrew from their grants.

Districts were awarded, or included in, a TIF grant through a competitive process, and the grants were designed to serve high-need schools. Therefore, TIF districts were not representative of all U.S. districts. An earlier report from this study (Max et al. 2014) showed that, compared with the average U.S. district, TIF districts were larger, were more likely to be urban and located in the South, and had a higher proportion of students who were racial or ethnic minorities and eligible for free or reduced-price lunch.

This report provides an overview of TIF implementation in all 2010 TIF districts and, within the evaluation districts, an in-depth analysis of implementation and the impacts of pay-for-performance on educator and student outcomes. Next, we describe the final sample of districts included in these analyses.

All TIF Districts in the Final Analysis Sample

In Chapter III of this report, we examine TIF implementation in all TIF districts (evaluation and non-evaluation) over the four years of TIF implementation—the 2011–2012 through the 2014–2015 school years. We describe the districts' reported implementation of the four required components of TIF and the challenges they encountered. As discussed later, this analysis relied on districts' responses to surveys we administered annually from 2012 to 2015. The final sample for this analysis consisted of evaluation and non-evaluation districts that participated in TIF and responded to the district survey (Table II.1).

Evaluation Districts in the Final Analysis Sample

The rest of this report focuses on the evaluation districts, from which we collected more detailed information than we collected from non-evaluation districts. This information—obtained from surveys, interviews, technical assistance documents, and administrative data—enabled us to describe the performance bonuses and performance ratings that educators actually earned, document districts' strategies for communicating key program features, analyze educators' understanding of and attitudes toward TIF, and estimate the impact of pay-for-performance on educator and student outcomes.

ED used the same criteria to award evaluation and non-evaluation TIF grants, but evaluation districts may differ from other TIF districts in important ways related to the evaluation requirements. The requirement to provide at least eight elementary or middle schools for the evaluation may have resulted in larger districts being part of the in-depth evaluation. In addition, the requirement for random assignment of pay-for-performance bonuses may have drawn in districts that were confident they could obtain educators' buy-in to randomly assign this required program component.

Evaluation and non-evaluation districts differed on several demographic and socioeconomic characteristics (Table II.2). Although we found few statistically significant differences, the relatively small sample size of 13 evaluation districts implied that only large differences would have been statistically significant. Therefore, we note differences that were larger than 10 percentage points or 10,000 students. On average, evaluation districts were larger than non-evaluation districts and had smaller percentages of students who were white (38 versus 50 percent). Evaluation districts were also more likely than non-evaluation districts to be in urban areas (69 versus 29 percent) and the West (46 versus 15 percent), and less likely to be in towns (8 versus 23 percent), rural areas (0 versus 30 percent), the Midwest (15 versus 29 percent), the South (23 versus 47 percent), and states with

right-to-work laws that tend to be associated with weaker unions (54 versus 67 percent). Evaluation and non-evaluation districts were equally likely to be in the Northwest and had similar percentages of schools eligible for Title I and students who were black, Hispanic, or received free or reduced-price lunch.

Table II.2. Comparison of TIF Evaluation Districts and Non-Evaluation Districts (Percentages Unless Otherwise Noted)

	Evaluation Districts	Non-Evaluation Districts
Student Racial and Ethnic Distribution		
White, non-Hispanic	38	50
Black, non-Hispanic	32	26
Hispanic	22	17
Student Socioeconomic Status		
Eligible for free or reduced-price lunch	63	67
Title 1-eligible schools (schoolwide)	72	80
Enrollment (average)		
Number of students	32,240	20,010
District Location		
Urban	69	29*
Suburban	23	19
Town	8	23
Rural	0	30*
Geographic Region		
Northeast	15	9
Midwest	15	29
South	23	47*
West	46	15*
Collective Bargaining		
In state with right-to-work laws[a]	54	67
Number of States	**8**	**24**
Number of Districts—Range[b]	**13**	**125-135**

Source: Common Core of Data for 2013–2014 school year.

Notes: The table is based on all 144 districts that implemented TIF in 2014–2015. Seven non-evaluation districts were not included in the 2013–2014 district-level data from the Common Core of Data. Common Core of Data school-level data are used to calculate socioeconomic indicators. Common Core of Data district-level data are used to calculate all other demographic characteristics.

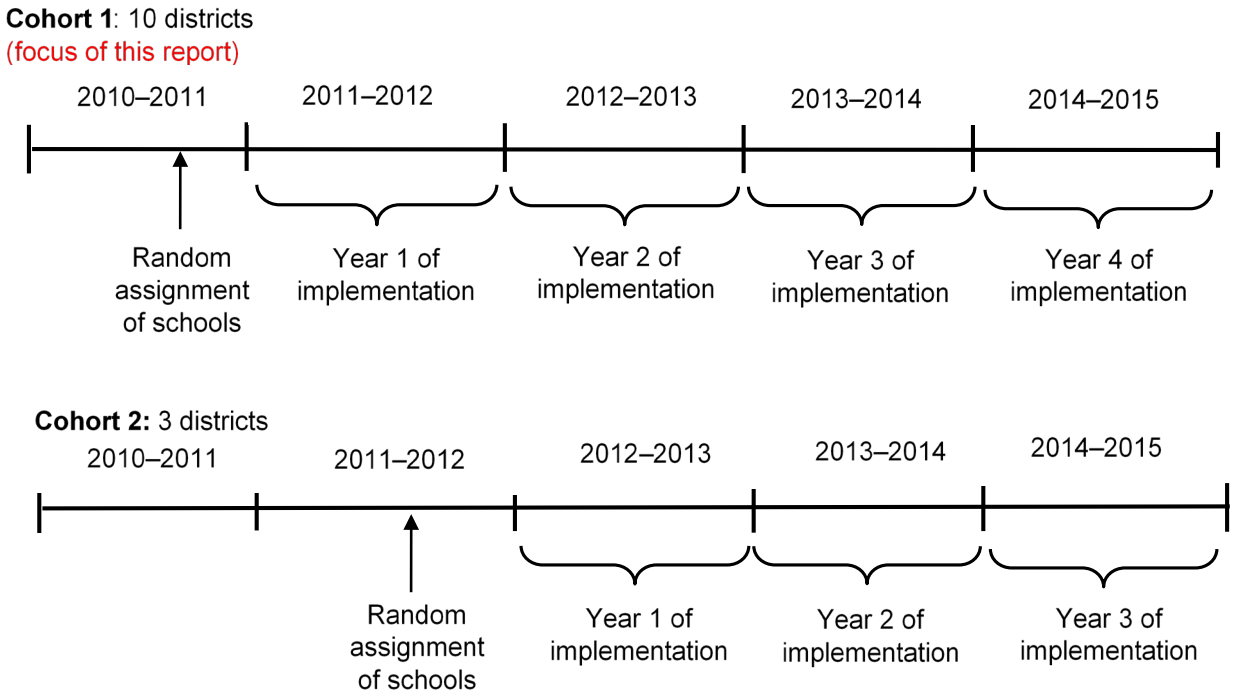
[a]The presence of right-to-work laws is a state-level indicator from the National Right-to-Work Legal Defense Foundation. Right-to-work laws prohibit unions from requiring nonmembers to pay fees, and such laws tend to be associated with weaker unions.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference between evaluation and non-evaluation districts is statistically significant at the .05 level, two-tailed test.

We classified evaluation districts into two cohorts—Cohorts 1 and 2—according to the year in which we randomly assigned their schools to a treatment or control group (Figure II.1). Cohort 1 consists of 10 districts in which we randomly assigned schools in spring and summer 2011. From these districts, we obtained data on four years of TIF implementation: 2011–2012 (Year 1), 2012–2013 (Year 2), 2013–2014 (Year 3), and 2014–2015 (Year 4). Cohort 2 consists of three districts in which we randomly assigned schools in spring and summer 2012 and obtained data on three years of TIF implementation, 2012–2013, 2013–2014, and 2014–2015, representing Years 1, 2, and 3 of this cohort's implementation of TIF.[7]

Figure II.1. Two Cohorts of Evaluation TIF Districts



The structure of the grants varied among the 10 Cohort 1 districts. Four of these districts received TIF grants directly from ED. The remaining 6 Cohort 1 districts were part of multidistrict grants administered by another grantee organization—such as a state education agency, university, association of charter schools, or nonprofit organization. In total, the 10 Cohort 1 districts represented eight distinct grantees.

This report primarily focuses on the 10 Cohort 1 evaluation districts—those for which data were available on four years of TIF implementation. As explained in Chapter I, because TIF is a comprehensive program for reforming educators' compensation and improving their effectiveness, it may take time for educators to fully understand the incentives available, the measures on which they are evaluated, and the improvements they have to make to earn bonuses. Focusing on Cohort 1 districts enabled us to examine the ways in which educators' understanding of

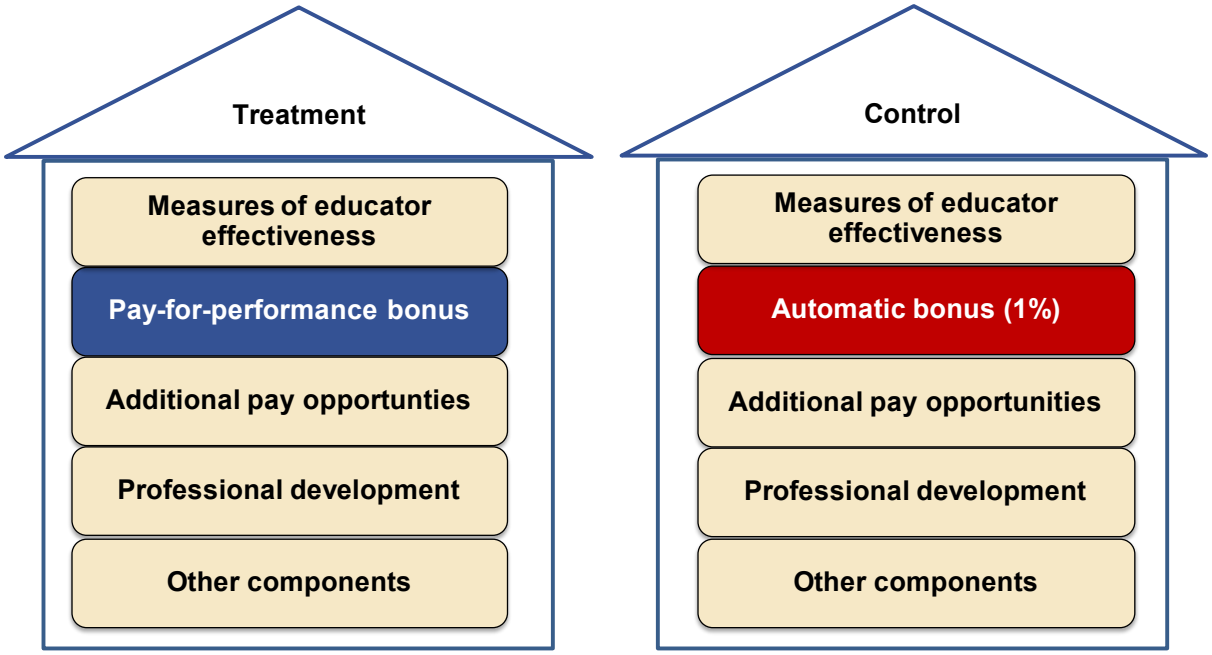
[7] Two Cohort 2 districts began putting some components of their TIF programs into place in 2011–2012, and Table II.1 includes these two districts in the counts of districts that implemented TIF in 2011–2012. However, because these districts were not ready for random assignment of schools until spring and summer 2012, we classified them as Cohort 2 districts and, for this report, specified 2014–2015 as Year 3 of the districts' implementation of TIF.

the program and the impacts of pay-for-performance bonuses evolved over a full four-year period while ensuring that the analysis included the same schools throughout this period. Unless otherwise noted, all findings in Chapters IV through VII are based on these 10 Cohort 1 districts.[8]

Experimental Design to Estimate the Impact of Pay-for-Performance

To ensure that the study's findings on the impacts of pay-for-performance could be attributed solely to the offer of pay-for-performance and not to other characteristics of districts, schools, or educators, we randomly assigned elementary and middle schools within each district to treatment and control groups. In Figure II.2, we illustrate the experimental design and highlight that treatment and control schools were expected to implement the same features of the district's performance-based compensation system, except for the pay-for-performance component. Educators (teachers and principals) at treatment schools were eligible to earn a pay-for-performance bonus; educators at control schools received an automatic bonus worth about 1 percent of their salary each year. The 1 percent bonus ensured that all educators in evaluation schools received some benefit from participating in the study: either the opportunity to earn a pay-for-performance bonus or the automatic bonus. Therefore, the impact of pay-for-performance estimated in this study potentially reflects two key differences between treatment and control schools: (1) bonuses in treatment schools were differentiated based on performance; and (2) bonuses in treatment schools were larger, on average, than in control schools.

Figure II.2. Random Assignment Design



[8] For key implementation features and outcomes, the appendices of this report provide findings from three years of implementation for Cohorts 1 and 2 together.

Evaluation districts chose the schools the evaluation would include. As with all schools that received TIF funds, schools in the evaluation needed to have at least half of their students receiving free or reduced-price lunch (Chapter I). In addition, because a primary objective of the study was to measure the impact of pay-for-performance on student achievement on state assessments, schools in the evaluation needed to include at least one grade level tested by state assessments (3rd to 8th grade).

Before random assignment, schools were paired based on having similar characteristics measured before the district's implementation of TIF—primarily student achievement, grade span, and school size. District staff either approved the pairs we constructed or directly specified the pairs based on their knowledge of the participating schools. One school from each pair was randomly assigned to the treatment group and the other school in the pair was assigned to the control group. We describe random assignment procedures in more detail in Appendix A.

We randomly assigned 183 elementary and middle schools to either the treatment or control group—138 schools assigned as part of Cohort 1 and 45 additional schools as part of Cohort 2 (Table II.3). Of the 138 Cohort 1 schools, our primary analysis sample consisted of 131 schools that implemented the TIF program for four years. This sample excluded seven schools (5 percent of all Cohort 1 schools) that (1) closed, (2) withdrew from their TIF grant and therefore declined to provide data to the evaluation, or (3) were paired with schools that closed or withdrew from their TIF grant.[9] Appendix A, Table A.1 describes this school attrition in more detail.[10]

Table II.3. Number of Schools in the Evaluation, by Cohort and Treatment Status

Cohort (# districts)	Timing of Random Assignment	Number of Treatment Schools	Number of Control Schools	Total Number of Schools
Cohort 1 (10 districts)	Spring/summer 2011	69	69	138
Cohort 2 (3 districts)[a]	Spring/summer 2012	23	22	45
Number of Schools		**92**	**91**	**183**
Final Analysis Sample (schools in Cohort 1 that implemented TIF for 4 years)		**65**	**66**	**131**

Source: Study authors' calculations.

[a]Counts of schools that were randomly assigned in spring or summer 2012 include a small number of schools (fewer than 3) from Cohort 1 districts to replace schools that closed.

[9] We randomly assigned schools using either matched pairs of schools or matched groups of schools (Appendix A). For the matched pairs of schools, we dropped the entire pair if one of the schools closed or withdrew from their TIF grant (and therefore declined to provide data to the evaluation). For the matched groups of schools, we dropped only the schools that closed or withdrew from their TIF grant.

[10] Of the 45 schools in Cohort 2, 39 implemented TIF for three years and were paired with schools that did so. Therefore, supplemental analyses that include Cohorts 1 and 2 together are based on 170 schools—131 schools from Cohort 1 and 39 schools from Cohort 2.

Baseline Characteristics of Treatment and Control Schools

The key advantage of this study's random assignment design is that, at the beginning of the study, the treatment and control groups were expected to include students and educators with similar characteristics. Because the two groups were expected to differ only in the opportunity for educators to receive pay-for-performance bonuses, differences in outcomes between the groups could be attributed to the impact of pay-for-performance.

At the beginning of the study, we found that treatment and control schools in the final analysis sample were similar on most of the measured characteristics of their students and educators. In the year before implementation—the year of random assignment before the first year of TIF implementation—the overall difference in student characteristics between treatment and control schools was not statistically significant ($p = 0.10$; Table II.4). On a few specific student characteristics, treatment and control schools differed slightly. Students in treatment schools had slightly lower achievement in math (by 0.04 standard deviations) than students in control schools. In addition, compared to control schools, a smaller percentage of students in treatment schools were white and a larger percentage were black, with differences of no more than 3 percentage points. Treatment and control schools had similar student achievement in reading before the implementation of TIF and similar proportions of students who received free or reduced-price lunch, had an Individualized Education Program, were over age for their grade, or were English language learners. As discussed later in this chapter, all analyses of the impacts of pay-for-performance on educator and student outcomes were adjusted to account for the slight preexisting differences in student achievement and racial or ethnic composition between treatment and control schools. Treatment and control schools had similar educator characteristics in Year 1, the first year of educator data available for all districts (Table II.5).[11,12]

The study schools' baseline characteristics confirm that the schools were both high-need and low-performing. As Table II.4 shows, in both the treatment and control schools, at least three-fourths of the students received free or reduced-price lunch, and the students' math and reading achievement was lower than the average achievement in their states by at least four-tenths of a standard deviation.

[11] Appendix A, Tables A.2 and A.3 show the characteristics of all study schools in Cohorts 1 and 2 at the beginning of the study. We found that treatment and control schools in this sample were similar on most of the measured characteristics of their students and educators.

[12] Appendix A, Table A.4 shows educators' characteristics within treatment and control schools in the pre-implementation year for 9 of 10 districts that provided educators' data for that year. In these districts, treatment and control schools were similar on most of the characteristics of their educators, with a few exceptions: teachers in treatment schools were 3 percentage points more likely than those in control schools to be white and 3 percentage points less likely to be black.

Table II.4. Characteristics of Students Enrolled in Treatment and Control Schools in the Pre-Implementation School Year (2010–2011) (Percentages Unless Otherwise Indicated)

	Treatment	Control	Difference
Achievement in the Pre-Implementation Year (average *z*-score)			
Math	-0.47	-0.43	-0.04*
Reading	-0.41	-0.40	-0.01
Race or Ethnicity			
White, non-Hispanic	27	30	-3*
Black, non-Hispanic	44	42	2*
Hispanic	23	22	1
Other	6	6	-1
Other Characteristics			
Female	49	49	-1
Eligible for free or reduced-price lunch	77	76	0
Disabled or has an Individualized Education Program	12	12	0
Over age for grade	13	13	0
English language learner	8	8	0
Grade Span			
Grades 3–5	64	64	0
Grades 6–8	36	36	0
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value			0.10
Number of Students—Range[a]	**12,299-21,816**	**12,540-22,037**	
Number of Schools—Range[a]	**41-65**	**42-66**	

Source: Student administrative data.

Note: The difference between the treatment and control estimates may not equal the difference shown in the table because of rounding.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

Table II.5. Characteristics of Educators in Treatment and Control Schools in Year 1 (Percentages Unless Otherwise Noted)

	Teachers			Principals		
	Treatment	Control	Difference	Treatment	Control	Difference
Demographic Characteristics						
Female	86	85	1	61	57	4
Race or ethnicity						
White, non-Hispanic	74	73	1	60	56	3
Black, non-Hispanic	20	21	-1	32	36	-4
Hispanic or Other	7	6	0	8	8	1
Age (average years)	42	41	0	49	48	1
Education						
Master's degree or higher	51	50	1	>90[a]	>90[a]	1
Experience in K–12 Education						
Total experience (average years)	12	11	0	16	15	2
Fewer than 5 years	24	25	-1	18	14	4
5–15 years	46	46	-1	34	40	-6
More than 15 years	30	28	2	48	46	2
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value			0.47			0.80
Number of Educators—Range[b]	**1,433-2,109**	**1,499-2,136**		**39-64**	**45-68**	
Number of Schools—Range[b]	**48-65**	**49-66**		**37-62**	**43-64**	

Source: Educator administrative data.

Notes: None of the differences are statistically significant at the .05 level, two-tailed test. The difference between the treatment and control estimates may not equal the difference shown in the table because of rounding.

[a]To reduce the risk of disclosing information on an individual, the exact percentage is not shown.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

Data Sources

The analyses in this report are based on data from eight sources. Table II.6 summarizes the data sources, along with response rates. Next, we describe each of these data sources in more detail.

Data for All 2010 TIF Districts

Common Core of Data. This publicly available database provided information on the characteristics of all TIF districts—including students' race and ethnicity, free or reduced-price lunch eligibility, average district enrollment, and geographic information—and student enrollment at each study school. We used data from the 2013–2014 school year to compare the characteristics of evaluation and non-evaluation districts. In addition, we used school-level enrollment data from the 2011–2012 through 2013–2014 school years when calculating the costs per student of pay-for-performance bonuses in the study's cost-effectiveness analysis.

Table II.6. Data Sources for This Report

Data Source	Type of Information	Response Rates (Percentages) 2011–2012	2012–2013	2013–2014	2014–2015
Data Collected from Evaluation and Non-Evaluation Districts					
1. Common Core of Data	Composition of student characteristics in districts	NA	NA	NA	NA
2. District survey	TIF program features, implementation experiences	89	95	91	94
Data Collected from Evaluation Districts Only					
3. District interviews	Detailed information on TIF implementation and program features	100	100	100	100
4. Principal survey	TIF program features, attitudes toward TIF program and job, hiring practices	98	95	92	95
5. Teacher survey	TIF program features, attitudes toward TIF program and job, time use	92	92	90	89
6. Technical assistance documents	Detailed information on implementation and program features	100	100	100	100
7. Student administrative data	Students' standardized test scores and background characteristics (grades 3 through 8)	100	100	100	100
8. Educator administrative data	Teachers' and principals' school assignments, background characteristics, performance ratings, and compensation from TIF	100	100	100	100

Notes: Response rates for the educator surveys are shown for treatment and control groups combined in Cohort 1 districts. None of the response rates differed between the treatment and control groups by more than 6 percentage points. See Appendix A for survey sample sizes by year.

NA is not applicable.

District survey. The district survey asked TIF districts to provide information on the components of their TIF programs, program communication strategies, and general experiences and challenges in implementation. The 2015 survey also asked districts about their plans for continuing TIF after the grant ends. We addressed these surveys to the person identified as overseeing or directing each district's TIF program. Districts' responses enabled us to describe programs in all TIF districts and to assess their implementation of the four required components of the TIF grant.

We administered the survey in 2012 (in the middle of the 2011–2012 school year), and near the end of the 2012–2013, 2013–2014, and 2014–2015 school years to all districts participating in TIF in those years. In 2015, 94 percent of TIF districts responded to the district survey (Appendix A, Table A.5). On average, districts that responded had smaller percentages of students who were eligible for free or reduced price lunch (65 versus 78 percent) and smaller percentages of schools that were eligible for Title I (79 versus 93 percent) than districts that did not respond. Districts that responded were also larger, on average, and more likely to be located in towns (23 versus 0 percent) than districts that did not respond (Appendix A, Table A.6).

Data for TIF Evaluation Districts Only

District interviews. Interviews with TIF program administrators in evaluation districts provided more in-depth information than that collected from the survey. Through these interviews, we probed for more details on how bonuses were determined, how the program was communicated to educators, the timing of bonus awards, types of challenges encountered in implementation, and plans for continuing TIF components. Information from the interviews enabled us to comprehensively describe implementation in evaluation districts and, when appropriate, to fill in missing information or supplement survey responses. This report used data from the first, second, third, and fourth years of interviews, which we conducted in the fall following each round of the district survey.

Principal and teacher surveys. We administered surveys to principals and teachers in the evaluation districts to learn about their understanding of and experiences with TIF program components, job satisfaction, attitudes toward TIF, and job-specific practices (such as principals' approaches to hiring teachers and teachers' allocation of time). We used educators' survey responses for three main purposes: (1) to describe educators' understanding of their TIF program; (2) to compare the experiences, attitudes, and classroom and school practices of educators in treatment and control schools; and (3) to examine how educators' understanding and attitudes may have changed over time.

In spring 2012, 2013, 2014, and 2015, we administered surveys to all principals and a sample of teachers within treatment and control schools that were participating in TIF in those years. Among full-time teachers, the teacher sample included all 4th-grade teachers; all 7th-grade math, English/language arts, and science teachers; and 77 percent of 1st-grade teachers in 2012 and 100 percent of 1st-grade teachers in 2013, 2014, and 2015. These groups represent elementary and middle school grades and subjects both with and without annual accountability testing.[13]

Response rates for principals were at least 92 percent in each year, and response rates for teachers were about 90 percent in each year (Table II.6). The response rates of treatment and control educators were generally similar (Appendix A, Table A.7).[14] For both principals and teachers, the largest treatment-control difference in response rates, which occurred in Year 3, was no more than 6 percentage points. We found few differences between the characteristics of respondents and nonrespondents to the teacher survey (Appendix A, Table A.10).[15] Among both teachers and principals, we found few differences between the characteristics of respondents from treatment and control schools (Appendix A, Tables A.11 and A.12).

[13] In 2013 and 2014, we also surveyed teachers from the prior-year sample who switched teaching assignments in the same school, moved to a different school in the same district, or left their district or the teaching profession. In 2015, we surveyed teachers from the prior-year sample who switched teaching assignments in the same school or moved to a different school in the same district. The final analysis sample did not include these teachers. In Appendix A, we explain in detail how we determined the teacher sample.

[14] Appendix A, Table A.8 provides response rates for Cohort 2, and Table A.9 shows the distribution of grade and subject assignments for the Cohort 1 teachers who responded to the survey and were included in the final analysis sample.

[15] We do not report comparisons of respondents and nonrespondents to the principal survey due to the small number of nonrespondents.

Technical assistance documents. The technical assistance team documented aspects of the evaluation districts' programs and implementation activities and experiences. The team conducted needs assessments in fall 2010 and spring 2011 for each evaluation district or grantee. The assessments examined evaluation districts' program design and planned implementation, progress in implementing the five core elements required by ED, and use of communication materials during the planning year to inform educators about the program.

The evaluation team reviewed the documents for all evaluation districts. When appropriate, the team used this information to report more detail on the evaluation districts' TIF programs and implementation experiences.

Student administrative data. We collected evaluation districts' administrative records on students enrolled in treatment and control schools. The data included information on students' background characteristics and their scores on state assessments in math and reading, enabling us to examine the impact of pay-for-performance on student achievement. Within Cohort 1 districts—those that completed four years of TIF implementation—the data covered all students in study schools in 2010–2011 to 2014–2015, representing the period from the year before implementation to Year 4 of implementation. We obtained similar data from Cohort 2 districts for 2011–2012 to 2014–2015.

Educator administrative data. We collected evaluation districts' administrative records on teachers and principals, including information on their assignments to schools, background characteristics, performance ratings determined by their TIF programs, and compensation received from TIF. These data enabled us to describe thoroughly the performance ratings, bonuses, and additional pay that educators received from TIF and to examine the impact of pay-for-performance on educators' effectiveness. Within Cohort 1 districts, all of these data covered Years 1 to 4 of implementation, and data on school assignments and background characteristics also covered the pre-implementation year. Similar data in Cohort 2 districts were available through the end of Year 3.[16]

Overview of Analytic Approach

In this section, we discuss the analytic approaches used in the rest of this report. Appendix B provides more technical details on the analytic methods.

Implementation of TIF in All Districts (Chapter III)

To describe implementation in all 2010 TIF districts, presented in Chapter III, we drew primarily from district survey responses. For each measure of program implementation included on the district survey, our basic analytic approach was to calculate means or percentages, as appropriate. We gave each district equal weight so that findings reflected the experiences of the average district that implemented a TIF program.

[16] Four Cohort 1 schools in Year 1, three in Year 2, four in Year 3, and four in Year 4 did not have full-time principals (Appendix A, Table A.13). The analysis of impacts on principals' outcomes measured from administrative data did not include these schools.

Implementation of TIF in Evaluation Districts (Chapter IV)

In Chapter IV, we describe the implementation of TIF in the 10 Cohort 1 districts that completed four years of program implementation. In addition to the district survey, we used information collected only from the evaluation districts: district interviews, technical assistance documents, administrative data on educators' performance ratings and compensation from TIF, and teacher and principal surveys.

To describe districts' program designs and implementation experiences, we used districts' responses to surveys and interviews to calculate means (or percentages, as appropriate), weighting each district equally. To describe actual bonus amounts and performance ratings, we used administrative data to calculate summary statistics (means, maximum levels, or percentages of educators receiving particular bonus amounts or ratings) separately for each district and then took an equally weighted average across all districts.

To describe educators' understanding of and experiences with TIF program components, we summarized educators' survey data separately by treatment status and year, giving each school equal weight. We compared the responses of treatment and control educators to determine whether they differed in their perceived eligibility for the component—pay-for-performance bonuses—that was supposed to differ between the two groups and whether they reported similar exposure to other components that were not supposed to differ. To ensure that any reported differences between the two groups were due solely to their differing eligibility for pay-for-performance rather than preexisting differences in the characteristics of their schools, we used a regression to adjust educators' reports for slight differences in baseline school characteristics in the same manner as in our impact analyses, described below.

Educators' understanding of program components could change as they gained more exposure to those components. We examined how educators' understanding changed with each additional year of TIF implementation. Separately for treatment and control schools, we compared average reports in Years 1, 2, 3, and 4 and conducted hypothesis tests to determine whether differences between each year and the prior year were statistically significant.

Impacts of Pay-for-Performance on Educator and Student Outcomes (Chapters V and VI)

We estimated the impacts of pay-for-performance on several outcomes within the Cohort 1 evaluation districts. In Chapter V, we present impacts on educators' attitudes (such as job satisfaction) and self-reported behaviors (such as principals' hiring practices). In the theory of change in Chapter I, these attitudes and behaviors are intermediate factors that shape the key outcomes of interest: educator effectiveness and student achievement. In Chapter VI, we report the impacts of pay-for-performance on those key outcomes.

Because the study used random assignment, any differences in educators' or students' outcomes between the treatment and control group can be attributed to pay-for-performance and not some other characteristic of the districts or schools. We estimated these differences using a linear regression that accounted for the random assignment design—in particular, the assignment of schools rather than individuals to the treatment and control groups, as well as the pairing of schools before random assignment. As shown earlier in this chapter, treatment and control schools differed slightly in average student achievement and students' racial or ethnic composition before TIF implementation. Therefore, all regressions in the impact analyses controlled for these baseline school

characteristics.[17] We estimated regressions separately by year and used weights for educators' or students' data to give each school equal weight, so that the estimates reflected the impact of pay-for-performance on the average study school.

Next, we discuss how we measured each type of outcome and determined the individuals whose outcomes were included in the impact analyses.

Educators' attitudes and behaviors. We measured educators' attitudes and self-reported behaviors directly from the survey responses of principals and teachers working in the study schools at the time of the survey administration. Analyses of teacher-reported outcomes were based on teachers who reported teaching 1st grade; 4th grade; or 7th-grade math, English/language arts, or science.

Educator effectiveness. We examined the impact of pay-for-performance on several measures districts used to evaluate educator effectiveness: (1) ratings based on the achievement growth of all students in a school (school achievement growth), which were used to evaluate both teachers and principals; (2) teachers' classroom observation ratings; (3) ratings based on the achievement growth of students in teachers' own classrooms (classroom achievement growth); and (4) observation ratings for principals. Using all full-time principals and teachers in the study schools, these analyses assessed whether the average educator performance ratings in these schools were any higher or lower as a result of pay-for-performance.[18] In the theory of change from Chapter I, pay-for-performance could lead to higher average ratings by either enabling schools to retain and recruit more effective educators or motivating educators to improve their performance.

Student achievement. We measured student achievement using students' scores on state assessments in math and reading.[19] Because student achievement was measured on different scales in different states and grades, we standardized all scores into z-scores by subtracting the statewide grade-specific mean and dividing by the statewide grade-specific standard deviation within each year. The analysis used all students in grades 3 through 8 who were tested in a study school in a given year. The tested students included those who had been enrolled in the same school at the time of random assignment and stayed in that school, as well as students who moved into a study school

[17] In this report, we present the average outcomes for the treatment group as regression-adjusted means. That is, we present the raw (unadjusted) average outcomes for the control group, and we compute the regression-adjusted treatment group mean as the sum of the control group mean and the estimated impact.

[18] Appendix B explains how educator performance ratings were standardized. Appendix A, Tables A.14 and A.15 show the percentages of educators who received performance ratings; Tables A.16 through A.18 show the characteristics of educators who did and did not receive performance ratings; and Tables A.19 through A.21 compare the characteristics of educators in treatment and control schools who received performance ratings. We found few differences between the characteristics of teachers with and without observation ratings, but teachers who received classroom achievement growth ratings in Year 4 were more likely to be female, younger, less educated, and less experienced than those who did not. Principals who received observation ratings in Year 4 were more likely to be black, younger, and less experienced than those who did not.

[19] One Cohort 1 district administered its own spring student assessment in all four years of TIF implementation to measure educator effectiveness because its state administered only fall assessments in most of those years. Two Cohort 1 districts and one Cohort 2 district used their state's assessment to measure educator effectiveness in all but the final year. These three districts switched to a commercial assessment that their state adopted in the final year to measure educator effectiveness. In both cases, our student achievement impact analysis used the same student assessments that the districts used to measure educator effectiveness.

after random assignment.[20] Therefore, this analysis measured the impact of pay-for-performance on the average student achievement of the study schools after one, two, three, and four years of TIF implementation, potentially reflecting changes in individual students' achievement and changes in the schools' student composition resulting from pay-for-performance.[21]

Because some students entered or left the study schools during the course of the study, implementing pay-for-performance at a school for a certain number of years did not mean that all students at the school were exposed to pay-for-performance for that duration. In fact, at the end of Year 4, our estimate is that 37 percent of students had been at their schools since the beginning of the study and were exposed to all four years of the program (Appendix B, Table B.6). The remaining students had been at their schools for three years (14 percent), two years (21 percent), or one year (28 percent). Although we could estimate the percentages of students who had been exposed to pay-for-performance for specified durations, our data were not sufficiently detailed to enable us to identify exactly which students were exposed for those durations (see Appendix B for details). Therefore, we could not measure the impacts of exposing students to the full duration of the pay-for-performance program. Instead, the impacts that we could conclusively measure, reported in Chapter VI, were the impacts of implementing pay-for-performance for specified numbers of years on the average student achievement of the study schools, based on the students at the schools who took each year's spring assessment. For simplicity, we refer to these impacts as impacts on student achievement.

Most evaluation districts (7 of the 10 Cohort 1 districts) administered new student assessments in Year 4. Within those districts, the new student assessments aligned to their states' recently developed college- and career-ready standards, whereas the assessments in previous years did not. Despite the adoption of new assessments, all student test scores—including those in Year 4—could be standardized into z-scores because, as described earlier, we applied the standardization approach separately in each year, state, and grade. Implicit in this approach is that we could not—and did not—compare students' test score levels across different years. All impact analyses entailed comparing test scores between treatment and control schools within the same year.

Factors Associated with Differences in Impacts (Chapter VI)

The impacts of pay-for-performance on student achievement could differ across districts, and even across treatment schools within districts. Such differences in impacts have the potential to shed light on whether particular factors, such as program characteristics or changes in teachers' behaviors, influenced the direction and magnitude of the achievement impacts. In Chapter VI (and Appendix G), we explore whether such factors were associated with student achievement impacts.

[20] There were no differences between treatment and control schools in percentages of students in grades 3 through 8 who had math and reading scores in Years 1, 2, 3, and 4 (Appendix A, Table A.22). Compared with students without scores, those with scores tended to have higher baseline achievement, were more likely to be Hispanic, female, and less likely to have an Individualized Education Program or be over age for their grade (Appendix A, Tables A.23 and A.24).

[21] In Years 1, 2, 3, and 4, students in the analysis sample from treatment and control schools had similar characteristics, suggesting that pay-for-performance did not induce changes in the schools' student composition (Appendix A, Tables A.25 and A.26). In Years 1 and 4, students from the analysis sample in treatment schools had lower baseline math achievement than students from the analysis sample in control schools, but this pattern simply mirrored the treatment-control difference in math achievement that we observed among students enrolled in the year before implementation (Table II.4).

Differences in student achievement impacts across districts provided an opportunity to examine whether the characteristics of districts' TIF programs and their implementation, as well as the context in which the implementation took place, were associated with the impacts of pay-for-performance. For example, some districts had pay-for-performance bonuses that were more differentiated for higher and lower performers than others, or had bonuses that were largely based on individual rather than group performance. To examine whether these and other characteristics were related to student achievement impacts, we compared impacts in subgroups of districts with and without a characteristic, or subgroups with high versus low levels of a characteristic. Differences in impacts between subgroups provided only suggestive evidence that a specified characteristic might have influenced impacts, given that the subgroups could also have differed on other characteristics.

Differences in student achievement impacts across schools presented an opportunity to explore whether pay-for-performance affected achievement by changing particular educator behaviors. For example, pay-for-performance bonuses might affect student achievement by increasing educators' effort on the job, encouraging educators to focus strategically on their performance ratings, or inducing teachers to change their classroom practices. If so, then treatment schools that experienced larger impacts on these educator behaviors should also tend to be those that experienced larger impacts on student achievement. To assess this possibility, we estimated the impacts of pay-for-performance on educator behaviors (measured by classroom observation ratings and responses on surveys) and student achievement separately for each treatment and control school pair. We then examined the association between impacts on educator behaviors and impacts on student achievement. Although these associations could suggest which behavioral changes might be responsible for impacts on achievement, they could also reflect the influence of other behaviors that the study did not measure.

Cost-Effectiveness (Chapter VI)

Given the impacts of pay-for-performance, policy decisions on whether to direct resources to pay-for-performance may depend, in part, on whether it is a cost-effective way of raising student achievement compared with other policies. To explore this question, we compared the cost-effectiveness of pay-for-performance with that of two other benchmark policies: (1) transfer incentives for high-performing teachers to work at low-performing schools and (2) class-size reduction. Both benchmark policies have been rigorously evaluated with large-scale random assignment studies that have found positive impacts in at least one grade and subject and have provided sufficiently detailed cost information for a cost-effectiveness analysis (Glazerman et al. 2013; Nye et al. 2000; Schanzenbach 2006; Harris 2009). Moreover, states or districts could conceivably implement either of those benchmark policies in place of pay-for-performance.

For each policy, we calculated the cost per student needed to achieve a given impact. Specifically, we measured each policy's cost-effectiveness ratio—the total per-student cost of the policy over a specified number of years divided by the total impact over the same number of years. We then compared the cost-effectiveness ratios of pay-for-performance and each benchmark policy and tested whether they differed statistically. Because a policy's cost-effectiveness could change with its duration, we compared pay-for-performance with a specific benchmark policy for each possible duration—one or two years, and in some cases three or four years—so long as both policies were implemented and evaluated for at least that length of time.

Educators' Attitudes and Districts' Plans Toward Continuing TIF Components (Chapter VII)

In Chapter VII, we describe the perspectives of multiple stakeholders—educators, evaluation districts, and all 2010 TIF districts—on the future of key TIF components after the end of the TIF grants. First, within the evaluation districts, we documented educators' opinions about how their districts should evaluate and compensate educators, because those opinions could influence their districts' decisions. To do so, we summarized educators' survey responses in Year 4 separately by treatment status, using the same approaches discussed earlier for analyzing educators' understanding (in Chapter IV) and attitudes and behaviors (in Chapter V). Second, we used district interview data to calculate the percentages of evaluation districts that planned to preserve, revise, or discontinue each of the four TIF components in the 2015–2016 school year and their reasons for doing so. Third, we used district survey data to calculate the percentages of all 2010 TIF districts that planned to continue key components and to describe the districts' funding sources for sustaining those components. Because pay-for-performance is of particular focus to this evaluation, we also compared the characteristics of districts that did and did not plan to continue pay-for-performance, using statistical tests to identify significant differences.

THIS PAGE IS INTENTIONALLY BLANK

III. PROGRAMS AND EXPERIENCES OF ALL 2010 TIF DISTRICTS

In this chapter, we broadly describe TIF program implementation. We first examine how many TIF districts implemented all four required components of the TIF grant (discussed in Chapter I). We then provide more detail on the implementation of each individual component to examine which components contributed to districts' ability (or inability) to implement all four required components. We conclude the chapter with details on challenges that districts reported in implementing TIF.

The findings presented in this chapter are from districts that were included in a 2010 TIF grant and responded to the district survey. Most districts undertook a planning year (2010–2011) to design the components and get stakeholder support, and then completed four years of program implementation. The districts completed surveys during each of those four years of implementation—specifically, in the middle of the 2011–2012 school year (Year 1) and the end of the 2012–2013 (Year 2), 2013–2014 (Year 3), and 2014–2015 (Year 4) school years. This chapter describes TIF implementation across all four years but draws examples primarily from Year 4, the final year of program implementation. Unless otherwise noted, the findings in Year 4 are similar to those in previous years (Max et al. 2014; Chiang et al. 2015; Wellington et al. 2016).

Key Findings on Programs and Experiences of All 2010 TIF Districts

- **Most districts implemented each individual component of a comprehensive, performance-based compensation system at the start of their programs.**
- **Districts were least likely to report providing teachers with professional development based on their actual performance and evaluating principals based on at least two observations.**
- **Few districts reported major challenges to program implementation, and the percentage of districts that reported such challenges decreased over the course of the grant.**
- **Only sustainability of the program was widely reported as a major challenge.**

TIF Required Components

The TIF grant required four components: (1) using student achievement growth and at least two formal observations to measure educator effectiveness, (2) offering a pay-for-performance bonus, (3) offering additional pay opportunities, and (4) providing professional development to support educators' understanding and use of the measures of effectiveness. Taken together, these components constitute a comprehensive, performance-based compensation system.

Implementation of TIF Required Components

Most districts implemented each individual component of TIF at the start of their programs, but were least likely to report offering the required professional development and evaluating principals using both student achievement growth and at least two observations. In each of the four years of implementation, nearly all the districts (over 90 percent) reported offering teachers and principals bonuses based on their performance (Table III.1). Likewise, many districts offered educators opportunities to earn additional pay for taking on extra roles or responsibilities (over 85 percent) and used both student achievement growth and classroom

observations to measure teacher effectiveness (about 80 percent). Fewer districts, but still over half, used both student achievement growth and observations of school practices to measure principal effectiveness (about 60 to 70 percent) and offered the required professional development to their teachers (59 to 74 percent).

Table III.1. TIF Districts' Reported Implementation of TIF Required Components for Teachers and Principals (Percentages)

	Year 1 (2011–2012)	Year 2 (2012–2013)	Year 3 (2013–2014)	Year 4 (2014–2015)
Teachers				
Requirement 1: Measures of Educator Effectiveness[a]	79	80	81	80
Requirement 2: Pay-for-Performance Bonus	94	98	100	99
Requirement 3: Additional Pay Opportunities[b]	86	91	88	88
Requirement 4: Professional Development	66	74	70	59
Implemented Requirements 1, 2, and 3	68	71	72	71
Implemented at Least Three of Four Requirements	85	90	88	82
Implemented All Requirements	46	52	50	45
Principals				
Requirement 1: Measures of Educator Effectiveness[a]	68	65	69	62
Requirement 2: Pay-for-Performance Bonus	94	99	97	98
Requirement 3: Additional Pay Opportunities[b]	86	91	88	88
Implemented Requirements 1, 2, and 3[c]	58	60	60	54
Number of Districts—Range[d]	**137-153**	**142-155**	**134-144**	**127-139**

Sources: District surveys (2012, 2013, 2014, and 2015).

[a]TIF districts were required to use student achievement growth and at least two observations by trained observers to evaluate teachers and principals.

[b]The TIF grant notice required that districts provide additional pay opportunities for educators, so these percentages are based on the percentage of TIF districts that reported offering these pay opportunities to either teachers or principals.

[c]The district survey did not include questions on professional development for principals.

[d]Sample sizes are presented as a range based on the data available for each row in the table.

Most TIF districts implemented at least three of the four required components, yet only about half implemented all of them for teachers. The percentage of districts that reported implementing all four required components for teachers was similar across the four years, never exceeding 52 percent (Table III.1). Nevertheless, in every year, most districts (more than 80 percent) reported implementing at least three of the four required components for teachers. Likewise, more than half of the districts implemented all required components for principals aside from professional development.[22]

[22] Professional development for principals is a requirement of TIF grants. However, given concerns about the length of the district survey, it did not include questions on whether districts implemented the required professional development for principals. The TIF grant notice also required pay for additional opportunities for educators. Most

Next, we provide an overview of districts' implementation of each individual required component. Findings presented here are primarily based on Year 4 and, unless otherwise noted, are similar to those in previous years (Max et al. 2014; Chiang et al. 2015; Wellington et al. 2016).

Requirement 1: Measures of Educator Effectiveness

TIF grantees were required to measure educator effectiveness based on student achievement growth and multiple observations by trained observers. These measures provide the basis for teachers and principals earning performance-based bonuses.

Most TIF districts reported meeting the requirement to use student achievement growth and at least two observations to measure teacher and principal effectiveness. In Year 4, 80 percent of TIF districts reported using student achievement growth and classroom observations to measure teacher effectiveness, and 62 percent reported meeting the requirement to measure principal effectiveness (Table III.1).

When implementing the required effectiveness measures, districts could choose how to evaluate teachers based on student achievement growth. For example, districts could evaluate teachers based on the achievement growth of the teachers' own students (classroom achievement growth); all students in the same grade, team, or subject area (achievement growth of student subgroups); all students in the school (school achievement growth); or some combination of these measures. Classroom achievement growth measures could give teachers more control over their own evaluation ratings, and achievement growth measures for larger groups could encourage collaboration among teachers.

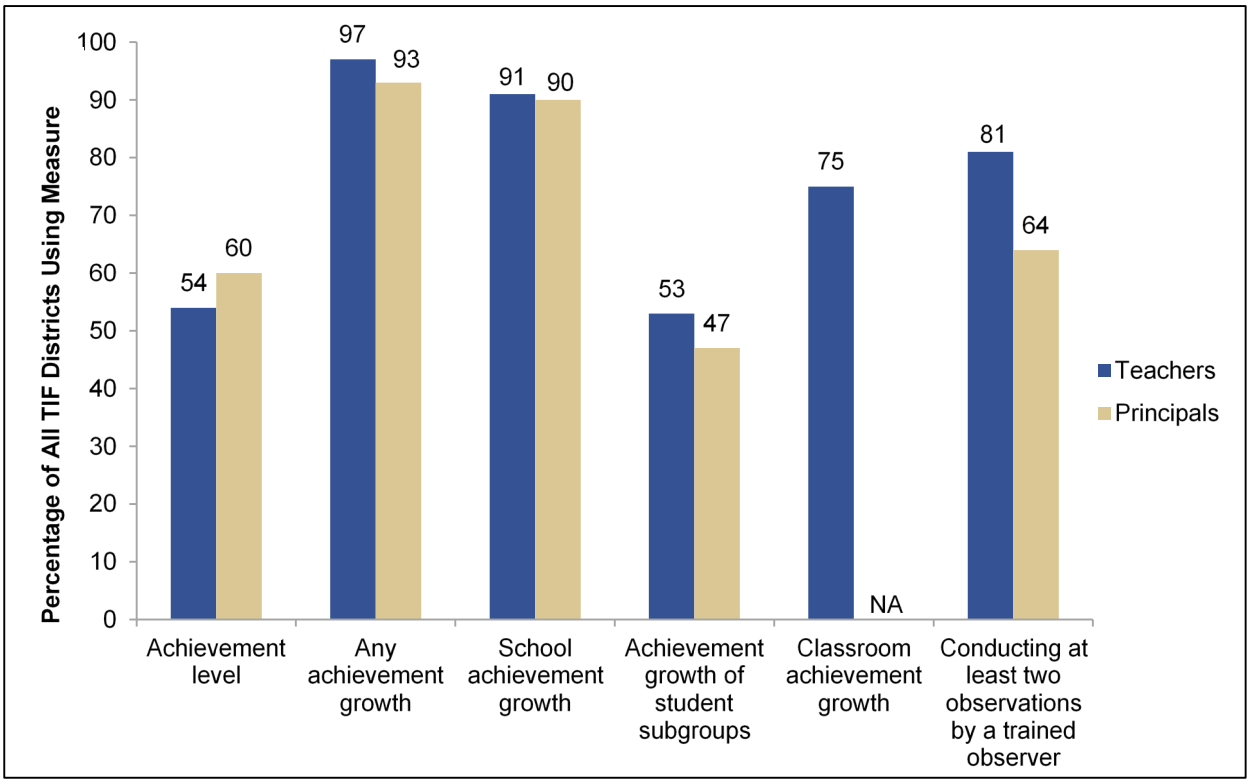
Nearly all TIF districts reported using school achievement growth to evaluate teachers. Most frequently, TIF districts reported evaluating teachers based on school achievement growth (91 percent in Year 4), followed by classroom achievement growth (75 percent) and achievement growth of student subgroups (53 percent; Figure III.1).

Most TIF districts reported using at least two formal observations to evaluate teachers. For example, 81 percent of districts in Year 4 reported using at least two formal observations by trained observers to evaluate teachers (Figure III.1). Districts planned to conduct, on average, four formal observations per teacher—more than the two required under the grant—lasting about 45 minutes each (Appendix C, Table C.1). Districts most frequently reported that principals conducted observations (97 percent).

Most TIF districts reported using student achievement growth and observations by trained observers to evaluate principals, but were more likely to use achievement growth than observations. Most frequently, districts reported using school achievement growth to evaluate principals (90 percent in Year 4; Figure III.1). Fewer districts (64 percent) reported conducting observations by trained observers. These districts planned to conduct, on average, four observations per principal, lasting about 45 minutes each (Appendix C, Table C.1). Districts most frequently reported that observations of principals were conducted by a central office administrator from the same district or the superintendent (53 and 55 percent in Year 4, respectively).

grantees met this requirement by offering additional pay opportunities to teachers. Therefore, if a district reported offering additional pay opportunities to either teachers or principals, it met this requirement.

Figure III.1. Measures of Student Achievement and Observations Used to Evaluate Teachers and Principals, All TIF Districts, Year 4 (Percentages)



Source: District survey, 2015.

Notes: Between 129 and 138 districts responded to the survey questions for teachers, and between 133 and 135 districts responded to the survey questions for principals. Teacher evaluation measures are those for teachers in tested grades and subjects.

Figure reads: In Year 4, 54 percent of all TIF districts reported using achievement levels to evaluate teachers, and 60 percent reported using achievement levels to evaluate principals.

NA is not applicable.

Requirement 2: Pay-for-Performance Bonuses

TIF districts were required to offer pay-for-performance bonuses to teachers and principals based purely on their performance, but districts could determine which types of teachers would be eligible for such bonuses and whether other school staff would also be eligible. The eligibility determination could affect educators' attitudes toward and responses to their TIF programs. For example, broadening eligibility for bonuses to all staff at a school might increase the staff's buy-in to the program and, if bonuses depend on school performance measures, encourage collaboration among staff. Alternatively, limiting eligibility to teachers of certain grades or subjects might enable districts to concentrate resources on improving classroom practices in high-priority academic areas.

Most TIF districts sought to make performance bonuses broadly available to a variety of school staff. Nearly all TIF districts reported that teachers and principals were eligible for pay-for-performance bonuses. For example, in Year 4, almost all (99 percent) of TIF districts reported that teachers were eligible for performance bonuses, and 98 percent reported that principals were eligible (Table III.2). Teachers' eligibility for performance bonuses almost never depended upon

teaching a grade or subject with annual, end-of-year state assessments. In fact, 99 percent of districts reported that teachers in grades or subjects without annual assessments (referred to as nontested) were eligible for performance bonuses. Moreover, districts tended not to restrict eligibility to teachers and principals. More than three-fourths (80 percent) of districts reported that assistant or vice principals were eligible for performance bonuses. Almost half of districts (48 percent) reported making nonteaching staff, such as counselors, librarians, or custodians, eligible for such bonuses.

Table III.2. Staff Eligibility for Pay-for-Performance Bonus, Year 4 (Percentages)

	All TIF Districts
Teachers	
Teachers in tested grades and subjects	99
Teachers in nontested grades and subjects	99
Principals	98
Other School Staff	
Assistant or vice principal	80
Other school administrators	37
Other teaching staff (for example, part-time teachers, substitutes, aides)	28
Nonteaching staff (for example, counselors, librarians, custodians)	48
Number of Districts—Range[a]	**118-139**

Source: District survey, 2015.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

Requirement 3: Additional Pay Opportunities

Nearly all TIF districts offered additional pay for teachers to take on roles and responsibilities, most often to support mentor and master or lead teacher opportunities. Of TIF districts reporting, 86 percent offered teachers additional pay for roles and responsibilities in Year 4 (Table III.3). Most often, districts offered additional pay for mentor (58 percent) and master or lead teachers (55 percent). About one-fifth of districts (18 percent) reported offering principals extra pay for assuming additional roles or responsibilities.

The TIF grant notice also encouraged, but did not require, districts to offer additional pay for educators to teach in high-need subject areas or to work in hard-to-staff schools. A minority of districts in Year 4 (27 percent) offered teachers additional pay for doing so (Appendix C, Table C.2). Of the districts reporting, 13 percent offered principals extra pay for working in a hard-to-staff school.

Requirement 4: Professional Development

The TIF grant notice required that districts provide professional development linked to the measures of educator effectiveness. This support included professional development to help educators understand the measures used to evaluate their performance, as well as to provide feedback based on their actual performance ratings to help improve their instructional practices.

Most TIF districts provided the required professional development to teachers, but were more likely to provide teachers with general information about how they were evaluated than to provide them with feedback based on their actual performance. Although most TIF districts (78 percent in Year 4) offered professional development to help teachers understand the performance measures used in the program, fewer districts (68 percent) offered the

more targeted professional development based on teachers' actual performance on those measures (Table III.4).[23]

Table III.3. Additional Pay Opportunities for Teachers and Principals, Year 4

	Percentage of TIF Districts that Offered Additional Pay	Average Maximum Pay in Districts Offering Additional Pay
Teachers		
Teachers Could Receive Additional Pay for Taking on Extra Roles or Responsibilities	86	NA
Roles and Responsibilities		
Mentor teacher	58	$3,453
Master or lead teacher	55	$6,836
Department chair or head	22	$1,830
Lead curriculum specialist	12	$3,263
Schoolwide committee or task force member	24	$895
Leadership team member	34	$1,630
Number of Districts—Range[a]	**136-137**	**16-70**
Principals		
Principals Could Receive Additional Pay for Taking on Extra Roles or Responsibilities in School or District	18	$4,897
Number of Districts	**136**	**23**

Source: District survey, 2015.

Note: Table reports on activities funded by TIF.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

NA is not applicable.

Table III.4. Planned Professional Development Activities for Teachers (Percentages)

	Year 1 (2011–2012)	Year 2 (2012–2013)	Year 3 (2013–2014)	Year 4 (2014–2015)
Focus of Professional Development				
Understanding performance measures of TIF program	87	94	87	78
Feedback based on TIF performance ratings	71	76	76	68
Number of Districts—Range[a]	**146-151**	**152**	**142**	**138-139**

Sources: District surveys (2012, 2013, 2014, and 2015).

[a]Sample sizes are presented as a range based on the data available for each row in the table.

[23] Surveys of district administrators did not ask about professional development for principals.

Challenges in Implementing and Sustaining TIF

The 2013, 2014, and 2015 district surveys included questions about challenges districts faced in implementing TIF. Our goal was to focus on topics that might shed light on the components that could make it difficult for districts to implement programs such as TIF in the future, and to examine whether districts find implementation less challenging over time. The survey asked district staff whether particular aspects of implementation were a major challenge, minor challenge, or not a challenge. For example, the survey asked about potential challenges related to (1) incorporating student achievement growth into teacher evaluations, (2) observing teachers' or principals' practices, (3) calculating pay-for-performance bonuses, (4) communicating the program to educators or other stakeholders, and (5) obtaining or maintaining support for the program. This section focuses on the activities that districts most often reported as a major challenge.[24]

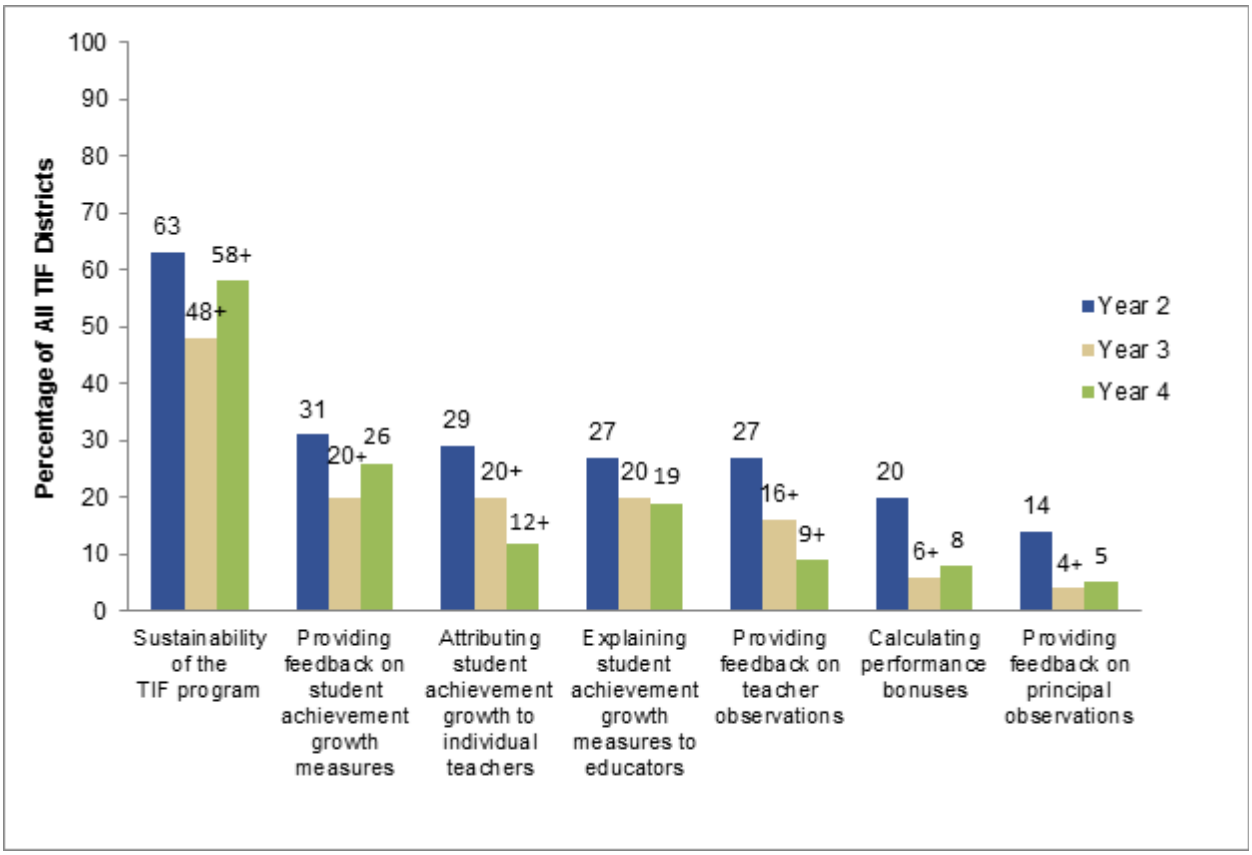
Few TIF districts reported that key activities related to implementing their programs were a major challenge. No aspect of TIF implementation was a major challenge to more than one-third of TIF districts in any year of the grant (Appendix C, Table C.3). For example, in Year 4, 26 percent of the districts reported that providing feedback on student achievement growth measures was a major challenge (Figure III.2). Fewer than 20 percent of districts in Year 4 reported that explaining such measures or providing feedback based on observations was a major challenge.

Over the course of the grant, fewer districts reported major challenges to implementation. With each additional year of program implementation, districts gained more experience evaluating educators, calculating bonuses, and providing feedback to educators on their performance. This could have reduced reports of challenges to implementing the program over time. Compared to Year 2, fewer districts in Year 3 reported major challenges to several activities related to implementation, such as providing feedback on student achievement growth measures and teacher observations (Figure III.2). Likewise, compared to Year 3, fewer districts in Year 4 reported major challenges with attributing student achievement growth to individual teachers (12 versus 20 percent) and providing useful and/or timely feedback from observations (9 versus 16 percent). In no case did significantly more districts report an item to be a major challenge in Year 4 than in Year 3 (Appendix C, Table C.3).

Only sustainability of the program was widely reported as a major challenge. Although few districts reported major challenges to implementation, many districts expressed concern about sustainability of the program. In each year, at least about half of the districts reported that sustainability of the TIF program was a major challenge (63 percent in Year 2, 48 percent in Year 3, and 58 percent in Year 4; Figure III.2). The increase in reported concern about sustainability between Years 3 and 4 might reflect that districts struggled to secure funding for their program after their grant ended. Districts that did not consider sustainability a major challenge may have either secured funding, expected to secure funding, or decided to discontinue their TIF programs. To address this question, the 2015 district survey asked about districts' plans for sustaining the TIF program after the grant period ended. Chapter VII presents these findings based on survey responses from all 2010 TIF districts and includes more detailed information based on survey responses and interviews of TIF administrators from the 10 evaluation districts.

[24] Appendix C, Table C.3 shows a full list of activities included in the surveys and the percentages of districts that reported these activities to be a major challenge, minor challenge, and not a challenge.

Figure III.2. Major Challenges in Implementing TIF (Percentages)



Sources: District surveys (2013, 2014, and 2015).

Notes: The sample was restricted to the 128 TIF districts that responded to the 2013, 2014, and 2015 district surveys. Further details about survey results, including results for activities that districts reported as a minor challenge or not a challenge, are available in Appendix C, Table C.3.

Figure reads: In Year 2, 63 percent of all TIF districts reported that sustainability of their TIF program was a major challenge. In Year 3, 48 percent reported that sustainability of their TIF program was a major challenge. In Year 4, 58 percent reported that sustainability of their TIF program was a major challenge.

+Difference with prior year is statistically significant at the s05 level, two-tailed test.

Summary

As a comprehensive program for reforming compensation and improving the effectiveness of educators, TIF programs were designed to have multiple, interrelated components. Our analysis of the implementation of the TIF grants among all 2010 TIF districts sought to determine whether they could put into place such a comprehensive system, and whether they faced particular challenges doing so.

The 2010 TIF districts implemented most of the required components of a comprehensive performance-based compensation system. In fact, the vast majority of districts implemented at least three of the four required components for teachers. Nevertheless, many districts did not implement all the required components. Failure to provide professional development that gave teachers feedback on their individual performance ratings was the districts' most common reason for not achieving full implementation of TIF for teachers.

Consistent with the finding that TIF districts implemented most of the required components, few districts reported major challenges to implementing TIF by the second year of their programs. In subsequent years, even fewer districts reported major challenges. Therefore, although most districts could avoid or overcome major challenges to implementing a comprehensive compensation reform program within two years, some required more time to surmount those challenges.

THIS PAGE IS INTENTIONALLY BLANK

IV. TIF IMPLEMENTATION IN EVALUATION DISTRICTS

In this chapter, we describe the implementation of TIF by the evaluation districts—those that received a grant to participate in the evaluation of TIF, including random assignment of the pay-for-performance component of the program. According to the theory of change presented in Chapter I, a series of steps needed to occur in implementing TIF for pay-for-performance to be able to improve educator effectiveness and student achievement. The components of the program had to provide incentives and supports for educators to improve their effectiveness, information about those components had to be communicated to educators, and educators had to receive and understand this information. This chapter examines whether and how each of these steps materialized in the evaluation districts' implementation of TIF. First, we examine districts' implementation of the four required components of TIF. We focus on aspects of the programs that could shape teachers' motivation to improve, such as whether performance measures provided educators with consistent information on their effectiveness and how districts structured pay-for-performance bonuses. Second, we examine how districts communicated information about TIF to educators, including information on the performance bonuses that educators received. In the final part of this chapter, we examine teachers' and principals' understanding of the TIF program in their districts. Describing how evaluation districts implemented the TIF grants provides useful context for interpreting findings presented later in this report on the program's impact on student outcomes.

The chapter is based on 10 evaluation districts that completed four years of TIF implementation. We refer to each year of implementation—2011–2012, 2012–2013, 2013–2014, and 2014–2015—as Years 1, 2, 3, and 4, respectively.[25] In these years, educators in treatment schools were eligible for pay-for-performance bonuses and educators in control schools were not. The information in this chapter is drawn from details we obtained from these districts through district, teacher, and principal surveys; interviews with district TIF administrators; administrative data provided by the districts; and technical assistance documents. Because Year 4 was the final year of TIF implementation, this chapter summarizes findings for the entire grant period. In some instances, we provide examples from all four years to describe how the evaluation districts implemented their TIF programs. In other instances, for simplicity, we draw examples primarily from Year 4. Unless otherwise noted, the findings in Year 4 are similar to those in previous years (Max et al. 2014; Chiang et al. 2015; Wellington et al. 2016).

[25] As discussed in Chapter II, evaluation districts were classified into two cohorts—Cohort 1 and Cohort 2—according to the year in which we randomly assigned their schools to a treatment or control group. The 10 districts examined in this chapter, whose schools were randomly assigned in spring and summer 2011, were classified as Cohort 1. Three additional districts, whose schools were randomly assigned in spring and summer 2012, were classified as Cohort 2. In Appendix D, we present key implementation findings from Years 1, 2, and 3 for Cohorts 1 and 2 combined.

Key Findings on TIF Implementation in Evaluation Districts

- **Most evaluation districts reported implementing all required components for teachers at the start of their programs. The only component not implemented by all of the districts was the professional development required by the grant.**
- **All evaluation districts reported using observations and school achievement growth to evaluate teachers, and most also chose to evaluate teachers based on the achievement growth of the students they taught.**
- **Most teachers received similar performance ratings and bonus amounts from one year to the next, with many teachers receiving higher ratings on classroom observations than on student achievement growth.**
- **The highest-performing teachers earned pay-for-performance bonuses that were significantly larger than the average bonus. Yet, most teachers received a bonus, which, on average, was no more than 5 percent of the average teacher salary.**
- **Most teachers understood that they were evaluated based on student achievement growth and classroom observations.**
- **Many teachers in schools that offered pay-for-performance bonuses did not understand that they were eligible for a bonus or underestimated how much they could earn from performance bonuses.**
- **Most teachers reported receiving professional development on how they were evaluated and how to improve their performance, but indicated they received only a few hours of professional development on each topic over the school year.**

Implementation of the Required Components of TIF

Our examination of the implementation of TIF programs in evaluation districts focuses on the four required components of TIF programs: (1) measures of educator effectiveness, (2) pay-for-performance bonuses, (3) additional pay opportunities, and (4) professional development. Together, these four required components constitute a comprehensive performance-based compensation system, and the grant required districts to implement all the individual components together. In this section, we report on TIF evaluation districts' success in implementing all components together and on their implementation of each component separately.

Implementation of All Required Components

Most evaluation districts reported implementing all required components for teachers at the start of their programs, and all districts reported meeting at least three of the four required components. The only component not implemented by all of the districts was the professional development required by the grant. In each year, 60 to 80 percent of evaluation districts implemented all four required components for teachers. Starting from the first year, all 10 evaluation districts reported using measures of effectiveness that included student achievement growth and at least two observations of classroom practices, offering bonuses based on how teachers performed on the effectiveness measures, and offering additional pay to take on extra roles or responsibilities (Table IV.1). However, only 6 to 8 of the 10 evaluation districts reported

providing the required professional development in each year (similar to all 2010 TIF districts), with no clear pattern of increasing or decreasing implementation over the course of the grant.

Table IV.1. Evaluation Districts' Reported Implementation of TIF Required Components for Teachers and Principals (Percentages)

	Year 1 (2011–2012)	Year 2 (2012–2013)	Year 3 (2013–2014)	Year 4 (2014–2015)
Teachers				
Requirement 1: Measures of Educator Effectiveness[a]	100	100	100	100
Requirement 2: Pay-for-Performance Bonus	100	100	100	100
Requirement 3: Additional Pay Opportunities[b]	100	100	100	100
Requirement 4: Professional Development	70	70	60	80
Implemented Requirements 1, 2, and 3	100	100	100	100
Implemented All Requirements	70	70	60	80
Principals				
Requirement 1: Measures of Educator Effectiveness[a]	70	100	100	100
Requirement 2: Pay-for-Performance Bonus	100	100	100	100
Requirement 3: Additional Pay Opportunities[b]	100	100	100	100
Implemented Requirements 1, 2, and, 3[c]	70	100	100	100
Number of Districts	**10**	**10**	**10**	**10**

Sources: District surveys (2012, 2013, 2014, and 2015) and district interviews (2012, 2013, 2014, and 2015).

[a]TIF districts were required to use student achievement growth and at least two observations by trained observers to evaluate teachers and principals.

[b]The TIF grant notice required that districts provide additional pay opportunities for educators, so these percentages are based on the percentages of TIF districts that reported offering these pay opportunities to either teachers or principals.

[c]The district survey did not include questions on professional development for principals.

After the first year of implementation, all evaluation districts also reported meeting three of the four required components for principals. In Year 1, 70 percent of districts reported evaluating principals using student achievement growth and at least two observations by trained observers, but all districts reported meeting this requirement in the following years (Table IV.1). All districts offered pay-for-performance bonuses to principals in all four years. Districts could meet the third requirement—additional pay opportunities—by providing opportunities to either teachers or principals; as discussed previously, all districts fulfilled this requirement in all four years. We were unable to assess whether districts implemented the fourth required component for principals—professional development—because we did not collect this information for principals.

Next, we describe implementation of each required component in more detail and compare the implementation over time.

Requirement 1: Measures of Educator Effectiveness

TIF grantees were required to measure educator effectiveness based on student achievement growth and multiple observations by trained observers. These measures provided the basis for

rewarding teachers and principals with performance bonuses. As discussed earlier, all evaluation districts reported evaluating teachers and principals using the criteria required by the grant.

However, districts could choose the achievement growth and observation measures they used. Therefore, in what follows, we first describe the performance measures that districts reported using to evaluate teachers and principals. We then use administrative data to document teachers' actual performance on those measures.[26]

One area of discretion involved how to evaluate teachers based on student achievement growth. For example, districts could evaluate teachers based on the achievement growth of the teachers' own students (classroom achievement growth); all students in the same grade, team, or subject area (achievement growth of student subgroups); all students in the school (school achievement growth); or some combination of these measures. Districts could measure student achievement growth using a value-added model or by calculating the change in students' achievement on a standardized test from one year to the next.

All evaluation districts reported using school achievement growth to evaluate teachers, and most also chose to evaluate teachers based on classroom achievement growth. More than half (60 to 70 percent) of the districts reported evaluating teachers based on classroom achievement growth (Table IV.2). Within the districts that used classroom achievement growth, about 40 to 60 percent of teachers received classroom achievement growth ratings (Appendix A, Table A.14), typically because they taught grades and subjects tested by state assessments. Fewer than half of the districts (30 to 40 percent) reported using achievement growth of student subgroups to evaluate teachers. To evaluate principals, all evaluation districts used school achievement growth, and 40 to 60 percent used achievement growth of student subgroups.

Among districts that used a particular type of achievement growth measure (such as school achievement growth), there were differences in how those measures were designed. For example, a review of technical assistance documents found that six evaluation districts used school achievement growth measures provided by the state and four districts used models developed by private vendors. In addition, among the districts that evaluated teachers based on classroom achievement growth, most of the districts (five of seven in Year 4) used a measure that was generated from a statistical model (for example, a value-added model) whereas the rest (two of seven in Year 4) rated teachers' performance based on the extent to which the teacher met goals for their students' achievement growth (often referred to as *student learning objectives* or SLOs).

Districts also had discretion in meeting the requirement to conduct observations of classroom or school practices. For example, districts could choose the rubrics they wanted to use to observe teachers and principals, the number of observations in a year (as long as there were at least two), and which staff to train as observers. In practice, three districts used the Teacher Advancement Program (TAP) teacher observation rubric, three used Danielson's Framework for Teaching rubric (or a modified version of it), and two districts used a modified version of Kim Marshall's observation rubric. The remaining two districts used an existing state or district teacher observation rubric. These observation rubrics captured a range of practices, such as the teacher's lesson preparation,

[26] These analyses focus on whether the districts reported evaluating educators using the measures required by the TIF notice. We did not explore whether districts used these measures because of their TIF grant, or whether they might have implemented these measures regardless of receiving a TIF grant.

classroom management, and instructional strategies. Some rubrics also captured practices such as the teacher's participation in professional responsibilities and creation of a respectful learning environment.

Table IV.2. Measures of Student Achievement and Observations of Practices Used to Evaluate Teachers and Principals, as Reported by Evaluation Districts (Percentages)

Performance Measure	Year 1 (2011–2012)	Year 2 (2012–2013)	Year 3 (2013–2014)	Year 4 (2014–2015)
Teachers				
Student Achievement				
Student achievement level	30	20	30	30
Any student achievement growth	100	100	100	100
School achievement growth	100	100	100	100
Achievement growth of student subgroups[a]	40	30	30	30
Classroom achievement growth	60	60	70	70
Observation Measure				
Conducting at least two observations by trained observer	100	100	100	100
Principals				
Student Achievement				
Student achievement level	60	50	30	30
Any student achievement growth	100	100	100	100
School achievement growth	100	100	100	100
Achievement growth of student subgroups[a]	60	50	60	40
Observation Measure				
Conducting at least two observations by trained observer	70	100	100	100
Number of Districts	**10**	**10**	**10**	**10**

Sources: District surveys (2012, 2013, 2014, and 2015).

Note: Teacher evaluation measures are those for teachers in tested grades and subjects.

[a]Examples of student subgroups include grouping students by grade, team, or subject area.

Other areas of discretion in conducting observations included specifying the frequency and length of observations and the types of staff who would conduct them. Districts typically chose to exceed the minimum requirement of two observations per year. On average in Year 4, for example, evaluation districts reported conducting four classroom observations per year, each lasting about 40 minutes (Appendix D, Table D.1). When deciding whether teachers would be observed by staff from their own school (such as principals and teacher leaders) or outside the school (such as district administrators), districts faced a number of considerations. For example, school staff might be more familiar with the context in which teachers worked and therefore provide better feedback, but they might also be less objective in their assessments because of their relationships with the teachers. In practice, many evaluation districts chose to use a mix of different types of observers. Nearly all districts (89 percent) reported that the principal or other administrators at the teacher's school conducted Year 4 classroom observations, although more than a third of the districts also reported having teacher leaders, peer observers, or district administrative staff conduct classroom observations.

Teachers might be more motivated to change their behavior based on their ratings if they believe the ratings are meaningful and accurate. If teachers receive notably different ratings from

different measures, or their ratings from the same measure fluctuate greatly from one year to the next, they could question whether the ratings accurately and consistently measure their performance. To examine whether educators might have received similar feedback on their effectiveness from their ratings on different measures, we examined the percentage of educators who received different combinations of ratings on observations and student achievement growth. However, different performance measures might be designed to evaluate different aspects of performance, so they do not necessarily have to produce identical ratings to be considered valid. Therefore, we also examined teachers' ratings from the same performance measure across years. Although ideally teachers' performance would improve over time, large fluctuations in yearly ratings could suggest that the ratings do not accurately or consistently measure educators' effectiveness.

Figure IV.1 depicts the percentages of teachers who received each possible combination of ratings based on classroom observations and school achievement growth in Year 4. The blue circles (those on the diagonal) show the percentages of teachers who received similar ratings on the two measures.[27] For example, 13 percent of teachers received a rating of "somewhat effective" on both classroom observations and school achievement growth. The orange circles (those above the diagonal) represent teachers who received a higher rating on classroom observations than on school achievement growth. For example, 4 percent of teachers received a rating of "highly effective" on classroom observations and a rating of "somewhat effective" on school achievement growth.

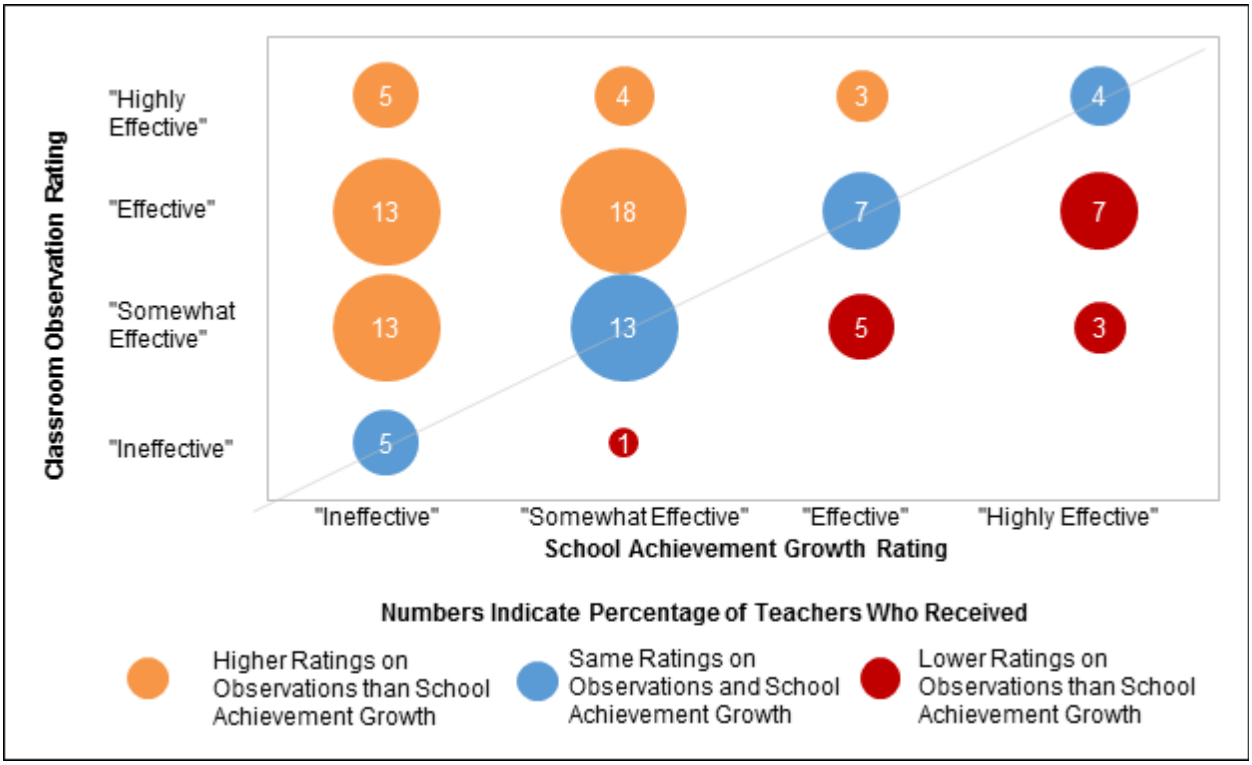
Many teachers and principals received higher ratings on observations than on school achievement growth. For example, in Year 4, fewer than one-third (29 percent) of all teachers received similar observation and school achievement growth ratings (represented by the blue circles in Figure IV.1). More than half (56 percent) received a higher rating on classroom observations than on school achievement growth (represented by the orange circles above the diagonal in Figure IV.1). A difference of one rating level between a teacher's ratings on the two measures—for example, earning a 4 versus 3 on a 1-to-4 rating scale—might be expected because these measures could be measuring different aspects of teacher effectiveness. For example, some of the practices measured by the observation rubrics, such as participation in professional responsibilities or creating a respectful learning environment, might not translate into large improvements in student achievement. But a difference of two rating levels could send a mixed message to teachers about their effectiveness. One-fifth (22 percent) of teachers received a classroom observation rating that was at least two levels above their school achievement growth rating, whereas only 3 percent received a school achievement growth rating at least two levels above their observation rating. These patterns were similar among principals. More than half of principals (57 percent) received a higher rating based on observations of their practices than on school achievement growth, and 43 percent received observation ratings that were at least two levels higher than their school achievement growth ratings (Appendix D, Table D.2).

Many teachers received higher ratings on classroom observations than on classroom achievement growth. Although it might be expected that school achievement growth ratings would differ from individual observation ratings because school achievement growth is based on the collective work of school staff, we found similar patterns among teachers evaluated on classroom

[27] To express ratings from different districts on a common scale, we transformed ratings to a 1-to-4 scale (see Appendix B for more information). Here, we describe ratings in terms of equal "quarters" on the 1-to-4 scale, with ratings from 1.00 to 1.75 described as "ineffective," ratings from 1.75 to 2.50 described as "partially effective," ratings from 2.50 to 3.25 described as "effective," and ratings from 3.25 to 4.00 described as "highly effective."

achievement growth.[28] For example, in Year 4, 26 percent of these teachers received similar observation and classroom achievement growth ratings, and 49 percent received a higher rating on observations than on classroom achievement growth (Appendix D, Table D.3). Overall, educators' ratings based on observations of their practices suggested they were more effective than their ratings based on student achievement growth suggested, regardless of the level (school or classroom) at which student achievement growth is measured.[29]

Figure IV.1. Comparison of Teachers' Ratings on Classroom Observations and School Achievement Growth in Year 4 (Percentages)



Source: Educator administrative data (N = 3,343 teachers).

Notes: Categories are study-constructed labels to represent quarters of a 1-to-4 rating scale. "Ineffective" = bottom quarter (1.00 to 1.75); "Somewhat Effective" = second quarter (1.75 to 2.50); "Effective" = third quarter (2.50 to 3.25); and "Highly Effective" = top quarter (3.25 to 4.00). The figure is based on teachers with ratings on both classroom observations and school achievement growth in Year 4. One district did not provide school achievement growth ratings for control schools in Year 4, so control schools from this district were excluded from this analysis.

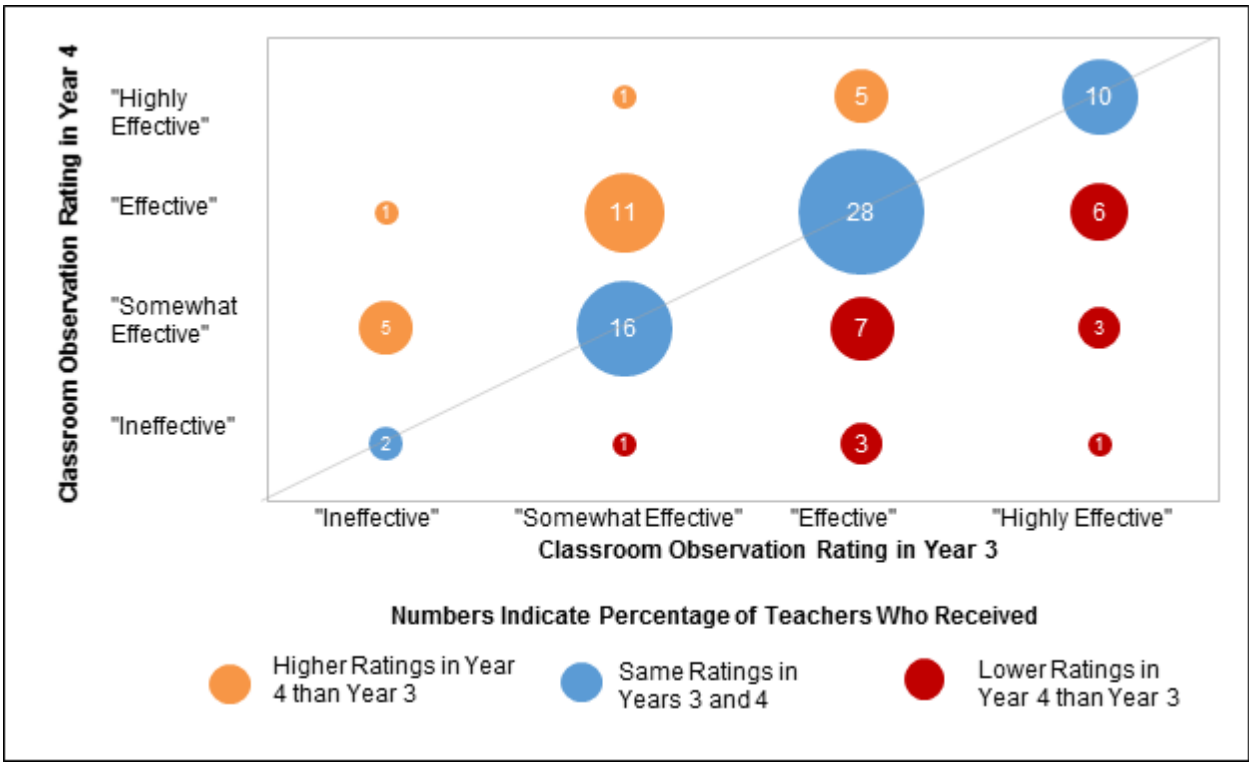
Figure reads: In Year 4, 5 percent of teachers received a school achievement growth rating of "ineffective" and a classroom observation rating of "highly effective".

[28] Within the seven districts that used classroom achievement growth in Year 4, about 60 percent of teachers (typically, those who taught grades and subjects in which annual state assessments were administered) received classroom achievement growth ratings (Appendix A, Table A.14).

[29] Our finding that classroom observation ratings and classroom achievement growth ratings frequently yielded different assessments of teachers' effectiveness is broadly consistent with prior studies showing that the two measures are not highly correlated (Kane and Staiger 2012; Lipscomb et al. 2015; Wayne et al. 2016).

Figure IV.2 illustrates the percentages of teachers who received each possible combination of ratings based on classroom observations for Years 3 and 4. Similar to Figure IV.1, the blue circles (those on the diagonal) show the percentages of teachers who received similar classroom observation ratings in Years 3 and 4. For example, 16 percent of teachers in both years received a rating of "somewhat effective" based on classroom observations. The red circles (those below the diagonal) represent teachers who received a lower rating based on classroom observations in Year 4 than Year 3. For example, 7 percent of teachers received a rating of "somewhat effective" in Year 4 and a rating of "effective" in Year 3.

Figure IV.2. Comparison of Teachers' Classroom Observation Ratings in Years 3 and 4 (Percentages)



Source: Educator administrative data (N = 2,649).

Notes: Categories are study-constructed labels to represent quarters of a 1-to-4 rating scale. "Ineffective" = bottom quarter (1.00 to 1.75); "Somewhat Effective" = second quarter (1.75 to 2.50); "Effective" = third quarter (2.50 to 3.25); and "Highly Effective" = top quarter (3.25 to 4.00). The figure is based on teachers with classroom observations ratings in both Years 3 and 4.

Figure reads: One percent of teachers received a classroom observation rating of "ineffective" in Year 3 and "effective" in Year 4.

On each performance measure, most teachers received similar ratings from one year to the next. For example, more than half of teachers received similar ratings, based on a 1-to-4 rating scale, in Year 4 as they did in Year 3. Specifically, 56 percent of teachers received a similar rating based on classroom observations, 51 percent received a similar rating based on school achievement growth, and 51 percent received a similar rating based on classroom achievement growth (Figure IV.2 and Appendix D, Tables D.4 and D.5). Teachers who received different ratings for the same measure in Years 3 and 4 were about equally likely to receive a higher or lower rating the following year. Also, when teachers earned different ratings across years, the ratings typically differed by just one level. On each measure, about one-third of teachers earned a rating in Year 4 that was one level

higher or lower than their rating in Year 3, whereas no more than 20 percent of teachers earned ratings that differed by two or more levels from one year to the next.

Requirement 2: Pay-for-Performance Bonuses

As discussed in Chapter I, grantees were required to offer bonuses to educators based on how they performed on the effectiveness measures. The goals of the bonuses were to incentivize educators and to reward them for being effective in their classrooms and schools. There were no additional requirements for earning the bonuses beyond performing well on the effectiveness measures.

The TIF grant notice provided guidance on ways to structure the bonuses, but districts had discretion in how they implemented that guidance. Therefore, the characteristics of these bonuses—for instance, the criteria for receiving them, their size, and the extent to which they differed across educators—were key factors that could determine their impact on educator and student outcomes. For example, when designing performance bonuses, districts faced the key decision of whether to offer separate bonuses for different performance measures or combine all of the performance measures into a single rating that determined educators' bonuses. Awarding separate bonuses for different performance measures could make it easier for educators to understand why they did or did not receive a bonus. However, it also had the potential to make earning a bonus less challenging because educators would have to perform well on only one measure to earn a bonus. Educators might even choose to focus on improving their performance only on the measure (or measures) that they believed they could change most easily.[30] Educators' responses to the bonuses could also depend on whether prior bonuses were large enough to catch their attention.

In what follows, we first use data from district surveys and interviews to describe how evaluation districts designed the bonuses, especially the factors that determined educators' bonus amounts. We then use administrative data on teachers and principals to describe the bonuses that educators actually received.

All evaluation districts offered teachers pay-for-performance bonuses, and all chose to offer separate bonuses for different performance measures. For example, in Year 4, all evaluation districts offered teachers bonuses based on school achievement growth, 70 percent of districts offered bonuses for classroom observations, 70 percent offered bonuses for classroom achievement growth, and 30 percent provided bonuses for achievement growth of student subgroups.[31] Most districts set an absolute maximum bonus that teachers could earn for each measure, but in some districts, the maximum bonus depended on the number of bonus recipients (Table IV.3). The features of districts' pay-for-performance bonus programs changed little across years.

[30] The 2012 TIF competition required grantees to assign educators one overall evaluation rating that combined information from observations and student achievement growth. (See https://www.federalregister.gov/articles/2012/06/14/2012-14269/applications-for-new-awards-teacher-incentive-fund#h-3.)

[31] In contrast, most (67 percent) of the Cohort 2 districts used a single, combined performance rating to determine bonuses (Appendix D, Table D.6). Appendix D, Table D.7 provides detailed information on teacher pay-for-performance programs for Cohorts 1 and 2.

As discussed in Chapter I, although districts had discretion to specify the structure of performance bonuses, the TIF grant notice encouraged districts to make performance bonuses:

- A *substantial* portion of teachers' compensation (giving an example of an average bonus worth 5 percent of the average educator salary)
- *Differentiated,* such that the highest performing teachers received significantly larger bonuses than the average (giving an example of a bonus in which at least some educators would receive a payout worth three times the average bonus)
- *Challenging* to earn (giving an example of a bonus that would be awarded only to educators performing significantly better than the average)

Table IV.3. Key Features of Evaluation Districts' Teacher Pay-for-Performance Bonus Programs in Year 4

	Districts									
Key Program Feature	1	2	3	4	5	6	7	8	9	10
Teachers could receive a bonus for multiple performance measures	X	X	X	X	X	X	X	X	X	X
Teachers could receive a bonus for school achievement growth	X	X	X	X	X	X	X	X	X	X
Teachers could receive a bonus for achievement growth of their students			X	X	X		X	X	X	X
Teachers could receive a bonus for the achievement growth of a student subgroup					X	X			X	
Teachers could receive a bonus for classroom observations	X	X	X	X		X	X			X
Student achievement growth was measured by a value-added model	X	X	X	X	X			X	X	X
A maximum bonus was specified for each performance measure		X			X	X	X	X	X	
Maximum bonus possible depended on the number of bonus recipients	X		X	X						X
Bonus amount for a performance measure could be affected by a factor aside from the teacher's rating on the measure[a]			X	X	X	X		X	X	X
District changed some aspect of its program from the 2013–2014 to the 2014–2015 school years			X						X	X

Sources: District interviews (2012, 2013, 2014, and 2015); grantees' Annual Performance Report (APR) documents; and technical assistance documents.

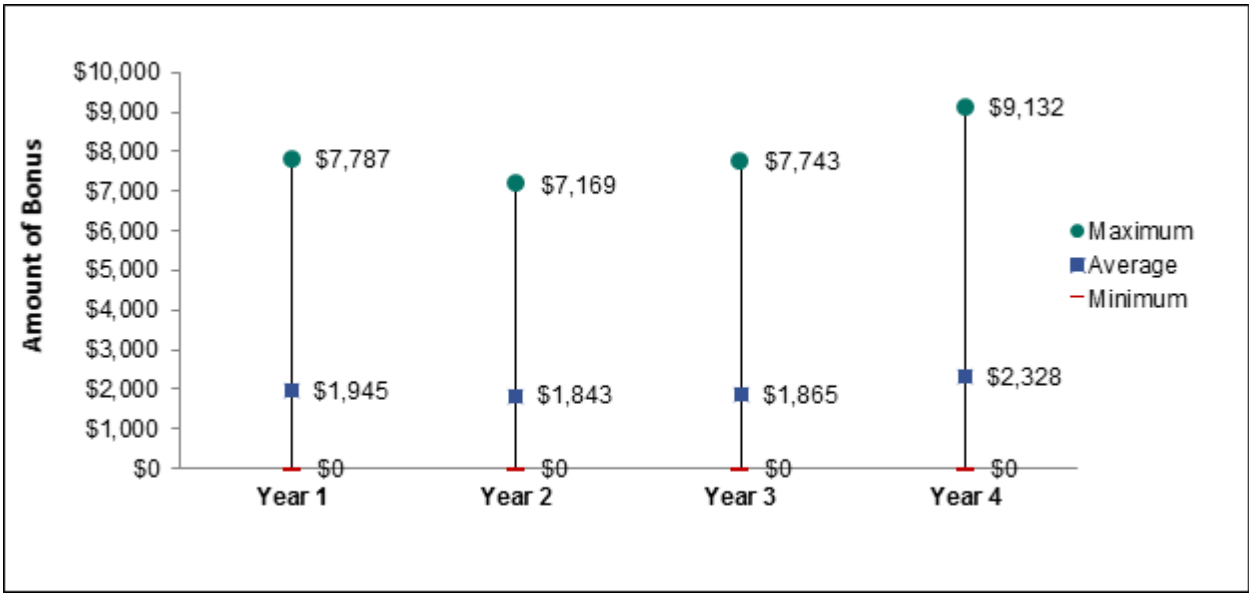
Notes: Grantees submit an APR to the U.S. Department of Education that describes how educators are evaluated. To ensure district confidentiality, the numbers assigned to districts in Table IV.3 do not correspond to the letters assigned to districts in other parts of the report.

[a]For example, a district could have required teachers to meet a minimum classroom observation rating to receive a bonus based on classroom achievement growth or taken teachers' attendance into account when determining the size of bonuses.

This guidance was intended to encourage districts to structure bonuses in a way that would motivate teachers to improve their effectiveness. To the extent that larger portions of teachers' compensation are linked to their performance, teachers might pay greater attention to the bonus program, increasing the likelihood that it triggers changes in their behavior. Bonuses with greater differentiation provide a stronger monetary incentive for teachers to achieve high performance ratings, because high ratings are linked to larger bonus amounts. Limiting the number of teachers who receive a bonus could increase teachers' motivation to improve if they believe that only high performers will receive a bonus.

On average, performance bonuses were no more than 5 percent of teachers' salaries, but the highest-performing teachers earned significantly larger bonuses. The average performance bonus across evaluation districts was about $2,000 in each year (Figure IV.3). This average bonus represented about 4 to 5 percent of the average teacher salary ($48,095 to $50,549 in Years 1 through 4).[32] Yet, the highest-performing teachers typically received bonuses that were notably larger than the average bonus. In each year, the maximum bonus was about $7,000 to $9,000 (for example, $9,132 in Year 4), or about four times the average bonus.

Figure IV.3. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools



Source: Educator administrative data (N = 2,151 in Year 1; N = 2,160 in Year 2; N = 2,236 in Year 3; and N = 2,083 in Year 4).

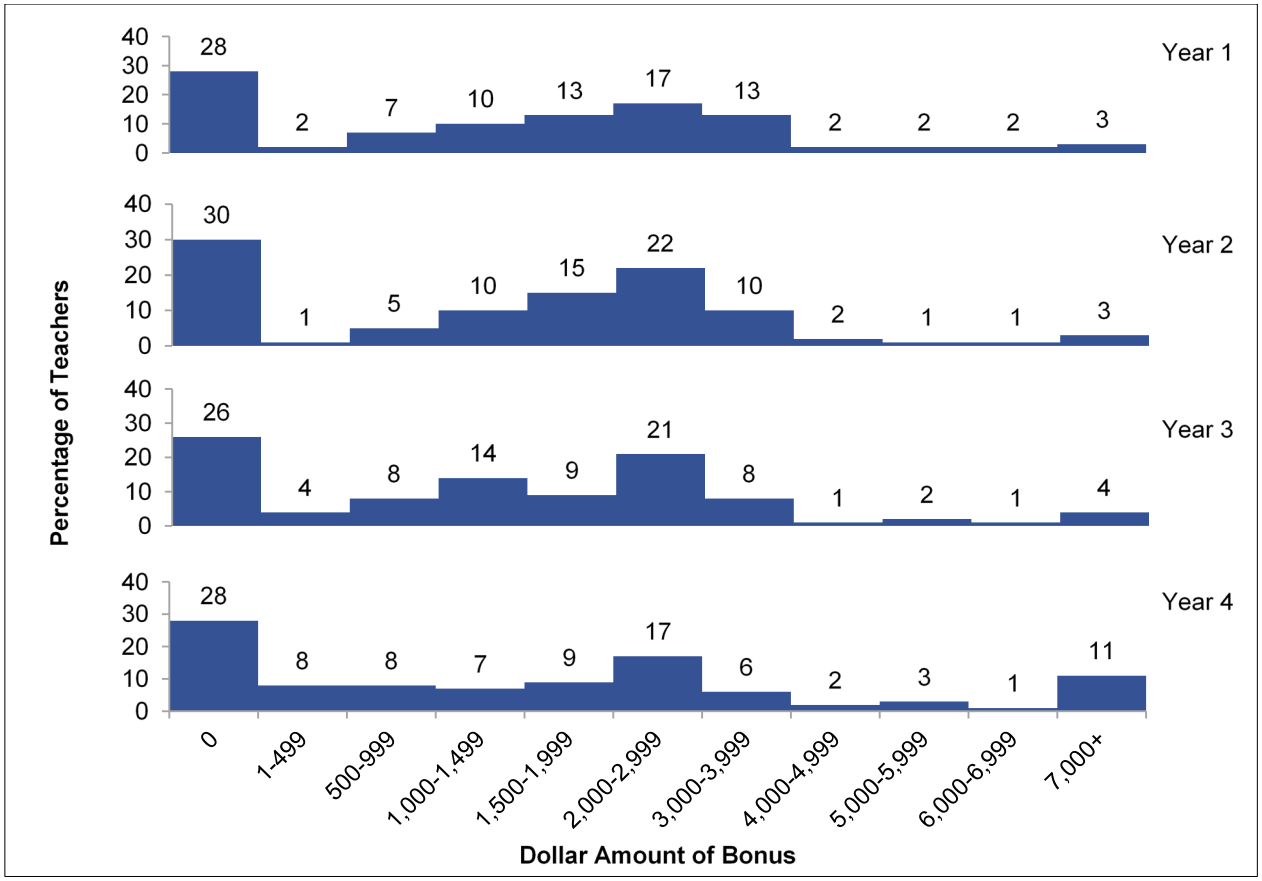
Notes: The statistics shown in this figure represent an equal-weighted average of the statistics from the 10 evaluation districts in Cohort 1. Findings were similar when districts were weighted by the number of schools (Appendix D, Figure D.2).

Figure reads: In Year 1, on average across the evaluation districts, the minimum pay-for-performance bonus was $0, the average pay-for-performance bonus was $1,945, and the maximum pay-for-performance bonus was $7,787.

[32] When findings for Years 1, 2, and 3 were based on both Cohorts 1 and 2, the average and maximum performance bonus amounts were similar to the average and maximum bonus amounts for Cohort 1 only (Appendix D, Figure D.1).

In each year, most teachers received a performance bonus. In each of the four years, about 70 percent of teachers in treatment schools received a bonus (Figure IV.4). Therefore, it is likely that even many teachers whose performance was below the average received a bonus.[33] Although most teachers received a bonus, large bonuses were more challenging to earn. Across districts, about half of teachers (for example, 49 percent in Year 4) received a performance bonus of at least $1,500, which is about three times the automatic 1 percent bonus that control teachers received (about $500). However, fewer than 15 percent of teachers in treatment schools received a performance bonus of at least $5,000, or about 10 percent of the average teacher salary among the evaluation districts.

Figure IV.4. Distribution of Pay-for-Performance Bonuses for Teachers in Treatment Schools



Source: Educator administrative data (N = 2,151 teachers in Year 1; N = 2,160 teachers in Year 2; N = 2,236 teachers in Year 3; and N = 2,083 teachers in Year 4).

Figure reads: In Year 1, 28 percent of teachers did not receive a pay-for-performance bonus, and 2 percent received a pay-for-performance bonus from $1 to $499.

[33] It is possible that pay-for-performance bonuses could have led teachers in treatment schools to improve their effectiveness substantially. If so, more than half of the treatment teachers could have qualified for bonuses even if the bonuses had been challenging to earn. However, if that had been the case, we would expect treatment teachers' observation and school achievement growth ratings to be significantly higher (statistically and meaningfully) than the ratings of teachers in control schools, but that did not occur (see Chapter VI).

The average and maximum performance bonus amounts for teachers and the percentage of teachers who received performance bonuses varied across districts. The averages across the 10 districts (Figures IV.3 and IV.4) masked differences across districts in the bonus amounts that teachers earned and the percentage of teachers who earned bonuses. For example, in Year 4:

- The maximum performance bonus amounts ranged from 2 to 12 times the average performance bonus. The maximum bonuses were less than $4,500 in two districts, $5,000 to $10,000 in six districts, and $15,000 or more in two districts.[34,35] The large variation in maximum bonus amounts across districts suggests that setting the range of performance bonuses was an important dimension on which the evaluation districts exercised discretion in designing their TIF program, and this led to substantially different maximum bonus amounts.
- The percentage of teachers who received performance bonuses ranged from 28 to 96 percent of teachers (Appendix D, Figure D.8). Only 2 of 10 districts awarded performance bonuses to fewer than half of teachers (Appendix D, Table D.8). In the other 8 districts, more than 70 percent of teachers received performance bonuses.[36]

Within districts, the distribution of bonuses was relatively similar across years, with two exceptions. Districts D and J had maximum bonuses that were $7,000 to $8,000 higher in Year 4 than in other years (Figure IV.5; Appendix D, Figures D.3, D.4, and D.5).

Because districts awarded separate bonuses for different performance measures, determining the amount of the bonus that was tied to each performance measure was a key decision that districts made to determine the structure of the incentives for teachers. For example, districts had to consider whether to tie larger bonuses to measures of individual performance (such as classroom observations and classroom achievement growth) or measures of school or team performance (such as school achievement growth and achievement growth of student subgroups). Larger bonuses for group performance measures might encourage collaboration, but larger bonuses for individual performance measures might enable teachers to feel more empowered to enhance the size of their own bonus. Among the measures, districts also had to consider whether larger bonuses for classroom observations or student achievement growth would provide stronger incentives for teachers to improve. Although student achievement growth was a more objective measure, teachers placed far less faith in student test scores than in their own principals to evaluate teacher effectiveness (Chapter V, Table V.3). As discussed next, the bonus structure differed substantially between districts that did and did not use classroom achievement growth and between teachers in

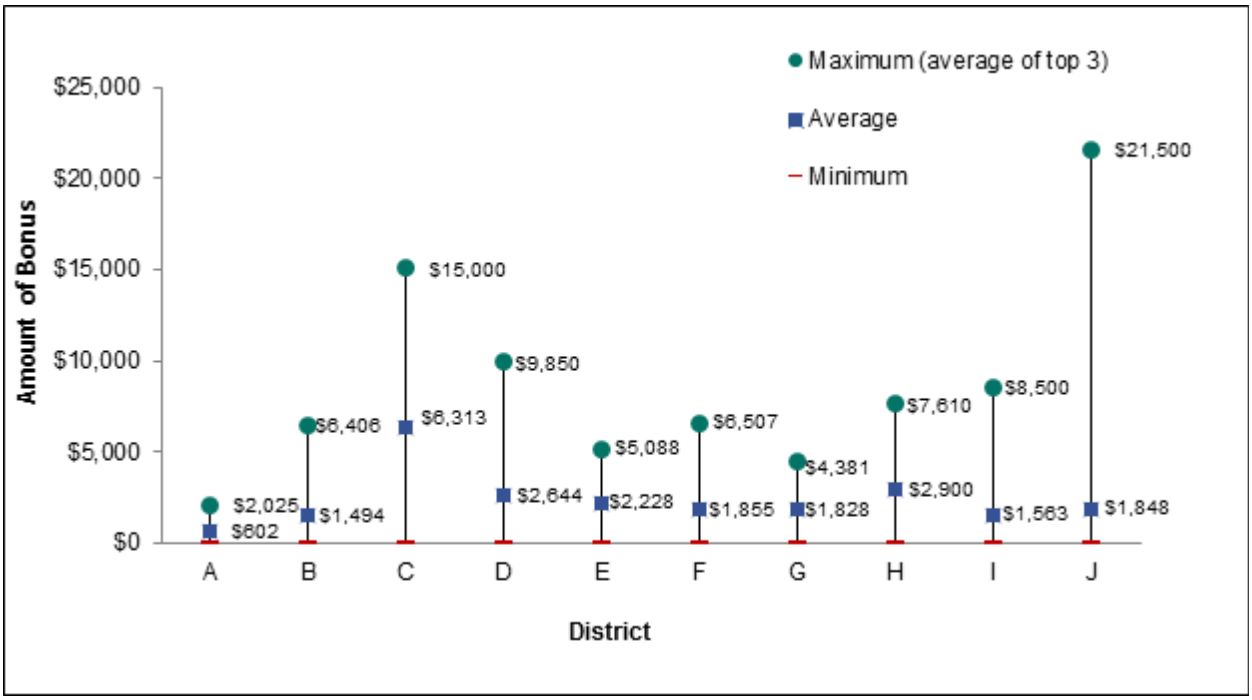
[34] In some districts, fewer than three teachers received the maximum bonus. To reduce the risk of disclosing information on individual teachers, we averaged the top three performance bonuses awarded to teachers in the district and used that value to represent the maximum bonus.

[35] Maximum performance bonus amounts varied to a similar extent across all Cohort 1 and Cohort 2 districts in Year 3. In particular, the maximum bonus amounts for the three districts in Cohort 2 ranged from about $4,300 to $6,000 (Appendix D, Figure D.7). To ensure districts' confidentiality, the lettering of the districts in this figure and in other parts of the report does not mirror the numbering of the districts in Table IV.3 or Appendix D, Tables D.6 and D.7.

[36] In Year 3, 11 of the 13 Cohort 1 and Cohort 2 districts awarded performance bonuses to more than half of their teachers (Appendix D, Table D.8).

tested and nontested grades and subjects within the districts that used classroom achievement growth.

Figure IV.5. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools in Year 4, by District



Source: Educator administrative data (N ranges from 68 teachers in District D to 374 in District J).

Note: To reduce the risk of disclosing information on an individual, the maximum bonus in this figure represents the average of the top three performance bonuses awarded to teachers in the district.

Figure reads: For District A in Year 4, the minimum pay-for-performance bonus was $0, the average pay-for-performance bonus was $602, and the average pay-for-performance bonus among the top three bonuses was $2,025.

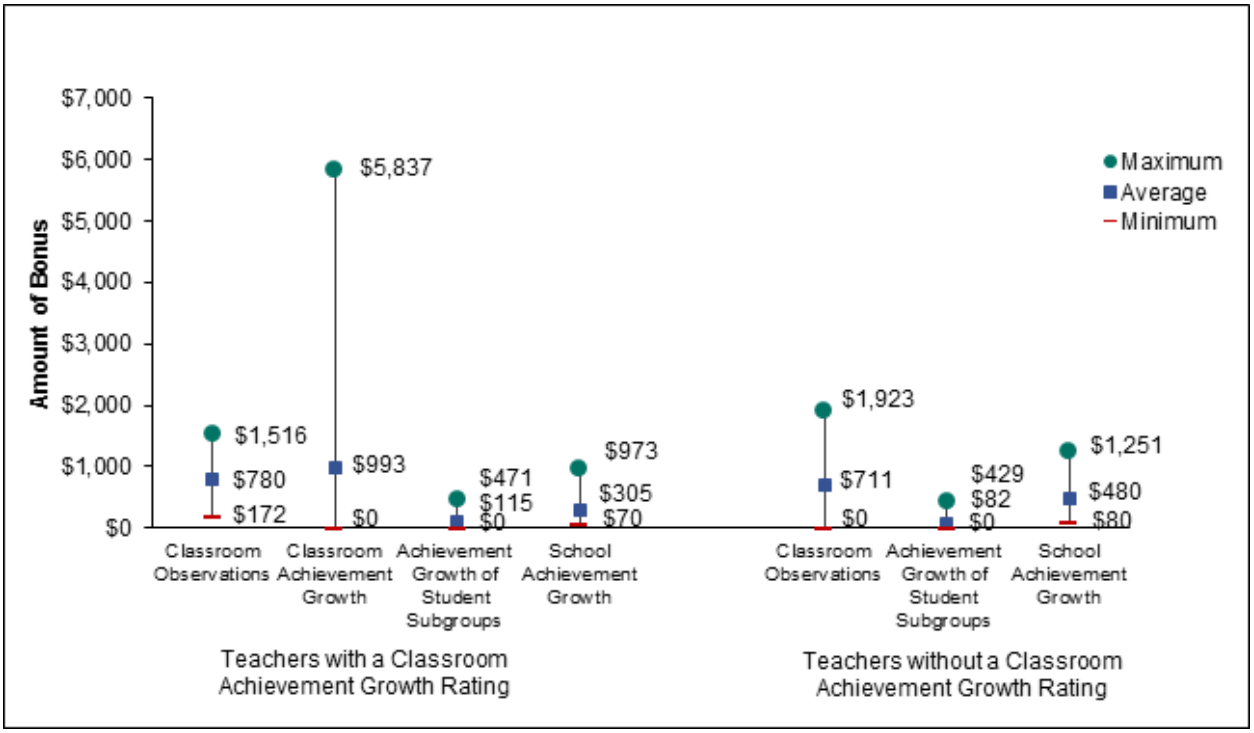
Because districts awarded separate bonuses for different performance measures, determining the amount of the bonus that was tied to each performance measure was a key decision that districts made to determine the structure of the incentives for teachers. For example, districts had to consider whether to tie larger bonuses to measures of individual performance (such as classroom observations and classroom achievement growth) or measures of school or team performance (such as school achievement growth and achievement growth of student subgroups). Larger bonuses for group performance measures might encourage collaboration, but larger bonuses for individual performance measures might enable teachers to feel more empowered to enhance the size of their own bonus. Among the measures, districts also had to consider whether larger bonuses for classroom observations or student achievement growth would provide stronger incentives for teachers to improve. Although student achievement growth was a more objective measure, teachers placed far less faith in student test scores than in their own principals to evaluate teacher effectiveness (Chapter V, Table V.3). As discussed next, the bonus structure differed substantially between districts that did and did not use classroom achievement growth and between teachers in tested and nontested grades and subjects within the districts that used classroom achievement growth.

Bonuses for teachers who were evaluated on classroom achievement growth were determined mostly by their individual performance on classroom observations and classroom achievement growth. For example, within the seven districts that used classroom achievement growth in Year 4, about 60 percent of teachers—typically, those who taught grades and subjects tested by state assessments—received classroom achievement growth ratings (Appendix A, Table A.14). Those teachers could potentially earn nearly $6,000 for their ratings on that measure alone—more than three times as much as the potential bonus for any other measure (Figure IV.6). However, few teachers received classroom achievement growth bonuses close to the maximum amount. On average, these teachers received a total bonus of $2,193 ($780 for classroom observations, $993 for classroom achievement growth, $115 for achievement growth of student subgroups, and $305 for school achievement growth). Therefore, these teachers on average earned about four times as large a bonus based on their individual performance ($1,773 for classroom observations and classroom achievement growth combined) than based on a group's performance ($420 for achievement growth of student subgroups and school achievement growth combined). They also earned, on average, more of their bonus based on the three measures of student achievement growth—classroom achievement growth, achievement growth of student subgroups, and school achievement growth—than on classroom observations ($1,413 versus $780).

Teachers in nontested grades and subjects earned smaller bonuses overall than their colleagues in tested grades and subjects. Most of their bonus was still determined by their individual performance, but in this case measured by classroom observations only. In those same seven districts, teachers who were not evaluated on classroom achievement growth—those in nontested grades and subjects—earned a smaller average bonus in total. Across the three main performance measures combined, teachers without classroom achievement growth ratings received, on average, a total bonus of $1,273, about 60 percent of the total average bonus earned by teachers who were assessed on classroom achievement growth in Year 4 ($2,193 for classroom achievement growth plus the other three performance measures combined; Figure IV.6). On average, teachers who were not evaluated on classroom achievement growth earned more for classroom observations ($711) than for the group performance measures based on student achievement growth—achievement growth of student subgroups and school achievement growth ($562).

In districts that did not award bonuses based on measures of classroom achievement growth, teachers' bonuses were determined mostly by group performance measures based on student achievement growth. In those three districts, teachers received, on average, a bonus of almost $3,700 across the three main performance measures—classroom observations, school achievement growth, and the achievement growth of student subgroups (Figure IV.7). Almost 70 percent ($2,525) of their overall bonus came from group performance measures based on student achievement growth ($1,872 for school achievement growth and $653 for achievement growth of student subgroups).

Figure IV.6. Minimum, Average, and Maximum Performance Bonus for Each Performance Measure, in Districts Using Classroom Achievement Growth Measures, Year 4



Source: Educator administrative data (N = 896 teachers with a classroom achievement growth rating, and N = 408 teachers without a classroom achievement growth rating).

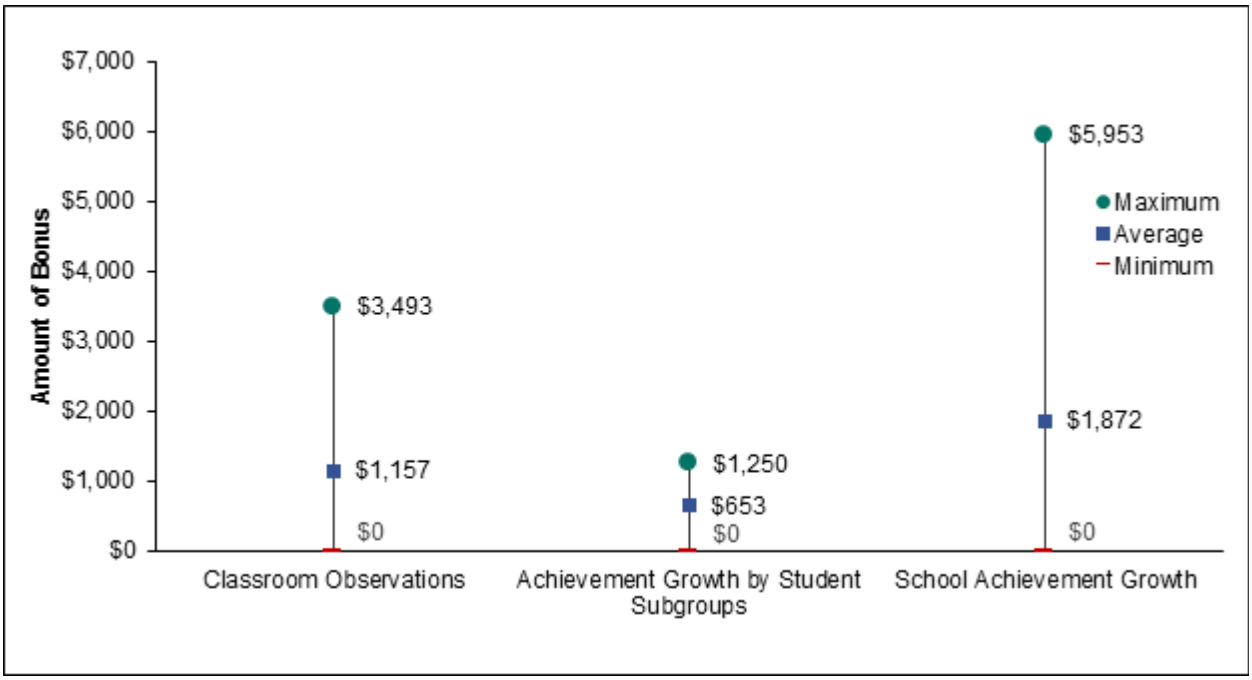
Notes: Seven districts used classroom achievement growth measures in Year 4. Figure is based on teachers in those districts who received classroom observation ratings.

Figure reads: On average across districts that used classroom achievement growth measures in Year 4, among teachers with a classroom achievement growth rating, the minimum bonus for classroom observation ratings was $172, the average was $780, and the maximum was $1,516. Across the same districts, among teachers without a classroom achievement growth rating, the minimum bonus for classroom observation ratings was $0, the average was $711, and the maximum was $1,923.

To explore whether the bonuses that teachers earned provided them with consistent messages about their performance over time, we examined the total bonus amounts that teachers received across years.

Half of teachers received similar performance bonus amounts from one year to the next. For example, 51 percent of teachers received a bonus in the same dollar amount range in Year 4 as they did in Year 3. Specifically, from Year 3 to Year 4, 20 percent of teachers continued not to earn a bonus, 11 percent continued to earn a bonus of $1,500 or less, 12 percent continued to earn a bonus ranging from $1,501 to $3,000, and 8 percent continued to earn a bonus above $3,000 (Appendix Table D.9).

Figure IV.7. Minimum, Average, and Maximum Performance Bonus for Each Performance Measure, in Districts Not Using Classroom Achievement Growth Measures, Year 4



Source: Educator administrative data (N = 457).

Notes: Three districts did not use classroom achievement growth measures in Year 4. Figure is based on teachers in those districts who received classroom observation ratings.

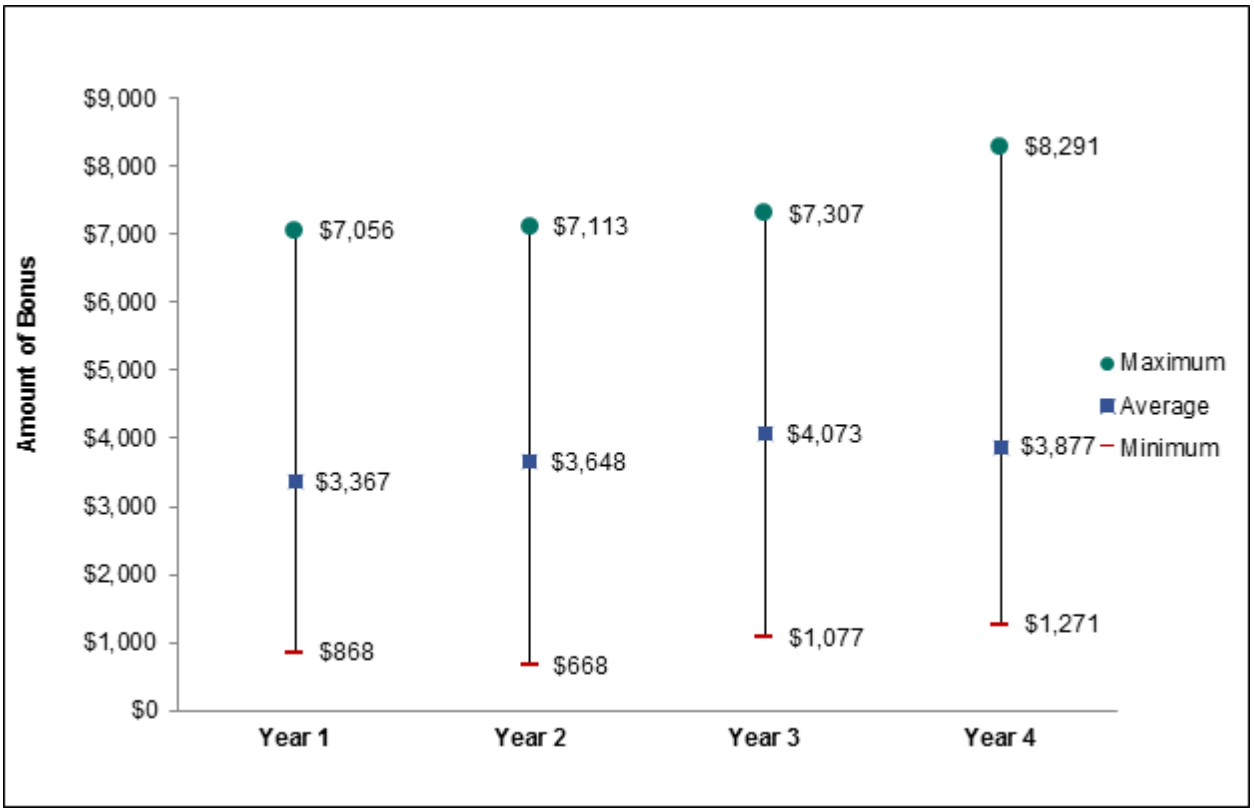
Figure reads: On average across districts that did not use classroom achievement growth measures in Year 4, the minimum bonus for classroom observation ratings was $0, the average was $1,157, and the maximum was $3,493.

We also examined the structure of bonuses awarded to principals. All evaluation districts provided principals the opportunity to earn a bonus based on school achievement growth, and 9 of 10 offered principals bonuses based on at least one other performance measure, such as an observation rating or the achievement growth of student subgroups.

The average performance bonuses for principals were no more than 5 percent of principals' salaries, the maximum bonuses were about twice the average bonus, and most principals received a bonus. Across evaluation districts, the average performance bonus was about $4,000 (Figure IV.8). This average bonus represented about 3 to 5 percent of the average principal salary ($89,267 to $91,311 in Years 1 through 4). The maximum bonus ranged from $7,056 in Year 1 to $8,291 in Year 4 and was consistently about twice the average bonus. More than 70 percent of principals in treatment schools received a bonus in each year (Appendix D, Figure D.12).[37] However, fewer than 15 percent received a large performance bonus of at least $8,000, or approximately 10 percent of the average principal salary among the evaluation districts.

[37] Only three districts had average bonuses that were at least 5 percent of principals' salaries in most years, one or two districts had maximum bonuses that were at least three times the average bonus, and one or two districts per year awarded bonuses to fewer than half their principals (Appendix D, Table D.10). When findings for Years 1, 2, and 3 were based on both Cohorts 1 and 2, the average and maximum performance bonus amounts were similar to the average and maximum bonus amounts for Cohort 1 only (Appendix D, Figure D.10).

Figure IV.8. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Principals in Treatment Schools



Source: Educator administrative data (N = 64 principals in Year 1; N = 67 principals in Year 2; N = 63 principals in Year 3; and N = 64 principals in Year 4).

Notes: The statistics shown in the figure represent an equal-weighted average of the statistics from the 10 evaluation districts in Cohort 1. When districts were weighted by the number of schools, average bonus amounts were similar to those shown in this figure, but maximum bonus amounts were about $1,200 higher than those shown in this figure (Appendix D, Figure D.11).

Figure reads: In Year 1, on average across the evaluation districts, the minimum pay-for-performance bonus was $868, the average pay-for-performance bonus was $3,367 and the maximum pay-for-performance bonus was $7,056.

As intended by the study design, the automatic 1 percent bonus provided to teachers and principals in control schools was small and generally did not vary substantially. The automatic bonus for educators in control schools ensured that all educators in evaluation schools had the opportunity to benefit monetarily from participating in the study. However, the automatic bonuses were purposefully designed to be small and fairly uniform in order for educators in treatment schools to be eligible for larger and more differentiated bonuses than educators in control schools. The average automatic bonus for teachers in control schools was about $400 to $500 each year, and the maximum automatic bonus was only slightly higher (about $700), with one exception (Appendix D, Figure D.13). In Year 4, one district gave substantially larger bonuses to teachers in control schools, which was reflected in the higher maximum bonuses to control teachers for that year.[38] For principals in control schools, the average automatic bonus was about $800 in each year, with a maximum automatic bonus of almost $1,000.

[38] One district awarded performance bonuses to teachers in control schools in Year 4 who had received high ratings on classroom achievement growth. The control teachers in this district did not know they could receive a

Requirement 3: Additional Pay Opportunities

Consistent with the goal of improving the teaching workforce in high-need schools, the TIF grant required that districts provide additional pay for effective educators to take on extra roles and responsibilities. Examples from the TIF notice included serving as a master or mentor teacher, whose roles typically include mentoring novice teachers, developing professional learning communities, and tutoring students. Using data from district surveys, district interviews, and administrative data, we examined the percentage of evaluation districts that provided additional pay opportunities, the types of roles and responsibilities offered, and the amount of the additional pay.

All evaluation districts met the TIF grant requirement to offer additional pay opportunities, most commonly in the form of master, mentor, or lead teacher opportunities. All districts reported offering additional pay for teachers to take on extra roles and responsibilities. Districts most commonly reported offering teachers additional pay for the roles of master and mentor teachers—for example, in Year 4, at least 50 percent of evaluation districts reported offering these roles (Table IV.4). During interviews, districts noted that master teachers might lead professional development sessions and mentor teachers might provide day-to-day coaching or modeling of lessons. Few districts—in fact, none in Year 4—offered additional pay opportunities to principals.

We compared the amount of money teachers could earn for these additional pay opportunities with the amount they could earn for pay-for-performance bonuses. According to the theory of change (Chapter I), pay-for-performance should encourage teachers to improve their practices in order to receive a bonus. However, if effective teachers could earn as much or more from becoming a master or mentor teacher, then teachers in treatment and control schools might have had similar incentives to improve to qualify for these additional pay opportunities. If so, these additional pay opportunities could have diminished the potential impacts of pay-for-performance. On the other hand, additional pay opportunities might be less attractive than a pay-for-performance bonus if the amount and type of additional work required for the additional pay do not appeal to teachers.

Although the amounts of additional pay for certain types of roles and responsibilities were substantial, few teachers actually received additional pay. For example, in Year 4, teachers could earn up to $6,100 in additional pay for serving as a master or lead teacher (Table IV.4). However, these large amounts of additional pay had no bearing on the compensation that most teachers received, because only a small fraction of teachers (17 percent) were chosen for and took on any extra responsibilities that were linked to additional pay (Appendix D, Table D.11). Therefore, on average, teachers received much less compensation from additional pay ($507 in Year 4) than from performance bonuses ($2,328 in Year 4; Figure IV.3).[39]

performance bonus until after the school year had ended, so the possibility of receiving one should not have affected their behavior. This analysis includes all bonuses given to teachers in control schools (1 percent automatic and performance bonuses).

[39] Average amounts of additional pay for roles and responsibilities did not differ between teachers in treatment and control schools (see Appendix D, Table D.12).

Table IV.4. Additional Pay Opportunities, as Reported by Evaluation Districts, Year 4

	Percentage of Districts that Offered Additional Pay	Average Maximum Pay in Districts Offering Additional Pay
Teachers Could Receive Additional Pay for Taking on Extra Roles or Responsibilities	100	NA
Roles and Responsibilities		
Mentor teacher	60	$2,133
Master or lead teacher	50	$6,100
Department chair or head	20	—
Lead curriculum specialist	0	—
Serving on a schoolwide committee or task force	30	—
Leadership team member	30	—
Additional Factors		
Teaching in a hard-to-staff school or high-need subject area	40	$3,250
Attending professional development activities or enrolling in graduate-level courses	10	—
Principals Could Receive Additional Pay for Taking on Extra Roles or Responsibilities in School or District	0	—
Number of Districts—Range[a]	**10**	**4-6**

Source: District survey, 2015.

Note: Table reports on activities funded by TIF.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

— is not reported because of small sample size.

NA is not applicable.

Requirement 4: Professional Development

The TIF grant required that districts provide professional development linked to the measures of educator effectiveness. This support included professional development to help educators understand the measures used to evaluate their performance as well as feedback based on their actual performance ratings to help improve their instructional practices.[40] To describe the professional development that districts offered, we used data from the district survey and interviews with district administrators.

Most districts provided the required professional development, but more districts offered professional development to help teachers understand how they were being evaluated than to help them improve their individual performance ratings. Nearly all evaluation districts (at least 90 percent in each year) provided general information to help teachers understand the performance measures used for their TIF program (Table IV.5). Fewer districts, but still over half (70 to 80 percent), offered targeted feedback to teachers based on their actual ratings. When districts did provide targeted feedback, it was more commonly aimed at helping teachers improve their observation ratings than their achievement growth ratings (Appendix D, Table D.13).

[40] Surveys of district administrators did not ask about professional development for principals.

Table IV.5. Professional Development Activities for Teachers Planned Under TIF, as Reported by Evaluation Districts (Percentages)

	Year 1	Year 2	Year 3	Year 4
Focus of Professional Development				
Understanding performance measures of TIF program	100	100	90	90
Feedback based on TIF performance ratings	70	70	70	80
Number of Districts	**10**	**10**	**10**	**10**

Sources: District surveys (2012, 2013, 2014, and 2015).

Communication of TIF Program

In addition to implementing the required components of TIF, districts had to effectively communicate information about those components to educators. In this section, we describe evaluation districts' reported communication about their TIF program, such as how and what information they communicated to educators. We focus on two types of information that districts had to communicate: general information about the program and specific information to individual teachers about the performance ratings and bonuses they had earned in the prior school year. Data for this section come from the district survey and interviews with district administrators. Examples of districts' communication approaches are primarily based on Years 3 and 4, the years in which district interviews included additional questions about these activities.

Districts differed in whether district or school staff had responsibility for communicating general information about TIF programs to teachers. Deciding who communicated information about TIF involved a trade-off. Using a centralized approach to communication (for example, having district or grantee staff explain the program to teachers) might help ensure uniformity and accuracy of information, whereas a decentralized approach might improve understanding by communicating through school staff who had closer relationships with the teachers (for example, asking principals to explain the program to their teachers).[41] In each year, some districts chose centralized approaches whereas others chose decentralized approaches. Before Year 4, slightly more districts (6 or 7 of 10) reported that communication about TIF came from district or grantee staff, rather than school staff (Appendix D, Table D.14; Chiang et al. 2015). In Year 4, half of the 10 districts reported that communication about TIF came from district or grantee staff, and half reported that communication came from school staff.

Districts used multiple approaches to explain their TIF program to educators. To ensure educators were aware of the TIF program, districts communicated key aspects of the program to educators each year. Districts communicated information through a variety of approaches. For example, most districts (at least 70 percent) reported using written materials, group presentations, and the district website to explain how teachers would be evaluated in Year 4 (Appendix D, Table D.14). During interviews, districts described using these communication approaches to review the observation rubrics and student achievement growth measures with teachers and to remind educators of the criteria for earning a performance bonus. Most districts also reported using group

[41] As discussed in Chapter II, 4 of the 10 districts received TIF grants directly from the U.S. Department of Education. The remaining districts were part of multidistrict grants administered by another grantee organization (such as a state education agency, university, association of charter schools, or nonprofit organization), and either grantee or district staff could have helped ensure uniformity of the information communicated to educators.

presentations (80 percent), the district website (80 percent), and written materials (70 percent) to inform educators of the potential amount they could earn in performance bonuses. On average, districts used about three approaches to communicate aspects of their TIF program to educators.

Districts differed in whether they assessed teachers' understanding of their eligibility to earn a performance bonus, and few districts did so in the final year. To ensure educators understood key program components, the technical assistance team encouraged districts to assess educators' understanding. This feedback could provide district administrators with valuable information on the effectiveness of their communication approaches. If necessary, districts could modify their communication activities to improve educators' understanding of their TIF program, including their understanding of their eligibility to earn a performance bonus. Despite being encouraged to assess educators' understanding, not all districts adopted this practice. In Year 3, near the middle of their grant period, six districts used a survey or focus group to determine if teachers understood their eligibility for a bonus, whereas four districts did not (Appendix D, Table D.14). By Year 4, the final year of the grant, only two districts assessed teachers' understanding of their bonus eligibility.

Some districts' plans to discontinue performance bonuses after the end of the TIF grant could have led them to stop assessing teachers' understanding of bonus eligibility in the final year. We found some, though not a very strong, relationship between districts' future plans and their assessment of teachers' understanding. As we discuss in Chapter VII, six districts planned to discontinue performance bonuses after the end of their grants. Of those six districts, none assessed teachers' understanding in the final year. Nevertheless, even among the remaining four districts that planned to continue or were considering continuing these bonuses, only two assessed teachers' understanding in the final year.

Few districts provided details to teachers about the number and size of the bonuses awarded in the previous year. In addition to knowing the criteria to earn a bonus, teachers also could find information about the actual bonuses awarded to other teachers helpful to predict the size and likelihood of their receiving a bonus for the current year. For example, information from the prior-year bonus awards, such as the average and maximum bonuses awarded and the percentage of teachers who received a bonus, could enable teachers to better assess whether they could earn a larger bonus than they had received in the past. Nevertheless, no more than 3 of the 10 districts reported informing educators about the percentage of teachers who received a bonus in Years 2 and 3, and no more than 2 districts reported providing information about the maximum or average bonuses awarded in these years (Table IV.6).

Not only is communicating general information about the TIF program important for promoting understanding and motivating educators to change their practices, communicating educators' individual performance ratings and bonuses can also be important.

Almost all districts used in-person meetings to inform teachers about their individual ratings on observations and student achievement growth. Teachers who receive in-person communication about their observation or student achievement growth rating might be more aware of their measured effectiveness. Letters or emails providing information on the teachers' performance ratings cannot guarantee that the teachers will read the information. Similarly, providing the information online does not mean teachers will access the information. Eighty to 90 percent of the districts reported that they used in-person meetings with either groups or individual teachers to inform them of both their observation rating and their student achievement growth

rating (Appendix D, Table D.15). By the end of the grant, fewer than half of the districts reported using other methods (such as an online system, letters, or emails) to communicate these ratings.

Table IV.6. Information Districts Provided to Teachers About Actual Pay-for-Performance Bonuses from Years 2 and 3 (Percentages)

	Bonuses from Year 2			Bonuses from Year 3		
	Treatment Teachers			Treatment Teachers		
	Those Who Got a Bonus	Those Who Did Not Get a Bonus	Control Teachers	Those Who Got a Bonus	Those Who Did Not Get a Bonus	Control Teachers
General Information on Performance Bonuses						
Maximum bonus anyone received in school or district	10	10	10	20	20	20
Average bonus received in school or district	10	10	10	10	10	10
Percentage of teachers in school or district who received a bonus	30	30	20	20	20	20
Explanation of how bonuses were calculated	100	100	80	80	80	70
Information on the Teacher's Individual Performance Bonus						
Whether individual received a bonus	100	60	NA	100	70	NA
Bonus amount	80	60[a]	NA	80	70	NA
Number of Districts	**10**	**10**	**10**	**10**	**10**	**10**

Sources: District interviews (2014 and 2015).

[a]This is $0 for treatment teachers who did not get a bonus.

NA is not applicable.

All districts informed bonus recipients that they earned a performance bonus, but fewer districts informed nonrecipients that they would not receive a bonus. For pay-for-performance to lead to improvements in teaching, it might be important for both bonus recipients and nonrecipients to know whether they got a bonus and the amount received. For teachers who earned a bonus, being aware that they earned a bonus could improve their overall job satisfaction and motivate them to work toward earning another bonus. Informing teachers who did not earn a bonus could help ensure that nonrecipients were aware of the missed opportunity to earn a bonus and motivate them to improve their teaching practices. All districts reported informing bonus recipients of their Year 2 and 3 awards; 60 to 70 percent reported informing nonrecipients that they did not earn a bonus (Table IV.6). In practice, because most teachers received a bonus, this resulted in about 10 to 25 percent of treatment teachers not being notified that they did not receive a performance bonus.

Most districts used letters to let teachers know whether they had earned an individual performance bonus and how much they earned. Some methods of communicating about performance bonuses, such as written correspondence, may better ensure uniformity of the message about an individual performance bonus. However, holding individual meetings with teachers to discuss their bonuses might enable teachers to better understand why they received (or did not

receive) a bonus. In general, evaluation districts chose uniform written correspondence, rather than individualized in-person meetings, to inform teachers of their individual bonuses. When informing bonus recipients of their individual performance bonuses from Years 1 through 3, most of the districts (at least 75 percent) chose to send letters, whereas fewer than half (30 to 37 percent) reported holding individual meetings with teachers (Appendix D, Table D.16).

Most districts did not notify teachers of the bonuses they earned before the start of the next school year. For information about bonuses to affect teachers' behavior, teachers must receive the information when there is still enough time to affect their school choice (for example, requesting a transfer to a school that offers or does not offer a bonus) or their teaching practices (for example, enrolling in professional development to learn how to perform better on the performance measures used to award bonuses). Evaluation districts differed in the timing of notifying teachers of their bonuses and paying those bonuses. For example, when awarding Year 3 bonuses, only three districts reported notifying and paying any teachers before the start of the subsequent (2014–2015) school year. In those districts, the early awards were based on observations of classroom or school practices, with awards based on achievement growth occurring later. The remaining seven districts reported notifying and paying teachers from October 2014 to June 2015.

Educators' Understanding of and Experiences with TIF

Although districts designed, implemented, and communicated the program components, teachers' and principals' understanding of and experiences with these components ultimately determined how the program had the potential to influence their behaviors and, consequently, student achievement (as described by the theory of change discussed in Chapter I). Examining educators' reports about program features was also important because it could identify ways in which their understanding of or exposure to the TIF program did or did not align with what grantees intended or what district officials reported, highlighting possible challenges in the implementation process.

This section examines educators' reported understanding of and experiences with TIF performance measures, pay-for-performance bonuses, additional pay opportunities, and professional development, drawing primarily on teachers' and principals' survey responses. Although pay-for-performance was the only component that was supposed to differ between treatment and control schools, educators' understanding of or experiences with all four required components could have differed between treatment and control schools (if, for example, information was communicated differently to the two groups of educators or they paid different amounts of attention to this information). Therefore, we describe the understanding and experiences of treatment and control educators separately. We also examine educators' evolving understanding of their TIF program, because that understanding might change as districts refine communication strategies and information becomes more widely disseminated. Because we administered Year 1 surveys before educators had received any performance bonuses and surveys in Years 2, 3, and 4 after bonuses had been awarded for one, two, and three years, changes in understanding might also result from educators' having received bonuses or heard about them.

Educators' Understanding of Performance Measures

For the program to change educators' behavior and ultimately student outcomes, educators must understand how they are being evaluated, as a first step toward figuring out how to improve their performance.

Most teachers were aware of being evaluated based on student achievement growth and classroom observations early in TIF implementation, and their awareness of these performance measures improved over time. At least 70 percent of teachers in Year 1 reported being evaluated on student achievement growth, and more than 70 percent reported being evaluated on at least two classroom observations (Table IV.7). Furthermore, the percentages of teachers who reported being evaluated on student achievement growth generally increased over time, with significant improvements between some years for both treatment and control teachers. By Year 4, 89 percent of treatment teachers and 78 percent of control teachers reported being evaluated on this measure. Likewise, the percentages of teachers who reported being evaluated on at least two classroom observations also generally grew over time, with more than 85 percent of teachers reporting being evaluated on this measure in Year 4. Similar to teachers, about 85 percent of principals in Year 4 reported being evaluated on student achievement growth; however, a smaller percentage of principals (less than 60 percent) reported being evaluated on at least two observations of their school practices (Table IV.8).

Educators in treatment schools tended to be more likely than educators in control schools to report being evaluated on student achievement growth. The study was designed so that educators in treatment and control schools should be evaluated in the same way. Consistent with this design, similar percentages of treatment and control teachers reported being evaluated on student achievement growth in Year 1 (about 70 percent). However, after Year 1, treatment teachers were more likely to report being evaluated on student achievement growth than control teachers (for example, 89 percent of treatment teachers and 78 percent of control teachers reported being evaluated on this measure in Year 4; Table IV.7). Treatment principals also were more likely than control principals to report being evaluated on student achievement growth in Years 2 and 3, but the two groups had similar understanding in Year 4 (Table IV.8). Overall, the evidence suggests that the offer of pay-for-performance could have led educators in treatment schools to be more aware of how they were evaluated.

Educators' Understanding of Their Eligibility for Pay-for-Performance Bonuses

The prospect of earning a performance bonus could motivate educators to improve their practices. To do so, however, they must have a correct understanding of their eligibility for bonuses. Based on the study design, we would expect that all teachers in treatment schools would report being eligible for a pay-for-performance bonus, whereas teachers in control schools would report only being eligible for an automatic 1 percent bonus.

Teachers' and principals' understanding of their bonus eligibility did not improve after the second year of implementation. Although treatment and control educators' understanding of their bonus eligibility improved significantly from Year 1 to Year 2, there was no further improvement after Year 2. About half (49 percent) of teachers in treatment schools in Year 1 reported being eligible for a performance bonus (Figure IV.9). In Years 2 through 4, about 60 percent of teachers in treatment schools reported being eligible for a performance bonus (61 percent in Year 2, 57 percent in Year 3, and 58 percent in Year 4).[42] Likewise, among principals in treatment schools, a higher percentage reported being eligible for a performance bonus in Year 2 (90 percent) compared to Year 1 (56 percent; Figure IV.10), but there were no further improvements in Years 3 or 4.

[42] When we restricted the sample to teachers who were in treatment schools in all four years, the results were similar. Therefore, the changes over time in the percentages of teachers who understood their eligibility likely reflected learning by individual teachers, rather than the arrival or departure of teachers with different levels of awareness.

Table IV.7. Teachers' Reports of the Measures Used to Evaluate Teachers (Percentages)

	Year 1		Year 2		Year 3		Year 4	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Student Achievement Measures								
Student achievement level	56	61	69 +	67	74	71	77	75
Any student achievement growth	71	70	78*+	72	84*	78+	89*	78
School achievement growth	62	63	73*+	68	79*	74	85*	75
Achievement growth of student subgroups[a]	55	56	66*+	60	67	68+	74	70
Classroom achievement growth	60	62	57	58	65	64	64	63
Classroom Observation Measure								
At least two classroom observations by trained observers	73	76	87 +	83+	91	88	88 +	86
Number of Teachers—Range[b]	**374-376**	**393-398**	**427-432**	**432-434**	**409-415**	**427-431**	**377-382**	**375-380**

Sources: Teacher surveys (2012, 2013, 2014, and 2015).

[a]Examples of student subgroups include grouping students by grade, team, or subject area.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference between treatment and control group is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Table IV.8. Principals' Reports of the Measures Used to Evaluate Principals (Percentages)

	Year 1		Year 2		Year 3		Year 4	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Student Achievement Measure								
Student achievement level	88	93	85*	69+	75	75	66*	82
Any student achievement growth	88	92	91*	67+	>94[a]*	81	85 +	84
School achievement growth	89	90	90*	65+	>94[a]*	81	82 +	84
Achievement growth of student subgroups[b]	83	90	82*	64+	78	69	72	70
Observation Measure								
At least two observations by trained observer	—	—	47	59	63+	51	52	58
Number of Principals—Range[c]	**58-62**	**58-60**	**62-63**	**57-58**	**57-58**	**57-59**	**57-59**	**60-62**

Source: Principal survey (2012, 2013, 2014, and 2015).

[a]To reduce the risk of disclosing information on an individual, the exact percentage is not shown.

[b]Examples of student subgroups include grouping students by grade, team, or subject area.

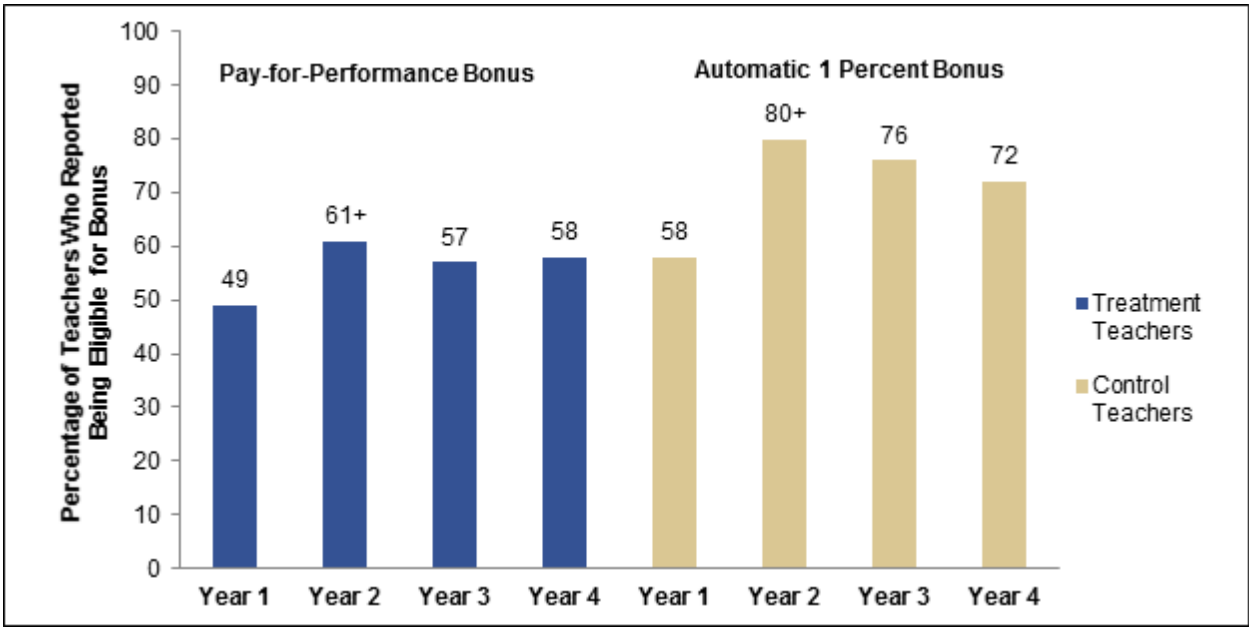
[c]Sample sizes are presented as a range based on the data available for each row in the table.

— is not available.

*Difference between treatment and control group is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Figure IV.9. Teachers' Bonus Eligibility, as Reported by Teachers (Percentages)



Sources: Teacher surveys (2012, 2013, 2014, and 2015).

Notes: A total of 368 treatment teachers in Year 1, 439 in Year 2, 420 in Year 3, and 391 in Year 4 responded to the question about eligibility for a pay-for-performance bonus. A total of 381 control teachers in Year 1, 445 in Year 2, 448 in Year 3, and 384 in Year 4 responded to the question about eligibility for an automatic 1 percent bonus. Appendix D provides details on the survey questions that asked teachers to report their bonus eligibility.

Figure reads: In Year 1, 49 percent of teachers in treatment schools reported being eligible for a pay-for-performance bonus, and 58 percent of control teachers reported being eligible for an automatic 1 percent bonus.

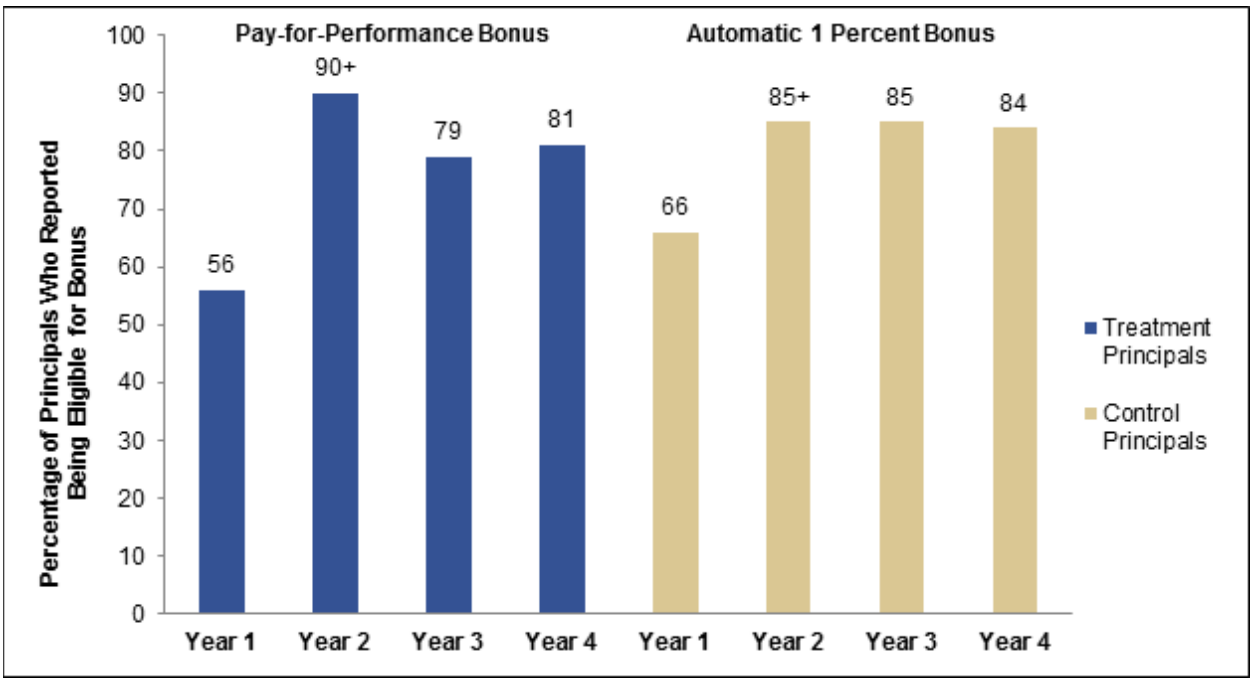
+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Many teachers and some principals in treatment schools did not understand they were eligible to earn a performance bonus. After the first year of TIF implementation, about 40 percent of treatment teachers were still unaware that they could potentially earn a performance bonus (57 to 61 percent of treatment teachers reported being eligible for a bonus in Years 2 through 4, implying that at least 39 percent of treatment teachers in each of these years did not report being eligible for a performance bonus; Figure IV.9). Although understanding of eligibility was better among principals than teachers, about 20 percent of principals in Years 3 and 4 still did not know they were eligible to earn a bonus based on their performance (Figure IV.10).[43,44]

[43] When analyses for Years 1 and 2 were based on Cohorts 1 and 2, similar but generally smaller percentages of teachers and principals reported being eligible for the correct type of bonus (Appendix D, Figures D.14 and D.15).

[44] Some educators thought they were eligible for the wrong bonus. In each year, about 20 percent of control teachers and 15 percent of control principals thought they were eligible for pay-for-performance bonuses (Appendix D, Table D.17).

Figure IV.10. Principals' Bonus Eligibility, as Reported by Principals (Percentages)



Sources: Principal surveys (2012, 2013, 2014, and 2015).

Notes: A total of 63 treatment principals in Year 1, 62 in Year 2, 57 in Year 3, and 60 in Year 4 responded to the question about eligibility for a pay-for-performance bonus. A total of 64 control principals in Year 1, 61 in Year 2, 61 in Year 3, and 62 in Year 4 responded to the question about eligibility for an automatic 1 percent bonus. Appendix D provides details on the survey questions that asked principals to report their bonus eligibility.

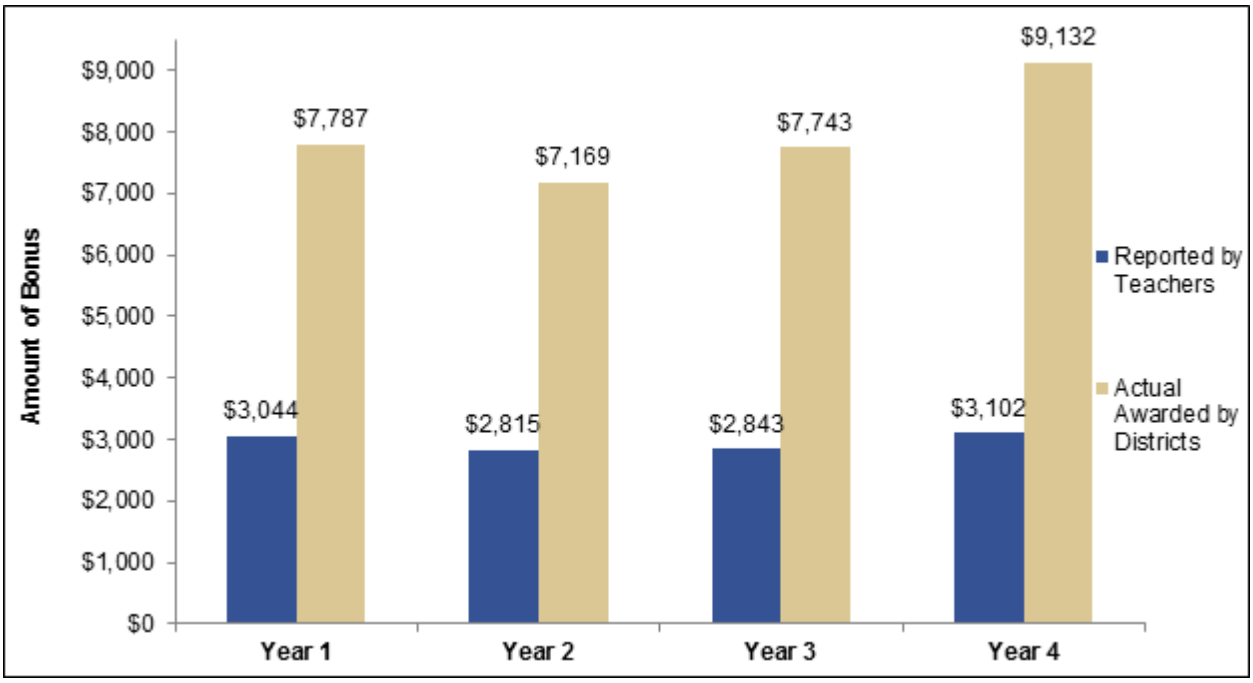
Figure reads: In Year 1, 56 percent of principals in treatment schools reported being eligible for a pay-for-performance bonus, and 66 percent of principals in control schools reported being eligible for an automatic 1 percent bonus.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Educators' Understanding of the Potential Amounts of Pay-for-Performance Bonuses

For performance bonuses to provide an incentive for teachers to change their behaviors, teachers not only must understand they are eligible for a bonus, but also must believe that the bonus they could earn is large enough to warrant changing their teaching practices or effort. Figure IV.11 shows, on average, the maximum performance bonus that teachers believed was available and the actual maximum performance bonus that districts awarded to teachers. Teachers' expectations in Year 1 would have been shaped primarily by how well districts communicated the design of the pay-for-performance component to their teachers. In Years 2, 3, and 4, however, teachers' expectations could also have been influenced by the actual bonuses awarded after Years 1, 2, and 3.

Figure IV.11. Reported and Actual Maximum Pay-for-Performance Bonus for Teachers in Treatment Schools



Sources: Teacher surveys (2012, 2013, 2014, and 2015) and educator administrative data.

Notes: Teachers' reports are based on data for teachers in tested grades and subjects, with each school receiving an equal weight. Districts' payouts are based on data for all teachers, with each district receiving an equal weight. Appendix D, Figure D.16 shows that our results are similar if districts are weighted by the number of schools when calculating districts' payouts.

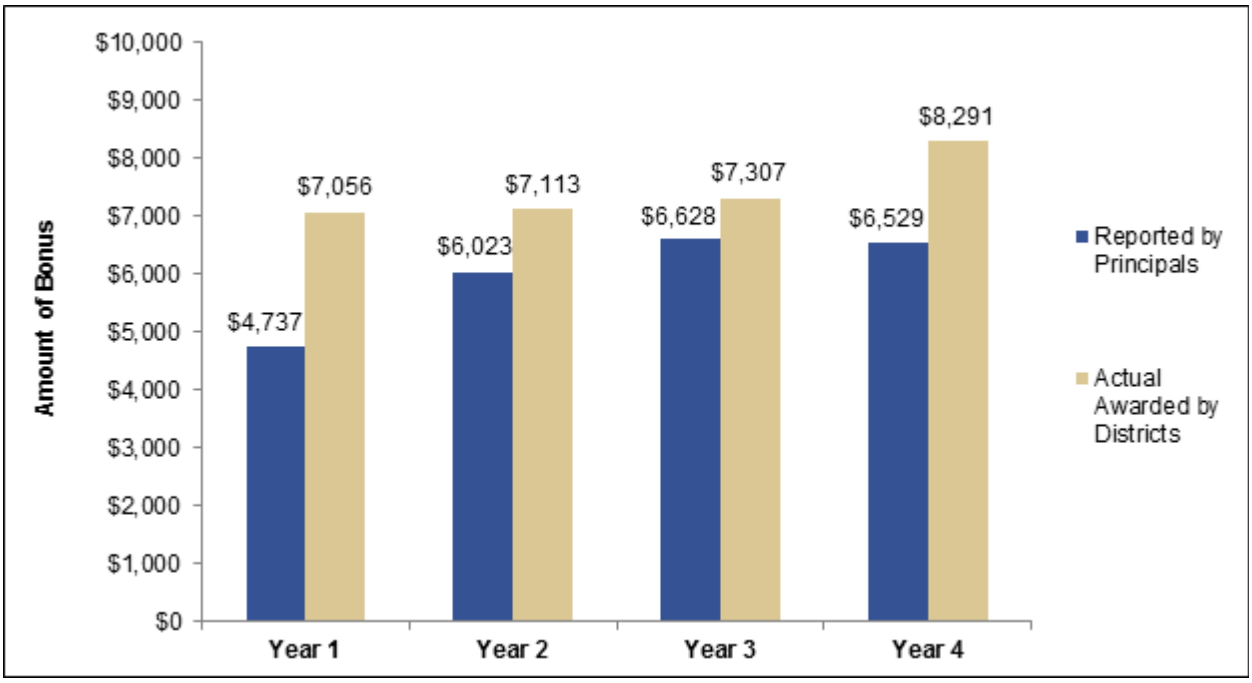
A total of 194 treatment teachers in tested grades and subjects responded to this survey question in Year 1, 215 in Year 2, 215 in Year 3, and 201 in Year 4. The maximum bonus amount was set to zero for all respondents who indicated they were ineligible for a bonus. For teachers who reported being eligible for the bonus but left the amount missing, bonus amounts were imputed through multiple imputation methods. This led to 24 additional responses for treatment teachers in Year 1, 14 additional responses in Year 2, 14 additional responses in Year 3, and 9 additional responses in Year 4. See Appendix B for additional discussion on the imputation methods. Appendix D, Table D.18 shows that our results are similar if we do not impute the missing bonus amounts.

Figure reads: In Year 1, on average, the actual maximum pay-for-performance bonus that evaluation districts awarded to teachers was $7,787, and the maximum pay-for-performance bonus teachers reported they could earn was $3,044.

Teachers underestimated how much they could earn for a performance bonus. In each year, teachers in treatment schools believed that the maximum bonus they could earn was no more than 40 percent of the actual maximum bonus districts awarded (Figure IV.11). For example, in Year 4, teachers in treatment schools, on average, reported that the maximum pay-for-performance bonus that teachers in their teaching position could receive was $3,102, whereas the actual maximum bonus awarded by districts was $9,132. If a teacher did not believe she was eligible for a performance bonus, we assumed that the teacher believed the maximum bonus she could earn was $0. Therefore, the maximum bonus reported by teachers, on average, might have been lower than the maximum reported by the district because of teachers' misunderstanding of their eligibility. However, even the teachers who believed they were eligible for a performance bonus underestimated the potential amount, reporting, on average, a maximum performance bonus (for example, $3,800 in Year 4) that was no more than half the size of the actual maximum bonus that districts awarded (results not shown).

Principals also underestimated the potential amount of performance bonuses they could receive, but their beliefs better aligned with actual bonus payouts than did teachers' beliefs. Across years, the maximum pay-for-performance bonus that principals reported they could receive ranged from 67 to 91 percent of the actual maximum bonus that districts awarded. For example, in Year 4, principals in treatment schools, on average, reported that the maximum bonus they could receive was $6,529—nearly 80 percent of the actual maximum bonus ($8,291) that districts awarded to principals (Figure IV.12). Principals who correctly reported their eligibility for a performance bonus believed the maximum bonus they could receive was about $7,500 in Year 4 (results not shown).[45]

Figure IV.12. Reported and Actual Maximum Pay-for-Performance Bonus for Principals in Treatment Schools



Sources: Principal surveys (2012, 2013, 2014, and 2015) and educator administrative data.

Notes: Principals' reported values are calculated giving each school an equal weight. Actual payouts are calculated giving each district an equal weight. When districts are weighted by the number of schools, actual maximum performance bonus amounts for principals are higher ($8,272 in Year 1, $8,309 in Year 2, $8,408 in Year 3, and $9,611 in Year 4), implying a somewhat wider gap between principals' reported maximum bonus amounts and the actual amounts (Appendix D, Figure D.17).

A total of 55 treatment principals responded to this survey question in Year 1, 60 in Year 2, 57 in Year 3, and 58 in Year 4. The maximum bonus amount was set to zero for all respondents who indicated they were ineligible for a bonus. For educators who reported being eligible for the bonus but left the amount missing, bonus amounts were imputed through multiple imputation methods. This led to 8 additional responses for treatment principals in Year 1, 2 additional responses in Year 2, 0 additional responses in Year 3, and 2 additional responses in Year 4. See Appendix B for additional discussion on the imputation methods. Appendix D, Table D.18 shows that our results are similar if we do not impute the missing bonus amounts.

Figure reads: In Year 1, on average, the actual maximum pay-for-performance bonus that evaluation districts awarded to principals was $7,056, and the maximum pay-for-performance bonus principals reported they could earn was $4,737.

[45] Findings for teachers and principals were similar when analyses for Years 1, 2, and 3 were based on Cohorts 1 and 2 (Appendix D, Figures D.18 and D.19).

One potentially important way in which teachers might learn about their eligibility for a bonus and details of the bonus program is by finding out that they received a bonus. Therefore, we explored whether teachers' misunderstanding of their bonus eligibility and potential bonus amounts could be due, in part, to being unaware of having received a bonus from the prior year.

Many treatment teachers were not aware they received a performance bonus. For example, 43 percent of the teachers who received a bonus based on their Year 3 performance reported that they did not receive one (Table IV.9). Almost all teachers (92 percent) who did not receive a bonus correctly reported not getting a performance bonus. As we show in the next section, understanding of bonus eligibility and potential bonus amounts was notably worse among teachers who did not report receiving a bonus from the prior year than among teachers who did.

Table IV.9. Actual and Reported Receipt of Pay-for-Performance Bonus from Year 3 for Teachers in Treatment Schools in Year 4

	Actual Bonus Receipt in Year 3	
Belief About Bonus Receipt in Year 3	Teachers Who Received a Bonus	Teachers Who Did Not Receive a Bonus
Percentage Who Reported Receiving a Bonus	57	8
Percentage Who Reported Not Receiving a Bonus	43	92
Total Percentage	**100**	**100**

Sources: Teacher survey (2015) and educator administrative data.

Notes: A total of 384 treatment teachers with bonus information from districts' administrative data responded to the survey question about whether they received a bonus based on their performance last year. Of those, 253 teachers received an actual Year 3 bonus and 131 did not.

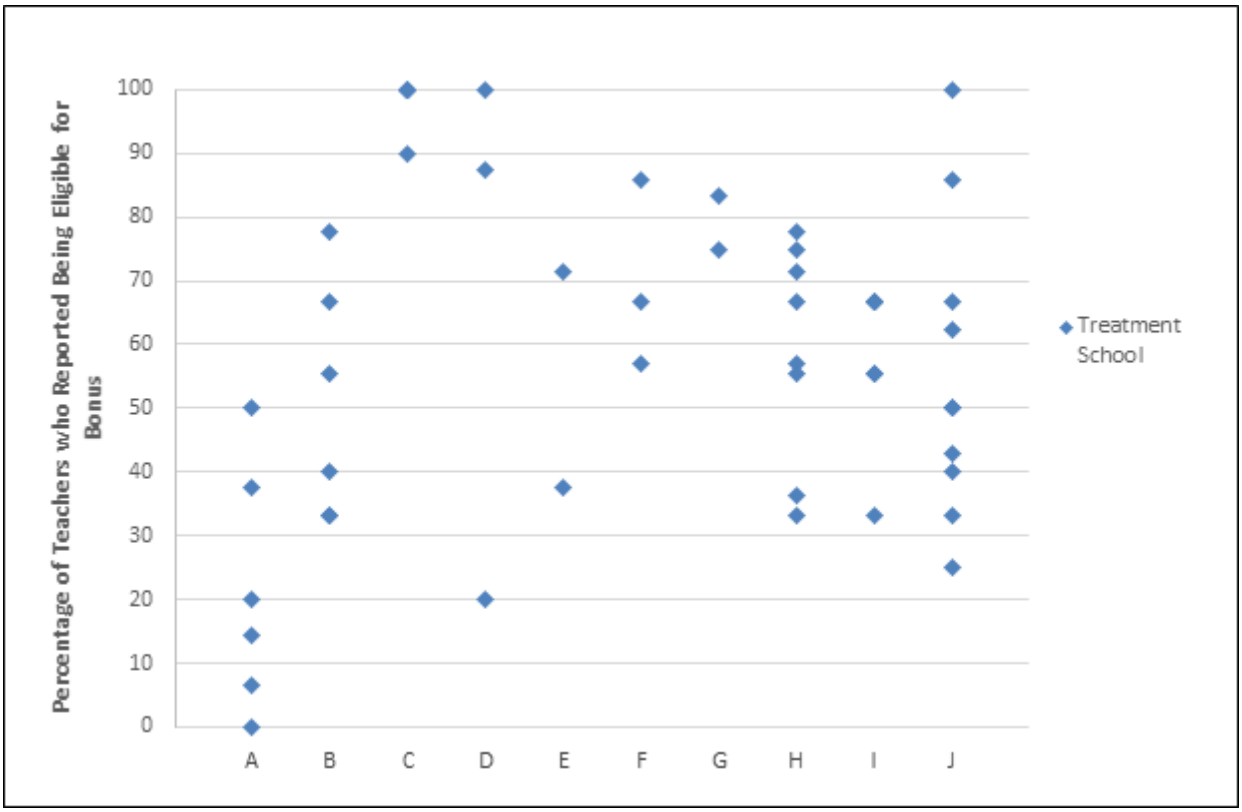
Examining Why Teacher Understanding Varies

Because teachers' understanding of their eligibility for a bonus and the potential size of the bonus can shape their behavior, we explored how teacher understanding varied across districts, across schools within the same district, and within the same school. If teacher understanding did not vary within a district, we might hypothesize that districtwide factors, such as including bonuses in teachers' regular paychecks or in separate bonus paychecks, were important in determining teachers' understanding. If teacher understanding varied within a district, but not within a school, we might conclude that school factors, such as whether the principal correctly understood and conveyed teachers' eligibility, influenced teachers' understanding. If teacher understanding varied within a school, variation in teachers' understanding could be explained by differences in teachers' characteristics, such as whether the teacher had ever received a bonus or attended TIF-related professional development sessions.

Most of the differences in teachers' understanding occurred among teachers in the same school. For example, Figure IV.13 displays the variation in treatment teachers' understanding of eligibility for pay-for-performance bonuses for each evaluation district in Year 4. Each diamond on the figure represents a treatment school and shows the percentage of teachers in that school reporting they were eligible for a performance bonus. Because all teachers in treatment schools were eligible for pay-for-performance bonuses, the ideal scenario would have been for all teachers to *report* being eligible, in which case all diamonds would have been at the top of the figure. In fact, as the figure shows, teacher understanding varied across districts, across schools within districts, and across teachers within schools. In many treatment schools, about half of the teachers reported being

eligible for pay-for-performance bonuses, and half reported they were not eligible. Statistically, we found that in each year at least 85 percent of the variation in treatment teachers' understanding of their eligibility for a pay-for-performance bonus occurred among teachers in the same school (Appendix D, Table D.19). Similarly, most of the variation (at least 70 percent) in teachers' understanding of the maximum bonus they could earn occurred among teachers in the same school (Appendix D, Table D.19).

Figure IV.13. Percentage of Teachers in Treatment Schools Who Reported They Were Eligible for a Pay-for-Performance Bonus, by District, Year 4



Source: Teacher survey, 2015 (N = 355 teachers in 51 treatment schools).

Notes: Schools within a district that have the same percentage of teachers who reported being eligible for a performance bonus are represented by a single diamond. Schools with fewer than three teachers who reported their bonus eligibility are not shown.

Figure reads: In Year 4, in the six schools in District A that offered teachers pay-for-performance bonuses, the percentages of teachers who reported being eligible for a bonus were 0, 7, 14, 20, 38, and 50.

We examined a variety of district, program, teacher, and school characteristics to determine whether differences in these factors could help explain differences in treatment teachers' understanding of their eligibility for a performance bonus and its potential amount. Because the maximum bonus amount awarded varied by district (ranging from $2,025 to $21,500; Figure IV.5), we expressed teachers' reports about the maximum bonus as a percentage of their district's actual maximum bonus awarded (with zero percent for teachers who did not believe they were eligible for a pay-for-performance bonus). The district and program characteristics we examined were whether the district (1) used district (rather than school) staff to communicate the TIF program to teachers, (2) expected at least 75 percent of teachers to attend TIF-required professional development, (3) used classroom achievement growth to determine performance bonuses, (4) awarded an average pay-for-performance bonus that was at least 4.5 percent of the average teacher salary, (5) paid

pay-for-performance bonuses through a separate bonus check (rather than teachers' regular paycheck), (6) told all treatment teachers the bonus amount they earned in the prior year (including $0 for those who did not receive one), and (7) held individual meetings with each teacher to discuss his or her bonus amount. Teacher characteristics we examined were whether the teacher (1) was a returning teacher to the school, (2) taught a tested grade/subject, (3) received or reported receiving a performance bonus for Year 3, (4) participated in TIF-related professional development, and (5) was or had a mentor teacher. We also examined one school factor—principals' understanding of teachers' eligibility.

Few district, program, teacher, or school characteristics consistently explained differences in teachers' understanding of their eligibility for a performance bonus or of how much they could earn. Treatment teachers' understanding of their eligibility for performance bonuses and of the maximum bonus they could earn were generally not associated with district, key program, teacher, or school characteristics (Appendix D, Tables D.20 and D.21). The few characteristics with a significant association in one year did not have significant associations in other years. For example, in Year 4, teachers in districts that used district staff to communicate the TIF program to teachers had worse understanding of their eligibility for performance bonuses than teachers in districts that used school staff for communication (49 versus 64 percent; Appendix D, Table D.20) but this relationship was not significant in either Years 2 or 3 (Chiang et al. 2015; Wellington et al. 2016). Similarly, the few characteristics that were associated with teachers' understanding of eligibility in Years 2 or 3 (for example, working in a district that offered a high average bonus or that informed both recipients and nonrecipients of their bonus amounts, being a mentor, or being a mentor as part of TIF) were not associated in other years (Chiang et al. 2015; Wellington et al. 2016).

Teachers who received or reported receiving a bonus based on the prior year's performance consistently had a better understanding of their eligibility for a performance bonus or of the maximum bonus they could earn. Because many treatment teachers were unaware that they received a performance bonus, we examined how teachers' understanding varied with both their actual and reported receipt of a bonus based on their prior year's performance. Both receiving an actual performance bonus and reporting to have received one were associated with improved understanding of eligibility for pay-for-performance bonuses and their potential size in Years 3 and 4 (Appendix D, Table D.21; Wellington et al. 2016). Teachers' belief of bonus receipt was, however, more strongly associated with understanding than actual bonus receipt. For example, about 90 percent of the teachers who believed they received a bonus the previous year correctly reported being eligible for a performance bonus, compared with about 70 percent of the teachers who actually received a bonus.

Educators' Understanding of and Experiences with Other TIF Components

Educators also reported their understanding of and experiences with the remaining two required components: additional pay opportunities and professional development to help them understand and improve their ratings on TIF performance measures. Educators' understanding of additional pay opportunities can shed light on how visible these opportunities were in the study schools. Educators' reported participation in TIF-related professional development can suggest the extent to which districts allocated resources and attention to this component. It might also illuminate whether educators received enough guidance to know how to improve their performance. As with implementing the performance measures used in TIF, evaluation districts were expected to implement these TIF components identically in treatment and control schools.

Table IV.10. Eligibility for Additional Pay Opportunities, as Reported by Teachers and Principals (Percentages)

	Year 1		Year 2		Year 3		Year 4	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Teachers								
Teachers Could Receive Additional Pay for Taking on Extra Roles or Responsibilities	57	56	89+	88+	89*	82+	87*	78
Roles or Responsibilities								
Mentor teacher	45	40	72+	74+	62*+	55+	60*	54
Master or lead teacher	40	39	55+	57+	62*+	54	61*	51
Department chair or head	18	20	22*	29+	24	28	26	26
Lead curriculum specialist	26	25	36+	38+	35	34	40	34
Schoolwide committee or task force member	11	11	18+	21+	18	22	19	18
Leadership team member	35	29	23+	27	20	18+	24	20
Additional Factors								
Teach in a hard-to-staff or high-need school	25	23	30+	31+	30	29	30*	26
Attend professional development activities or enroll in graduate level courses	30	28	25	24	30	27	29	26
Number of Teachers—Range[a]	**240–376**	**234–393**	**431–435**	**438–444**	**398–417**	**425–447**	**357–389**	**370–384**
Principals								
Principals Could Receive Additional Pay for Taking on Extra Roles or Responsibilities	<5[b]*	14	22+	16	23	21	18	21
Number of Principals	**63**	**63**	**63**	**61**	**58**	**61**	**59**	**61**

Sources: Teacher and principal surveys (2012, 2013, 2014, and 2015).

[a]Sample sizes are presented as a range based on the data available for each row in the table.

[b]To reduce the risk of disclosing information on an individual, the exact percentage is not shown.

*Difference between treatment and control group is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Most teachers were aware of the opportunity to take on additional roles and responsibilities. After Year 1, more than 75 percent of teachers reported that opportunities for earning extra pay for additional roles and responsibilities were available at their school. Although there was a significant improvement in teachers' understanding of this TIF component from Year 1 to Year 2 for both treatment and control teachers, there was no improvement in subsequent years. In fact, fewer teachers in control schools in Year 3 (82 percent) reported they could earn extra pay for taking on additional roles or responsibilities than in Year 2 (88 percent), and these percentages were similar in Year 4. As a result, teachers' awareness of additional pay opportunities in Years 3 and 4 were higher in treatment schools than in control schools (Table IV.10).

Principals were less likely than teachers to report being offered additional pay opportunities. Fewer than 25 percent of principals reported that opportunities for earning extra pay for additional roles and responsibilities were available for them in each year (Table IV.10). As discussed earlier, none of the evaluation districts reported offering extra pay for principals to accept additional responsibilities in Year 4, but 10 percent of the districts reported offering principals extra pay for attending professional development or enrolling in graduate courses (not shown). Some principals might have interpreted their eligibility for earning extra pay for these other factors as extra pay for additional roles or responsibilities.

Most teachers reported receiving the professional development required under the TIF grant but indicated they received only a few hours of it. In each year, more than half of teachers reported that they received or expected to receive professional development focused on understanding performance measures used in TIF (ranging from 77 percent of treatment teachers in Year 1 to 53 percent of treatment teachers in Year 4; Appendix D, Table D.22). Most teachers (50 to 60 percent) also reported that they received or expected to receive feedback based on their performance ratings. Of those who expected to receive any professional development on these two topics, the expected amount of time on each topic per year ranged from two to six hours (Appendix D, Table D.23).

Summary

According to the theory of change presented in Chapter I, some key steps had to occur in the implementation of TIF for pay-for-performance to improve educator effectiveness and student achievement. This chapter examined whether and how each of these steps materialized in the evaluation districts' implementation of TIF. Describing how evaluation districts implemented the TIF grant provides useful context for interpreting findings presented later in this report on the program's impacts on educator and student outcomes.

Some findings from this chapter support the possibility that the structure of the performance bonuses that the evaluation districts offered had the potential to motivate educators to improve their effectiveness. For example, performance bonuses for the highest-performing teachers were about four times as large as the average bonuses, providing a clear monetary incentive for teachers to earn high ratings on the effectiveness measures. Teachers who could earn bonuses based on the achievement growth of the students they taught could earn the largest bonuses. In addition, teachers' performance ratings were consistent from one year to the next, which might have provided teachers with confidence in how they were evaluated.

On the other hand, several other factors might have dampened the potential for pay-for-performance to improve educator effectiveness. For example, many teachers in treatment schools believed they were ineligible for a performance bonus or underestimated how much they could earn from these bonuses. In addition, most educators received a bonus, and the average bonuses were not large. Therefore, even if educators had perfect understanding of their eligibility and the amount they could earn, certain aspects of the bonus structure might not have provided educators with an incentive to change their behavior.

If educators were motivated to change their practices, they still might have found it difficult to determine how to do so. Throughout the grant period, although nearly all evaluation districts provided general professional development to help teachers understand the performance measures on which they were being evaluated, fewer districts provided targeted professional development to help teachers improve their individual ratings on those measures. Furthermore, even among teachers who expected to receive professional development on these topics, the expected amount of time on each topic was no more than six hours over the school year.

V. IMPACTS OF PAY-FOR-PERFORMANCE ON EDUCATORS' ATTITUDES AND BEHAVIORS

The ways in which pay-for-performance programs affect educators' attitudes (such as job satisfaction) and behaviors (such as principals' approaches to recruiting teachers) can shape how pay-for-performance affects student outcomes. As the theory of change in Chapter I shows, pay-for-performance bonuses can increase student achievement by motivating educators to improve their practices and by attracting and retaining more effective educators. However, if the presence of pay-for-performance discourages useful collaboration, lowers morale, or makes a school less appealing to effective educators, it could have a negative effect on the work environment and on student achievement.

Key Findings on the Impacts of Pay-for-Performance on Educators' Attitudes and Behaviors

- **Most teachers were satisfied with their jobs and the TIF program.**
- **Although initially less satisfied with their jobs and TIF, teachers in treatment schools were as satisfied as, and sometimes more satisfied than, teachers in control schools by the third year of implementation.**
- **By the third year of TIF implementation, principals in treatment schools were more likely to use components of TIF to recruit teachers than principals in control schools.**

In this chapter, we use data from teacher and principal surveys to estimate the impacts of pay-for-performance on educators' self-reported attitudes and behaviors. Educators in treatment schools were eligible for pay-for-performance bonuses, but educators in control schools were not. Because both treatment and control schools offered all the other required components of the TIF program, any differences in responses between educators in treatment and control schools can be attributed to the impacts of pay-for-performance.[46]

The analysis in this chapter is based on 10 evaluation districts that completed four years of TIF implementation. We refer to these four years (2011–2012, 2012–2013, 2013–2014, and 2014-2015) as Years 1, 2, 3, and 4, respectively.[47] Examining impacts over four years provided an opportunity to see whether these impacts evolved over time. As discussed in Chapter IV, educators' understanding

[46] As discussed in Chapter IV, some educators in the study schools misunderstood their eligibility for pay-for-performance or the potential amounts they could earn. The impacts reported in this chapter reflect the impact of pay-for-performance given educators' actual beliefs. This study was not designed to assess the impacts of pay-for-performance bonuses in the scenario in which all educators correctly understood their eligibility and the amount they could earn in a bonus. In addition, for all of the outcomes reported in this chapter, the impact findings could reflect pay-for-performance having changed an individual educator's attitudes and behaviors or having enabled schools to attract or retain more educators with particular attitudes and behaviors.

[47] As discussed in Chapter II, evaluation districts were classified into two cohorts—Cohort 1 and Cohort 2—according to the year in which we randomly assigned their schools to a treatment group or a control group. The 10 districts examined in this chapter, whose schools were randomly assigned in spring and summer 2011, were classified as Cohort 1. Three additional districts, whose schools were randomly assigned in spring and summer 2012, were classified as Cohort 2. In Appendix E, Tables E.1 through E.4, we present three years of impacts on educators' satisfaction and attitudes for Cohorts 1 and 2 combined.

of key TIF components increased in the first few years of TIF implementation, which could have led to changes in their attitudes and behaviors. Moreover, whereas educators responded to surveys in Year 1 before they had received any performance bonuses, we administered surveys in Years 2, 3, and 4 after districts had awarded one, two, and three years of bonuses, respectively. Therefore, these data provide an opportunity to examine whether educators' initial impressions of performance-based compensation changed over the course of repeated rounds of bonuses.

Impact of Pay-for-Performance on Educators' Attitudes

This section presents the impacts of pay-for-performance on educators' satisfaction with and attitudes toward their jobs and the TIF program.

Job Satisfaction

Most teachers and principals in treatment and control schools were satisfied with their jobs. The percentage of teachers who were somewhat or very satisfied with their job overall ranged from 68 to 77 percent across years and across the treatment and control groups (Table V.1).[48] On each specific aspect of their professional opportunities, evaluation system, and school environment, at least half of teachers reported being satisfied. Teachers tended to be most satisfied with their opportunities to enhance their skills, feedback on their performance, the quality of interactions with colleagues, and colleagues' efforts. They tended to be least satisfied with school morale and their opportunities to earn extra pay, reporting satisfaction rates ranging from 50 to 63 percent. Among principals, the percentage who reported being satisfied with each aspect of their professional opportunities, evaluation system, and school environment was consistently above 60 percent (Table V.2). The only exception was that fewer than 60 percent of control principals were satisfied with their opportunities to earn extra pay in Years 3 and 4.

Although initially less satisfied with their jobs, teachers in treatment schools were as satisfied as, and sometimes more satisfied than, teachers in control schools by the third year of implementation. In Years 1 and 2, treatment teachers tended to report being less satisfied than control teachers. For example, in Year 2, treatment teachers reported being less satisfied than control teachers with the use of student achievement scores to assess their performance (61 versus 69 percent) and recognition of their accomplishments (61 versus 66 percent; Table V.1). Opportunities to earn extra pay were the only aspect of their job on which treatment teachers were more satisfied than control teachers in Year 2. However, in Year 3, treatment teachers were no longer less satisfied than control teachers on any dimension and, in fact, reported being more satisfied on some dimensions—school morale, the quality of interactions with colleagues, and opportunities to earn extra pay. In Year 4, treatment teachers continued to report that they were at least as satisfied as control teachers on each aspect of their job. Larger percentages of treatment teachers than control teachers were satisfied with their opportunities to earn extra pay (58 versus 51 percent) and with the feedback on their performance (87 versus 77 percent).

[48] Teachers' satisfaction rates in the study schools appeared to resemble the satisfaction rates found in high-need schools nationwide. For example, in a nationally representative sample of teachers who were administered the 2011–2012 School and Staffing Survey (SASS), 73 percent of teachers who taught in high-need schools (in which at least half of students were eligible for free or reduced-price lunch) agreed that they were generally satisfied with being a teacher at their school. The satisfaction questions in the SASS were not identical to those in this study's teacher survey, so differences in reported satisfaction rates on the two sets of questions could be due, in part, to differences in wording.

Table V.1. Teachers' Satisfaction with Professional Opportunities, Evaluation System, and School Environment (Percentages Who Are Somewhat or Very Satisfied)

	Year 1			Year 2			Year 3			Year 4		
Satisfaction Dimension	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
Opportunities for Pay and Development												
Opportunities for professional advancement	67	75	-9*	73	74	-2	74	74	0	71	77	-6
Opportunities to enhance skills	76	78	-2	81	81	0	79	81	-3	81	80	0
Opportunities to earn extra pay	62	57	5	63	54	10*	62	50	12*	58	51	7*
Key Features of Evaluation System												
Use of classroom observations to assess skills	69	76	-7	—	—	—	—	—	—	74	75	-2
Use of student achievement scores to assess performance	66	67	0	61	69	-8*	66	66	1	59	60	-1
Feedback on my performance	—	—	—	76	80	-4	79	81	-2	87+	77	10*
School Environment												
Recognition of accomplishments	55	61	-7*	61+	66	-6*	69	62	7	67	63	4
Quality of interaction with colleagues	75	81	-6*	82+	82	0	84	79	5*	80	80	0
Colleagues' efforts	84	85	-1	84	83	0	84	83	1	82	82	0
School morale	50	55	-5	60+	59	1	63	53	10*	62	57	5
Job Satisfaction												
Overall job satisfaction	68	73	-5	74	74	0	76	75	1	77	70	7
Number of Teachers—Range[a]	**379-383**	**392-399**		**439-443**	**446-449**		**422-426**	**455-459**		**390-393**	**392-396**	

Sources: Teacher surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

— is not available.

*Impact is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Table V.2. Principals' Satisfaction with Professional Opportunities, Evaluation System, and School Environment (Percentages Who Are Somewhat or Very Satisfied)

	Year 1			Year 2			Year 3			Year 4		
Satisfaction Dimension	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
Opportunities for Pay and Development												
Opportunities for professional advancement	—	—	—	86	89	-3	94	90	3	>95[a]	>80[a]	13*
Opportunities to enhance skills	92	95	-3	87	85	2	96	89	7	93	84	10
Opportunities to earn extra pay	72	66	6	64	64	0	84+	54	30*	74	48	27*
Key Features of Evaluation System												
Use of observations to assess skills	—	—	—	61	85	-25*	96+	85	10	94	80	13
Use of student achievement scores to assess performance	—	—	—	65	82	-17*	86	81	5	80	75	4
Feedback on my performance	83	87	-4	66	80	-14	83	82	1	87	82	5
School Environment												
Recognition of accomplishments	77	82	-5	64	75	-12	81+	69	11	87	67	20*
Quality of interaction with colleagues	>85[a]	>94[a]	-7	86	90	-4	95	89	7	93	87	6
Colleagues' efforts	>90[a]	>94[a]	-5	89	85+	4	94	94+	1	93	90	2
School morale	71	87	-16*	74	82	-8	90+	84	6	87	87	0
Number of Principals—Range[b]	**62-63**	**59-61**		**62-63**	**60-61**		**57-58**	**61-62**		**61**	**61**	

Sources: Principal surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]To reduce the risk of disclosing information on an individual, the exact percentage is not shown.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

— is not available.

*Impact is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Pay-for-performance could affect some groups of teachers differently, so we examined impacts separately within subgroups. We separated teachers based on (1) grade–subject assignments (those in tested grades and subjects with annual accountability tests and those in nontested grades and subjects); and (2) years of teaching experience (fewer than 5, 5 to 15, or more than 15). These groupings stemmed from several hypotheses. Teachers in tested grades and subjects could have felt more pressure from the TIF program than did teachers in nontested grades and subjects because they were evaluated on their own students' achievement growth or because the school's ability to receive a school-based award depended in part on their students' achievement. On the other hand, as shown in Chapter IV, teachers who were evaluated on their own students' achievement growth could earn higher bonuses than other teachers in the same districts. Similarly, teachers in nontested grades and subjects may have felt they had less control over their rated performance and bonuses. Compared to bonuses for teachers in tested grades and subjects, bonuses for those in nontested grades and subjects were more heavily determined by school achievement growth (Chapter IV, Figure IV.6). This could have led these teachers to be less supportive of pay-for-performance or their TIF program. In addition, we separated teachers by their level of experience to assess whether those who had taught longer under a different evaluation and compensation system might have been less receptive to the new system.

The results of the subgroup analyses should be interpreted carefully. The impact within each subgroup, which is based purely on the study's random assignment design, captures the effect of pay-for-performance on outcomes within that subgroup. However, a difference in impacts between two subgroups simply indicates whether impacts were larger or smaller in one subgroup than in another. It does not necessarily indicate whether the characteristic that distinguishes the two subgroups *caused* the difference in impacts, because characteristics other than the one being considered also might have differed between these subgroups. Nevertheless, because the subgroup analyses can identify the groups that respond most to pay-for-performance, the analysis can inform best practices for designing or targeting future pay-for-performance programs.

Veteran teachers initially responded less favorably to pay-for-performance than less experienced teachers, but by Year 3 teachers' responses to pay-for-performance had no consistent relationship with their level of experience. In Years 1 and 2, the negative impacts of pay-for-performance on teachers' satisfaction were most pronounced among veteran teachers—those with more than 15 years of experience (Max et al. 2014; Chiang et al. 2015). However, from Year 3 onward, we found no consistent pattern of veteran teachers responding more or less favorably to pay-for-performance compared to less experienced teachers (Wellington et al. 2016; Appendix E, Table E.5). For example, in Year 4, the impact on teachers' satisfaction with their opportunities to earn extra pay tended to be the least positive among veteran teachers. However, the impact on teachers' satisfaction with the feedback on their performance tended to be the most positive for veteran teachers. Across all four years, for the dimensions on which pay-for-performance changed teachers' satisfaction, the impacts were similar for teachers in tested and nontested grades and subjects.

Beginning in the third year of TIF implementation, principals in treatment schools tended to be more satisfied with their jobs than principals in control schools. In Years 1 and 2, fewer principals in treatment schools than control schools were satisfied with their professional opportunities, key features of their evaluation system, and their school environment, although most of those differences were not statistically significant (Table V.2). By Year 3, this overall pattern reversed, and more treatment principals were satisfied on these dimensions than control principals. These differences were again not significant, except that significantly more principals in treatment schools than control schools reported being satisfied with their opportunities to earn extra pay (84 versus 54 percent). The pattern of greater satisfaction among treatment principals continued and

became stronger in Year 4. In Year 4, more treatment principals than control principals reported being satisfied with their opportunities for professional advancement (greater than 95 percent versus greater than 80 percent), opportunities to earn extra pay (74 versus 48 percent), and recognition of accomplishments (87 versus 67 percent).

Educators' Attitudes Toward TIF

Most teachers and principals had favorable attitudes toward TIF. In every year, at least two-thirds of teachers agreed or strongly agreed with the statement that they were glad to participate in the TIF program (Table V.3). Likewise, at least half of teachers felt that TIF was fair. Among principals, about half or more felt that TIF contributed to greater collegiality and professionalism among the staff at their school (Table V.4).

Teachers in treatment schools initially had less favorable attitudes toward TIF than teachers in control schools, but by the third year they had similar attitudes toward most aspects of the program. Treatment teachers generally felt less positively about TIF than control teachers in the first two years of implementation, particularly in the second year. For example, in Year 2, treatment teachers were more likely than control teachers to report that TIF reduced their freedom to teach (39 versus 30 percent) and harmed collaboration (29 versus 21 percent), and less likely to believe that test scores measured what students learned (34 versus 41 percent; Table V.3). However, beginning in Year 3, these differences were smaller and no longer statistically significant. In fact, there were a few exceptions in which treatment teachers in Years 3 or 4 felt more favorably toward TIF than control teachers (for example, 40 versus 33 percent in Year 3 believed that TIF increased their job satisfaction; 59 versus 51 percent in Year 4 reported that TIF caused teachers to work more effectively). The only consistent difference in attitudes, which we found in all four years, was that treatment teachers were more likely than control teachers to feel increased pressure to perform because of TIF.

In the final two years, the impacts of pay-for-performance on teachers' attitudes toward TIF were not consistently related to the teachers' level of experience. For example, in Year 3, the impact of pay-for-performance on teachers' reporting that TIF increased their job satisfaction was the most positive for veteran teachers—those with more than 15 years of experience (Wellington et al. 2016). On the other hand, in Year 4, the impact on teachers' reporting that TIF enhanced teacher effectiveness was the *least* positive for veteran teachers (Appendix E, Table E.7). These inconsistent patterns in the final two years of implementation contrast with those from the first two years, when there was a more consistent pattern that veteran teachers' attitudes toward TIF responded most negatively to pay-for-performance (Max et al. 2014; Chiang et al. 2015).

For the aspects of TIF on which pay-for-performance changed teachers' attitudes, the impacts were similar for teachers in tested and nontested grades and subjects throughout the four years of implementation (Max et al. 2014; Chiang et al. 2015; Wellington et al. 2016; Appendix E, Table E.7). The only exception was that the positive impact in Year 4 on teachers' belief that TIF enhanced teacher effectiveness occurred only in nontested grades and subjects.

Principals in treatment and control schools had similar attitudes toward their TIF programs. We asked principals about their attitudes toward several aspects of TIF, such as the clarity with which the program had been communicated, the fairness of the evaluation system, and the program's effects on school staff. In all years and on nearly all aspects of the TIF program, the attitudes of treatment and control principals were similar (Table V.4). In fact, of the 32 comparisons we made between the attitudes of treatment and control principals, we found only one statistically significant difference.

Table V.3. Teachers' Attitudes Toward TIF Program (Percentages Who Agree or Strongly Agree)

	Year 1			Year 2			Year 3			Year 4		
Statement	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
Overall Attitudes Toward TIF												
I am glad that I am participating in the TIF program	66	65	1	66	71+	-5	69	68	1	73	72	0
The TIF program is fair	53	58	-5	54	59	-6	57	60	-3	60	57	3
My job satisfaction has increased due to the TIF program	28	33	-5	38+	38	0	40	33	7*	44	40	4
Attitudes Toward Evaluation and Compensation Approaches Used by TIF												
My principal is a good judge of teacher talent	67	73	-6	74+	74	0	79	77	2	78	76	2
Standardized student test scores in my district measure what students have learned	35	33	2	34	41+	-7*	29	34	-5	34	36	-2
Teachers who do the same job should receive the same pay	57	58	-1	62	66+	-4	68	66	2	65	63	2
The process used to determine how bonuses are determined was adequately explained to me	67	59	8*	66	62	4	71	63	8*	69	67	2
Attitudes Toward Effects of TIF on Teachers' Effort and Practices												
I feel increased pressure to perform due to the TIF program	65	53	11*	66	51	15*	64	53	11*	62	51	11*
The TIF program has harmed the collaborative nature of teaching	23	24	-1	29	21	8*	26	25	1	32	27	5
The TIF program has caused teachers to work more effectively	49	46	3	51	56+	-5	57	51	5	59	51	8*
I have less freedom to teach the way I would like to teach due to the TIF program	34	35	-1	39	30	10*	37	35	2	35	31	5
Number of Teachers—Range[a]	**373-380**	**382-398**		**395-435**	**383-442**		**382-421**	**383-447**		**349-387**	**341-383**	

Sources: Teacher surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Impact is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Table V.4. Principals' Attitudes Toward TIF Program (Percentages Who Agree or Strongly Agree)

	Year 1			Year 2			Year 3			Year 4		
Statement	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
Attitudes Toward Effects of TIF on Teachers' Behavior												
The TIF program contributes to greater collegiality and professionalism among the staff at this school	48	55	-7	55	68+	-13	63	57	5	57	66	-9
Teachers at this school are more comfortable with frequent formal observations of their teaching because of the TIF program	53	63	-10	58	68	-10	71	67	4	64	69	-6
Attitudes Toward Fairness of TIF												
The evaluation system omits important aspects of school administration that should be considered	28	30	-2	55+	48	7	42	47	-4	56	56	-1
This school has less chance of earning a bonus because of the characteristics of our student population	23	20	3	40+	24	15	29	31	-2	35	29	6
Attitudes Toward Stakeholders' Buy-in and Sustainability of TIF												
I played an important role in implementing the TIF program at my school	82	84	-2	86	84	2	91	77	14*	77	87	-10
The TIF program has been clearly communicated to me	85	89	-4	>90[a]	>90[a]	-4	>90[a]	>90[a]	7	91	92	-1
Parents and the school community believe the TIF program is important	40	48	-8	50	43	6	40	53	-13	45	52	-7
The TIF program is likely to continue for the foreseeable future	84	87	-3	71	73	-2	57+	51+	6	36+	41	-5
Number of Principals—Range[b]	**61-64**	**60-64**		**58-62**	**58-60**		**57-58**	**58-61**		**59-61**	**58-62**	

Sources: Principal surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]To reduce the risk of disclosing information on an individual, the exact percentage is not shown.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

*Impact is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Association Between Receiving Bonuses and Teachers' Attitudes

Findings from Years 1 and 2 suggested that pay-for-performance tended to have a negative effect on teachers' attitudes toward their job and the TIF program. However, in subsequent years, teachers in treatment schools reported attitudes toward their job and the TIF program that were either similar to or more favorable than those of control teachers. One possible explanation for this shift in attitudes could be that treatment teachers' attitudes improved because they experienced multiple years of performance bonuses. As noted in Chapter IV, about 70 percent of treatment teachers received performance bonuses in each year.

To explore this possibility, we examined whether treatment teachers' attitudes varied by whether they received a bonus based on the prior year's performance and by whether they believed they received one. As discussed in Chapter IV, teachers' reports of bonus receipt did not always align with actual bonus receipt, so teachers' recollection of receiving a bonus might be just as important, if not more important, in predicting their satisfaction as actually receiving a bonus. This descriptive analysis can suggest whether teachers may have had more favorable attitudes toward their job and TIF if they received (or believed they received) a monetary reward for their performance. However, this analysis does not provide conclusive evidence about the effects of receiving bonuses on teachers' attitudes because teachers who did and did not receive (or did and did not believe that they received) bonuses could have differed on many other characteristics that influenced their attitudes.

On average, bonus recipients and nonrecipients did not differ in their attitudes toward their job or the TIF program. Among treatment teachers in Year 4, we found no significant difference in any measure of job satisfaction between teachers who had and had not been awarded a bonus based on the prior year's performance (Appendix E, Table E.6). We also found no significant differences in Year 2 (Chiang et al. 2015) and Year 3 (Wellington et al. 2016). Likewise, bonus recipients and nonrecipients generally had similar attitudes toward TIF in Year 4 (Table V.5) and in prior years (Chiang et al. 2015; Wellington et al. 2016). On the other hand, teachers who *reported* receiving a bonus, in general, tended to report more favorable attitudes toward their TIF program and their job than teachers who did not, although those differences were generally not statistically significant (Table V.5 and Appendix E, Table E.6)—similar to Year 3 findings (Wellington et al. 2016).

Impact of Pay-for-Performance on Principals' Recruitment Efforts

Principals can shape the effectiveness of their schools' teaching staff, in part, through the ways in which they recruit teachers. Therefore, as shown in the theory of change in Chapter I, one potentially important way in which pay-for-performance could affect student achievement is by triggering changes in principals' approaches to recruiting teachers. In this section, we examine whether pay-for-performance led to changes in principals' strategies for and success at recruiting teachers.[49]

[49] In Appendix E, we report impacts on other principal behaviors that more indirectly affect teachers' motivation and retention, including principals' approaches to assigning teachers to grades and subjects and providing nonmonetary benefits to their teachers (Appendix E, Tables E.9 and E.10). We also examined whether pay-for-performance affected how teachers reported spending their time (Appendix E, Table E.11). We found no consistent evidence that performance bonuses affected these other principal behaviors or teachers' time on school-related activities.

Table V.5. Treatment Teachers' Attitudes Toward TIF Program, by Bonus Receipt and Report of Bonus Receipt, Year 4 (Percentages Who Agree or Strongly Agree)

	Actual Year 3 Bonus Receipt			Report of Year 3 Bonus Receipt		
Statement	Received a Bonus	Did Not Receive a Bonus	Difference	Reported Receiving Bonus	Reported Not Receiving a Bonus	Difference
Overall Attitudes Toward TIF						
I am glad that I am participating in the TIF program	77	72	5	74	69	5
The TIF program is fair	65	56	9	61	56	5
My job satisfaction has increased due to the TIF program	48	50	-2	51	41	10
Attitudes Toward Evaluation and Compensation Approaches Used by TIF						
My principal is a good judge of teacher talent	70	76	-6	72	73	-1
Standardized student test scores in my district measure what students have learned	25	29	-4	39	26	13*
Teachers who do the same job should receive the same pay	55	60	-5	64	64	-1
The process used to determine how bonuses are determined was adequately explained to me	74	68	6	72	66	5
Attitudes Toward Effect of TIF on Teachers' Effort and Practices						
I feel increased pressure to perform due to the TIF program	60	66	-6	61	59	2
The TIF program has harmed the collaborative nature of teaching	27	32	-5	36	33	3
The TIF program has caused teachers to work more effectively	65	59	5	63	55	8
I have less freedom to teach the way I would like to teach due to the TIF program	29	41	-12	36	37	-1
Number of Teachers—Range[a]	**233-247**	**112-136**		**145-150**	**196-229**	

Sources: Teacher survey (2015) and educator administrative data.

Notes: Pay-for-performance bonus receipt information comes from Year 3 educator administrative data. The difference between those who received (or reported receiving) a bonus and those who did not may not equal the difference shown in the table due to rounding.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

Table V.6. Principals' Reports of Incentives Used to Recruit Teachers (Percentages Who Reported They Were Always or Often Used)

	Year 1			Year 2			Year 3			Year 4		
Incentives	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
TIF-Related Incentives												
Opportunities to earn performance-based pay	25	14	11	34	17	17	43	10	33*	43	15	28*
Opportunities for career advancement	26	21	5	29	28	0	35	17	18*	34	15	20*
Opportunities for professional development	62	62	0	68	57	10	77	56	21*	73	56	17
The TIF program	47	29	18*	44	40+	4	55	39	16	50	35	15
Other Job-Related Incentives												
Salary	24	22	2	23	22	1	30	18	12	34	13	21*
The level of teacher involvement in school decision making	48	57	-10	55	52	2	54	56	-2	70	57	13
Collegiality of teaching staff	77	86	-9	81	88	-7	84	74+	10	79	82	-3
The school culture and/or educational philosophy	85	86	-1	81	92	-11	86	79+	7	82	84	-2
The school's reputation	73	72	1	66	77	-11	72	66	6	68	74	-5
The school's location or neighborhood	38	41	-2	30	28+	1	35	46	-11	43	39	4
The level of student achievement at the school	51	51	0	45	44	1	50	37	12	48	47	1
Number of Principals—Range[a]	**60-63**	**62-64**		**60-63**	**60-61**		**55-58**	**59-61**		**59-61**	**60-61**	

Sources: Principal surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Impact is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Table V.7. Principals' Reports of Teaching Vacancies and Hiring Experiences (Averages Unless Otherwise Noted)

	Year 1			Year 2			Year 3			Year 4		
	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
Classrooms with teacher vacancies	3	4	-2*	4	5	-1	5	5	0	5	5	0
Applications school reviewed for positions	27	28	-2	33	33	0	49+	40+	9	59+	58+	2
Applicants school interviewed	10	14	-4*	11	17+	-5*	17+	17	0	21	18	4
Offers school made	3	5	-2*	4	5	-1	6	5	0	6	6	0
Offers that were accepted	3	5	-2*	4	4	-1	5	5	0	4	5	0
Interview ratio (number of applicants interviewed per classroom vacancy)	3	4	0	3	4	-1	4	4	-1	5+	3	1*
Acceptance rate (percentage of offers accepted out of offers made)	76	81	-5	82	84	-3	85	82	3	85	89	-3
Number of Principals—Range[a]	**57-62**	**60-63**		**60-63**	**58-61**		**56-58**	**55-59**		**57-60**	**58-60**	

Sources: Principal surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Impact is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

To understand the possible impact of pay-for-performance on teacher recruitment, we asked principals whether and how they used TIF to recruit teachers to their school. Nearly all principals in the study had input into hiring decisions at their schools (Appendix E, Table E.8), so pay-for-performance had the potential to influence the principals' approaches to recruiting teachers. Although all study principals might use opportunities offered through their TIF program to recruit teachers, principals in treatment schools might recruit teachers differently because TIF offered teachers the possibility of earning higher bonuses in their schools than in control schools. In theory, offering larger bonuses might help principals recruit more higher-performing teachers.

By the third year of TIF implementation, principals in treatment schools were more likely to use components of TIF to recruit teachers than principals in control schools. In the first two years, we found few differences between treatment and control principals in the use of incentives to recruit teachers (Table V.6). However, beginning in Year 3, larger percentages of treatment principals than control principals emphasized key components of TIF when recruiting teachers, and these differences largely persisted into Year 4. For example, when recruiting teachers to teach at their schools in Year 4, treatment principals were more likely to emphasize opportunities for earning performance-based pay (43 versus 15 percent) and career advancement (34 versus 15 percent) than control principals. Larger percentages of treatment principals than control principals also reported emphasizing opportunities for professional development and the TIF program as recruitment incentives, although the differences were not statistically significant. Treatment and control principals reported using other job-related incentives similarly, with one exception. In Year 4, more treatment principals than control principals reported using salary as a recruitment incentive.

Pay-for-performance had little impact on principals' success in filling teacher vacancies. Principals of treatment and control schools reported having similar recruitment experiences in terms of interviews per vacancy and acceptances per offer made (Table V.7). The only exception was that treatment principals in Year 4 reported interviewing a larger number of applicants per vacancy than control principals (5 versus 3). Although, in general, treatment schools did not find it any easier or harder to fill teacher vacancies than control schools, it is still possible that the effectiveness of the teachers who filled those vacancies differed. We will examine this possibility in Chapter VI.

Summary

The ways in which pay-for-performance affects educators' attitudes and behaviors can shape how it affects student outcomes. The goal of pay-for-performance is to increase student achievement by motivating educators to improve their performance and by attracting and retaining more effective teachers. However, if the presence of pay-for-performance discourages useful collaboration, lowers morale, or makes a school less appealing to effective educators, it may not accomplish this goal.

The findings from this chapter suggest that, beginning in the third year and continuing through the fourth year of implementation, the impact of pay-for-performance on educators' satisfaction was unlikely to hinder educators' effectiveness and could even have enhanced it. Most teachers and principals reported being satisfied with key aspects of their job and TIF program. Although findings from the first couple of years of implementation suggested that pay-for-performance caused educators to be less satisfied, by the third year of implementation educators in treatment schools were as satisfied as, and sometimes more satisfied than, educators in control schools with aspects of their job and their TIF program. These findings suggest that educators might initially resist pay-for-performance initiatives, but after a few years of firsthand experience with the program, they might become more accepting of performance bonuses.

THIS PAGE IS INTENTIONALLY BLANK

VI. IMPACTS OF PAY-FOR-PERFORMANCE ON EDUCATOR EFFECTIVENESS AND STUDENT ACHIEVEMENT

A central objective of the TIF grants is to improve student achievement in high-need schools by increasing the effectiveness of the educators working in those schools. Our evaluation was designed to rigorously assess whether the pay-for-performance component of grantees' TIF programs accomplished this goal. In this chapter, we present findings on whether pay-for-performance led to changes in educator effectiveness and student achievement across the four years of TIF implementation.

Key Findings on the Impacts of Pay-for-Performance on Educator Effectiveness and Student Achievement

- **The impacts of pay-for-performance on teachers' and principals' performance ratings based on student achievement growth were mixed, with positive impacts in the first and fourth years of implementation but not in other years.**
- **Pay-for-performance led teachers to earn slightly higher classroom observation ratings by the third year of implementation, but had no impact on the observation ratings of principals.**
- **Pay-for-performance led to slightly higher student achievement in reading and math by the second year of implementation.**
- **The impacts of pay-for-performance on student achievement differed across districts and schools, but few characteristics of the districts or schools explained these differences in impacts.**
- **Pay-for-performance was generally more cost-effective than reducing class size and, at the end of two years, was about as cost-effective as offering incentives for high-performing teachers to transfer to low-performing schools.**

As shown in the theory of change from Chapter I, a main principle of TIF is that increasing educator effectiveness is the key to improving student achievement. Therefore, the first section of this chapter reports the impacts of pay-for-performance on educator effectiveness, as measured by the educators' TIF performance ratings. Those ratings were largely based on measures of student achievement growth in classrooms and schools and on observations of classroom or school practices. Because those ratings determined performance bonus amounts, pay-for-performance was designed to motivate educators to improve their performance on those measures. However, those measures might not capture all aspects of educator performance that matter for student achievement. Therefore, in the second section of this chapter, we directly examine whether pay-for-performance bonuses led to improved student achievement on reading and math assessments. In light of these impacts, the final section of this chapter considers whether pay-for-performance was more or less cost-effective than various alternative policies that districts could implement instead of pay-for-performance.

Our analyses in this chapter compare the outcomes of educators and students in treatment schools with those of educators and students in control schools. Educators in treatment schools were eligible for pay-for-performance bonuses, but educators in control schools were not. Because both treatment and control schools offered all the other required components of the TIF program,

any differences in outcomes between treatment and control schools can be attributed to the impact of pay-for-performance.[50] As discussed in Chapter IV, some educators in the study schools misunderstood their eligibility for pay-for-performance or the potential amounts they could earn. The impacts reported in this chapter reflect the impact of pay-for-performance given educators' beliefs. This study was not designed to assess the impacts of pay-for-performance bonuses had all educators correctly understood their eligibility and the amount they could earn in a bonus. Data for this chapter come from districts' administrative records on educators and students.

The chapter is based on 10 evaluation districts that completed four years of TIF implementation. We refer to these four years (2011–2012, 2012–2013, 2013–2014, and 2014-2015) as Years 1, 2, 3, and 4, respectively.[51] Examining impacts over four years provided an opportunity to see whether these impacts evolved over time. For example, impacts could have grown over those four years for several reasons. Educators' understanding of key TIF components increased over the course of the grant (Chapter IV), and educators could have been increasingly motivated to improve after seeing repeated rounds of performance bonuses awarded. In addition, unlike in the first two years, teachers in treatment schools in the final two years were no longer less satisfied with their jobs or with TIF than teachers in control schools (Chapter V). Improved satisfaction among treatment teachers could, in turn, have led to better performance. Finally, even if educators had been motivated by pay-for-performance from its outset, it could have still taken time for educators to change their practices or decisions on where to work in response to the opportunity to earn performance bonuses.

Impact of Pay-for-Performance on Educator Performance Ratings

Pay-for-performance was designed to raise educators' performance on the measures used in TIF. Specifically, by linking bonuses to those performance measures, pay-for-performance was supposed to have motivated educators to improve their ratings on those measures and encouraged educators who would score well on them to work in schools offering performance bonuses. In this section, we assess whether pay-for-performance had its intended effect of raising educator performance ratings.

[50] Appendix F provides supplemental information on the number of schools used for the analyses in this chapter and information needed for calculating effect sizes (see Tables F.1 and F.2). In addition, as discussed in Chapter IV, some educators in the study schools misunderstood their eligibility for pay-for-performance or the potential amounts they could earn. The impacts reported in this chapter reflect the impact of pay-for-performance given educators' actual beliefs. This study was not designed to assess the impacts of pay-for-performance bonuses in the scenario in which all educators correctly understood their eligibility and the amount they could earn in a bonus. Findings from the cost-effectiveness analysis presented in this chapter would be identical if we had taken into account that not all teachers understood their eligibility for a bonus. Since both the costs and the impacts per student would be larger by the same factor if we had adjusted the impacts and costs by the fraction of students whose teachers correctly understood their bonus eligibility, pay-for-performance would require the same per-unit cost to achieve a specified impact.

[51] As discussed in Chapter II, evaluation districts were classified into two cohorts—Cohort 1 and Cohort 2—according to the year in which we randomly assigned their schools to a treatment or control group. The 10 districts examined in this chapter, whose schools were randomly assigned in spring and summer 2011, were classified as Cohort 1. Three additional districts, whose schools were randomly assigned in spring and summer 2012, were classified as Cohort 2. In Appendix F, we present three years of impacts on educator effectiveness and student achievement for Cohorts 1 and 2 combined.

As discussed in Chapter IV, districts had to evaluate teachers and principals based on student achievement growth and at least two observations of classroom or school practices. However, districts had flexibility in how they implemented this requirement. For example, they could choose to evaluate teachers based on the achievement growth of the teachers' own students (classroom achievement growth); all students in the same grade, team, or subject area; all students in the school (school achievement growth); or some combination of these measures.

We examined the impact of pay-for-performance on four measures of educator effectiveness obtained from district administrative records: (1) school achievement growth ratings, which were used to evaluate teachers and principals; (2) classroom achievement growth ratings for teachers; (3) classroom observation ratings for teachers; and (4) observation ratings for principals. Different districts selected or designed these measures in different ways, but all of the measures placed educators into three to five performance categories—such as effective or highly effective—or on a numeric scale in which an increase of one point was similar to advancing one performance level. To express ratings from different districts on a common scale, we expressed each rating as a score on a 1-to-4 rating scale, with 1 being the lowest and 4 being the highest possible rating an educator could receive on the district's measure of performance (see Appendix B for details). Thus, an increase from 3 to 4 on the rating scale can roughly be interpreted as a change from being classified as effective to being classified as highly effective.

We examined each performance measure separately for two reasons. First, the different measures may capture different aspects of effectiveness. For example, classroom observations could have identified aspects of teachers' instruction that mattered for classroom climate but not for students' math or reading achievement. Second, as discussed in Chapter IV, districts awarded separate bonuses for different performance measures, so educators could have focused on improving their performance on the measures that they could influence most easily or that were tied to the largest bonuses.

The following findings reflect the impacts of pay-for-performance bonuses on average educator performance ratings in schools that offered those bonuses. For simplicity, we refer to these findings as impacts on teachers' or principals' ratings. As we discuss later in this chapter, average ratings in schools could change for a variety of reasons, including improvements in educators' practices and the hiring or departure of higher- or lower-performing educators.

Districts' Measures of Student Achievement Growth in Schools and Classrooms

School and classroom achievement growth were the two most common student achievement growth measures that districts used to evaluate educators. School achievement growth, which all evaluation districts included in their educator evaluation systems, combines the contributions of all staff at a school, so impacts on school achievement growth might reflect how teachers, principals, or other school staff responded to pay-for-performance. In addition, 6 of the 10 districts also used measures of classroom achievement growth in all four years of implementation to evaluate a subset of their teachers, and one additional district began doing so in Year 3. In those districts, teachers who received classroom achievement growth ratings were typically those who taught grades and subjects tested using annual state assessments.

The impacts of pay-for-performance on teachers' and principals' performance ratings based on student achievement growth were mixed, with positive impacts in Years 1 and 4 but not in other years. In Years 1 and 4, educators in treatment schools had school achievement

growth ratings that were 0.36 and 0.39 points higher on a 1-to-4 rating scale than those of educators in control schools (Table VI.1).[52,53] In the other years, the difference in school achievement growth ratings between the two groups was either somewhat smaller (0.27 points in Year 2) or much smaller (0.04 points in Year 3), and neither difference was statistically significant (*p*-value = 0.07 in Year 2 and 0.75 in Year 3).

Table VI.1. Student Achievement Growth Ratings (Points on 1-to-4 Scale)

Performance Measure and Year	Treatment	Control	Impact	*p*-value	**Number of Teachers**	**Number of Schools**
School Achievement Growth (Based on 9 or 10 Districts)						
Ratings in Year 1 (9 districts)	2.62	2.25	0.36*	0.04	**NA**	**123[a]**
Ratings in Year 2 (10 districts)	2.55	2.27	0.27	0.07	**NA**	**130**
Ratings in Year 3 (10 districts)	2.41	2.37	0.04	0.75	**NA**	**131**
Ratings in Year 4 (9 districts)	2.59	2.20	0.39*	0.03	**NA**	**121[b]**
Classroom Achievement Growth (Based on 6 or 7 Districts)[c]						
Ratings in Year 1 (6 districts)	2.26	2.08	0.18*	0.04	**1,076**	**72**
Ratings in Year 2 (6 districts)	2.23	2.17	0.06	0.25	**1,332**	**72**
Ratings in Year 3 (7 districts)	2.53	2.53	0.00	0.97	**2,040**	**90**
Ratings in Year 4 (7 districts)	2.71	2.53	0.18*	0.00	**1,957**	**90**

Source: Educator administrative data.

Note: The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]School achievement growth ratings for one district in Year 1 were not included because the district did not place educators into performance categories or onto a numeric scale.

[b]One district did not provide school achievement growth ratings for control schools in Year 4, so all schools in the district were excluded from this analysis.

[c]In Years 1 and 2, six districts evaluated teachers based on classroom achievement growth. In Years 3 and 4, seven districts evaluated teachers based on classroom achievement growth. One district expanded the number of teachers evaluated based on classroom achievement growth in Year 2 and then changed the classroom achievement growth measures used to evaluate some of its teachers in Year 3.

*Impact is statistically significant at the .05 level, two-tailed test.

NA is not applicable.

[52] Appendix F, Tables F.3 and F.4 show findings from alternative ways of estimating impacts on school achievement growth ratings and classroom observation ratings in Year 4. Prior reports provide findings based on these alternative approaches in Years 1 and 2 (Chiang et al. 2015) and Year 3 (Wellington et al. 2016).

[53] When we included both Cohorts 1 and 2 in the analysis, the impacts of pay-for-performance on educators' achievement growth ratings differed in two ways from the main findings. First, the impacts of pay-for-performance on school and classroom achievement growth ratings in Year 1 were no longer statistically significant. Second, the impact on classroom achievement growth ratings in Year 3 was negative and statistically significant (Appendix F, Table F.5).

The patterns of findings were similar for classroom achievement growth ratings. Among teachers who were evaluated on classroom achievement growth, those in treatment schools earned ratings that were 0.18 points higher than those of teachers in control schools in Years 1 and 4 (Table VI.1). However, the two groups earned similar classroom achievement growth ratings in Years 2 and 3.

To explore why the impacts on school and classroom achievement growth varied from year to year, we considered several possible explanations, but none were consistent with the evidence. First, we examined whether changes in the group of schools included in the analyses could explain these patterns. Findings were generally similar when the analyses in all four years were based on the same group of schools (Appendix F, Table F.7). Second, we considered whether any aspects of implementation were shared by only Years 1 and 4—the years in which we found positive impacts on achievement growth ratings. This did not seem to be the case. In fact, the bonus structure was generally stable across all four years, educators' understanding of TIF measures to evaluate their performance grew over the course of the grant, their understanding of their eligibility for a bonus and how much they could earn was similar in Years 2 through 4, and the impacts of pay-for-performance on educator satisfaction became more favorable from Year 2 to 3 (Chapters IV and V). In addition, different districts accounted for the positive impacts in Years 1 and 4, ruling out some common aspect of implementation within districts that might account for the pattern in these years.[54]

Observation Ratings for Teachers and Principals

In all districts, both teachers and principals received ratings based on formal observations of their practices. Trained observers rated teachers on their classroom practices and rated principals on the practices they implemented in their schools.

Pay-for-performance led teachers to earn slightly higher classroom observation ratings by Year 3. Impacts of pay-for-performance on classroom observation ratings grew over the four years of implementation and became statistically significant by Year 3. In Year 3, treatment teachers had classroom observation ratings that were 0.05 points higher on a 1-to-4 point rating scale than those of control teachers (Table VI.2).[55] This impact grew to 0.09 points in Year 4. Nevertheless, these impacts remained small, never amounting to more than 10 percent of the increment between two performance levels.

The small impacts on observation ratings that did emerge could have resulted from a number of possible factors. One possible explanation is that the observed classroom practices in treatment schools were truly, but just slightly, better than those in control schools. These differences in practices could have emerged because individual teachers in treatment schools improved their practices in response to the opportunity to earn pay-for-performance bonuses or because the offer of those bonuses led more effective teachers to stay at or join those schools. Later in this chapter we discuss whether there is evidence for these types of effects. Another possible explanation is that the

[54] District-specific impacts in Year 1 had a weak, negative correlation with district-specific impacts in Year 4 for both school achievement growth ratings (-0.25) and classroom achievement growth ratings (-0.21).

[55] The impact on classroom observation ratings was not statistically significant when Cohorts 1 and 2 were included in the analysis for Years 1, 2, and 3 (Appendix F, Table F.6).

observers in treatment schools could have been more lenient in their ratings than those in control schools. Teachers were typically observed by principals at their schools (Chapter IV), and principals' awareness of their teachers' eligibility for performance bonuses may have (consciously or unconsciously) affected the ratings they gave. Given the available data, we cannot separate true differences in classroom practices from differences in the leniency of the observers.

Table VI.2. Observation Ratings for Teachers and Principals (Points on 1-to-4 Scale)

Performance Measure and Year	Treatment	Control	Impact	*p*-value	**Number of Educators**	**Number of Schools**
Teachers' Classroom Observation Ratings						
Ratings in Year 1	2.94	2.91	0.03	0.21	**3,599**	**131**
Ratings in Year 2	2.98	2.93	0.05	0.08	**3,598**	**131**
Ratings in Year 3	2.97	2.91	0.05*	0.02	**3,622**	**131**
Ratings in Year 4	2.88	2.80	0.09*	0.02	**3,544**	**131**
Observation Ratings for Principals[a]						
Ratings in Year 1	3.08	3.18	-0.10	0.20	**105**	**105**
Ratings in Year 2	3.14	3.01	0.13	0.19	**118**	**117**
Ratings in Year 3	3.37	3.32	0.05	0.52	**119**	**118**
Ratings in Year 4	3.25	3.21	0.04	0.63	**122**	**122**

Source: Educator administrative data.

Notes: The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]Analyses of observation ratings for principals included fewer than 131 schools because (1) one district did not provide observation ratings for principals in Year 1; and (2) in each year, some principals had missing observation scores.

*Impact is statistically significant at the .05 level, two-tailed test.

Pay-for-performance had no impact on the observation ratings of principals. In contrast to the findings for teachers, principals in treatment and control schools had similar observation ratings in all four years (Table VI.2).

Educator Performance Ratings for Returning and Newly Hired Teachers

Because pay-for-performance raised teachers' classroom observation ratings in Years 3 and 4 and classroom achievement growth ratings in Year 4, we examined whether those impacts reflected increased effectiveness primarily among returning or newly hired teachers.[56] Examining impacts on returning and newly hired teachers can suggest the ways in which pay-for-performance leads to greater teacher effectiveness.

[56] We did not undertake a similar analysis of the positive impacts on classroom achievement growth ratings in Year 1 because we did not have the data to classify teachers as returning or newly hired for all districts. Among principals, we examined impacts of pay-for-performance on the performance ratings earned by returning and newly hired principals, but the findings were generally imprecise because of small numbers of principals (see Appendix F, Tables F.12 and F.13).

According to the theory of change from Chapter I, pay-for-performance could increase teacher effectiveness in three possible ways. First, it could help schools keep more effective teachers. Second, it could enable schools to recruit more effective teachers. Third, it could motivate teachers to become more effective—for instance, by adopting better classroom practices. The effectiveness of returning and newly hired teachers should reflect different combinations of these three influences, so positive impacts on one group but not the other could suggest which of these influences might be strongest.

In particular, if pay-for-performance enabled schools to keep more effective teachers, then returning teachers ought to have higher performance ratings in treatment schools than control schools. In fact, from Year 1 to 4, a slightly higher percentage of teachers stayed in treatment schools than control schools (51 versus 49 percent; Appendix F, Table F.8). If this increased retention was concentrated among effective teachers, then we might expect returning teachers in treatment schools to outperform their counterparts in control schools.[57]

If pay-for-performance enabled schools to recruit more effective teachers, then newly hired teachers should have higher ratings in treatment schools than control schools. As discussed in Chapter V, in Years 3 and 4, more than 40 percent of treatment principals (compared with no more than 15 percent of control principals) reported using pay-for-performance as a recruitment tool for hiring teachers. Therefore, many treatment principals could have believed the offer of performance bonuses would attract better teachers.

In addition, pay-for-performance could have led both returning and newly hired teachers to teach more effectively than they otherwise would have. As a result of changes in teaching practices, the ratings of either returning or newly hired teachers could be higher in treatment schools than control schools. However, returning teachers may have been more motivated and able to change their practices in response to performance bonuses. First, returning teachers had at least one year of experience with the program and the bonuses it awarded, and therefore may have better understood the program. For example, as shown in Chapter IV, returning teachers in treatment schools were 14 percentage points more likely to understand their eligibility for pay-for-performance bonuses than were new teachers in treatment schools (although this difference was not statistically significant). Second, returning teachers should have received at least one year of feedback on their performance, which could help them to improve in response to the opportunity to earn bonuses.

In short, stronger impacts on returning teachers could suggest that retention of more effective teachers was a key way in which pay-for-performance influenced teacher effectiveness, or that returning teachers' prior experience with the program made them more willing or able to change their practices. On the other hand, stronger impacts on newly hired teachers could suggest that recruitment of more effective teachers was an important effect of pay-for-performance.

We classified returning teachers as those who had stayed in their school since the previous year and newly hired teachers as those who were new to their school in the current year. For example, in Year 4, returning teachers had stayed in their school from Year 3 to 4, and newly hired teachers were new to their school in Year 4. Findings that defined returning teachers as those who stayed in their school since Year 1 were similar (Appendix F, Table F.11). Because returning and new teachers at

[57] Pay-for-performance could have also altered other characteristics of the schools' staff, such as their demographic and professional characteristics. However, we found little evidence that pay-for-performance led to changes in those characteristics in Year 4 (Appendix F, Table F.10) or in previous years (Wellington et al. 2016; Chiang et al. 2015).

the same school earned the same school achievement growth ratings, these analyses focused only on the two measures—classroom observations and classroom achievement growth—that captured individual teachers' performance.

Pay-for-performance raised the performance ratings of returning teachers in Years 3 and 4. For each outcome on which pay-for-performance had positive impacts across all teachers, we found similar, positive impacts on returning teachers. On a 1-to-4 scale, returning teachers in treatment schools outperformed those in control schools by 0.06 and 0.07 points on their classroom observation ratings in Years 3 and 4 (though the impact in Year 4 was not quite statistically significant; p-value = 0.07), and by 0.18 points on their classroom achievement growth ratings in Year 4 (Table VI.3). Impacts on newly hired teachers were not statistically significant in any year but, in Year 4, were similar in magnitude to the impacts we found for returning teachers.

These findings leave open the possibility that pay-for-performance could have influenced teacher effectiveness through any of the ways discussed earlier. Given the consistent evidence that pay-for-performance led to higher performance ratings among returning teachers, it is likely that pay-for-performance either helped schools retain more effective teachers or motivated returning teachers to become more effective. Because the results in Year 4 also suggest that newly hired teachers in treatment schools may have performed better than those in control schools (though the differences were not significant), it is possible that pay-for-performance began enabling schools to recruit more effective teachers or motivate new teachers to become more effective by the end of the TIF grant.

Impact of Pay-for-Performance on Student Achievement

Improving student achievement is the ultimate objective of the TIF grants. The grants were designed to accomplish this objective by enhancing educator effectiveness, and the analyses in the previous section suggest that pay-for-performance had varying impacts on educators' performance ratings across years. However, the presence or absence of impacts on those performance ratings does not definitively show whether pay-for-performance affected student achievement. There is no guarantee that the performance measures used by TIF districts accurately captured aspects of teaching or leadership that might be important for student achievement. Even when student achievement information was incorporated directly into particular measures, such as measures of school achievement growth, districts differed considerably in how they converted that information into ratings, which could weaken the connection between those ratings and student achievement.

In this section, we directly examine the impact of pay-for-performance on student achievement in the study schools, using administrative data on students' reading and math scores from state assessments. In contrast to the educator performance ratings, which used student achievement information differently in different districts and measures, the analysis in this section used the same method for all districts to compare student achievement in treatment and control schools. Moreover, this analysis enabled us to examine the impacts of pay-for-performance separately on math and reading achievement, whereas educator performance ratings often combined information on student achievement or classroom practices across subjects.

Table VI.3. Impacts of Pay-for-Performance on the Performance Ratings of Returning and Newly Hired Teachers (Points on 1-to-4 Scale)

	Returning Teachers		Newly Hired Teachers				
Performance Measure and Year	Impact	*p*-value	Impact	*p*-value	**Number of Returning Teachers**	**Number of Newly Hired Teachers**	**Number of Schools**
Year 2							
Classroom observation ratings	0.05	0.12	-0.02	0.72	**2,883**	**715**	**131**
Classroom achievement growth ratings	0.07	0.24	-0.03	0.83	**1,015**	**317**	**72**
Year 3							
Classroom observation ratings	0.06*	0.03	0.01	0.84	**2,853**	**769**	**131**
Classroom achievement growth ratings	0.01	0.89	-0.03	0.73	**1,594**	**446**	**90**
Year 4							
Classroom observation ratings	0.07	0.07	0.09	0.21	**2,766**	**778**	**131**
Classroom achievement growth ratings	0.18*	0.01	0.16	0.18	**1,566**	**391**	**90**

Source: Educator administrative data.

Note: Returning teachers were those who had stayed in their school since the previous school year, and newly hired teachers were those who were new to their school in the current year. For example, in Year 4, returning teachers had stayed in their school from Year 3 to 4, and newly hired teachers were new to their school in Year 4.

*Impact is statistically significant at the .05 level, two-tailed test.

As discussed in Chapter II, we standardized student test scores into z-scores that reflected how well each student scored compared with the average student in his or her state, grade, and year. This approach enabled us to estimate impacts in the same units—standard deviations of student test scores—in all four years, despite the fact that in Year 4 seven of the ten districts administered new student assessments that aligned to their states' college- and career-ready standards.

However, adopting new assessments could affect the impact findings in at least two ways. First, in districts that changed assessments, student outcomes in Year 4 came from a different assessment than the one used for measuring schools' baseline average student achievement. The likely result was that controlling for baseline achievement in the impact analyses improved precision to a lesser extent in Year 4 than in earlier years (Appendix B). Second, to the extent that teachers were less knowledgeable about how to improve their students' achievement on the new assessments than on the old ones, the impacts on student achievement in Year 4 could underestimate the impacts that would have materialized if the new assessments had been in place for several years.

To measure the impacts of pay-for-performance, we compared the average student achievement of treatment and control schools at the end of each year, based on the students at these schools who took that year's spring assessment (Chapter II). Achievement differences between treatment and control schools measured the extent to which schools' average student achievement was higher or lower due to having implemented pay-for-performance since the beginning of the study. For example, the difference at the end of Year 4 measured the impact of pay-for-performance on the average student achievement of the study schools after four years of implementation.

Because some students entered or left the study schools during the course of the study, implementing pay-for-performance at a school for a certain number of years did not mean that all students at the school were exposed to pay-for-performance for all of those years. In fact, at the end of Year 4, our estimate is that 37 percent of students had been at their schools since the beginning of the study and were exposed to all four years of the program (Appendix B, Table B.6). The remaining students had been at their schools for three years (14 percent), two years (21 percent), or one year (28 percent).

Although we could estimate the percentages of students who had been exposed to pay-for-performance for specified durations, our data generally did not enable us to identify exactly which students were exposed for durations of longer than one or two years (see Appendix B for details). Therefore, we could not measure the impacts of exposing students to the full four-year duration of the pay-for-performance program. In supplemental analyses, we measured impacts on students who had been exposed to one or two years of pay-for-performance, using only those students who were enrolled in the study schools when the schools were randomly assigned to the treatment and control groups, and we obtained similar results to those presented in this section (Appendix F, Table F.20). Because those supplemental analyses could not span the full duration of the pay-for-performance program, they are not the focus of this chapter.

Throughout the rest of this section, we report the impacts of implementing pay-for-performance for one, two, three, or four years on the average student achievement of the study schools. For simplicity, we refer to these impacts as impacts on student achievement.

Overall Impacts

We first compared average student achievement in all treatment schools and control schools that participated in the study. The large number of participating schools—65 treatment and 66 control—provided an opportunity to measure the overall impacts of pay-for-performance with a great deal of precision (see Appendix B for details on the levels of precision).

Pay-for-performance led to slightly higher student achievement in reading and math by the second year of implementation. At the end of the first year, student reading achievement was higher by 0.03 standard deviations in treatment schools than in control schools, and the total difference remained at 0.03 to 0.04 standard deviations across the subsequent three years (Table VI.4). The differences in reading achievement at the end of Years 1 through 3 were statistically significant, but the difference at the end of Year 4 was not quite significant (p-value = 0.08). In math, student achievement was higher by 0.04 standard deviations as a result of implementing pay-for-performance for two years, and remained higher by 0.04 to 0.06 standard deviations thereafter, although only the impact of implementing pay-for-performance for three years was statistically significant. In all years and subjects, the average achievement of students in both treatment and

control schools was below the statewide mean, as indicated by the negative *z*-score values in Table VI.4.[58]

There are several ways to interpret the magnitudes of the impacts on student achievement. First, each impact can be expressed as a difference in percentiles of achievement. In the first two years of TIF implementation, the average student in control schools earned a math score at about the 33rd percentile of student achievement statewide (Figure VI.1).[59] After treatment schools had implemented pay-for-performance for two years, their students earned an average math score at about the 35th percentile—a difference of 2 percentile points. Thereafter, math scores of students in treatment schools remained 1 to 2 percentiles ahead of those of students in control schools, though only the difference at the end of Year 3 was statistically significant. In reading, a difference of 2 percentile points between students in treatment and control schools emerged at the end of Year 1, and the difference remained at 1 to 2 percentile points after subsequent years of implementation.

Table VI.4. Student Achievement in Math and Reading (Student *z*-Score Units)

Year and Subject	Treatment	Control	Impact	*p*-value	**Number of Students**	**Number of Schools**
Year 1						
Math	-0.43	-0.45	0.02	0.33	**40,535**	**131**
Reading	-0.37	-0.40	0.03*	0.04	**40,256**	**131**
Year 2						
Math	-0.38	-0.43	0.04	0.07	**40,454**	**131**
Reading	-0.36	-0.39	0.03*	0.02	**40,122**	**131**
Year 3						
Math	-0.36	-0.42	0.06*	0.02	**39,770**	**131**
Reading	-0.33	-0.37	0.04*	0.02	**39,538**	**131**
Year 4						
Math	-0.34	-0.37	0.04	0.13	**38,939**	**131**
Reading	-0.34	-0.38	0.04	0.08	**38,929**	**131**

Source: Student administrative data.

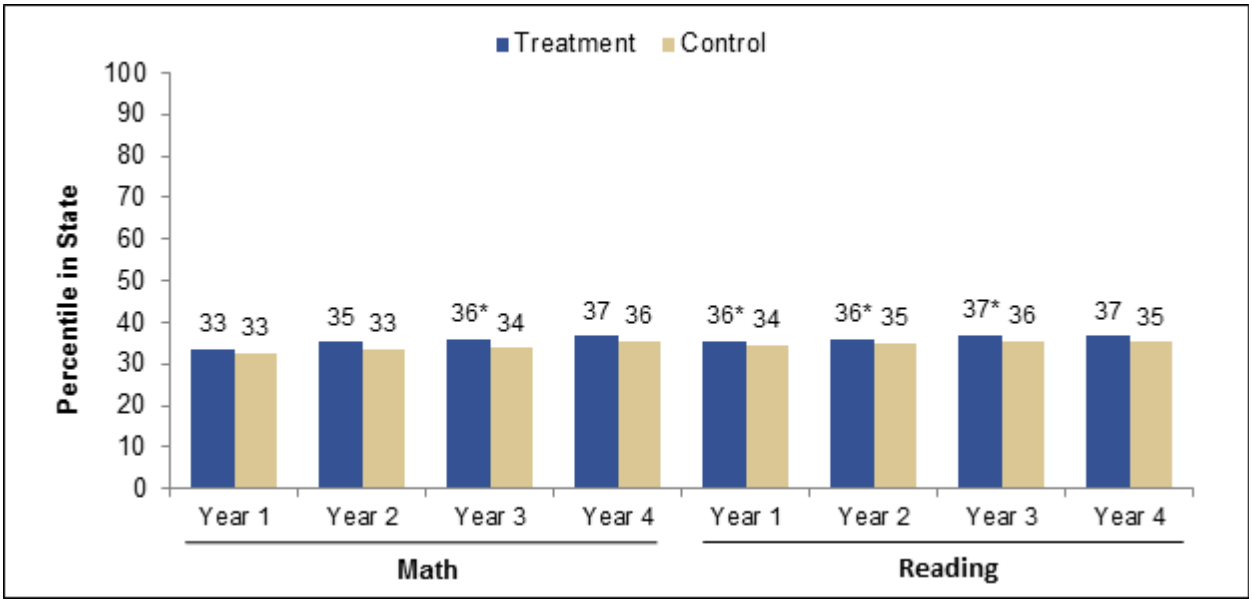
Note: The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

*Impact is statistically significant at the .05 level, two-tailed test.

[58] A variety of alternative analytic approaches produced impact findings that were similar to the main findings (Appendix F, Tables F.14 and F.15). One approach that did not account for preexisting differences between treatment and control schools produced estimates of impacts that were close to zero, but this approach was vulnerable to error by allowing these preexisting differences to obscure the impacts. In addition, when both Cohorts 1 and 2 were included in the analysis, the impacts after one and two years were similar in magnitude, though occasionally different in statistical significance, from the impacts based on Cohort 1 only (Appendix F, Table F.16). After three years, the impacts based on Cohorts 1 and 2 were not statistically significant and were smaller in magnitude than the significant impacts based on Cohort 1 only.

[59] This approximation is based on a normal distribution for student achievement.

Figure VI.1. Average Student Achievement in Treatment and Control Schools (Percentiles)



Source: Student administrative data (N = 40,535 students for Year 1 math; N = 40,454 students for Year 2 math; N = 39,770 for Year 3 math; N = 38,939 for Year 4 math; N = 40,256 students for Year 1 reading; N = 40,122 students for Year 2 reading; N = 39,538 for Year 3 reading; N = 38,929 for Year 4 reading).

Figure reads: At the end of Year 1, students in treatment schools earned an average math score at the 33rd percentile in their state, and students in control schools also earned an average math score at the 33rd percentile.

*Difference in student achievement between treatment and control schools (when expressed in student *z*-score units, based on Table VI.4) is statistically significant at the .05 level, two-tailed test.

The impacts can also be expressed in weeks of learning for the average student nationwide (Hill et al. 2008). After two to four years of pay-for-performance, the difference in student math achievement between treatment and control schools—0.04 to 0.06 standard deviations—was equivalent to about three to four weeks of learning in math. Likewise, across all four years, the difference in student reading achievement—0.03 to 0.04 standard deviations—amounted to about three to four weeks of learning in reading.[60]

Pay-for-performance could affect elementary and middle school grades differently, so we examined impacts separately by grade span. Effective teaching might require different sets of instructional skills at different grade levels, with subject-matter knowledge, for example, assuming greater importance when teaching more advanced content. If teachers could improve some skills more readily than others in response to incentives, then pay-for-performance could lead to different impacts in different grade levels. Across all four years and in both math and reading, we found that the impacts of pay-for-performance were generally larger in middle school grades than in elementary

[60] We converted the impacts into weeks of learning in the following manner. Using six nationally normed math assessments, Hill et al. (2008) found that students in grades 3 through 8 grew, on average, about 0.5 standard deviations per year in math achievement. Therefore, an impact of 0.04 to 0.06 standard deviations on student math achievement was equivalent to 8 to 12 percent of a year of learning, or about 3 to 4 weeks of learning in a typical 36-week school year. Likewise, using seven nationally normed assessments in reading, Hill et al. (2008) found that students in these grades grew an average of 0.36 standard deviations per year in reading achievement. Therefore, an impact of 0.03 to 0.04 standard deviations on student reading achievement was equivalent to 8 to 11 percent of a year of learning, amounting again to about 3 to 4 weeks of learning.

school grades, but these differences in impacts were usually not statistically significant (Appendix F, Table F.17). The only exception was that implementing pay-for-performance for four years led to a significantly greater improvement in reading achievement in middle school grades than in elementary school grades.

Differences in Student Achievement Impacts Across Districts

The findings shown in Table VI.4 represent an average impact of pay-for-performance across the 10 districts in the study. However, these districts differed in many ways, including the design and implementation of their pay-for-performance programs and the political or institutional contexts in which the implementation took place. These differences raise the possibility that the impacts of pay-for-performance could have also differed among districts.

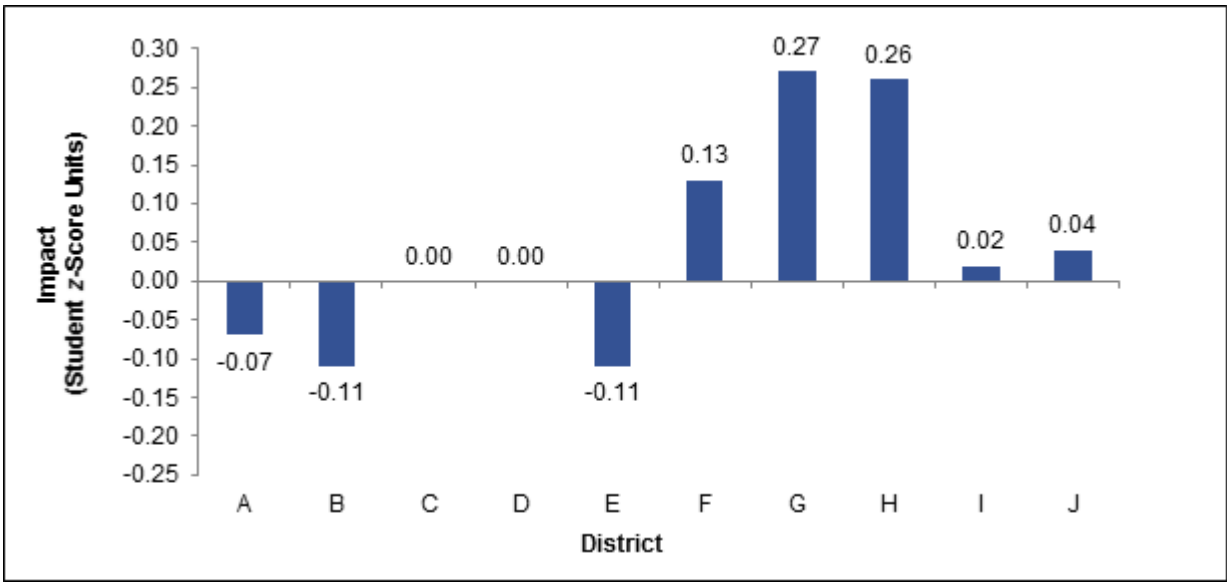
The impacts of pay-for-performance on math and reading achievement differed substantially across districts. Although, on average, pay-for-performance had a positive impact on math and reading achievement, impacts varied across districts by a statistically significant degree and a substantial magnitude. For example, district-specific impacts of pay-for-performance on math achievement after four years ranged from -0.11 to 0.27 standard deviations. The impacts were positive in 5 of the 10 districts, negative in 3 districts, and about zero (within 0.01 standard deviations) in the other 2 (Figure VI.2).[61] Impacts on reading achievement after four years also varied across districts, ranging from -0.23 to 0.22 standard deviations (Figure VI.3). The impacts were positive in 7 districts and negative in the remaining 3 districts. Differences in impacts across districts in prior years were also substantial in size and statistically significant (Chiang et al. 2015; Wellington et al. 2016).

Differences in impacts between particular districts were often, but not always, similar across years. For example, in math, four of the five districts that experienced positive impacts of pay-for-performance after four years had also experienced positive impacts after three years. In reading, five of the seven districts that experienced positive impacts after four years had also experienced positive impacts after three years.[62] Despite this overall pattern of consistency in impacts over time, there were exceptions in which impacts within some districts changed substantially. For example, in one district (District H in Figures IV.2 and IV.3), impacts in both math and reading doubled in size, from 0.10 standard deviations after three years to over 0.20 standard deviations after four years. In another district (District E), the impact on reading achievement was close to zero (-0.01 standard deviations) after three years but substantial and negative (-0.23 standard deviations) after four years. Although impacts remained consistent in some districts while they changed in other districts, impacts differed substantially across districts overall in every year, as noted earlier.

[61] Within each district, the small number of schools meant that only very large impacts would have been statistically significant. Therefore, we do not report the statistical significance of district-specific impacts and instead focus on the overall variation in impacts across all 10 districts. Appendix F, Figures F.1 and F.2 show that impacts after three years of pay-for-performance also varied across all 13 districts in Cohorts 1 and 2.

[62] One statistical measure of the consistency in impacts over time is the correlation in impacts between particular years (which can range from a perfect negative correlation of -1 to a perfect positive correlation of 1). We found moderate to high correlations in impacts between consecutive years, depending on the years examined. The year-to-year correlation in impacts ranged from 0.68 to 0.91 in math and 0.57 to 0.80 in reading.

Figure VI.2. Impact of Pay-for-Performance on Student Achievement in Math After Four Years of Implementation, by District (Student *z*-Score Units)

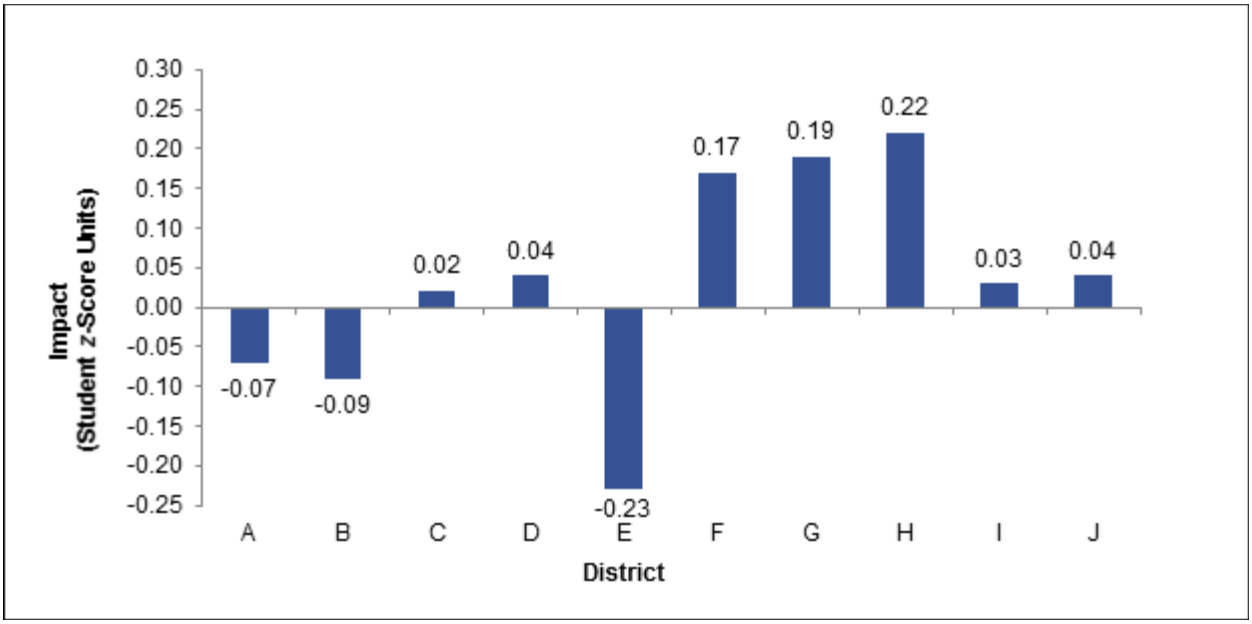


Source: Student administrative data (N = 38,939).

Note: An F-test of the null hypothesis that impacts are equal across districts has a *p*-value of less than 0.01.

Figure reads: In District A, pay-for-performance lowered student math achievement by 0.07 student *z*-score units after four years.

Figure VI.3. Impact of Pay-for-Performance on Student Achievement in Reading After Four Years of Implementation, by District (Student *z*-Score Units)



Source: Student administrative data (N = 38,929).

Note: An F-test of the null hypothesis that impacts are equal across districts has a *p*-value of less than 0.01.

Figure reads: In District A, pay-for-performance lowered student reading achievement by 0.07 student *z*-score units after four years.

We sought to explain why impacts differed across districts. In particular, as discussed in Chapter IV, both the design and implementation of TIF programs also differed across districts. Therefore, we examined whether impacts were systematically larger or smaller in districts that designed or implemented their programs in particular ways. We focused on program characteristics that might influence educators' motivation or capacity to respond to pay-for-performance bonuses, including (1) the use of student achievement growth in teachers' own classrooms to measure teacher effectiveness and award bonuses, (2) the size of the average bonus, (3) the amount of differentiation of bonuses, (4) the timing of awarding bonuses based on the prior year, and (5) teachers' understanding of their pay-for-performance eligibility (see Appendix G for a detailed discussion of how we selected and measured these characteristics). In addition, we examined whether differences in impacts were related to two features of the districts' institutional and political contexts—district size and the presence of right-to-work laws. The impacts of pay-for-performance on student achievement may be smaller in larger districts if, for example, teachers in larger districts have less input into district policies, which in turn could make teachers feel less invested in new policies such as pay-for-performance. To the extent that teachers' unions promote skepticism of performance-based compensation, political factors that weaken unions, such as right-to-work laws, could lead more teachers to respond favorably to performance bonuses.

Few characteristics of TIF districts and their programs explained differences in the impacts of pay-for-performance on student achievement. Although some characteristics of the design or implementation of TIF programs were associated with achievement impacts in particular subjects within particular years, none were consistently associated with impacts across multiple years. For example, after four years of pay-for-performance, impacts in one subject or the other tended to be more positive in districts that used classroom achievement growth, awarded larger average bonuses, and awarded bonuses earlier (Appendix G, Tables G.2 and G.3). However, given that we examined a large number of relationships between program characteristics and student achievement impacts, there was an increased likelihood that these significant relationships occurred just by chance. Moreover, none of the program characteristics that were associated with impacts after four years had previously been associated with impacts after two or three years (Chiang et al. 2015; Wellington et al. 2016). In fact, in previous years, only one other program characteristic—the amount of differentiation of bonuses—was associated with impacts (in a negative manner), and this association appeared in only one year and subject (Chiang et al. 2015). Of the characteristics examined, only the contextual factors—district size and the presence of right-to-work laws—were related to impacts in multiple years and subjects. In particular, student achievement impacts tended to be more positive in smaller districts and those with right-work-laws (Appendix G, Table G.4).

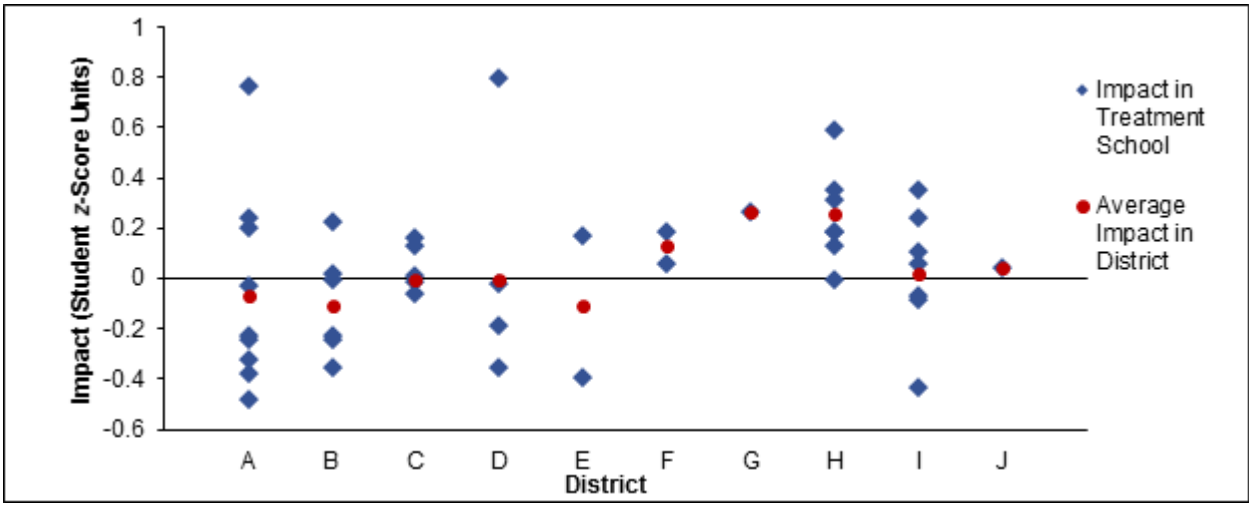
Differences in Student Achievement Impacts Across Schools

Within each of the 10 districts in the study, treatment schools could have also differed in the degree to which pay-for-performance affected student achievement. Although treatment schools within a district participated in programs with the same design and possibly the same implementation, pay-for-performance may have affected teachers' and principals' behaviors differently across schools, leading to differences in impacts on student achievement. For example, in schools whose teachers were more motivated by earning a bonus, pay-for-performance may have had stronger impacts on teaching practices and, therefore, student achievement. The same pay-for-performance program could also affect schools differently depending on the schools' context, such as how well or poorly the schools were performing before the program began. We examined

whether impacts on student achievement differed across schools and, if so, assessed potential reasons for those differences.

The impacts of pay-for-performance on math and reading achievement differed across treatment schools, even within the same district. After four years of implementation, most districts had at least some treatment schools that experienced positive impacts of pay-for-performance on student achievement and some that experienced negative impacts (Figures VI.4 and VI.5).[63] Statistically, most of the variation in impacts across schools occurred within the same district (76 percent in math and 63 percent in reading) rather than across districts. We found similar variation in impacts across schools after three years of pay-for-performance (Wellington et al. 2016).

Figure VI.4. Impact of Pay-for-Performance on Student Math Achievement After Four Years of Implementation, by Treatment School and by District (Student *z*-Score Units)



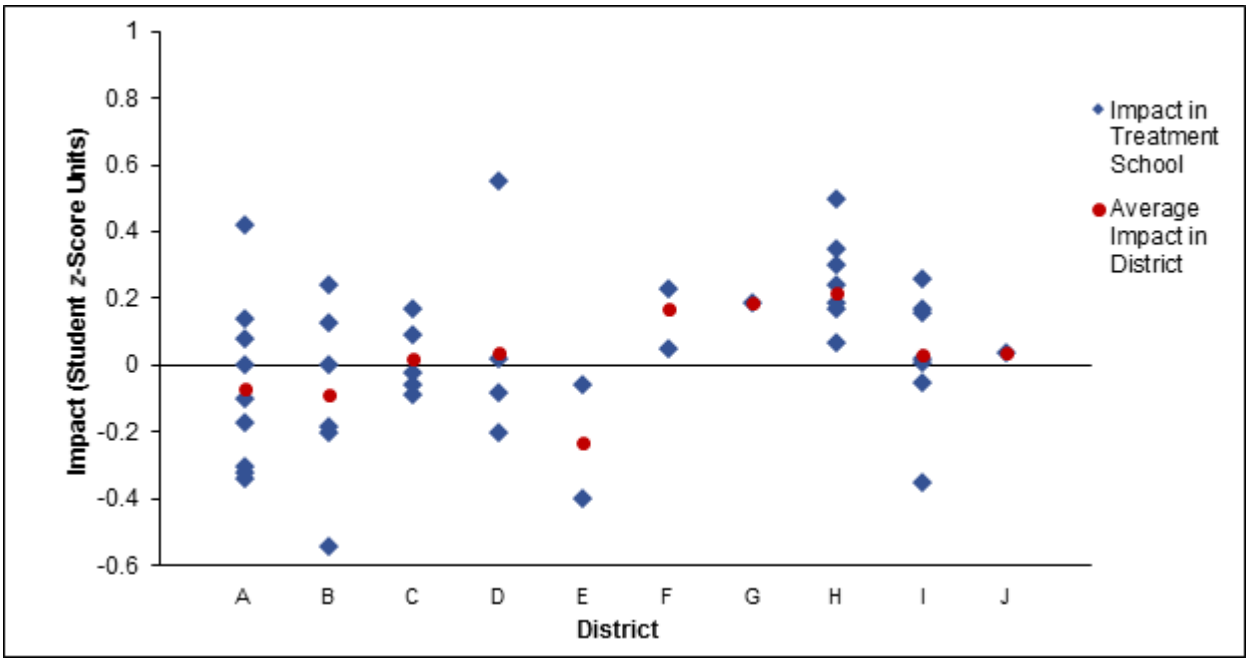
Source: Student administrative data (N = 38,939).

Notes: The impact of pay-for-performance on a treatment school is the difference in achievement between that school and the control school with which it was paired during random assignment. Treatment schools that were assigned together during random assignment (as a single group) are represented by a single diamond (see Appendix A for details on the random assignment process).

Figure reads: After four years, within the nine treatment schools in District A, pay-for-performance raised student math achievement by 0.20, 0.24, and 0.77 standard deviations in three of the schools, and lowered student math achievement by 0.03, 0.23, 0.24, 0.32, 0.38, and 0.48 standard deviations in the other six schools.

[63] We measured the impact on each treatment school as the difference in student achievement between that school and the control school with which it was paired during random assignment (Chapter II).

Figure VI.5. Impact of Pay-for-Performance on Student Reading Achievement After Four Years of Implementation, by Treatment School and by District (Student z-Score Units)



Source: Student administrative data (N = 38,929).

Notes: The impact of pay-for-performance on a treatment school is the difference in achievement between that school and the control school with which it was paired during random assignment. Treatment schools that were assigned together during random assignment (as a single group) are represented by a single diamond (see Appendix A for details on the random assignment process).

Figure reads: After four years, within the nine treatment schools in District A, pay-for-performance raised student reading achievement by 0.08, 0.14, and 0.42 standard deviations in three of the schools, had no effect in one school, and lowered student reading achievement by 0.1, 0.17, 0.3, 0.32, and 0.34 standard deviations in the other five schools.

Given that impacts on student achievement differed across schools, we sought to determine whether those differences were related to two types of factors: (1) differences in impacts on teachers' and principals' behaviors and (2) differences in schools' baseline student achievement. If schools with larger impacts of pay-for-performance on certain behaviors also had larger impacts on student achievement, this pattern would provide suggestive (correlational) evidence that pay-for-performance might affect student achievement by influencing those behaviors. If schools with higher or lower baseline student achievement tended to experience larger impacts, this information could suggest ways to target pay-for-performance policies to the types of schools that would benefit the most.

The educator behaviors we examined were based on the theory of change for how pay-for-performance might affect student achievement (Chapter I). In an effort to earn pay-for-performance bonuses, principals and teachers may act strategically, shifting attention toward activities that improve measures on which those bonuses are based; they may increase their effort on the job; or they may adopt different teaching practices known to be more effective. To measure these behaviors, we used educators' responses to survey questions on topics that could reflect strategic behavior, effort, and changes in practices (see Appendix G for a list of all of the survey questions used). For each treatment school, we measured the extent to which pay-for-performance (1) promoted strategic behavior (for example, principals' assigning teachers to grades and subjects based

primarily on their ability to improve test scores); (2) increased teacher effort (for example, increased time on instructional activities outside of school hours); and (3) changed teaching practices (for example, teachers' reporting that TIF improved the collaborative nature of teaching). In addition, we used impacts on observation ratings (from administrative data) as an overall measure of impacts on teaching practices.

Changes in teachers' reported behaviors and observation ratings due to pay-for-performance did not explain differences across schools in impacts on student achievement. After four years of pay-for-performance, only one of the 18 relationships we examined between impacts on educator behaviors and impacts on student achievement (nine measures of behaviors and two subjects) was statistically significant (Appendix G, Table G.6). Given the large number of relationships examined, there was an increased likelihood that the single significant finding could have occurred just by chance. In fact, the single relationship that was significant after four years of implementation was not significant after three years (Wellington et al. 2016). Likewise, only one other relationship was significant after three years, and that relationship was not significant after four years. Overall, we found no consistent evidence of any relationship between impacts on educator behaviors and impacts on student achievement.

In addition, we examined whether differences in student achievement across schools before the first year of implementation could explain differences in the degree to which pay-for-performance improved student achievement. For example, schools with lower baseline student achievement may benefit more from increases in teacher effectiveness because they have more room to improve. However, having lower baseline student achievement may indicate that the school faces more challenges—such as a dysfunctional school culture, challenging student home environments, or high student mobility—that render student achievement less responsive to policy interventions.

Differences across schools in average student achievement before implementing pay-for-performance did not explain differences in the impacts of pay-for-performance on student achievement. The relationship between schools' baseline student achievement and the subsequent impacts on student achievement were small and not statistically significant in either math or reading (Appendix G, Table G.7).

Cost-Effectiveness of Pay-for-Performance

Districts have limited budgets and must choose how to allocate their funds among many different policies. Before implementing a pay-for-performance program, districts would likely want to compare the benefit of this policy—for example, improved student achievement—to its cost and consider whether alternative policies could achieve these benefits in a more cost-effective manner. Findings from Chapter IV indicate that the costs of the pay-for-performance bonuses in TIF were not substantial. The average bonus for teachers and principals in treatment schools was about $2,000 and $4,000 in each year, respectively. Setting aside potential administrative expenses from implementing pay-for-performance (discussed in Appendix H), the cost of pay-for-performance bonuses amounted to an additional $100 per student per year beyond the cost of the automatic 1 percent bonuses for educators in control schools. These findings raise the question of whether the impact of pay-for-performance, even if small, was worth the expense. To explore this question, we conducted a cost-effectiveness analysis to determine how costly it would have been for districts to achieve the same impact if they had instead implemented either of two alternative policies to improve student achievement: (1) transfer incentives for high-performing teachers to move to low-

performing schools and (2) class-size reduction. In this section, we describe the main findings from the cost-effectiveness analysis. Appendix H provides a comprehensive discussion of the cost-effectiveness analysis approach and findings.

Both transfer incentives and class-size reduction are policies that researchers have rigorously evaluated and districts could conceivably implement instead of pay-for-performance. Compared to pay-for-performance, transfer incentives target a similar population of schools (high-poverty) and are aimed at improving student achievement through a similar channel (improving teacher quality). In a random assignment study of a transfer incentive program that gave bonuses to high-performing teachers to teach in low-performing schools for two years, researchers found positive impacts on elementary students' achievement in the second year (Glazerman et al. 2013). Class-size reduction, on the other hand, is designed to improve student achievement through channels other than teacher quality, but it is a frequently discussed policy and is commonly used as a benchmark for cost-effectiveness studies of educational interventions (for example, Levin et al. 1987; Yeh 2009; Borman et al. 2002; Glazerman et al. 2013). A random assignment study of class-size reduction that included students for four consecutive years (from kindergarten through grade 3) found that students in small classes had higher achievement than students in regular-sized classes in all four years (for example, Finn and Achilles 1990; Krueger 1999; Nye et al. 2000; Schanzenbach 2006).

For each policy, we used information from either this evaluation (for pay-for-performance) or other published studies (for transfer incentives and class-size reduction) to calculate the policy's cumulative cost per student and cumulative impact on student achievement from the start of the policy through the end of each year of implementation. This information enabled us to calculate the per-student cost needed to achieve an impact of 0.04 standard deviations in math and reading averaged together—the actual average impact of pay-for-performance. When comparing policies, we assumed that districts would specify a particular length of time for implementing a policy and would like to identify which policy would be the most cost-effective way to raise student achievement over that specified time. Therefore, we compared policies based on the same duration of implementation. We could compare pay-for-performance and class-size reduction after one to four years of each policy, but could compare pay-for-performance and transfer incentives after only one or two years, the duration of the transfer incentives study.

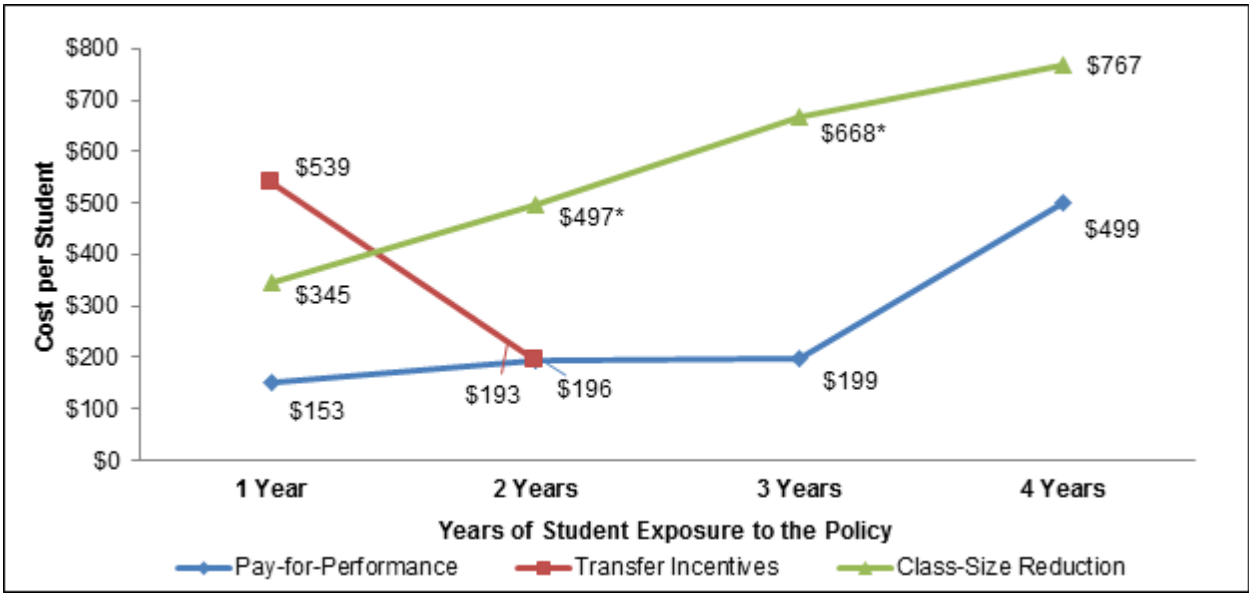
This analysis did not cover the full range of alternative policies that districts could implement instead of pay-for-performance. For example, rather than financing pay-for-performance, districts might consider channeling resources into adopting new curricula, improving educational technology, or enhancing professional development. Our objective was not to compare pay-for-performance to a comprehensive set of alternative policies, which would require an extensive review of available information on costs and impacts. Instead, we analyzed a two policies—transfer incentives and class-size reduction—that serve as convenient but well-researched benchmarks for interpreting how costly it would be to raise achievement through pay-for-performance.

At the end of two years, pay-for-performance and transfer incentives were similar in their cost-effectiveness. To raise student achievement by 0.04 standard deviations after two years, transfer incentives would require $193 per student, nearly identical to the $196 per-student cost of pay-for-performance (Figure VI.6). Transfer incentives were initially less cost-effective than pay-for-performance—though not by a statistically significant difference—due to the high start-up costs associated with recruiting high-performing teachers to transfer and the small size of the impacts observed in the first year. However, because larger impacts of transfer incentives emerged in the

second year while the start-up costs did not recur, transfer incentives became as cost-effective as pay-for-performance at the end of two years.

Pay-for-performance was generally more cost-effective than class-size reduction. In each policy duration that we studied (one to four years), pay-for-performance required about one-third to two-thirds the cost that class-size reduction would have needed to achieve the same impact on student achievement (Figure VI.6). For example, after four years, raising student achievement by 0.04 standard deviations required spending $499 per student on pay-for-performance, but would have required spending $767 per student on class-size reduction. Differences in cost-effectiveness between pay-for-performance and class-size reduction were statistically significant in two of the four durations that we examined (two years and three years).

Figure VI.6. Cumulative Cost per Student Needed to Achieve the Actual Impact of Pay-for-Performance, by Years of Student Exposure to the Policy



Sources: Impacts and costs of pay-for-performance are based on educator and student administrative data and the Common Core of Data; impacts and costs of transfer incentives are based on Glazerman et al. (2013); impacts of class-size reduction are based on Nye et al. (2000) with standard errors approximated from Schanzenbach (2006); costs of class-size reduction are based on Harris (2009).

Notes: Figure shows the per-student cost of obtaining a 0.04 standard deviation impact on test scores. Impacts on math and reading scores were averaged, giving equal weight to both subjects. All costs were adjusted for inflation and expressed in May 2016 dollars. All costs and impacts were discounted at a 3 percent discount rate to their value in the first year of the policy.

Figure reads: To raise student achievement by 0.04 standard deviations at the end of one year, transfer incentives would require $539 per student, class-size reduction would require $345 per student, and pay-for-performance would require $153 per student.

*Difference in cost-effectiveness from pay-for-performance is statistically significant at the .05 level, two-tailed test.

The three policies included in our analysis were implemented with the goal of improving student achievement overall. Therefore, our main analysis focused on their cost-effectiveness at raising math and reading achievement averaged together. However, the cost-effectiveness of a policy could vary by subject area, and policymakers might conceivably be interested in implementing a policy in only one subject area (for example, a hard-to-staff subject). When we examined the cost-effectiveness of these policies separately for math and reading, the results were generally consistent

with the main results. Compared to pay-for-performance, transfer incentives would require a somewhat higher cost ($91 more per student) to raise math achievement by 0.04 standard deviations after two years and a somewhat lower cost ($96 less per student) to raise reading achievement, but neither of those differences was statistically different (Appendix H, Table H.3). Across all subjects and policy durations, class-size reduction required a higher cost (an extra $122 to $557 per student) than pay-for-performance to achieve the same impact, but published information was not available to test whether those differences were statistically significant.

Although districts may want to select the most cost-effective ways to produce sustained, long-term gains in student achievement, our analysis cannot assess the cost-effectiveness of these policies beyond the two or four years in which they were implemented and studied. If differences in cost-effectiveness between policies were guaranteed to be stable, then findings from the short policy durations for which evidence is available could provide insights on the policies' cost-effectiveness over the longer term. However, we found that these differences were not stable. As shown in Figure VI.6, pay-for-performance and class-size reduction became less cost-effective over time as impacts did not grow at the pace that costs did, whereas transfer incentives became more cost-effective from the first to second year. Because the policies' cost-effectiveness was still changing at the time that the evaluations ended, we cannot project which policies would be most cost-effective over the longer term.

Summary and Connections to Other Key Findings from the Evaluation Districts

A primary objective of TIF grants is to raise student achievement in high-need schools. Within the evaluation districts, this study found that the pay-for-performance component of TIF had small, positive impacts on student achievement by the second year of implementation. From that year onward, reading and math achievement was higher by 1 to 2 percentile points in schools that offered performance bonuses than in schools that did not.

To draw lessons from these impact findings for future policies on performance-based compensation, it is useful to consider possible explanations for why performance bonuses within the TIF program had any positive impacts on student achievement and why those impacts were small. According to the rationale behind pay-for-performance bonuses (Chapter I), they can improve student achievement only if educators (1) face a bonus structure that provides meaningful incentives for improvement, (2) understand key components of the program, (3) feel motivated to adjust their practices or their choice of where to work to earn these bonuses, and (4) know how to change their practices in ways that improve student achievement. As we summarize next, some, but not all, of these factors were in place within the evaluation districts.

First, the structure of the bonuses provided a mix of stronger and weaker incentives for teachers to improve (Chapter IV). The highest-performing teachers could earn a performance bonus worth about four times the average bonus, which provided an incentive for teachers to demonstrate high performance. However, in each year, the criteria for earning any bonus resulted in about 70 percent of teachers earning bonuses within schools that offered them. Therefore, even some teachers who performed worse than the typical teacher earned a bonus. If failing to receive a bonus represented a clear signal about having room for improvement, the bonus structure gave a minority of teachers this type of encouragement to improve.

Second, educators' misunderstanding of these bonuses may have hampered the degree to which this policy could influence educators' behavior (Chapter IV). Across all four years of implementation, many teachers were unaware they were eligible for a performance bonus or underestimated the amount they could earn. These teachers perceived limited or no monetary incentives to improve their performance even though their districts had actually structured the bonuses to provide stronger incentives.

Third, some teachers perceived schools that offered performance bonuses to be a more appealing place to work, potentially enhancing their motivation to remain at these schools and improve their practices (Chapter V). By the third year of TIF implementation, teachers' satisfaction with key aspects of their jobs was either unchanged or improved as a result of pay-for-performance. However, we have no evidence that these favorable impacts on teachers' job satisfaction contributed to improvements in student achievement. In fact, positive impacts on student achievement emerged in the first two years—a period when pay-for-performance was actually *lowering* satisfaction.

Fourth, it is unclear whether teachers knew how to change their classroom practices in ways that could improve student achievement. By the third year of implementation, pay-for-performance led to small increases in teachers' classroom observation ratings (Chapter VI). However, schools that experienced larger impacts of pay-for-performance on observation ratings did not experience larger impacts on student achievement. Therefore, we found no indication that changes in teachers' measured practices were the source of the improvements in student achievement. The disconnect between changes in measured practices and changes in achievement could have been due to a number of factors, which this study did not have the data to examine. One possibility is that the amount of targeted professional development teachers received—no more than six hours over the school year (Chapter IV)—was insufficient to promote changes in practices that were substantial enough to improve student achievement. Another possibility is that the observation measures encouraged teachers to focus on aspects of instruction that were not related to student achievement.

Although the impacts of pay-for-performance were small, the costs of the bonuses were also low enough such that this policy was at least as cost-effective as some alternative policies that have been evaluated. Specifically, a cost-effectiveness analysis suggests that pay-for-performance was more cost-effective than class-size reduction (through four years of program implementation) and about as cost-effective as providing transfer incentives for high-performing teachers to move to low-performing schools (at the end of two years). However, the available evidence cannot predict the policies' cost-effectiveness beyond the limited number of years in which these policies were implemented and evaluated.

VII. EDUCATORS' ATTITUDES AND DISTRICTS' PLANS TOWARD CONTINUING KEY TIF COMPONENTS

Although the 2014–2015 school year was the final year of the 2010 TIF grants, TIF grantees were expected to sustain their programs beyond the years covered by the grant. Districts' plans to continue implementing aspects of their TIF programs into the 2015–2016 school year may depend on educators' support for the program and districts' financial considerations. In this chapter, we present views about the future of the TIF components from several perspectives. First, we present educators' attitudes toward continuing components of their TIF programs based on survey responses from teachers and principals in the evaluation districts. Second, we report on the evaluation districts' plans to continue implementing the four TIF requirements based on interviews with the districts' TIF administrators. Finally, we present all 2010 TIF districts' plans to continue implementing key components of TIF based on responses to the district survey. All data on educators' attitudes and districts' plans toward continuing TIF components come from their responses near the end of the final year (2014–2015) of the grant.

Key Findings About Educators' Attitudes and Districts' Plans Toward Continuing TIF

- **Most teachers supported continuing the use of classroom observations to measure their performance, but only about half supported continuing the use of student achievement growth.**
- **Most principals supported continuing to measure educators' performance based on both student achievement growth and observations of classroom or school practices.**
- **Most teachers and principals thought districts should continue to offer performance bonuses.**
- **Many 2010 TIF districts reported planning to continue three components of TIF—multiple measures of educator performance, additional pay opportunities, and professional development—after their TIF grants ended, although slightly fewer than half of these districts planned to offer pay-for-performance bonuses.**

Educators' attitudes and districts' plans toward continuing TIF components may have been influenced by the federal education policies that were in place at the end of the 2014–2015 school year. At that time, federal education policy was governed by the *No Child Left Behind Act* (NCLB; Public Law 107-110), which required states to hold schools accountable for meeting a number of performance targets. However, more than 40 states had received enhanced flexibility from the law's accountability requirements in exchange for undertaking various reforms. Two reforms that these states committed to implement were similar to TIF components: (1) evaluating educators based on student achievement growth and observations and (2) using the results of these evaluations to guide professional development. All evaluation districts and nearly all 2010 TIF districts (98 percent) were in states that had received enhanced flexibility, which may have influenced districts' plans for 2015–2016 and the attitudes of their educators. Since then, the enactment of the *Every Student Succeeds Act* (ESSA; Public Law 114-95), the replacement for NCLB, has given states more latitude over how to design their educator evaluation systems beginning in the 2016–2017 school year, which may alter districts' plans over the longer term.

Educators' Attitudes Toward Continuing TIF Evaluation Measures and Pay-for-Performance

Educators' attitudes toward key TIF components could affect districts' decisions to continue implementing all or part of the TIF program. This section reports the opinions of teachers and principals in evaluation districts on the measures that should be used to evaluate educators and whether educators should receive pay-for-performance bonuses for the 2015–2016 school year.

Most teachers supported continuing the use of classroom observations to measure their performance, but only about half supported continuing the use of student achievement growth. Three-quarters of teachers supported using multiple observations by trained observers to measure their effectiveness in the 2015–2016 school year (Table VII.1). Fewer teachers (48 percent) supported the use of student achievement growth. On both measures, opinions were similar among teachers in treatment and control schools.

Table VII.1. Teachers' Attitudes Toward Continuing TIF Performance Measures and Pay-for-Performance (Percentages Who Agree or Strongly Agree)

	All Teachers	Treatment	Control	Impact
For the 2015–2016 School Year, the District Should:				
Use student achievement growth to measure teacher performance	48	50	45	5
Use multiple observations conducted by trained observers to measure teacher effectiveness	75	77	73	5
Provide additional pay based on teacher performance	66	70	61	8*
Number of Teachers—Range[a]	**773-778**	**385-387**	**388-391**	

Source: Teacher survey (2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Impact is statistically significant at the .05 level, two-tailed test.

Most principals supported continuing to measure educators' performance based on student achievement growth and observations of classroom or school practices. Over 90 percent of principals agreed that their districts should use multiple observations to measure their own performance (93 percent) and that of their teachers (92 percent) in 2015–2016 (Table VII.2). At least 75 percent responded that student achievement growth should be used. These views were similar among principals of treatment and control schools.

Most teachers and principals thought districts should continue to offer performance bonuses. Overall, about two-thirds of teachers supported continuing to reward teachers based on their performance in the 2015–2016 school year. Treatment teachers, those who could benefit from performance bonuses, were more likely to support continuing the bonuses than teachers in control schools (70 versus 61 percent; Table VII.1). More than 80 percent of principals believed that their district should continue to award bonuses to teachers (86 percent) and principals (81 percent; Table VII.2). Views on performance bonuses were similar among principals of treatment and control schools.

Table VII.2. Principals' Attitudes Toward Continuing TIF Performance Measures and Pay-for-Performance (Percentages Who Agree or Strongly Agree)

	All Principals	Treatment	Control	Impact
For the 2015-2016 School Year, the District Should:				
Use student achievement growth to measure teacher performance	84	86	82	5
Use multiple observations to measure teacher performance	92	87	93	-6
Use student achievement growth to measure principal performance	76	74	76	-1
Use multiple observations to measure principal performance	93	91	95	-4
Award performance bonuses to teachers	86	86	85	1
Award performance bonuses to principals	81	82	79	3
Number of Principals—Range[a]	**118-122**	**57-60**	**61-62**	

Source: Principal survey (2015).

Note: None of the impacts are statistically significant at the .05 level, two-tailed test.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

Evaluation Districts' Plans to Continue Key TIF Components

Districts might plan to retain parts of the TIF program in their existing form, revise aspects of it, or discontinue particular components entirely. As noted in the previous section, educators in evaluation districts felt favorably about performance measures and thought districts should continue to offer performance bonuses. Evaluation districts might have taken educators' preferences into consideration, although other factors likely influenced their decisions to continue aspects of TIF as well (such as financial considerations). During interviews at the end of the TIF grant, we asked TIF district administrators about their plans for continuing key components of the TIF program.

Almost all evaluation districts reported planning to implement three key components of TIF—measures of educator performance, additional pay opportunities, and professional development—in the 2015–2016 school year. At least 80 percent of districts expected to continue to evaluate educators' performance with measures broadly similar to those used in TIF, provide extra pay for teachers who took on additional roles or responsibilities, and provide teachers with professional development based on their own performance (Table VII.3). Yet, most districts reported they expected to revise how they implemented these program components. For example, half of districts expected to revise the measures used to evaluate educators. Examples of such revisions included replacing the observation rubric used for TIF with the district observation tool and decreasing the number of observations from four to two observations per year. Eight of the 10 districts expected to offer extra pay for those taking on additional roles or responsibilities, but 6 of these districts expected to decrease the number of positions or amount of pay offered. Although all districts expected to provide professional development similar to what they offered under the TIF grant, 7 of the 10 expected to revise what they provided, such as offering professional development only based on what teachers had found helpful or on topics that the district believed were effective.

Table VII.3. Evaluation Districts' Reported Plans to Retain, Revise, or Discontinue TIF Program Components in 2015–2016 (Percentages)

	Retain in Existing Form	Revise	Discontinue
Evaluation Measures	40	50	10
Pay-for-Performance Bonus[a]	0	35	65
Extra Pay for Additional Roles or Responsibilities	20	60	20
Professional Development	30	70	0
Number of Districts		**10**	

Source: District interviews, 2015.

Notes: One evaluation grantee that consisted of multiple charter schools indicated that the decision to continue, revise, or discontinue TIF components was up to each charter school. In each row of the table, this grantee is included in the response category that reflects the TIF administrator's expectation for what most of the charter schools would be doing.

[a]One district indicated that pay-for-performance bonuses might be revised or discontinued, so this district was assumed to have a 50 percent chance of revising these bonuses and a 50 percent chance of discontinuing them.

Most evaluation districts expected to discontinue pay-for-performance bonuses in the 2015–2016 school year, primarily because they lacked funding or believed those bonuses were not effective. Nearly two-thirds (65 percent) of the evaluation districts expected to discontinue performance bonuses completely (Table VII.3).[64] The two most common reasons cited for discontinuing the bonuses were the belief that they did not improve teacher effectiveness (four districts) and lack of funding (four districts).[65] The remaining districts (35 percent) expected to offer performance bonuses under a revised structure—for example, offering smaller bonuses than those awarded through their TIF grants. None of the districts expected to offer pay-for-performance bonuses under the same structure as TIF.

All 2010 TIF Districts' Plans to Continue Key TIF Components

In addition to describing the evaluation districts' plans for continuing key components of TIF, we also sought to understand the plans of the 2010 TIF districts more broadly. To do so, we used responses from the district survey. The district survey asked all 2010 TIF districts about their plans for continuing the components of TIF (without distinguishing whether continuing a component would involve retaining it in its existing form or revising it), and asked about planned funding sources for sustaining TIF.

Many TIF districts reported planning to continue key components of their TIF program after their grant ended, although slightly fewer than half of districts planned to offer pay-for-

[64] We found no differences in characteristics (including student population, location, and state employment policies) of evaluation districts that reported planning to offer pay-for-performance bonuses compared with those that did not plan to do so (not shown).

[65] The four districts that cited insufficient improvement in teacher effectiveness as a reason for discontinuing bonuses did, in fact, tend to experience below-average impacts of pay-for-performance on student achievement after four years of implementation. In math, impacts in all four districts were less favorable than the average impact. In reading, impacts in three of the four districts were less favorable than the average impact.

performance bonuses to educators. At least 80 percent of TIF districts reported they planned to continue to evaluate teachers and principals based on student achievement growth and multiple observations of classroom or school practices in the 2015–2016 school year (Table VII.4). Most districts also reported they would continue to offer teachers extra pay for taking on additional roles or responsibilities (74 percent) and provide teachers with professional development and feedback based on their actual performance ratings to help improve their instructional practices (90 percent). However, slightly fewer than half of the districts reported they planned to offer teachers (47 percent) or principals (45 percent) bonuses based on performance.

Table VII.4. All 2010 TIF Districts' Reported Plans to Continue TIF Components in 2015–2016 (Percentages)

	Districts' Plans for	
	Teachers	Principals
Performance Measures		
Include a measure of student achievement growth in the performance evaluation	83	80
Include multiple observations in the performance evaluation	89	80
Offer Pay-for-Performance Bonus		
Based solely on performance (for example, as measured by student achievement growth, observations, or a combination of measures)	47	45
Offer Additional Pay		
For taking on additional roles or responsibilities	74	20
For teaching in a hard-to-staff school or subject	22	11
Professional Development		
Provide targeted professional development to improve performance identified by the performance evaluation	90	—
Number of Districts—Range[a]	**134-136**	**131-133**

Source: District survey, 2015.

Note: Table reports on districts' implementation plans for at least some of their TIF schools.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

— is not available.

Districts that choose to continue to implement TIF components, including pay-for-performance bonuses, may have to acquire additional funding or obtain support from key stakeholders. For districts that indicated they planned to continue or expand aspects of their TIF program, the most common source of planned funding was federal funding—either a new TIF grant (28 percent) or other federal funding (58 percent; Table VII.5).

Districts' decisions to continue offering pay-for-performance bonuses could have depended on several characteristics of the districts' socioeconomic, institutional, and political context. To explore these factors, we compared the characteristics of districts that planned to continue offering pay-for-performance bonuses and those that did not. We focused on three types of characteristics: (1) socioeconomic disadvantage, which could contribute to low student achievement and motivate districts to continue reforms aimed at raising achievement; (2) district size, given that larger districts may have a greater administrative capacity to carry out pay-for-performance, but may also face the challenge of satisfying a wider range of stakeholders; and (3) the presence of right-to-work laws, which tend to weaken teachers' unions, a potential source of opposition to pay-for-performance. Of these characteristics, only one was associated with districts' decisions to continue offering pay-for-performance bonuses. Compared with districts that did not plan to continue offering pay-for-

performance bonuses, districts that planned to do so were more likely to be in states with right-to-work laws (88 versus 44 percent; Table VII.6).

Table VII.5. All 2010 TIF Districts' Plans for Funding TIF Program Components in 2015–2016 (Percentages)

	All TIF Districts that Plan to Continue or Expand TIF
Funding Source to Continue or Expand	
New TIF grant	28
Other federal funding	58
Revised salary schedule	38
New source of district or state funding	34
Outside funding (such as philanthropy)	27
Number of Districts—Range[a]	**85-88**

Source: District survey, 2015.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

Table VII.6. Comparison of Characteristics of Districts that Plan and Do Not Plan to Continue Offering Teachers Pay-for-Performance Bonuses in 2015–2016 (Percentages Unless Otherwise Noted)

	Plan to Continue Pay-for-Performance Bonuses	Do Not Plan to Continue Pay-for-Performance Bonuses
Student Socioeconomic Status		
Eligible for free or reduced-price lunch	67	64
Title 1-eligible schools (schoolwide)	80	77
Enrollment (average)		
Number of students	29,840	16,468
Collective Bargaining		
In state with right-to-work laws[a]	88	44*
Number of States	**15**	**24**
Number of Districts—Range[b]	**57-64**	**70-72**

Sources: District survey (2015) and Common Core of Data for 2013–2014 school year.

Notes: Table reports on districts' plans to offer pay-for-performance bonuses based solely on performance (for example, as measured by student achievement growth, observations, or a combination of measures). The table is based on all 148 districts that implemented TIF in 2014–2015. Seven non-evaluation districts were not included in the 2013–2014 district-level data from the Common Core of Data. Common Core of Data school-level data are used to calculate socioeconomic indicators. Common Core of Data district-level data are used to calculate enrollment.

[a]The presence of right-to-work laws is a state-level indicator from the National Right-to-Work Legal Defense Foundation. Right-to-work laws prohibit unions from requiring nonmembers to pay fees, and such laws tend to be associated with weaker unions.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference between districts that plan to continue and those that do not plan to continue to offer pay-for-performance bonus is statistically significant at the .05 level, two-tailed test.

Summary

The TIF grant expected grantees to sustain their programs beyond the five years of their grants, but districts' plans to do so may also depend on their experiences implementing TIF. We found that most districts implemented each of the key components of TIF and encountered few major challenges (Chapters III and IV). Some evaluation districts saw improvements in student achievement in schools that offered performance bonuses compared to those that did not (Chapter VI). As discussed in this chapter, educators in the evaluation districts also generally had favorable attitudes toward continuing key components of their TIF program, including pay-for-performance bonuses. All of these factors could have provided favorable conditions for districts to continue implementing some components of their TIF program.

Although most 2010 TIF districts and most evaluation districts reported that they planned to continue to implement three of the key components of their TIF program—using TIF measures to evaluate educators, offering additional pay opportunities, and providing professional development based on teachers' individual performance—fewer than half of these districts reported that they planned to continue offering pay-for-performance bonuses. The evaluation districts reported that lack of funding was a common reason for discontinuing these bonuses, which is consistent with the fact that most 2010 TIF districts regarded sustainability of their TIF program to be a major challenge (Chapter III).

THIS PAGE IS INTENTIONALLY BLANK

REFERENCES

Angrist, Joshua D., and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press, 2009.

Balch, Ryan, and Matthew Springer. "Performance Pay, Test Scores, and Student Learning Objectives." *Economics of Education Review*, vol. 44, 2015, pp. 114-125.

Bayonas, Holli. "Guilford County Schools Mission Possible Program: Year 3 (2008–09) External Evaluation Report." Greensboro, NC: University of North Carolina at Greensboro, SERVE Center, 2010.

Borman, Geoffrey, Gina M. Hewes. "The Long-Term Effects and Cost-Effectiveness of Success for All." *Educational Evaluation and Policy Analysis*, vol. 24, no. 4, Winter 2002, pp. 243–266.

Caulkins, Jonathan P., Susan S. Everingham, C. Peter Rydell, James Chiesa and Shawn Bushway. "An Ounce of Prevention, a Pound of Uncertainty: The Cost-Effectiveness of School-Based Drug Prevention Programs." Santa Monica, CA: RAND Corporation, 1999.

Cellini, Stephanie Riegg and James Edwin Kee, "Cost-Effectiveness and Cost-Benefit Analysis," Chapter 21 of *Handbook of Practical Program Evaluation*, Third Edition, edited by Joseph S. Wholey, Harry P. Hatry, and Kathryn E. Newcomer. San Francisco, CA: Jossey-Bass, 2010: 493-530.

Chiang, Hanley, Alison Wellington, Kristin Hallgren, Cecilia Speroni, Mariesa Herrmann, Steven Glazerman, and Jill Constantine. "Evaluation of the Teacher Incentive Fund: Implementation and Impacts of Pay-for-Performance After Two Years." Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, September 2015.

Dee, Thomas, and James Wyckoff. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT." *Journal of Policy Analysis and Management*, vol. 34, no. 2, 2015, pp. 267–297.

Donald, Stephen G., and Kevin Lang. "Inference with Difference-in-Differences and Other Panel Data." *Review of Economics and Statistics*, vol. 89, no.2, 2007, pp. 221–233.

Freedman, David A. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics*, vol. 40, no.2, 2008, pp. 180–193.

Finn, Jeremy D. and Charles M. Achilles. "Answers and Questions about Class Size: A Statewide Experiment." *American Educational Research Journal*, vol. 27, 1990, pp. 557-577.

Fryer, Roland. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." *Journal of Labor Economics*, vol. 31, no. 2, 2013, pp. 373-427.

Fryer, Roland, Steven Levitt, John List, and Sally Sadoff. "Enhancing the Efficacy of Teacher Incentives Through Loss Aversion: A Field Experiment." Working paper no. 18237. Cambridge, MA: National Bureau of Economic Research, July 2012.

Fulbeck, Eleanor. "Teacher Mobility and Financial Incentives: A Descriptive Analysis of Denver's ProComp." *Educational Evaluation and Policy Analysis*, vol. 36, no. 1, March 2014, pp. 67–82.

Glazerman, Steven, Allison McKie, and Nancy Carey. "An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year One Impact Report." Washington, DC: Mathematica Policy Research, April 2009.

Glazerman, Steven, and Allison Seifullah. "An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year Two Impact Report." Washington, DC: Mathematica Policy Research, May 2010.

Glazerman, Steven, and Allison Seifullah. "An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) After Four Years." Washington, DC: Mathematica Policy Research, March 2012.

Glazerman, Steven, Ali Protik, Bing-ru Teh, Julie Bruch, and Jeffrey Max. "Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment." NCEE 2013-4003. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, November 2013.

Goldhaber, Dan, and Joe Walch. "Strategic Pay Reform: A Student Outcomes-Based Evaluation of Denver's ProComp Teacher Pay Initiative." *Economics of Education Review*, vol. 31, no.6, 2012, pp. 1067–1083.

Goodman, Sarena, and Lesley Turner. "Does Whole School Performance Pay Improve Student Learning? Evidence from the New York City Schools." *Education Next*, vol. 11, no. 2, Spring 2011.

Harris, Douglas N. "Toward Policy-Relevant Benchmarks for Interpreting Effect Sizes: Combining Effects With Costs." *Educational Evaluation and Policy Analysis*. March 2009, Vol. 31, No. 1, pp. 3–29.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives*, vol. 2, no. 3, 2008, pp. 172–177.

Imberman, Scott, and Michael Lovenheim. "Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System." *Review of Economics and Statistics*, vol. 97, no. 2, 2015, pp. 364–386.

Kane, Thomas J., and Douglas O. Staiger. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*. Seattle, WA: Bill and Melinda Gates Foundation, 2012.

Krueger, Alan B. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics*, May 1999, pp. 497-532.

Levin, Henry, Gene V. Glass, and Gail R. Meinster. "Cost-Effectiveness of Computer-Assisted Instruction." *Evaluation Review*, vol. 11, no. 1, February 1987, pp. 50-72.

Levin, Henry, Patrick McEwan. *Cost-Effectiveness Analysis: Methods and Applications*. Second Edition. Sage, Thousand Oak, CA: 2001.

Liang, Kung-Yee, and Scott L. Zeger. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, vol. 73, no. 1, 1986, pp. 13–22.

Lipscomb, Stephen, Jeffrey Terziev, and Duncan Chaplin. "Measuring Teachers' Effectiveness: A Report from Phase 3 of Pennsylvania's Pilot of the Framework for Teaching." Cambridge, MA: Mathematica Policy Research, 2015.

Marsh, Julie A., Matthew G. Springer, Daniel F. McCaffrey, Kun Yuan, Scott Epstein, Julia Koppich, Nidhi Kalra, Catherine DiMartino, and Art Peng. *A Big Apple for Educators: New York City's Experiment with Schoolwide Performance Bonuses*. Final evaluation report. Santa Monica, CA: RAND Corporation, 2011.

Max, Jeffrey, Jill Constantine, Alison Wellington, Kristin Hallgren, Steven Glazerman, Hanley Chiang, and Cecilia Speroni. "Evaluation of the Teacher Incentive Fund: Implementation and Early Impacts of Pay-for-Performance After One Year." Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, September 2014.

Nye, Barbara, Larry V. Hedges, Spyros Konstantopoulos. "The Effects of Small Classes on Academic Achievement: The Results of the Tennessee Class Size Experiment." *American Educational Research Journal*. Spring 2000, vol. 37, No. 1, pp. 123-151.

Puma, Michael, Robert Olsen, Stephen Bell, and Cristofer Price. "What to Do When Data Are Missing in Group Randomized Controlled Trials." NCEE 2009-0049. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, October 2009.

Rubin, Donald. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, 1987.

Schafer, Joseph, and John Graham. "Missing Data: Our View of the State of the Art." *Psychological Methods*, vol. 7, no. 2, 2002, pp. 147–177.

Schanzenbach, Diane W. "What Have Researchers Learned from Project STAR?" Brookings Papers on Education Policy, No. 9 (2006/2007), pp. 205-228.

Schenker, Nathaniel, and Jeremy Taylor. "Partially Parametric Techniques for Multiple Imputation." *Computation Statistics & Data Analysis*, vol. 22, no. 4, 1996, pp. 425–446.

Shifrer, Dara, Ruth Turley, and Holly Heard. "Houston Independent School District's ASPIRE Program: Estimated Effects of Receiving Financial Awards." Working paper. Houston, TX: Houston Education Research Consortium, 2013.

Slotnik, William, Maribeth Smith, Barbara Helms, and Zhaogang Qiao. "It's More than Money: Teacher Incentive Fund—Leadership for Educators' Advanced Performance. Charlotte-Mecklenburg Schools." Boston, MA: Community Training and Assistance Center, February 2013.

Sojourner, Aaron, Elton Mykerezi, and Kristine West. "Teacher Pay Reform and Productivity: Panel Data Evidence from Adoptions of Q-Comp in Minnesota." *Journal of Human Resources*, vol. 49, no. 4, 2014, pp. 945–981.

Springer, Matthew, Jessica Lewis, Michael Podgursky, Mark Ehlert, Lori Taylor, Omar Lopez, and Art (Xiao) Peng. "Governor's Educator Excellence Grant (GEEG) Program: Year Three Evaluation Report." Nashville, TN: National Center on Performance Incentives, 2009a.

Springer, Matthew, Jessica Lewis, Michael Podgursky, Mark Ehlert, Timothy Grownberg, Laura Hamilton, Dennis Jansen, Brian Stecher, Lori Taylor, Omar Lopez, and Art (Xiao) Peng. "Texas Educator Excellence Grant (TEEG) Program: Year Three Evaluation Report." Nashville, TN: National Center on Performance Incentives, 2009b.

Springer, Matthew, Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel McCaffrey, Matthew Pepper, and Brian Stecher. "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT)." Society for Research on Educational Effectiveness (SREE), 2011.

Springer, Matthew, John Pane, Vi-Nhuan Le, Daniel McCaffrey, Susan Burns, Laura Hamilton, and Brian Stecher. "Team Pay for Performance: Experimental Evidence from the Round Rock Pilot Project on Team Incentives. *Educational Evaluation and Policy Analysis*, vol. 34, no. 4, 2012, pp. 367–390.

Springer, Matthew, Dale Ballou, and Art Peng. "Estimated Effect of the Teacher Advancement Program on Student Test Score Gains: Findings from an Independent Appraisal." *Education Finance and Policy*, vol. 9, no. 2, 2014, pp. 193–230.

Springer, Matthew, and Lori Taylor. "Designing Incentives for Public School Teachers: Evidence from a Texas Incentive Pay Program." *Journal of Education Finance*, vol. 41, no. 3, 2016, pp. 344-381.

Springer, Matthew, Walker Swain, and Luis Rodriguez. "Effective Teacher Retention Bonuses: Evidence from Tennessee." *Educational Evaluation and Policy Analysis*, vol. 38, no. 2, 2016, pp. 199–221.

Wayne, Andrew J., Michael S. Garet, Seth Brown, Jordan Rickles, Mengli Song, and David Manzeske. "Early Implementation Findings From a Study of Teacher and Principal Performance Measurement and Feedback: Year 1 Report." Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, November 2016.

Wellington, Alison, Hanley Chiang, Kristin Hallgren, Cecilia Speroni, Mariesa Herrmann, and Paul Burkander. "Evaluation of the Teacher Incentive Fund: Implementation and Impacts of Pay-for-Performance After Three Years." Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, August 2016.

Yeh, Stuart S. "Class size Reduction or Rapid Formative Assessment? A Comparison of Cost-Effectiveness." *Educational Research Review*, vol. 4, no. 1, 2009, pp. 7-15.

APPENDIX A

SUPPLEMENTAL INFORMATION ON STUDY SAMPLE, DESIGN, DATA, AND METHODS FOR CHAPTER II

THIS PAGE IS INTENTIONALLY BLANK

This appendix provides more detailed information about characteristics of TIF districts, the study design, the teacher survey sample, survey response rates, and sample sizes for analyses using educator and student administrative data.

As discussed in Chapter II, evaluation districts were classified into two cohorts—Cohort 1 and Cohort 2—according to the year in which we randomly assigned their schools to a treatment group or a control group. The 10 districts whose schools were randomly assigned in spring and summer 2011 were classified as Cohort 1. Three additional districts, whose schools were randomly assigned in spring and summer 2012, were classified as Cohort 2. By the time of this report, the 2010 TIF grants had ended. Cohort 1 completed four years of implementation, 2011–2012, 2012–2013, 2013–2014, and 2014–2015, referred to as Years 1, 2, 3, and 4. Cohort 2 districts completed three years of implementation, 2012–2013, 2013–2014, and 2014–2015, referred to as Years 1, 2, and 3 for this cohort.

Random Assignment of Schools to the Treatment and Control Groups

To randomly assign schools within a district to the treatment and control groups, we used a matched-pair randomization approach designed to maximize the balance between the treatment and control groups on observable characteristics. Specifically, we used two approaches: (1) creating matched pairs of schools, and (2) creating matched groups of schools.

Matched pairs of schools. We randomly assigned most of the schools (72 of 138 Cohort 1 schools, and 42 of 45 Cohort 2 schools) to treatment and control groups within matched pairs of schools. One school in each pair was randomly selected to be in the treatment group; the other school was assigned to the control group. Within each district, pairs were constructed so the schools that were paired together would (1) have identical sets of grades represented; (2) be similar in average student achievement; and (3) be similar on other characteristics, such as school size, percentage of students eligible for free or reduced-price lunch, and racial/ethnic composition. District staff either approved the pairs that we constructed or directly specified the pairs based on their knowledge of the participating schools. Because pairing reduced the chance that randomization would produce treatment and control groups with large baseline differences, it enhanced precision for estimating the impacts of pay-for-performance bonuses.

Matched groups of schools. For the remaining schools (66 of 138 Cohort 1 schools, and 3 of 45 Cohort 2 schools), we randomly assigned groups of schools to treatment and control groups within matched pairs of groups. This was analogous to the matched-pairs procedure described previously, except that we assigned groups of schools within matched pairs of groups rather than assigning individual schools within matched pairs of individual schools. We used this approach when the randomization had to satisfy constraints that could not be met with paired random assignment of individual schools. For example, some districts requested that certain schools be assigned to the same treatment status if they were expected to be consolidated in the future or were in the same feeder pattern (for instance, grouping a middle school with the elementary schools from which its students typically came). Moreover, in some districts, all participating schools in the district were grouped into two groups that were well matched on average baseline characteristics; this was done to address concerns that several individual schools would not have had suitable matches if pairs of individual schools had been constructed. As with the pairing of individual schools described earlier, the pairing of groups of schools was designed to minimize the chance that randomization would produce treatment and schools that were dissimilar on baseline characteristics.

School Attrition

For our primary analysis in Chapters IV through VI, we focus on Cohort 1 schools that had implemented TIF for four full years (Year 1 is 2011–2012, Year 2 is 2012–2013, Year 3 is 2013–2014, and Year 4 is 2014–2015). Of the 138 Cohort 1 schools that were randomly assigned, 7 schools were dropped from all analyses to keep a constant analysis sample of 131 schools each year. After the first year of TIF implementation, these 7 schools either closed, chose to drop out of the study, were consolidated, or were matched to a school that closed, dropped out, or was consolidated. The results based on Cohorts 1 and 2 (shown in later appendices) include schools that have implemented TIF for at least three years. These supplemental analyses of Years 1 and 2 are based on a constant analysis sample of 170 Cohorts 1 and 2 schools, out of the total 183 schools that were randomly assigned.

As Table A.1 shows, school attrition was low, ranging from 4.3 to 5.8 percent in analyses for Cohort 1 and 6.6 to 7.7 percent for Cohorts 1 and 2. Difference in the attrition rate between treatment and control schools was also small (largest differential attrition was 1.5 percent).

Table A.1. School Attrition, Cohorts 1 and 2 (Percentages Unless Otherwise Noted)

	Overall	Treatment	Control	Differential Attrition
Cohort 1				
Number of Schools Randomly Assigned	138	69	69	NA
Analyses of Student, Educator Administrative Data[a] and Teacher Survey Data				
Number of schools in Year 1 analyses	131	65	66	NA
Number of schools in Year 2 analyses	131	65	66	NA
Number of schools in Year 3 analyses	131	65	66	NA
Number of schools in Year 4 analyses	131	65	66	NA
Attrition rate, Year 1	5.1	5.8	4.3	1.5
Attrition rate, Year 2	5.1	5.8	4.3	1.5
Attrition rate, Year 3	5.1	5.8	4.3	1.5
Attrition rate, Year 4	5.1	5.8	4.3	1.5
Analyses of Principal Survey Data				
Number of schools in Year 1 analyses	130	65	65	NA
Number of schools in Year 2 analyses	131	65	66	NA
Number of schools in Year 3 analyses	131	65	66	NA
Number of schools in Year 4 analyses	131	65	66	NA
Attrition rate, Year 1	5.8	5.8	5.8	0.0
Attrition rate, Year 2	5.1	5.8	4.3	1.5
Attrition rate, Year 3	5.1	5.8	4.3	1.5
Attrition rate, Year 4	5.1	5.8	4.3	1.5
Cohorts 1 and 2				
Number of Schools Randomly Assigned	183	92	91	NA
Analyses of Student, Educator Administrative Data[a] and Teacher Survey Data				
Number of schools in Year 1 analyses	170	85	85	NA
Number of schools in Year 2 analyses	170	85	85	NA
Number of schools in Year 3 analyses	170	85	85	NA
Attrition rate, Year 1	7.1	7.6	6.6	1.0
Attrition rate, Year 2	7.1	7.6	6.6	1.0
Attrition rate, Year 3	7.1	7.6	6.6	1.0
Analyses of Principal Survey Data				
Number of schools in Year 1 analyses	169	85	84	NA
Number of schools in Year 2 analyses	170	85	85	NA
Number of schools in Year 3 analyses	170	85	85	NA
Attrition rate, Year 1	7.7	7.6	7.7	-0.1
Attrition rate, Year 2	7.1	7.6	6.6	1.0
Attrition rate, Year 3	7.1	7.6	6.6	1.0

Notes: The primary analyses in the main body of the report are based on schools that implemented the program for four years (Cohort 1). Supplemental analyses are based on study schools that implemented the program for at least three years (Cohorts 1 and 2) and are reported in the appendices.

[a]Includes analyses of educator performance ratings.

NA is not applicable.

Baseline Characteristics of Treatment and Control Schools

By virtue of random assignment, treatment and control schools should have similar characteristics at the time of randomization. In Chapter II, we examined whether random assignment produced treatment and control groups that were equivalent at the beginning of the study (the 2010–2011 school year) for the Cohort 1 schools in our main analyses. Tables A.2 and A.3 show similar information for study schools in Cohorts 1 and 2 within the analysis sample.

We lacked baseline data on educators for one of the 10 Cohort 1 districts; therefore, in Chapter II, we showed educator characteristics at the beginning of Year 1. Of the 131 Cohort 1 schools in the final analysis sample, 19 were in the district that did not provide pre-implementation information. Table A.4 shows pre-implementation characteristics for the 112 schools in the nine Cohort 1 districts that provided us with educator characteristics in the pre-implementation year.

Table A.2. Characteristics of Students Enrolled in Treatment and Control Schools in the Pre-Implementation Year, Cohorts 1 and 2 (Percentages Unless Otherwise Noted)

	Treatment	Control	Difference
Achievement in the Pre-Implementation Year (average *z*-score)			
Math	-0.54	-0.51	-0.03
Reading	-0.49	-0.47	-0.02
Race or Ethnicity			
White, non-Hispanic	25	27	-2*
Black, non-Hispanic	47	46	1
Hispanic	22	20	2*
Other	6	7	0
Other Characteristics			
Female	48	49	-1
Eligible for free or reduced-price lunch	80	79	1
Disabled or has an Individualized Education Program	14	14	0
Over age for grade	13	13	0
English language learner	9	8	0
Grade Span			
Grades 3–5	63	63	0
Grades 6–8	37	37	0
Test of Whether Characteristics jointly Predict Treatment Status: *p*-value			0.16
Number of Students—Range[a]	**18,895-29,138**	**18,725-28,759**	
Number of Schools—Range[a]	**59-85**	**59-85**	

Source: Student administrative data.

Notes: The pre-implementation year is 2010–2011 for Cohort 1 and 2011–2012 for Cohort 2. The difference between the treatment and control estimates may not equal the difference shown in the table because of rounding.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

Table A.3. Characteristics of Educators in Treatment and Control Schools in Year 1, Cohorts 1 and 2 (Percentages Unless Otherwise Noted)

	Teachers			Principals		
	Treatment	Control	Difference	Treatment	Control	Difference
Demographic Characteristics						
Female	85	84	1	63	59	5
Race or ethnicity						
White, non-Hispanic	76	75	1	60	55	4
Black, non-Hispanic	18	19	-1	33	37	-5
Hispanic or Other	7	6	0	8	7	0
Age (average years)	42	42	1*	50	47	2*
Education						
Master's degree or higher	59	59	0	95	94	1
Experience in K–12 Education						
Total experience (average years)	12	12	0	16	14	2
Fewer than 5 years	22	23	-1	19	13	6
5-15 years	45	46	-1	32	43	-10
More than 15 years	33	31	2	49	45	4
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value			0.24			0.43
Number of Educators—Range[a]	**2,182-2,907**	**2,162-2,833**		**48-84**	**54-87**	
Number of Schools—Range[a]	**67-85**	**67-85**		**46-82**	**52-83**	

Source: Educator administrative data.

Notes: Year 1 is 2011–2012 for Cohort 1 and 2011–2012 for Cohort 2. The difference between the treatment and control estimates may not equal the difference shown in the table because of rounding. The number of principals exceeds the number of schools in the analysis sample because a few schools had more than one principal.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

Selection of the Teacher Survey Sample

As discussed in Chapter II, we surveyed a subset of the teachers in all of the study schools that were randomized in spring and summer 2011 (Cohort 1 schools) or in spring and summer 2012 (Cohort 2 schools). Here, we describe the rationale for the specific grades and subjects included in our sample and our methods for selecting the teachers to whom we administered the 2012, 2013, 2014, and 2015 teacher surveys.

Teaching Assignments Targeted by the Surveys

For the teacher surveys, we targeted teachers who taught 1st grade, 4th grade, 7th-grade math, 7th-grade English/language arts, or 7th-grade science in the study schools. We decided to focus on specific grades and subjects, rather than all elementary and middle school grades and subjects, to minimize the chance that the grades and subjects represented in the teacher sample would differ substantially between the treatment and control schools that were compared in the analysis. In other words, we wanted any treatment-control differences in teacher-reported outcomes to be attributable to pay-for-performance, rather than to an imbalance in grades or subjects.

Table A.4. Characteristics of Educators in Treatment and Control Schools in the Pre-Implementation Year, Cohort 1 (Percentages Unless Otherwise Noted)

	Teachers			Principals		
	Treatment	Control	Difference	Treatment	Control	Difference
Demographic Characteristics						
Female	87	85	2	64	60	4
Race or ethnicity						
White, non-Hispanic	76	73	3*	70	61	9
Black, non-Hispanic	17	20	-3*	25	32	-8
Hispanic or Other	7	7	0	5	6	-1
Age (average years)	43	43	0	48	48	0
Education						
Master's degree or higher	58	59	-2	>88[a]	>88[a]	8
Experience in K–12 Education						
Total experience (average years)	13	13	0	16	15	1
Fewer than 5 years	20	19	1	11	10	1
5-15 years	46	47	0	43	43	0
More than 15 years	34	34	0	46	47	-1
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value			0.03			0.00
Number of Educators—Range[b]	**729-1,812**	**770-1,790**		**25-54**	**28-56**	
Number of Schools—Range[b]	**27-56**	**27-56**		**24-53**	**26-54**	

Source: Educator administrative data.

Notes: One district did not provide data for the pre-implementation year. The difference between the treatment and control estimates may not equal the difference shown in the table because of rounding. The number of principals exceeds the number of schools in the analysis sample because a few schools had more than one principal.

[a]To reduce the risk of disclosing information on an individual, the exact percentage is not shown.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

We chose these grades and subjects so that they would encompass different groups of teachers who were thought to face different incentives from pay-for-performance—in particular, teachers in tested grade/subject combinations (4th grade, 7th-grade math, and 7th-grade reading)—and those in nontested grade/subject combinations (1st grade and 7th-grade science). Teachers in nontested grades/subjects might be eligible for bonuses based heavily on performance measures that they could affect only indirectly (such as student achievement growth in other grades and subjects within the same school). On the other hand, teachers in tested grades/subjects could have a more direct influence on performance ratings—and, therefore, bonus amounts—that were linked to the achievement growth of students in their own classrooms.

The set of targeted grades was also designed to include both elementary and middle school grades because of their different classroom structures. Elementary school teachers typically teach self-contained classrooms and are responsible for all core subjects, whereas middle school teachers typically work in a departmentalized setting in which they are responsible for one subject (such as math *or* reading). Among the tested elementary grades, we chose to target 4th grade because it is typically the earliest grade at which student achievement growth on state assessments can be calculated and is more likely than grade 5 to have self-contained classes. Among the tested middle

school grades and subjects, we chose 7th-grade math and reading because they are more likely than 8th-grade subjects to be assessed by end-of-grade tests that are uniform across all students (rather than end-of-course tests that depend on the course in which students are enrolled) but are more likely than 6th-grade classes to be departmentalized.

We chose 1st grade and 7th-grade science as the nontested grades and subjects in our target population, for several reasons. First grade has full-day classes and is less likely than grades 2 and 3 to have standardized testing. Science is a well-defined subject that is not tested annually, and retaining certified science teachers is an important policy goal.

Sampling Approach

Although the 2012, 2013, 2014, and 2015 surveys focused on teachers in the targeted grades and subjects described above, there were some differences in the sampling approach used each year. Specifically, in 2013, 2014, and 2015 we sampled (1) all teachers in targeted grades and subjects (as opposed to a subset of them), and (2) teachers who were surveyed in the prior year, even if they were no longer teaching a targeted grade and subject.

Sampling approach for teachers in targeted grades and subjects. Within each study school and year, we used administrative data provided by the evaluation districts to identify teachers who were assigned to any of the targeted grades and subjects. In 2012, we sampled all 4th-grade teachers; all 7th-grade math, English/language arts, and science teachers; and 77 percent of 1st-grade teachers. Because our analysis of impacts on student achievement focuses on tested grades and subjects, our sampling approach for the teacher survey was designed to give greater emphasis to tested grades and subjects than to nontested ones. Therefore, we selected all teachers who taught any of the tested grades and subjects targeted by the survey and selected a subset of teachers who taught the nontested grades and subjects targeted by the survey. Specifically, for each nontested grade and subject (1st grade or 7th-grade science) in each study school, we randomly selected three teachers from the teachers assigned to that combination of school, grade, and subject. If no more than three teachers were assigned to that combination, all such teachers were chosen. In practice, this approach led to the selection of all 7th-grade science teachers in the sampling frame—because of the small numbers of such teachers in each school—and 77 percent of the 1st-grade teachers in the sampling frame.[66] In 2013 and 2014, we surveyed all teachers in targeted grades and subjects, including 100 percent of 1st-grade teachers, which led to an increase in the total number of teachers in these targeted teaching assignments.

Sampling approach for teachers previously surveyed. In 2013 and 2014, we also sampled those teachers who were surveyed in the prior year but were no longer teaching a targeted grade and subject. This included teachers who left the district or teaching profession. In 2015, we surveyed teachers from the prior year who switched teaching assignments in the same schools or moved to a different school in the same district, but not teachers who left the teaching profession or the district. If pay-for-performance had an impact on teachers' school choice or career decisions, this subset of teachers would have allowed us to document reasons why teachers switch schools or leave the teaching profession.

[66] Due to an error in the sampling algorithm, we inadvertently sampled all 1st-grade teachers in three districts' study schools.

We wanted to survey teachers from two groups of teachers: (1) teachers in the targeted grades and subjects, and (2) teachers we had surveyed the year before but were no longer teaching a targeted grade or subject. However, because some teaching rosters were not sufficiently detailed (for example, describing teachers' grades as a range of grades) or were inaccurate, our sample included 97 teachers in 2012, 113 in 2013, 120 in 2014, and 128 in 2015 who reported they were not teaching in the targeted grades and subjects, although we had believed they were. We excluded these teachers from the teacher survey analyses. We did not need to replace these ineligible teachers because we had already selected all teachers identified by the administrative data as teaching the grades and subjects targeted by the survey. Similarly, some teachers we surveyed in 2013, 2014, and 2015 because we had surveyed them in the prior year, reported they were teaching a targeted grade and subject, although based on administrative data we thought they were not. We included these teachers' responses in our Years 2, 3, and 4 teacher survey analyses.

Survey Response Rates and Analysis of Missing Outcomes in Survey Data

In this section, we report the response rates for each of the three surveys (district, teacher, and principal surveys) and years used in this report. Because of the high response rate (more than 89 percent across all surveys), the potential for nonresponse bias is minimal. Nonetheless, we assessed the extent to which the respondents are similar to nonrespondents and, for educator surveys, whether respondents are similar across treatment and control schools.

Table A.5 shows the response rates for the 2015 district survey, and Table A.6 compares district characteristics of respondents and nonrespondents on such dimensions as district location and size.

Table A.5. District Survey Response Rates Overall and by Evaluation Status, 2014–2015 School Year, Cohorts 1 and 2

	Overall	Non-Evaluation Districts	Evaluation Districts
All Districts			
Number of districts	148	135	13
Number of respondents	139	126	13
Response Rate (respondents over total)	94	93	100

Source: District survey, 2015.

Notes: Table excludes 11 districts that were sent a survey but were found not to be implementing TIF at the time of the survey administration. The difference in response rates between non-evaluation and evaluation districts was not statistically significant at the .05 level.

Table A.6. District Characteristics by Districts' Response Status, 2014–2015 School Year, Cohorts 1 and 2 (Percentages Unless Otherwise Noted)

	Respondents	Nonrespondents
Student Racial or Ethnic Distribution		
White, non-Hispanic	50	37
Black, non-Hispanic	26	37
Hispanic	18	11
Student Socioeconomic Status		
Eligible for free or reduced-price lunch	65	78*
Title 1 eligible schools (schoolwide)	79	93*
Enrollment (averages)		
Total enrollment	22,233	3,482*
District Location[a]		
Urban	33	25
Suburban	19	25
Town	23	0*
Rural	26	50
District Census Bureau Region		
Northeast	9	11
Midwest	27	44
South	47	22
West	17	22
Number of Districts—Range[b]	**130-139**	**8-9**

Source: District survey (2015) and Common Core of Data for 2013–2014 school year.

Notes: Seven TIF non-evaluation districts are not included in the 2013–2014 district-level data from the Common Core of Data. Common Core of Data school-level data are used to calculate socioeconomic indicators. Common Core of Data district-level data are used to calculate all other demographic characteristics. The difference between respondents and nonrespondents was not statistically significant at the .05 level.

[a]District location indicates the physical location of the district agency.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

Tables A.7 and A.8 show teacher and principal sample sizes and response rates. Table A.7 reports the total number of surveyed teachers in 1st grade, 4th grade, and 7th-grade math, English/language arts, and science and principals in Cohort 1 schools, along with their response rates and the final analyses samples. Table A.8 shows response rates for teachers (those in targeted grades and subjects) and principals in Cohort 2.

Table A.9 presents the distribution of grade and subject assignments for the Cohort 1 teachers who responded to the survey and were included in the final analysis samples.

Table A.7. Teacher and Principal Response Rates for the Final Analyses Samples, Cohort 1

	Year 1 (2012 Survey)			Year 2 (2013 Survey)			Year 3 (2014 Survey)			Year 4 (2015 Survey)		
	Total	Treatment	Control	Total	Treatment	Control	Total	Treatment	Control	Total	Treatment	Control
Teachers												
Number of Sampled Teachers[a]	950	467	483	946	467	479	1,012	502	510	961	485	476
Number of respondents	870	423	447	869	430	439	913	437	476	859	420	439
Response rate (percentage)	92	91	93	92	92	92	90	87	93*	89	87	92*
Number of Teachers in the Final Analysis Sample[b]	**786**	**384**	**402**	**899**	**446**	**453**	**888**	**427**	**461**	**795**	**397**	**398**
Principals												
Number of Sampled Principals	130	65	65	131	65	66	131	65	66	131	65	66
Number of respondents	128	64	64	125	63	62	121	58	63	124	61	63
Response rate (percentage)	98	98	98	95	97	94	92	89	95	95	94	95
Number of Principals in the Final Analysis Sample[c]	**128**	**64**	**64**	**124**	**63**	**61**	**120**	**58**	**62**	**123**	**61**	**62**

Sources: Teacher and principal surveys (2012, 2013, 2014, and 2015).

[a]The teacher sample for the final analysis included 1st grade, 4th grade, and 7th-grade math, English/language arts, and science teachers.

[b]The final analysis sample excludes teachers who reported working part-time or teaching grades and subjects other than the targeted 1st grade, 4th grade, and 7th-grade math, English/language arts, and science. In addition, it includes teachers who were not in our original sample of teachers in targeted grades and subjects but who responded to the survey and self-identified as teaching in those targeted grades and subjects.

[c]The analysis sample in Year 2 excludes a few respondents who did not identify themselves as principals in the survey.

*Difference in response rates between treatment and control groups is statistically significant at the .05 level, two-tailed test.

Table A.8. Teacher and Principal Response Rates for the Final Analyses Samples, Cohort 2

	Year 1 (2013 Survey)			Year 2 (2014 Survey)			Year 3 (2015 Survey)		
	Total	Treatment	Control	Total	Treatment	Control	Total	Treatment	Control
Teachers									
Number of Sampled Teachers[a]	244	128	116	293	147	146	283	142	141
Number of respondents	223	114	109	245	129	116	247	121	126
Response rate (percentage)	91	89	94	84	88	79	87	85	89*
Number of Teachers in the Final Analysis Sample[b]	**223**	**121**	**102**	**253**	**135**	**118**	**241**	**119**	**122**
Principals									
Number of Sampled Principals	39	20	19	39	20	19	39	20	19
Number of respondents	33	17	16	35	19	16	36	19	17
Response rate (percentage)	85	85	84	90	95	84	92	95	89
Number of Principals in the Final Analysis Sample[c]	**33**	**17**	**16**	**35**	**19**	**16**	**36**	**19**	**17**

Sources: Teacher and principal surveys (2013, 2014, and 2015).

Note: None of the differences in response rates between treatment and control groups were statistically significant at the .05 level, two-tailed test.

[a]The teacher sample for the final analysis included 1st grade, 4th grade, and 7th-grade math, English/language arts, and science teachers.

[b]The final analysis sample excludes teachers who reported working part-time or teaching grades and subjects other than the targeted 1st grade, 4th grade, and 7th-grade math, English/language arts, and science. In addition, it includes teachers who were not in our original sample of teachers in targeted grades and subjects but who responded to the survey and self-identified as teaching in those targeted grades and subjects.

[c]The analysis sample excludes a few respondents who did not identify themselves as principals in the survey.

Table A.9. Teacher Respondents, by Teaching Assignment and Treatment Status, Cohort 1

	Year 1			Year 2			Year 3			Year 4		
Grade Taught	Total	Treatment	Control	Total	Treatment	Control	Total	Treatment	Control	Total	Treatment	Control
1st Grade Only	224	107	117	302	157	145	312	144	168	280	136	144
4th Grade Only	220	109	111	219	104	115	213	102	111	182	88	94
7th-Grade English/Language Arts and/or Math Only	201	98	103	197	96	101	172	83	89	163	82	81
7th-Grade Science Only	64	35	29	59	33	26	54	25	29	56	25	31
More than One Targeted Grade or Subject	77	35	42	122	56	66	137	73	64	114	66	48
Total	**786**	**384**	**402**	**899**	**446**	**453**	**888**	**427**	**461**	**795**	**397**	**398**

Sources: Teacher surveys (2012, 2013 2014, and 2015).

Notes: Targeted grades and subjects for the survey were 1st grade, 4th grade, and 7th-grade math, English/language arts, and science. Counts are for teachers in those targeted grades and subjects who responded to the survey and are included in the final analysis sample.

We matched administrative data to survey respondents to compare (1) the characteristics of respondents and nonrespondents, and (2) the characteristics of educators in treatment and control schools. Tables A.10 through A.12 present our nonresponse analyses for the teacher and principal surveys. Table A.10 compares the characteristics of teachers who responded to the survey to those who did not. Because there were few principal nonrespondents, we do not report a similar analysis for the principal survey. Tables A.11 and A.12 compare the characteristics of respondents in treatment and control schools for teachers and principals, respectively. Because we did not receive administrative data on educator characteristics for all survey respondents, the sample sizes in Tables A.10 through A.12 are smaller than the number of teacher and principal survey respondents.

Table A.10. Characteristics of Teacher Survey Respondents and Nonrespondents, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 1		Year 2		Year 3		Year 4	
	Respondents	Nonrespondents	Respondents	Nonrespondents	Respondents	Nonrespondents	Respondents	Nonrespondents
Demographic Characteristics								
Female	89	80	89	81	89	86	87	89
Race or ethnicity								
White, non-Hispanic	72	67	66	71	66	69	75	67
Black, non-Hispanic	22	26	28	20	25	21	18	21
Hispanic or Other	6	7	7	9	8	9	8	12
Age (average years)	40	43	42	43	42	43	41	42
Education								
Master's degree or higher	46	34	53	41	48	46	51	43
Experience in K–12 Education								
Total experience (average years)	12	14	11	11	13	12	10	11
Fewer than 5 years	24	23	26	33	22	31	36	31
5-15 years	43	34*	44	39	40	36	40	43
More than 15 years	33	43*	30	28	38	33	24	26
Number of Teachers—Range[a]	**589-830**	**45-75**	**824-1,102**	**54-91**	**861-1,089**	**67-109**	**795-998**	**92-120**

Sources: Teacher surveys (2012, 2013, 2014, and 2015) and educator administrative data.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference between respondents and nonrespondents is statistically significant at the .05 level, two-tailed test.

Table A.11. Characteristics of Teacher Survey Respondents by Treatment Status, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 1		Year 2		Year 3		Year 4	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Demographic Characteristics								
Female	89*	85	90	87	90	87	88	87
Race or ethnicity								
White, non-Hispanic	74	68	71	68	73	75	71	75
Black, non-Hispanic	19	23	21*	26	21	18	24*	19
Hispanic or Other	8	9	8	6	5	7	5	6
Age (average years)	40	39	41	41	40	40	41	41
Education								
Master's degree or higher	44*	52	49	55	49	49	50	52
Experience in K–12 Education								
Total experience (average years)	11	10	11	10	10	10	9	10
Fewer than 5 years	27	27	25	29	29	31	33	38
5-15 years	46	50	48	47	44	46	46	39
More than 15 years	26	23	26	24	26	23	21	23
Number of Teachers—Range[a]	**246-365**	**292-389**	**327-441**	**339-448**	**337-418**	**358-448**	**312-392**	**308-388**

Sources: Teacher surveys (2012, 2013, 2014, and 2015) and educator administrative data.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

Table A.12. Characteristics of Principal Survey Respondents by Treatment Status, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 1		Year 2		Year 3		Year 4	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Demographic Characteristics								
Female	59	67	61	68	66	63	64	68
Race or ethnicity								
White, non-Hispanic	66	60	61	52	60	54	58	53
Black, non-Hispanic	27	33	34	38	35	34	36	35
Hispanic or Other	7	7	6*	11	5	12	6	11
Age (average years)	49	48	48	49	47	49	47	49
Education								
Master's degree or higher	95	92	101	95	95	95	91	94
Experience in K–12 Education								
Total experience (average years)	16	15	16*	14	18	16	18	16
Fewer than 5 years	19	13	18	19	6	14	8	14
5-15 years	31	38	26*	43	39	42	36	43
More than 15 years	50	48	56*	39	55	44	56	43
Number of Principals—Range[a]	**37-60**	**39-60**	**46-61**	**38-56**	**41-57**	**42-59**	**44-58**	**46-61**

Sources: Principal surveys (2012, 2013, 2014, and 2015) and educator administrative data.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

Sample Sizes and Analysis of Missing Outcomes in Educator Administrative Data

We used districts' administrative records for all analyses of educator effectiveness. In this section, we describe the samples and the characteristics of educators included in these analyses.

All analyses of educator effectiveness were restricted to educators who worked full-time in the study schools. The 131 Cohort 1 schools included 4,303 full-time teachers in Year 1, 4,402 full-time teachers in Year 2, 4,521 full-time teachers in Year 3, and 4,215 full-time teachers in Year 4. The number of full-time principals was not the same as the total number of study schools because a few schools did not have a full-time principal or had more than one full-time principal. Table A.13 shows the number of full-time principals listed in the administrative data and the number of schools in those principals worked.

Table A.13. Number of Full-Time Principals Listed in the Administrative Data and the Number of Schools in Which They Worked, Cohort 1

	Treatment	Control
Principals Included in the Analyses of Principal Outcomes		
Year 1 (2011–2012)		
All principals at the beginning of the year	66	70
Full-time principals at the beginning of the year (*eligible to be included in analysis*)	64	69
Year 2 (2012–2013)		
All principals at the beginning of the year	68	71
Full-time principals at the beginning of the year (*eligible to be included in analysis*)	67	70
Year 3 (2013–2014)		
All principals at the beginning of the year	65	70
Full-time principals at the beginning of the year (*eligible to be included in analysis*)	63	69
Year 4 (2014–2015)		
All principals at the beginning of the year	64	65
Full-time principals at the beginning of the year (*eligible to be included in analysis*)	64	63
Schools Included in the Analyses of Principal Outcomes		
Year 1 (2011–2012)		
All Cohort 1 schools	65	66
Schools with principals at the beginning of the year	64	66
Schools with full-time principals at the beginning of the year	62	65
Year 2 (2012–2013)		
All Cohort 1 schools	65	66
Schools with principals at the beginning of the year	65	65
Schools with full-time principals at the beginning of the year	64	64
Year 3 (2013–2014)		
All Cohort 1 schools	65	66
Schools with principals at the beginning of the year	64	66
Schools with full-time principals at the beginning of the year	62	65
Year 4 (2014–2015)		
All Cohort 1 schools	65	66
Schools with principals at the beginning of the year	64	66
Schools with full-time principals at the beginning of the year	64	63

Source: Educator administrative data.

Note: The number of principals in the analysis might differ from the total number of schools because a few schools did not have a full-time principal or had more than one full-time principal.

We assessed educator effectiveness using several districts' measures used to evaluate and determine TIF performance bonuses, including classroom observation ratings and achievement growth ratings. Table A.14 (teachers) and Table A.15 (principals) describe the sample sizes using different measures of educator effectiveness. In Years 1, 2, 3, and 4, all 131 Cohort 1 schools provided classroom observations ratings for at least some teachers. One district (with 19 schools) did not provide principal observation ratings for Year 1; all 10 Cohort 1 districts provided principal observation ratings for Years 2, 3, and 4. Not all schools within a district, however, provided principal observation ratings.

Table A.14. Teachers Who Had Performance Ratings, Cohort 1 (Percentages)

	Treatment	Control	Difference	*p*-Value	**Number of Teachers**	**Number of Schools**
Year 1						
Had Classroom Observation Rating	86	86	0	0.75	**4,303**	**131**
Had Classroom Achievement Growth Rating[a]	38	39	-1	0.30	**2,854**	**72**
Year 2						
Had Classroom Observation Rating	84	83	1	0.40	**4,402**	**131**
Had Classroom Achievement Growth Rating[a]	44	43	1	0.56	**2,923**	**72**
Year 3						
Had Classroom Observation Rating	84	83	1	0.56	**4,521**	**131**
Had Classroom Achievement Growth Rating[a]	59	58	1	0.52	**3,576**	**90**
Year 4						
Had Classroom Observation Rating	85	85	0	0.79	**4,215**	**131**
Had Classroom Achievement Growth Rating[a]	61	62	-1	0.43	**3,253**	**90**

Source: Educator administrative data.

Note: None of the differences were statistically significant at the .05 level, two-tailed test.

[a]Percentages are based only on teachers in districts that evaluated teachers using classroom achievement growth. In Year 1 and Year 2, 6 of 10 districts evaluated teachers based on classroom achievement growth. In Years 3 and 4, seven districts evaluated teachers based on classroom achievement growth.

Table A.15. Principals Who Had Observation Ratings, Cohort 1 (Percentages)

Outcome	Treatment	Control	Difference	*p*-Value	**Number of Principals**	**Number of Schools**
Year 1						
Had Observation Rating[a]	100	95	6	0.17	**108**	**108**
Year 2						
Had Observation Rating	96	85	10*	0.01	**137**	**128**
Year 3						
Had Observation Rating	97	87	9	0.06	**132**	**127**
Year 4						
Had Observation Rating	97	95	2	0.49	**127**	**127**

Source: Educator administrative data.

Notes: The number of principals exceeds the number of schools in the analysis sample because a few schools had more than one principal. The difference between the treatment and control estimates may not equal the difference shown in the table because of rounding.

[a]Percentages are based on 9 of 10 districts that provided data on observation scores for both treatment and control principals in Year 1.

*Difference is statistically significant at the .05 level, two-tailed test.

To help contextualize our findings, in Chapter II, we examined the extent to which educators who received a rating score (and thus were included in the analyses of educator effectiveness) are different from those who did not. We also assessed whether there were differences in the characteristics of treatment and control educators who received ratings. Tables A.16 through A.21 present these findings for the teacher and principal analyses samples. Table A.18 compares characteristics of principals with and without observation ratings in Years 2, 3, and 4 only, because of the small number of principals in Year 1 who did not receive an observation rating. Analyses for Tables A.17 and A.20 are based only on teachers in the 6 of 10 districts that evaluated teachers using classroom achievement growth in Years 1 and 2 and on teachers in the 7 of 10 districts that evaluated teachers using classroom achievement growth in Years 3 and 4.

Table A.16. Characteristics of Teachers With and Without Classroom Observation Ratings, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 1		Year 2		Year 3		Year 4	
	Teachers with Observation Ratings	Teachers Without Observation Ratings	Teachers with Observation Ratings	Teachers Without Observation Ratings	Teachers with Observation Ratings	Teachers Without Observation Ratings	Teachers with Observation Ratings	Teachers Without Observation Ratings
Demographic Characteristics								
Female	85	82	86	85	86	85	86	87
Race or ethnicity								
White, non-Hispanic	66	64	65	67	64	66	66	64
Black, non-Hispanic	29	30	30	29	31	29	28	31
Hispanic or Other	5	6	5	4	5	6	5	5
Age (average years)	40	41	41	42	41	41	42	44*
Education								
Master's degree or higher	41	43	43	44	40	44*	41	40
Total Experience in K–12 Education (average years)	10	11	11	11	10	10	11	13*
Fewer than 5 years	29	32	31	31	33	36	30	25
5-15 years	47	40*	44	40	44	43	40	39
More than 15 years	24	27	25	29*	23	22	30	36
Number of Teachers—Range[a]	**2,566-3,566**	**366-679**	**2,741-3,583**	**355-764**	**2,872-3,605**	**552-807**	**2,819-3,542**	**456-668**
Number of Schools—Range[a]	**97-131**	**64-98**	**99-131**	**72-105**	**105-131**	**83-102**	**105-131**	**87-111**

Source: Educator administrative data.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference between teachers with and without ratings is statistically significant at the .05 level, two-tailed test.

Table A.17. Characteristics of Teachers With and Without Classroom Achievement Growth Ratings, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 1		Year 2		Year 3		Year 4	
	Teachers with Classroom Achievement Growth Ratings	Teachers Without Classroom Achievement Growth Ratings	Teachers with Classroom Achievement Growth Ratings	Teachers Without Classroom Achievement Growth Ratings	Teachers with Classroom Achievement Growth Ratings	Teachers Without Classroom Achievement Growth Ratings	Teachers with Classroom Achievement Growth Ratings	Teachers Without Classroom Achievement Growth Ratings
Demographic Characteristics								
Female	87	84	86	85	87	83*	88	83*
Race or ethnicity								
White, non-Hispanic	63	65	64	66	65	66	66	67
Black, non-Hispanic	32	28*	28	28	28	27	27	26
Hispanic or Other	5	6	7	7	7	7	7	8
Age (average years)	39	40	38	41*	39	41*	39	43*
Education								
Master's degree or higher	36	39	38	40	36	46*	38	44*
Total Experience in K–12 Education (average years)	9	10	8	10*	8	10*	8	11*
Fewer than 5 years	34	33	39	34*	38	34*	39	31*
5-15 years	47	42	47	44	46	41*	44	39*
More than 15 years	19	25	15	22*	16	25*	17	30*
Number of Teachers—Range[a]	**618-1,059**	**1,327-1,738**	**927-1,317**	**1,187-1,552**	**1,478-2,037**	**1,138-1,443**	**1,533-1,956**	**1,106-1,292**
Number of Schools—Range[a]	**55-72**	**55-72**	**58-72**	**58-72**	**72-90**	**71-87**	**72-90**	**72-90**

Source: Educator administrative data.

Notes: Analyses are based on districts that evaluated teachers using classroom achievement growth. In Year 1 and Year 2, 6 of 10 districts evaluated teachers based on classroom achievement growth. In Years 3 and 4, seven districts evaluated teachers based on classroom achievement growth.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference between teachers with and without ratings is statistically significant at the .05 level, two-tailed test.

Table A.18. Characteristics of Principals With and Without Observation Ratings, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 2		Year 3		Year 4	
	Principals with Observation Ratings	Principals Without Observation Ratings	Principals with Observation Ratings	Principals Without Observation Ratings	Principals with Observation Ratings	Principals Without Observation Ratings
Demographic Characteristics						
Female	75	72	73	81	79	60
Race or ethnicity						
White, non-Hispanic	54	60	54	66	48	100*
Black, non-Hispanic	42	40	41	34	44	0*
Hispanic or Other	4	0	5	0	9	0
Age (average years)	47	52	45	55*	45	55*
Education						
Master's degree or higher	92	100	95	89	100	100
Total Experience in K–12 Education (average years)	13	14	12	13	16	25*
Fewer than 5 years	23	35	16	13	9	0
5-15 years	48	13*	60	68	40	20
More than 15 years	30	52	24	19	52	80*
Number of Principals—Range[a]	**83-117**	**12-18**	**84-119**	**5-12**	**98-122**	**5**
Number of Schools—Range[a]	**82-116**	**12-15**	**84-118**	**5-11**	**90-122**	**4-5**

Source: Educator administrative data.

Notes: The number of principals exceeds the number of schools in the analysis sample because a few schools had more than one principal. Findings for Year 1 are suppressed due to small sample sizes of principals without observation ratings.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

Table A.19. Characteristics of Teachers with Classroom Observation Ratings, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 1		Year 2		Year 3		Year 4	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Demographic Characteristics								
Female	87*	85	86*	85	86	85	86	85
Race or ethnicity								
White, non-Hispanic	74	73	73	72	73	72	74	73
Black, non-Hispanic	20	21	20	22	21	20	20	19
Hispanic or Other	7	6	7	6	6	7	6*	8
Age (average years)	42	41	42	41	42	42	41	41
Education								
Master's degree or higher	51	49	49	51	48	48	48	49
Total Experience in K–12 Education (average years)	12	11	11	11	11	11	10	10
Fewer than 5 years	23	25	26	29	27	30	33	36
5-15 years	47	47	46	45	46	44	43	41
More than 15 years	30	28	27	27	28	26	24	22
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value	0.05		0.03		0.28		0.01	
Number of Teachers—Range[a]	**1,249-1,779**	**1,317-1,787**	**1,320-1,772**	**1,421-1,811**	**1,425-1,790**	**1,447-1,815**	**1,409-1,760**	**1,410-1,782**
Number of Schools—Range[a]	**48-65**	**49-66**	**49-65**	**50-66**	**52-65**	**53-66**	**52-65**	**53-66**

Source: Educator administrative data.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

Table A.20. Characteristics of Teachers with Classroom Achievement Growth Ratings, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 1		Year 2		Year 3		Year 4	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Demographic Characteristics								
Female	89	86	88	87	89	87	89	88
Race or ethnicity								
White, non-Hispanic	62	64	63	62	65	64	65	63
Black, non-Hispanic	32	30	30	32	30	30	30	29
Hispanic or Other	5	6	7	6	5	6	5	8
Age (average years)	40*	38	40*	38	41	40	40	40
Education								
Master's degree or higher	37	38	38	37	35	35	38	35
Total Experience in K–12 Education (average years)	10*	8	9*	8	10	10	10	9
Fewer than 5 years	32	36	36*	42	32	34	36	37
5-15 years	45	48	43	47	42	44	42	42
More than 15 years	24*	16	21*	11	26*	22	23	21
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value	0.00		0.00		0.05		0.46	
Number of Teachers—Range[a]	**286-523**	**332-536**	**433-644**	**494-676**	**744-1,018**	**734-1,019**	**758-974**	**775-982**
Number of Schools—Range[a]	**27-36**	**28-36**	**29-36**	**29-36**	**36-45**	**36-45**	**36-45**	**36-45**

Source: Educator administrative data.

Notes: Analyses are based only on teachers in the districts that evaluated teachers using classroom achievement growth. In Year 1 and Year 2, 6 of 10 districts evaluated teachers based on classroom achievement growth. In Year 3, seven districts evaluated teachers based on classroom achievement growth.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

Table A.21. Characteristics of Principals with Observation Ratings, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 1[a]		Year 2		Year 3		Year 4	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Demographic Characteristics								
Female	59	62	59	65	60	64	58	70
Race or ethnicity								
White, non-Hispanic	65	62	59	54	63	49	60	48
Black, non-Hispanic	24	29	33	34	31	39	34	40
Hispanic or Other	10	10	8	13	6	12	6	12
Age (average years)	49	47	48	48	49	48	47	48
Education								
Master's degree or higher	>90[b]	>90[b]	>90[b]	>90[b]	>90[b]	>90[b]	93	93
Total Experience in K–12 Education (average years)	16	15	15	14	18	15	18	16
Fewer than 5 years	15	11	18	15	8	16	12	16
5-15 years	38	41	37	47	38	41	35	40
More than 15 years	47	48	45	38	54	43	54	44
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value	0.70		0.42		0.03		0.03	
Number of Principals—Range[c]	**32-53**	**34-52**	**44-61**	**39-56**	**44-60**	**40-59**	**47-62**	**43-60**
Number of Schools—Range[c]	**32-53**	**34-52**	**44-61**	**38-55**	**44-60**	**40-58**	**47-62**	**43-60**

Source: Educator administrative data.

Notes: The number of principals exceeds the number of schools in the analysis sample because a few schools had more than one principal. None of the differences are statistically significant at the .05 level, two-tailed test.

[a]Characteristics are based on 9 of 10 districts that provided data on observation scores for both treatment and control principals in Year 1.

[b]To reduce the risk of disclosing information on an individual, the exact percentage is not shown.

[c]Sample sizes are presented as a range based on the data available for each row in the table.

Sample Sizes and Analysis of Missing Outcomes in Student Administrative Data

Chapter VI estimates the impact of pay-for-performance on students' math and reading scores on state standardized exams. Table A.22 shows the total number of students with available scores who were in the sample for those analyses. Tables A.23 and Table A.24 describe the characteristics of students with and without test scores in math and reading, respectively.

Table A.22. Students Who Had Test Scores, Cohort 1 (Percentages)

	Treatment	Control	Difference	Number of Students	Number of Schools
Year 1					
Math	93	92	1	**44,442**	**131**
Reading	92	92	0	**44,442**	**131**
Year 2					
Math	92	92	0	**44,567**	**131**
Reading	91	92	-1	**44,567**	**131**
Year 3					
Math	93	93	0	**43,058**	**131**
Reading	92	93	0	**43,058**	**131**
Year 4					
Math	92	92	1	**42,593**	**131**
Reading	92	92	1	**42,593**	**131**

Source: Student administrative data.

Notes: None of the differences were statistically significant at the .05 level, two-tailed test. The difference between the treatment and control estimates may not equal the difference shown in the table because of rounding.

Our primary analysis in Chapter VI estimates the impact of pay-for-performance on students enrolled in study schools in a given year. As such, our impact estimates measure the impact of pay-for-performance on participating schools, not the impact on individual students. Therefore, this impact can be the result of changes in teacher productivity, changes in teacher composition (because of school mobility), or changes in student composition. Although we cannot disentangle how much of an effect on achievement might result from changes in students or teachers, Tables A.25 and A.26 show that average student characteristics were similar between treatment and control schools across years, suggesting that pay-for-performance did not induce changes in the schools' student composition.

Table A.23. Characteristics of Students Who Did and Did Not Have Math Test Scores, Cohort 1 (Percentages Unless Otherwise Noted)

Characteristic	Year 1		Year 2		Year 3		Year 4	
	Had Test Scores	Did Not Have Test Scores	Had Test Scores	Did Not Have Test Scores	Had Test Scores	Did Not Have Test Scores	Had Test Scores	Did Not Have Test Scores
Achievement in Pre-Implementation Year (average *z*-score)[a]								
Math	-0.43	-0.80*	-0.44	-0.80*	-0.66	-0.96*	-0.66	-1.15*
Reading	-0.38	-0.72*	-0.39	-0.69*	-0.59	-1.03*	-0.62	-0.99*
Race or Ethnicity								
White, non-Hispanic	28	29	28	29	27	29	27	29
Black, non-Hispanic	42	44*	42	44*	41	42	41	42
Hispanic	24	19*	24	20*	25	22*	26	23*
Other	6	7	6	7	6	7	6	6
Other Characteristics								
Female	50	43*	49	45*	50	44*	50	44*
Eligible for free or reduced-price lunch	77	79	77	78	81	81	81	76*
Disabled or has an Individualized Education Program	12	31*	13	34*	11	32*	14	33*
Over age for grade	12	24*	12	22*	11	24*	11	21*
English language learner	8	8	7	7	12	11	16	15
Grade Span								
Grades 3–5	65	66	64	65	67	66	67	63*
Grades 6–8	35	34	36	35	33	34	33	37*
Number of Students—Range[b]	**23,527-40,565**	**1,495-3,877**	**20,861-40,465**	**1,130-4,102**	**13,637-39,771**	**563-3,287**	**8,074-38,933**	**498-3,660**
Number of Schools—Range[b]	**83-131**	**79-128**	**83-131**	**71-122**	**105-131**	**59-123**	**68-131**	**47-129**

Source: Student administrative data.

[a]These averages are only calculated for students who were tested in the pre-implementation year, so they exclude students in grade 3 in Year 1; students in grades 3 and 4 in Year 2; students in grades 3 through 5 in Year 3; and students in grades 3 through 6 in Year 4.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference between students with and without math test scores is statistically significant at the .05 level, two-tailed test.

Table A.24. Characteristics of Students Who Did and Did Not Have Reading Test Scores, Cohort 1 (Percentages Unless Otherwise Noted)

Characteristic	Year 1		Year 2		Year 3		Year 4	
	Had Test Scores	Did Not Have Test Scores	Had Test Scores	Did Not Have Test Scores	Had Test Scores	Did Not Have Test Scores	Had Test Scores	Did Not Have Test Scores
Achievement in Pre-Implementation Year (average *z*-score)[a]								
Math	-0.43	-0.84*	-0.44	-0.83*	-0.64	-0.97*	-0.64	-1.20*
Reading	-0.38	-0.72*	-0.38	-0.75*	-0.59	-0.97*	-0.60	-1.07*
Race or Ethnicity								
White, non-Hispanic	28	28	28	28	27	28	27	29
Black, non-Hispanic	42	43	42	44*	41	41	41	42
Hispanic	24	21*	24	22	25	24	26	23*
Other	6	8	6	7	6	7	6	7
Other Characteristics								
Female	50	43*	50	44*	50	43*	50	45*
Eligible for free or reduced-price lunch	77	78	77	80	81	81	81	76*
Disabled or has an Individualized Education Program	12	31*	13	34*	11	31*	14	33*
Over age for grade	12	23*	12	22*	11	24*	11	21*
English language learner	8	9	7	7	12	13	16	16
Grade Span								
Grades 3–5	64	68	64	66	67	67	67	64*
Grades 6–8	36	32	36	34	33	33	33	36*
Number of Students—Range[b]	**23,363-40,269**	**1,542 4,173**	**20,825-40,132**	**1,166-4,435**	**13,641-39,541**	**561-3,517**	**8,102-38,942**	**470-3,651**
Number of Schools—Range[b]	**83-131**	**80-129**	**83-131**	**79-130**	**104-131**	**67-126**	**67-131**	**48-129**

Source: Student administrative data.

[a]These averages are only calculated for students who were tested in the pre-implementation year, so they exclude students in grade 3 in Year 1; students in grades 3 and 4 in Year 2; students in grades 3 through 5 in Year 3; and students in grades 3 through 6 in Year 4.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference between students with and without reading test scores is statistically significant at the .05 level, two-tailed test.

Table A.25. Characteristics of Students in the Math Analysis Sample, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 1		Year 2		Year 3		Year 4	
Characteristic	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Achievement in the Pre-Implementation Year (average *z*-score)[a]								
Math	-0.46*	-0.41	-0.46	-0.44	-0.64	-0.72	-0.72*	-0.57
Reading	-0.39	-0.38	-0.39	-0.40	-0.56	-0.67	-0.62	-0.54
Race or Ethnicity								
White, non-Hispanic	27	29	27	29	26	29	27	28
Black, non-Hispanic	42	41	42	41	42	40	41	41
Hispanic	24	23	25	23	26	24	26	25
Other	6	6	6	7	6	7	7	6
Other Characteristics								
Female	49	50	49	49	49	50	49	50
Eligible for free or reduced-price lunch	77	78	77	76	81	82	82	80
Disabled or has an Individualized Education Program	12	12	13	13	11	11	14	15
Over age for grade	12	11	12	11	11	10	11	11
English language learner	8	9	7	8	12	12	16	16
Grade Span								
Grades 3–5	64	64	64	64	67	66	67	66
Grades 6–8	36	36	36	36	33	34	33	34
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value	0.01*		0.15		0.04*		0.25	
Number of Students—Range[b]	**11,596-20,213**	**11,848-20,322**	**10,162-19,997**	**10,693-20,457**	**6,636-19,759**	**7,015-20,011**	**3,915-19,160**	**4,157-19,779**
Number of Schools—Range[b]	**41-65**	**42-66**	**41-65**	**42-66**	**52-65**	**53-66**	**34-65**	**34-66**

Source: Student administrative data.

[a]These averages are only calculated for students who were tested in the pre-implementation year, so they exclude students in grade 3 in Year 1; students in grades 3 and 4 in Year 2; students in grades 3 through 5 in Year 3; and students in grades 3 through 6 in Year 4.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference between students in treatment and control schools is statistically significant at the .05 level, two-tailed test.

Table A.26. Characteristics of Students in the Reading Analysis Sample, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 1		Year 2		Year 3		Year 4	
Characteristic	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Achievement in the Pre-Implementation Year (average *z*-score)[a]								
Math	-0.45*	-0.41	-0.46	-0.43	-0.64	-0.71	-0.70*	-0.57
Reading	-0.39	-0.38	-0.39	-0.39	-0.56	-0.67	-0.60	-0.54
Race or Ethnicity								
White, non-Hispanic	27*	29	27	29	26	29	27	28
Black, non-Hispanic	43	42	42	41	42	41	41	41
Hispanic	24	23	25	23	26	24	26	25
Other	6	6	6	7	6	7	7	6
Other Characteristics								
Female	50	50	50	50	49	50	49	50
Eligible for free or reduced-price lunch	77	78	77	76	81	82	82	80
Disabled or has an Individualized Education Program	12	12	13	13	11	11	14	15
Over age for grade	12	11	12	11	11	10	11	11
English language learner	8	9	7	8	12	12	16	15
Grade Span								
Grades 3–5	64	64	64	64	67	66	67	66
Grades 6–8	36	36	36	36	33	34	33	34
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value	0.03*		0.18		0.07		0.15	
Number of Students—Range[b]	**11,492-20,028**	**11,803-20,228**	**10,123-19,763**	**10,696-20,359**	**6,635-19,611**	**7,021-19,927**	**3,927-19,159**	**4,172-19,770**
Number of Schools—Range[b]	**41-65**	**42-66**	**41-65**	**42-66**	**51-65**	**53-66**	**33-65**	**34-66**

Source: Student administrative data.

[a]These averages are only calculated for students who were tested in the pre-implementation year, so they exclude students in grade 3 in Year 1; students in grades 3 and 4 in Year 2; students in grades 3 through 5 in Year 3; and students in grades 3 through 6 in Year 4.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference between students in treatment and control schools is statistically significant at the .05 level, two-tailed test.

THIS PAGE IS INTENTIONALLY BLANK

APPENDIX B

SUPPLEMENTAL INFORMATION ON ANALYTIC METHODS FOR CHAPTER II

THIS PAGE IS INTENTIONALLY BLANK

In this appendix, we provide the rationale for and technical details of the methods used in the report. First, we describe how we standardized educator performance ratings and student test scores across districts. Second, we discuss the technical approach for describing the distribution of performance ratings and TIF payouts in evaluation districts. Third, we provide details on the analytic methods used to estimate impacts of pay-for-performance on educator and student outcomes. Fourth, we discuss the interpretation of the student achievement impacts, focusing on the extent to which students were exposed to pay-for-performance by the end of each year of implementation. Fifth, we describe the methods used to explore differences in impacts across teacher and student subgroups, districts, and schools. Sixth, we discuss our approach to estimating year-to-year changes in average teacher perceptions of TIF. Seventh, we specify the methods used to impute educators' beliefs about maximum pay-for-performance bonus amounts if they reported being eligible for pay-for-performance but did not answer survey questions about bonus amounts. Finally, we summarize the level of precision in the study by reporting minimum detectable impacts for key outcomes examined in the impact analyses.

As discussed in Chapter II, evaluation districts were classified into two cohorts—Cohort 1 and Cohort 2—according to the year in which we randomly assigned their schools to a treatment or control group. The 10 districts whose schools were randomly assigned in spring and summer 2011 were classified as Cohort 1. Three additional districts, whose schools were randomly assigned in spring and summer 2012, were classified as Cohort 2. Cohort 1 completed four years of implementation—2011–2012, 2012–2013, 2013–2014, and 2014–2015—referred to as Years 1, 2, 3, and 4. Cohort 2 districts completed three years of implementation—2012–2013, 2013–2014, and 2014–2015—referred to as Years 1, 2, and 3 for this cohort.

Standardizing Outcomes

The two key outcomes discussed in Chapter VI—educator performance ratings and student achievement—were measured using scales or assessments that varied across districts. This section discusses the methods we used to standardize these outcomes for the analysis.

Educator Performance Ratings

We measured educator effectiveness with several measures that districts used in their TIF programs to evaluate educators and determine performance bonuses. As we noted in Chapter I, districts had to evaluate teachers and principals based on student achievement growth and at least two observations of classroom or school practices. However, districts had flexibility in how they implemented this requirement. For example, districts could choose to evaluate teachers based on the achievement growth of the teachers' own students (classroom achievement growth); all students in the same grade, team, or subject area (achievement growth of student subgroups); all students in the school (school achievement growth), or some combination of these measures. Our analysis used four measures: (1) school achievement growth ratings, which were used to evaluate both teachers and principals; (2) teachers' classroom observation ratings; (3) teachers' classroom achievement growth ratings; and (4) principals' observation ratings.

Each of these performance measures either placed educators into three to five performance categories—such as "effective" or "highly effective"—or placed educators onto a numeric scale (typically ranging from 1 to 4 or 1 to 5) in which a one-unit increase was analogous to advancing by a performance level. To express ratings from different districts on a common scale, we transformed the data in two steps. First, if the districts used performance categories but did not already express

the performance categories as numbers, we ordered the categories and denoted them with consecutive whole numbers, with 1 as the lowest-performing category. This step resulted in all performance ratings being placed on a district-specific numeric scale that had a defined minimum and maximum possible rating. Second, because the range of the scale varied across districts, a one-unit increase would have a different meaning in different districts unless the rating scales were rescaled to have a common range. Therefore, we rescaled all ratings into a common 1-to-4 rating scale with the following formula:

$$(1)\quad \tilde{R}_{jd} = 3 \times \left(\frac{R_{jd} - R_{\min,d}}{R_{\max,d} - R_{\min,d}} \right) + 1,$$

where $\tilde{R}_{jd}$ was the rescaled rating of educator j in district d, R_{jd} was the rating on the district's original numeric scale, and $R_{\min,d}$ $R_{\max,d}$ were the minimum and maximum ratings that educators in district d could theoretically receive. Using this formula, an educator who received the lowest rating on the district's scale would receive a rescaled rating of 1, and an educator who received the highest rating on the district's scale would receive a rescaled rating of 4. As another example, an educator who received a 3 on a district scale that ranged from 1 to 5 would have a rescaled rating of 2.5.

At an early stage of the analysis, we explored, but ultimately rejected, an alternative approach to standardizing educator performance ratings across districts. The alternative approach standardized performance ratings into z-scores by subtracting district-specific means of the ratings and dividing by district-specific standard deviations of the ratings. We concluded that placing performance ratings on a 1-to-4 scale, as described above, would be preferable to converting the ratings into z-scores for several reasons. First, in some districts, estimates of standard deviations would be based on small sample sizes and would therefore not be very reliable. For example, in the smallest evaluation districts that had four to six study schools, only four to six distinct data points would be available for calculating the standard deviation of a school achievement growth rating. Second, some measures produced very little variation in ratings within particular districts, implying that even a small impact (on the original scale) would be misleadingly represented as a huge effect size in z-score units. Third, the 1-to-4 rating scale corresponded more closely to the information that educators actually received and to which they would potentially respond.

Student Achievement

We measured student achievement with students' scores on state assessments in math and reading. Because student achievement was measured on different scales in different states and grades, we standardized scores into z-scores by subtracting the statewide grade-specific mean and dividing by the statewide grade-specific standard deviation. In one district that used a commercial assessment in all four years, the means and standard deviations used to construct z-scores came from a nationally representative sample of students who took the assessment in the same grade, because statewide grade-specific means and standard deviations were not available for this assessment. For three other districts that used a commercial assessment in Year 4 only, grade-specific means and standard deviations were available for students who took the assessment in the same state, and we used this information to construct z-scores. Despite these exceptions, for simplicity we refer to z-scores as placing students within a statewide distribution of scores throughout this report.

We used the following method to eliminate outliers. First, we dropped all scores that were below the minimum or above the maximum values specified by the state assessment's technical manual. Second, we dropped all scores that were more than 5 standard deviations above or below the statewide grade-specific mean. Finally, we recoded scores by giving scores that were between 3.5 and 5 standard deviations above the statewide grade-specific mean the value of 3.5. Similarly, scores that were between -3.5 and -5 standard deviations were given the value of -3.5. Table B.1 shows the percentage of scores that were dropped or recoded, by subject and treatment status. These exclusions and modifications together affected no more than three-tenths of 1 percent of all scores.

Describing the Average Distribution of Performance Ratings and Payouts

In Chapter IV, we described the distribution of performance ratings and payouts (including performance bonuses, automatic 1 percent bonuses, and additional pay) that educators received from their TIF programs within the 10 Cohort 1 districts. We described these distributions with descriptive statistics, including minimum, average, and maximum bonus amounts; percentage of bonus amounts in specific dollar amount ranges; and percentage of performance ratings in specific ranges of the performance scale. Next, we specify how we weighted the data when calculating these descriptive statistics.

We calculated each descriptive statistic in two steps. In the first step, we calculated the descriptive statistic separately within each district. Within each district, we weighted the educator data so that each school contributed equally to the statistic for that district. Specifically, we assigned weights to educators with nonmissing values of the variable so that the sum of their weights was equal across all schools in the district. An educator *j* in school *s* was weighted by weight $W_{js} = 1/N_s$ where N_s was the number of individuals with nonmissing values of the variable in school *s*. In the second step, we took an equal-weighted average of the descriptive statistic across the 10 districts. In supplemental findings (reported in Appendix D), we modified the second step to take a weighted average of the descriptive statistic across the 10 districts, with each district weighted by the number of treatment and control schools in the final analysis sample. Those supplemental findings effectively gave each school the same weight to provide comparable results to the impact analyses, which, as described next, gave equal weight to schools as well.

Table B.1. Test Scores That Were Dropped or Recoded, Cohort 1 (Percentages)

	Year 1		Year 2		Year 3		Year 4	
Type of Exclusion or Recoding	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Math								
Dropped because score was below the minimum score or above the maximum score specified by the technical manual	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dropped because score was more than 5 standard deviations above or below the statewide mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Recoded to 3.5 standard deviations above or below the statewide mean because the score was between 3.5 and 5 standard deviations above or below the statewide mean	0.2	0.1	0.2	0.2	0.1*	0.2	0.1	0.1
Number of Students with Test Scores	**20,217**	**20,323**	**19,997**	**20,458**	**19,759**	**20,011**	**19,162**	**19,780**
Reading								
Dropped because score was below the minimum score or above the maximum score specified by the technical manual	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dropped because score was more than 5 standard deviations above or below the statewide mean	0.1	0.1	0.1	0.1	0.1*	0.0	0.0	0.0
Recoded to 3.5 standard deviations above or below the statewide mean because the score was between 3.5 and 5 standard deviations above or below the statewide mean	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1
Number of Students with Test Scores	**20,039**	**20,238**	**19,777**	**20,374**	**19,625**	**19,932**	**19,159**	**19,770**

Source: Student administrative data.

*Difference between students in treatment and control schools is statistically significant at the .05 level, two-tailed test.

Estimating Impacts of Pay-for-Performance on Educator and Student Outcomes

In this section, we describe the estimation model we used to estimate impacts of pay-for-performance on educator and student outcomes, which we presented in Chapters V and VI. For simplicity, we refer primarily to impacts on educator and student outcomes, but we used the same analytic methods to estimate differences between treatment and control schools in educators' understanding and experiences with TIF implementation, which we presented in Chapter IV.

Main Estimation Model

To estimate the impact of pay-for-performance on educator and student outcomes, we used a regression model that reflected the random assignment design—specifically, the assignment of clusters of educators or students rather than individual educators or students, and the pairing of these clusters before random assignment. We estimated the following model:

$$(2)\quad Y_{js} = \beta T_s + X'_{js}\delta + Z'_s\gamma + B'_s\pi + \varepsilon_{js},$$

where Y_{js} was the outcome for individual (student or educator) j in school s; T_s was an indicator equal to 1 for treatment schools and zero for control schools; X'_{js} was a vector of individual characteristics; Z'_s was a vector of school characteristics; B'_s was a vector of indicators for the random assignment block (matched pair of schools or matched groups of schools); δ, γ, and π were coefficient vectors to be estimated; and ε_{js} was a random error term. The coefficient β represented the average impact of pay-for-performance.

We estimated equation (2) using ordinary least squares (OLS) and employed Huber-White sandwich standard errors (Liang and Zeger 1986) that accounted for the clustering of educator and student outcomes at the level of the random assignment unit (schools or groups of schools). These standard errors were robust to any arbitrary form of correlation among outcomes in the same cluster.

We estimated equation (2) separately by year of TIF implementation. As discussed in Chapter II, to estimate impacts on teachers' attitudes and behaviors, we used teachers who reported teaching 1st grade, 4th grade, or 7th-grade math, English/language arts, or science in a study school at the time of each year's teacher survey. To estimate impacts on principals' attitudes and behaviors, we used principals who led a study school at the time of each year's principal survey. To estimate impacts on educator performance ratings, we used full-time educators who worked in a study school during the specified year. To estimate impacts on student achievement, we used students who were tested in a study school at the time of their state's spring assessment.

As shown in equation (2), we estimated a single average impact from data that were pooled across districts instead of calculating a weighted average of district-specific impacts. This approach avoided using district-specific estimates whose standard errors could be biased downward because of small numbers of clusters within each district (Donald and Lang 2007).

Covariates

We controlled for several individual and school covariates in the impact equations to improve precision and adjust for slight preexisting differences between treatment and control schools from the pre-implementation year (2010–2011 for Cohort 1 and 2011–2012 for Cohort 2). For all educator and student outcomes, the school covariates included (1) the school-level averages of math and reading test scores in the pre-implementation year, based on all students in grades 3 to 8 who were tested in the school in the pre-implementation year; and (2) the fractions of the school's enrolled students in grades 3 to 8 who were black, Hispanic, or other race/ethnicity in the pre-implementation year. We chose these covariates because, as shown in Chapter II (Table II.4), there were slight differences between treatment and control schools in average student achievement and racial/ethnic composition in the pre-implementation year.

For some outcomes, we also included individual covariates—those that measured the individual characteristics of educators or students in the analysis samples. These individual covariates allowed for further improvements in precision. The choice of whether to control for individual covariates depended on whether differences in sample composition between treatment and control schools were regarded as random errors (from sampling or random assignment) to be controlled for or whether such differences might actually reflect part of the impact of pay-for-performance. For three categories of outcomes—educators' attitudes, educators' self-reported behaviors, and educator performance ratings—we did not control for individual covariates because pay-for-performance could, in theory, affect those outcomes by way of changing the composition of the educator workforce. For one key outcome, student achievement, and one supplemental outcome, educator retention, we controlled for the characteristics of individuals in the analysis samples, as discussed next.

When estimating impacts on student achievement, we sought to compare students in treatment and control schools who were, on average, equivalent on observed background characteristics. As discussed in Chapter II, we found no evidence that pay-for-performance affected the composition of the student population in the study schools, so we regarded the slight differences in characteristics between students in treatment and control schools as random error to be controlled for. We controlled for students' math and reading test scores from the pre-implementation year; indicators for gender, race/ethnicity (indicators for blacks, Hispanics, and students with other race/ethnicity), being old for grade, being an English language learner, having an Individualized Education Program, and receipt of free or reduced-price lunch; and fixed effects for combinations of states and assessment grades. Appendix A, Tables A.25 and A.26 show the means of student characteristics (based on nonmissing values) in the math and reading analysis samples, respectively.

In supplemental analyses, we estimated the impact of pay-for-performance on educator retention (Appendix F, Tables F.8 and F.9). Our main measures of educator retention captured whether educators who worked in study schools in Year 1 continued working in the same schools in subsequent years. When estimating impacts on educator retention between Year 1 and subsequent years, we sought to compare treatment and control educators who were, on average, equivalent at the starting point (Year 1) of the analysis period. As Table II.5 shows, treatment and control educators were, indeed, similar in observed characteristics in Year 1, so we regarded any remaining slight differences between the groups as random error to be controlled for. We controlled for dichotomous indicators for gender, race/ethnicity (indicators for whites and blacks), having earned a master's degree or higher, and experience in K–12 education (indicators for 5 to 15 years and more

than 15 years), as well as the educator's age in years. Table II.5 shows the means of these variables (based on nonmissing values) in the analysis sample.

Weights

We weighted educator and student outcomes so that each school contributed equally to the average impact estimate. Specifically, we assigned weights to individuals with nonmissing outcomes so that the sum of their weights was equal across all schools. An individual *j* in school *s* was weighted by weight $W_{js} = 1/N_s$, where N_s was the number of individuals with nonmissing values for the outcome in school *s*.

Handling Missing Data

When estimating impacts on an outcome, our analysis sample included only individuals who had nonmissing values of the outcome variable, and we dropped individuals who had missing values of the outcome variable. Simulations have suggested that, for randomized controlled trials, this approach may have only a small amount of bias (0.05 standard deviations or less) when outcome data are missing at random among individuals with the same covariate values (Puma et al. 2009).

Individuals were not excluded from the analysis samples if they had missing covariate values, as long as they had nonmissing values of the outcome variable. For each covariate, we replaced missing values with a placeholder value (zero). In addition, for each covariate, we constructed an additional binary indicator for whether an individual originally had a missing value for that covariate, and we controlled for this binary indicator in the impact regressions. Simulations by Puma et al. (2009) have shown that this approach to handling missing covariate data is likely to keep estimation bias at less than 0.05 standard deviations.

Tables B.2 through B.5 show the percentages of individuals who were missing covariate values. Although there were some statistically significant differences between treatment and control schools in the percentages of students with missing covariate values, those differences did not exceed 2 percentage points. We found no significant treatment-control differences in the percentages of teachers or principals with missing covariate values, with two exceptions: treatment teachers were more likely than control teachers to have missing values for age, and treatment principals were more likely than control principals to have missing values for experience in K–12 education.

Table B.2. Students in the Math Analysis Sample with Missing Covariate Data (Percentages)

	Year 1		Year 2		Year 3		Year 4	
Missing Data on:	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Achievement in the Pre-Implementation Year[a]								
Math	34	34	57	56	78*	77	89*	87
Reading	35	34	58	57	79*	77	89*	88
Race or Ethnicity								
Missing race characteristics	0	0	0	0	0	0	0*	0
Other Characteristics								
Female	0	0	0	0	0	0	1	0
Eligible for free or reduced-price lunch	37*	37	37	37	20	20	22	20
Disabled or has an Individualized Education Program	0*	0	16	16	16	16	20	21
Over age for grade	1	1	1	1	5	6	9*	10
English language learner	0*	0	16	16	18	17	23	23
Number of Students	**20,213**	**20,322**	**19,997**	**20,457**	**19,759**	**20,011**	**19,160**	**19,779**
Number of Schools	**65**	**66**	**65**	**66**	**65**	**66**	**65**	**66**

Source: Student administrative data.

Note: Some treatment and control values differ by a statistically significant degree but are shown as equal in the table because of rounding.

[a]This characteristic is only defined for students who were tested in the pre-implementation year, so it is missing for students in grade 3 in Year 1; students in grades 3 and 4 in Year 2; students in grades 3 through 5 in Year 3; and students in grades 3 through 6 in Year 4.

*Difference between students in treatment and control schools is statistically significant at the .05 level, two-tailed test.

Table B.3. Students in the Reading Analysis Sample with Missing Covariate Data (Percentages)

	Year 1		Year 2		Year 3		Year 4	
Missing Data on:	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Achievement in the Pre-Implementation Year[a]								
Math	34	34	57	56	78*	77	89*	87
Reading	34	34	57	56	79*	77	89*	87
Race or Ethnicity								
Missing race characteristics	0*	0	0	0	0	0	0*	0
Other Characteristics								
Female	0	0	0	0	0	0	1	0
Eligible for free or reduced-price lunch	37*	37	37	37	20	20	22	20
Disabled or has an Individualized Education Program	0*	0	16	16	16	16	20	21
Over age for grade	1	1	1	1	5	6	9*	10
English language learner	0*	0	16	16	18	17	23	24
Number of Students	**20,028**	**20,228**	**19,763**	**20,359**	**19,611**	**19,927**	**19,159**	**19,770**
Number of Schools	**65**	**66**	**65**	**66**	**65**	**66**	**65**	**66**

Source: Student administrative data.

Note: Some treatment and control values differ by a statistically significant degree but are shown as equal in the table because of rounding.

[a]This characteristic is only defined for students who were tested in the pre-implementation year, so it is missing for students in grade 3 in Year 1; students in grades 3 and 4 in Year 2; students in grades 3 through 5 in Year 3; and students in grades 3 through 6 in Year 4.

*Difference between students in treatment and control schools is statistically significant at the .05 level, two-tailed test.

Table B.4. Teachers in the Educator Retention Analysis Sample in Year 1 with Missing Covariate Data, Cohort 1 (Percentages)

Missing Data on:	Treatment	Control	Difference
Sex	2	1	0
Race or Ethnicity	3	3	0
Age	4	3	1*
Education	32	32	1
Experience in K–12 Education	14	12	2
Number of Teachers	**2,151**	**2,152**	
Number of Schools	**65**	**66**	

Source: Educator administrative data.

Note: The difference between the treatment and control estimates may not equal the difference shown in the table because of rounding.

*Difference is statistically significant at the .05 level, two-tailed test.

Table B.5. Principals in the Educator Retention Analysis Sample in Year 1 with Missing Covariate Data, Cohort 1 (Percentages)

Missing Data on:	Treatment	Control	Difference
Education	38	35	3
Experience in K–12 Education	24	18	6*
Number of Principals	**64**	**69**	
Number of Schools	**62**	**65**	

Source: Educator administrative data.

Note: The difference between the treatment and control estimates may not equal the difference shown in the table because of rounding.

*Difference is statistically significant at the .05 level, two-tailed test.

Interpreting the Student Achievement Impacts

In this section, we provide additional information for interpreting the main impacts of pay-for-performance on student achievement that we reported in Chapter VI.

The main impacts on student achievement measured the impacts of pay-for-performance on schools' average test scores. In each year, our impact analysis included all students in grades 3 through 8 who were tested in a study school in that year. Because of random assignment, treatment and control schools ought to have been equivalent in average student achievement in the absence of pay-for-performance. Therefore, if the average test scores of students who were tested in treatment schools differed from the average test scores of students who were tested in control schools, we could conclude that pay-for-performance had affected schools' average student achievement. However, to interpret this analysis properly, it is important to recognize two key points.

First, positive impacts on schools' average student achievement could materialize either because pay-for-performance led to increased learning by individual students, or because it changed the types of students who enrolled in the treatment schools compared to the control schools. As discussed in Chapter II and Appendix A, we compared the background and demographic characteristics of students in grades 3 to 8 attending treatment and control schools in each year of the study and found that the students were similar on these characteristics (Tables A.25 and A.26). This suggests that pay-for-performance did not lead to changes in the types of students enrolled in the study schools. However, there remains the possibility that pay-for-performance could have changed the unmeasured characteristics of the student population, resulting in changes in average student achievement.

Second, in Years 2 through 4, the impact of pay-for-performance on student achievement reflected a mix of student exposure to pay-for-performance. For example, at the end of Year 2, each treatment school had some proportion of students who were exposed to pay-for-performance for two years (those who were enrolled at the school in both Years 1 and 2), while the remainder had one year of exposure (those who moved to the school between Years 1 and 2). Therefore, rather than addressing the question of how individual students' test scores responded to multiple years of exposure to pay-for-performance, our analyses in Years 2 through 4 addressed the question of how schools' average test scores responded to multiple years of implementing pay-for-performance.

Estimating impacts of specific durations of student exposure to pay-for-performance was generally not feasible beyond two years. To determine whether a specific duration of exposure to pay-for-performance led to increased student learning, we would need to use an alternative analysis sample that would differ from the main analysis sample in two key ways. First, it would include only students whom we could verify were enrolled in the study schools at the time of random assignment. As a result, any subsequent differences in achievement between the treatment and control samples could not be due to the entry of different types of students into the two samples and must, instead, be due to differences in learning. Second, the analysis sample would include only students who could have remained at the same school—that is, who would not have progressed beyond the school's grade span—for the whole specified duration.

Because our data included only students in grades 3 through 8, students whom we could verify were enrolled in the study schools at the time of random assignment mostly exited those schools within one to two years. In a typical elementary school—for example, spanning grades K–5—we could identify students in grades 3 through 5 who were enrolled at the time of random assignment (in the pre-implementation year). However, under normal grade progression, only third and fourth graders in the pre-implementation year could be exposed to Year 1 of pay-for-performance (when they were in grades 4 and 5, respectively); the fifth graders in the pre-implementation year were expected to exit the school before the start of Year 1. Only third graders in the pre-implementation year could be exposed to both Years 1 and 2 of pay-for-performance. No students identified in the pre-implementation year could be exposed to a third year of pay-for-performance, because all of them were expected to exit the elementary school before the beginning of the third year. Had data on grades K–2 been available, we would have been able to identify younger students enrolled in the pre-implementation year who were subsequently exposed to more than two years of pay-for-performance, but these data were not available. In typical middle schools (for example, spanning grades 6 through 8) the situation was similar, with progressively fewer cohorts who were enrolled at the time of random assignment being exposed to pay-for-performance, and no cohorts being exposed for more than two years.

In supplemental analyses, we estimated the impacts of one and two years of exposure to pay-for-performance on students' achievement, using only students who were enrolled in study schools at the time of random assignment and were young enough to stay within their schools' grade span for one and two years. Appendix F provides details on the approach and findings from these analyses. However, because these analyses could not examine longer durations of exposure and were based on narrow sets of grade levels within each school, we did not designate this approach as the main approach. Instead, as discussed earlier, our main approach estimated the impacts of implementing pay-for-performance for one, two, three, and four years on the study schools' average student achievement, using *all* students tested in the schools at the end of each year.

At the end of each year of implementation, estimating the percentages of students who had been exposed to particular durations of pay-for-performance was feasible. As discussed earlier, the impacts on schools' average student achievement in Years 2 and beyond reflected a mix of student exposure to pay-for-performance. Among students in the analysis sample, estimating the distribution of student exposure to pay-for-performance—that is, the percentages of students who had been exposed for particular durations—could provide context for interpreting the main impacts. In what follows, we describe our approach to estimating the distribution of student exposure.

The challenge with estimating the distribution of student exposure was that for many students, we could not directly observe how long they had been enrolled at their school because our data covered only grades 3 to 8. For example, at the end of Year 2, we could not identify which third graders in treatment schools had been enrolled at the same school in the previous year, and thus had been exposed to two years of pay-for-performance. Similarly, at the end of Year 3, we could not identify which third and fourth graders in treatment schools had been enrolled at the same school since Year 1, and thus had been exposed to three years of pay-for-performance. These limitations, stemming from the grade range of our data, were largely the same limitations (described earlier) that restricted our ability to estimate the impacts of specific durations of exposure to pay-for-performance.

Our solution consisted of three steps. First, we calculated the year-to-year stay rate—the percentage of students who remained at the same school since the previous year—using grades for which data were available. Second, we assumed that this year-to-year stay rate also pertained to grades for which data were not available. Third, we repeatedly applied this year-to-year stay rate to infer the percentage of students who remained at the same school for two, three, and four years. Consider the example of elementary students at the end of Year 4. We could directly calculate the percentage of fourth and fifth graders who had been at the same school in Year 3, but could not do so for third graders (because they would have been in second grade in Year 3, outside of our data). Nevertheless, we assumed that the same percentage—for example, 80 percent—applied to third graders as well. Therefore, in this hypothetical example, we would conclude that 80 percent of the analysis sample in treatment elementary schools at the end of Year 4 had been at the same school in Year 3 and, therefore, had been exposed to at least two years of pay-for-performance. Of this 80 percent, we would then assume that 80 percent had been at the same school in Year 2, under the assumption that year-to-year stay rates were stable and the choice to remain at a school was independent across years. This assumption would imply that 64 percent of the treatment analysis sample at the end of Year 4 had been at the same school in Year 2 and had been exposed to at least three years of pay-for-performance. Repeating this approach one more time, we would assume that 51 percent (80 percent of 64 percent) of the treatment analysis sample at the end of Year 4 had been at the same school in Year 1, and thus had been exposed to pay-for-performance for all four years.

One complicating factor is that not all students at a school could possibly have been at the school in previous years under normal grade progression. Consider, for example, a typical middle school with grades 6 through 8. Under normal grade progression, most seventh and eighth graders would have been at the same school in the previous year (unless they had made a move or transfer), whereas no sixth graders would have. If we calculated that 80 percent of current seventh and eighth graders had been at the same school in the previous year, it would be inappropriate to assume that this percentage also pertained to current sixth graders; it would be far more accurate to assume that no sixth graders had stayed from the previous year (setting aside the students who repeated sixth grade).

Therefore, when estimating the percentage of students who had stayed at their schools for at least a specified duration, it was important to make a distinction between potential stayers—those who could have remained at the school for that duration based on normal grade progression—and students who could not. To calculate the distribution of student exposure to pay-for-performance, the three key steps were to (1) calculate the year-to-year stay rate, λ, for potential stayers, (2) identify the fraction of students at each school who were potential stayers for each possible duration up to four years, and (3) multiply the fraction of students who were potential stayers by the estimated fraction of potential stayers who actually stayed (that is, multiply by the appropriate power of λ). Details of this approach and findings from these calculations follow.

Step 1. Calculate the year-to-year stay rate for potential stayers. To calculate λ for potential stayers, we first needed to identify students who could potentially have been at the same school in the previous year. For simplicity, we focused on students at the end of Year 4 who could potentially have been at the same school in Year 3, and included only those students whom we could potentially observe in our data in the prior year—that is, those who were currently in grades 4 through 8. We flagged these students with the following indicator variable:

$$(3)\quad ps_{js} = 1[grade_{js4} - 1 \in [\max(lowgrade_{s3}, 3),\ \min(highgrade_{s3}, 7)]],$$

where $1[\cdot]$ was the indicator function equal to 1 if the term in brackets was true and zero otherwise, j indexed students, s indexed schools, $grade_{js4}$ was student j's grade in school s in Year 4, $lowgrade_{s3}$ was the lowest grade offered in school s in Year 3 (based on the Common Core of Data), and $highgrade_{s3}$ was the highest grade offered in school s in Year 3. Therefore, ps_{js} indicated that student j in school s was a potential stayer and should have been in a grade covered by our data in Year 3. We than calculated $\hat{\lambda}$, our estimate of the probability of having been in the same school in the prior year conditional on being a potential stayer, as the weighted proportion of students with $ps_{js} = 1$ who were in fact in the same school in the prior year, using weights that gave every school equal weight. Among students who could potentially have been in the same school in the prior year, we estimated that 77 percent were actually in the same school in the prior year.

Step 2. Identify the fraction of students at each school who were potential stayers for each possible duration up to four years. We next estimated the proportion of students in Year 4 who were potential stayers for at least two, three, and four years, using students in all grades 3 through 8. For t = 2, 3, and 4, we defined ps_{jst} to be an indicator for whether a student at the end of Year 4 could potentially have

been at the school for at least t years. In other words, ps_{js2} flagged students at the end of Year 4 who could have been at the same school in both Years 3 and 4; ps_{js3} flagged students at the end of Year 4 who could have been at the same school in Years 2 through 4; and ps_{js4} flagged students at the end of Year 4 who could have been at the same school in all of the Years 1 through 4. Formally,

$$(4)\quad ps_{jst} = 1[grade_{js4} - t + 1 \in [lowgrade_{s,4-t+1}, highgrade_{s,4-t+1}]],$$

where $lowgrade_{s,4-t+1}$ and $highgrade_{s,4-t+1}$ were the lowest and highest grades offered by school s in Year $(4 - t + 1)$ of TIF implementation based on the Common Core of Data. For example, the variable ps_{js2} was equal to 1 if the Year 4 grade of student j in school S minus 1 was in the grade range offered by the school in Year 3. For each t, we took a weighted average of ps_{jst} across students, again giving each school equal weight.

Step 3. Multiply the fraction of students who were potential stayers by the estimated fraction of potential stayers who actually stayed. We estimated the proportion of students exposed to pay-for-performance for at least t years as

$$(5)\quad pe_t = \overline{ps}_t * \hat{\lambda}^{t-1}.$$

As noted above, equation (5) assumed that year-to-year stay rates were stable over time and the choice to stay in a school was independent across years.

After estimating the fraction of students who were exposed for *at least t* years using equation (5), we calculated the fraction of students who were exposed for *exactly t* years. Specifically, the fraction who were exposed for exactly *t* years was equal to the fraction who were exposed for at least *t* years minus the fraction who were exposed for at least (*t* + 1) years.

The results of this analysis are reported in Table B.6. At the end of the first year of implementation, all students in our analysis within treatment schools had, by definition, been exposed to pay-for-performance for one year only. At the end of each subsequent year of implementation, we estimated that 28 percent of students were completing their first year at their school, and therefore received exactly one year of exposure to pay-for-performance. In Year 2, the remainder of the students had, by definition, been exposed to pay-for-performance for two years. With each subsequent year of implementation, the proportion of students exposed to pay-for-performance for all years of implementation diminished.

Table B.6. Distribution of Student Exposure to Pay-for-Performance Among Students Enrolled in Treatment Schools at the End of Each Year of Implementation

	Year 1	Year 2	Year 3	Year 4
Students Exposed to Pay-for-Performance for Exactly:				
1 year	100	28	28	28
2 years		72	21	21
3 years			51	14
4 years				37

Source: Student administrative data.

Exploring Differences in Impacts Across Subgroups, Districts, and Schools

We examined not only the average impacts of pay-for-performance on the full sample of educators and students in the study, but also the degree to which impacts differed across subgroups of educators or students defined by their characteristics, and across individual districts and schools. We also explored the extent to which differences in districts' and schools' characteristics were associated with differences in impacts. This section describes the estimation models for these analyses.

Estimating the Impacts of Pay-for-Performance on Educator and Student Subgroups

We estimated the impacts of pay-for-performance within various types of subgroups. In Chapter V, we assessed how the impacts of pay-for-performance on educators' attitudes differed by teachers' teaching assignment and level of experience. In Chapter VI, we examined the impacts of pay-for-performance on the performance ratings of returning and new teachers. In Appendix F, we examined the impacts of pay-for-performance on student achievement by grade span.

In each type of subgroup analysis, the full sample of students or educators could be partitioned into either two or three mutually exclusive subgroups. For example, suppose that teachers could be partitioned into three subgroups (such as those with low, moderate, and high levels of teaching experience), identified by the binary indicators $Group1_j$, $Group2_j$, and $Group3_j$, respectively. We estimated the following model:

$$(6)\quad Y_{js} = \beta_1 T_s + \gamma_2 Group2_j + \gamma_3 Group3_j + \beta_2 (T_s \times Group2_j) + \beta_3 (T_s \times Group3_j) + X'_{js}\delta + Z'_s\gamma + B'_s\pi + \varepsilon_{js}.$$

In equation (6), the impact of pay-for-performance on teachers in groups 1, 2, and 3 were represented by the parameters β_1, $(\beta_1 + \beta_2)$, and $(\beta_1 + \beta_3)$. All other variables in equation (6) were the same as those defined in equation (2). We tested the statistical significance of the estimates of β_2 and β_3 to determine whether impacts on groups 2 and 3, respectively, differed from the impact on group 1. For scenarios in which individuals were partitioned into two (rather than three) subgroups, equation (6) was identical except that it did not include indicators and interaction terms involving $Group3_j$.

Measuring Variation in Student Achievement Impacts Across Districts

To assess whether the impacts of pay-for-performance on student achievement varied across districts (Chapter VI), we estimated a modified version of equation (2) for student achievement outcomes as follows:

$$(7)\quad Y_{js} = \sum_{d=1}^{10} \beta_d (T_s \times I_s^{(d)}) + X'_{js}\delta + Z'_s\gamma + B'_s\pi + \varepsilon_{js},$$

where $I_s^{(d)}$ was an indicator for district *d*, β_d represented the impact of pay-for-performance in district *d*, and all other variables were the same as those in equation (2). Equation (7) produced a district-specific impact estimate for each of the 10 Cohort 1 districts. An *F*-test for the joint equality of the 10 impact estimates determined if impacts varied across districts to a statistically significant degree.

Explaining Variation in Student Achievement Impacts Across Districts

Because we found that impacts differed significantly across districts (Chapter VI), we assessed whether differences in district characteristics—including TIF program characteristics and district contextual factors—were related to those differences in impacts. We estimated the relationships between district characteristics and district-level impacts in two ways (see Appendix G for detailed findings from these analyses). First, on each characteristic, we categorized districts into two subgroups that differed according to the presence or absence of the characteristic, or according to whether districts had high or low levels of the characteristic. We estimated the difference in impacts between the two subgroups using equation (6), except that the subgroup indicator ($Group2_j$) was omitted because it was redundant with random assignment block indicators.

Second, for district characteristics that could be expressed on a continuous scale (such as the average size or amount of differentiation in teachers' pay-for-performance bonuses), we examined the relationship between the continuous measure of the characteristic and the student achievement impacts. To do so, we estimated the following regression:

$$(8)\quad Y_{js} = \beta_1 T_s + \beta_2 (T_s \times C_s) + X'_{js}\delta + Z'_s\gamma + B'_s\pi + \varepsilon_{js},$$

where C_s was the continuous measure of the district characteristic, and all other variables were the same as those in equation (2). Equation (8) did not include C_s by itself as an independent variable because it was redundant with the random assignment block indicators. The coefficient of interest, β_2, measured the difference in student achievement impacts associated with a one unit increase in the district characteristic.

Measuring Variation in Student Achievement Impacts Across Treatment Schools

In Chapter VI, we also examined differences across treatment schools in the impacts of pay-for-performance on student achievement. We sought to determine how much of the variation in impacts occurred among treatment schools in the same district rather than across districts. For this analysis, we defined the impact of pay-for-performance on each treatment school as the impact in

the random assignment block (matched pair of schools or matched groups of schools) to which the treatment school belonged. Therefore, the key step in this analysis was to estimate the impact of pay-for-performance for each random assignment block. To do so, we used a modified version of equation (2) for student achievement outcomes, in which the treatment indicator was replaced by a vector of interaction terms between the treatment indicator and indicators for each of the 44 random assignment blocks:

$$(9)\quad Y_{js} = \sum_{b=1}^{44} \beta_{bd}(T_s \times B_s^{(bd)}) + X_{js}^{'}\delta + Z_s^{'}\gamma + B_s^{'}\pi + \varepsilon_{js}.$$

In equation (9), $B_s^{(bd)}$ was an indicator for random assignment block *b* in district *d*, β_{bd} represented the impact of pay-for-performance in block *b*, and all other variables were the same as those in equation (2).

After estimating the block-specific impacts, we calculated the percentage of the variation in those impacts that occurred across districts versus across random assignment blocks in the same district. To do this, we estimated the following random-effects model:

$$(10)\quad \hat{\beta}_{bd} = u_d + \omega_{bd},$$

where $\hat{\beta}_{bd}$ was the estimated block-specific impact of pay-for-performance on student achievement from equation (9), u_d was a district-specific random effect, and ω_{bd} was a block-specific random error term. We used maximum likelihood to estimate $Var(\omega_{bd})$, the variance of impacts across random assignment blocks within the same district, and expressed it as a fraction of the total variance of impacts across all random assignment blocks, $Var(\hat{\beta}_{bd})$.

Explaining Variation in Student Achievement Impacts Across Treatment Schools

Exploring the role of educator behaviors. In Appendix G, we examined the degree to which differences across schools in the impacts of pay-for-performance on educator behaviors could account for differences in impacts on student achievement. To measure educator behaviors, we used educators' responses to survey questions on topics that could reflect strategic behavior, effort, and changes in teaching practices. We also used teachers' observation ratings (from administrative data) as a direct measure of teaching practices.

We used two steps to examine these associations. First, we estimated the impacts of pay-for-performance on educator behaviors and student achievement in each random assignment block. The regression model for estimating block-specific impacts of pay-for-performance on student achievement was provided earlier (in equation [9]). We used the same regression model to estimate block-specific impacts of pay-for-performance on each measure of educator behavior.

Second, using random assignment blocks as the unit of analysis, we estimated subsequent regression models in which block-specific impacts on student achievement were the dependent variable and block-specific impacts on a specific educator behavior were the independent variable. In these regressions, we weighted each block by the number of treatment and control schools, and

we used standard errors that were robust to heteroskedasticity. The regression coefficient captured the relationship between impacts on a particular educator behavior and impacts on student achievement.

Using variation across blocks, rather than districts, to examine associations between impacts on educator behaviors and impacts on student achievement had advantages and disadvantages. On the one hand, the number of random assignment blocks (44 in Cohort 1) greatly exceeded the number of districts (10 in Cohort 1), so impacts on educator behaviors and student achievement varied more across blocks than across districts. In fact, most of the variation in impacts on student achievement occurred among blocks in the same district (76 percent in math and 63 percent in reading) rather than across districts. Thus, the greater number of blocks than districts improved our ability to detect associations, if those associations existed.

On the other hand, treatment and control schools in the same block might have differed on preexisting characteristics related to both educator behaviors and student achievement, generating an association between behaviors and achievement that was not due to pay-for-performance. For example, at the time of random assignment, suppose a treatment school had a more effective principal than the control school to which it was paired. Both before and after the start of the study, the treatment school might demonstrate greater teacher effort and higher student achievement than the control school as a result of the more effective principal. This would lead us to believe that pay-for-performance raised student achievement by way of raising teacher effort, even though neither the difference in effort nor the difference in achievement was due to pay-for-performance. Given this potential for finding associations that do not truly reflect the influence of pay-for-performance, these block-level analyses can, at best, produce suggestive evidence about whether pay-for-performance affected student achievement by way of affecting educator behaviors.

This disadvantage would also be present, but to a smaller degree, if the analysis had been based on variation across districts rather than blocks. Treatment and control schools in the same district might be imbalanced on factors related to both educator behaviors and student achievement, but the imbalance would, on average, be smaller due to the larger number of schools. Nevertheless, this analysis focused on variation across blocks rather than districts to maximize the potential for detecting associations between educator behaviors and student achievement, as described earlier.

The issue of potential treatment-control imbalances does not apply to our main estimates of the impacts of pay-for-performance on educator and student outcomes. Our main impact estimates are averages of impacts across all blocks (unlike in this analysis, which compares blocks with more positive and more negative differences in educator behaviors to assess whether the blocks also differ in impacts on student achievement). When averaging across a large number of blocks, these imbalances tend to offset each other. This is evidenced by comparing characteristics of students in treatment and control schools in the pre-implementation school year (Chapter II, Table II.4), and comparing characteristics of educators in treatment and control schools (Chapter II, Table II.5 and Appendix A, Table A.4). In addition, our main impact estimates control for preexisting differences in characteristics between treatment and control schools.

Exploring the role of baseline achievement. In Appendix G, we also examined whether baseline differences in average student achievement across schools were related to subsequent differences in the impacts of pay-for-performance on student achievement. To measure baseline achievement, we took the average of math and reading z-scores within each block in the pre-

implementation year, weighting the scores so that each school and subject within a block contributed equally to the block-level average. We then followed a procedure similar to how we examined associations between impacts on educator behaviors and impacts on student achievement. Specifically, we estimated a block-level regression in which block-specific impacts on student achievement were the dependent variable and the block-level measure of baseline achievement was the independent variable. As before, we weighted each block by the number of treatment and control schools and used standard errors that were robust to heteroskedasticity.

Estimating Average Changes in Educator Survey Responses

We used the following approach to examine whether average educator perceptions of TIF in the study schools changed from one year to the next (that is, from Years 1 to 2, Years 2 to 3, and Years 3 to 4) as bonuses were awarded and educators gained more experience with program components. First, for each school *s* and year *t*, we calculated the average response of educators (indexed by *j*) to the survey item:

$$(11)\quad \bar{Y}_{st} = \frac{1}{N_{st}} \sum_{j=1}^{N_{st}} Y_{jst},$$

where N_{st} was the number of educators in school *s* in year *t*. Second, we restricted the sample to schools (indexed from 1 to *N*) that had nonmissing values of $\bar{Y}_{st}$ in the two years being compared (Years 1 versus 2, Years 2 versus 3, or Years 3 versus 4). This analysis was not restricted just to teachers who responded to the survey in both years, because such a restriction would not have allowed the analysis to capture changes in average perceptions that resulted from the entry of new teachers in Year 2, 3, or 4 who might have had different perceptions than the teachers they replaced. Finally, using both years of data, we estimated the following regression, separately by treatment status:

$$(12)\quad \bar{Y}_{st} = \delta Later_t + \sum_{h=1}^{N} \varphi_h I_s^{(h)} + \omega_{st},$$

where $Later_t$ was an indicator for the later year (for example, Year 2 when comparing Years 1 and 2) and $I_s^{(h)}$ was an indicator for school *h*. The coefficient δ represented the average within-school change in the outcome from the earlier to the later year.

Method for Imputing Missing Values of Educator-Reported Bonus Amounts

For one set of survey items—those that asked educators to report the maximum bonus amounts for which they were eligible—we used a different approach to handling missing data than the approach used for other variables. The reason is that the occurrence of nonresponse in this set of survey items depended upon another variable: whether the educator reported being eligible for the bonus. For simplicity, we refer to a concrete example—teachers' reports of the maximum pay-for-performance bonus amounts for which they were eligible—but the same logic applies to other types of bonuses, as well as to the principal survey. Teachers were asked to report the maximum pay-for-performance bonus amount only if they indicated, in a preceding question, that they were

eligible for pay-for-performance. Among teachers who reported being eligible, there was a mix of missing and nonmissing responses to the subsequent question about maximum bonus amounts. On the other hand, among teachers who reported being ineligible, the maximum bonus amount was *always* nonmissing in the analysis because it was defined to be zero.

Consequently, among the full set of teachers who answered the eligibility question, only those who reported being eligible for pay-for-performance could have had a missing report of the maximum bonus amount. This meant that the subset of teachers who had nonmissing values for the maximum bonus amounts was disproportionately made up of teachers who reported being ineligible, and had a maximum bonus amount of zero. Therefore, if only respondents to the bonus amount question were included in the analysis without further corrections for missing data, the average reported maximum bonus amount would have been biased toward zero.

Our solution was to use multiple imputation (MI) to substitute imputed values for missing values of educator-reported bonus amounts among educators who reported being eligible for a specified type of bonus. Because MI accounts for statistical uncertainty in the imputation process, it offers the key analytic advantage of yielding appropriate standard errors for estimates that use the imputed values (Rubin 1987; Schafer and Graham 2002; Puma et al. 2009).

For teachers' reports of maximum bonus amounts, we conducted MI using five steps. First, we estimated an imputation model—separately for each year—in which the reported maximum bonus amount was modeled as a linear function of treatment status, the school covariates listed in the previous section, and random assignment block indicators. We estimated the imputation model using only teachers who reported being eligible for the specified bonus *and* reported a nonmissing bonus amount. Second, we used the estimated coefficients and standard errors from the imputation model to form a posterior distribution for the true coefficients of the imputation model. We made a random draw from this posterior distribution, producing a specific set of coefficients. Third, we used the specific set of coefficients drawn in the previous step to generate predicted values of the perceived bonus amount for all teachers who reported being eligible, including respondents and nonrespondents to the question about bonus amounts. Fourth, for each nonrespondent to the bonus amount question, we identified the three respondents who had the closest predicted values to that of the nonrespondent. Fifth, we randomly selected one of these three respondents, and the reported maximum bonus amount of the selected respondent served as the imputed value for the nonrespondent.

Steps 2 through 5 are known as predictive mean matching. In this method, there are no clear rules for choosing the number of respondents with whom a nonrespondent should be matched in step 4. Schenker and Taylor (1996) found that matching each nonrespondent with three respondents performed well in simulations. We followed this approach.

We repeated the second through fifth steps 40 times to generate 40 imputed values for each missing value of a teacher-reported bonus amount among teachers who reported being eligible for the specified bonus. We then used these imputed values along with the original, nonmissing values of reported bonus amounts to estimate the analysis model, equation (2), on the full set of teachers who answered the eligibility question. Following standard procedures, we used Rubin's (1987) rules for calculating standard errors of the estimated coefficients in equation (2).

We used the same approach to impute principal-reported maximum bonus amounts. However, unlike for teachers, we did not control for random assignment block indicators in the imputation model due to the small number of principal respondents per block. Instead, we controlled for district indicators.

Minimum Detectable Impacts

The impact estimation methods described earlier in this appendix were intended, in part, to maximize the precision of the impact estimates. To summarize the level of precision in this study, Table B.7 shows, for each key outcome in this study, the realized value of the minimum detectable impact (MDI) based on the study's actual data, sample definitions, and estimation approach. The MDI was the smallest true impact for which the study had an 80 percent probability of obtaining an estimate that was statistically significant at the 5 percent level. For each outcome, we calculated the MDI as 2.8 multiplied by the standard error of the impact estimate.

As noted in Chapter VI, the impact of pay-for-performance on student achievement in reading in Year 4 was similar in magnitude to the impact in previous years, but in Year 4 this impact was not statistically significant. The reason is that the impact estimate in Year 4 was less precise than it was in previous years. Consistent with this finding, the MDI was highest for reading in Year 4 (Table B.7). The larger MDI in Year 4 may be in part due to the possibility that the baseline measures of schools' average student achievement—key covariates in the regression models for estimating impacts—had less explanatory power for predicting student achievement in Year 4 than in previous years as many districts switched to new assessments in Year 4 (see Chapters II and VI for a discussion of these new assessments). To explore this possibility, in each year we tested the null hypothesis that baseline school-level achievement was unrelated to student achievement in the outcome year. Consistent with this possibility, the *F*-statistic for this test was lower in Year 4 ($F(89,2)=13.81$) than in both Year 2 ($F(89,2)=34.39$) and Year 3 ($F(89,2)=23.98$), though not Year 1 ($F(89,2)=13.38$). These results indicate that baseline school-level achievement had a weaker relationship with student achievement outcomes in Year 4 than in two out of the three previous years.

Table B.7. Realized Values of Minimum Detectable Impacts

Outcome	Units	Minimum Detectable Impact	Standard Error of Impact Estimate
School Achievement Growth Ratings, Year 1	Points on 1-to-4 scale	0.48	0.17
School Achievement Growth Ratings, Year 2	Points on 1-to-4 scale	0.42	0.15
School Achievement Growth Ratings, Year 3	Points on 1-to-4 scale	0.39	0.14
School Achievement Growth Ratings, Year 4	Points on 1-to-4 scale	0.48	0.17
Teachers' Classroom Observation Ratings, Year 1	Points on 1-to-4 scale	0.06	0.02
Teachers' Classroom Observation Ratings, Year 2	Points on 1-to-4 scale	0.08	0.03
Teachers' Classroom Observation Ratings, Year 3	Points on 1-to-4 scale	0.06	0.02
Teachers' Classroom Observation Ratings, Year 4	Points on 1-to-4 scale	0.11	0.04
Teachers' Classroom Achievement Growth Ratings, Year 1	Points on 1-to-4 scale	0.22	0.08
Teachers' Classroom Achievement Growth Ratings, Year 2	Points on 1-to-4 scale	0.14	0.05
Teachers' Classroom Achievement Growth Ratings, Year 3	Points on 1-to-4 scale	0.14	0.05
Teachers' Classroom Achievement Growth Ratings, Year 4	Points on 1-to-4 scale	0.17	0.06
Observation Ratings for Principals, Year 1	Points on 1-to-4 scale	0.22	0.08
Observation Ratings for Principals, Year 2	Points on 1-to-4 scale	0.28	0.10
Observation Ratings for Principals, Year 3	Points on 1-to-4 scale	0.20	0.07
Observation Ratings for Principals, Year 4	Points on 1-to-4 scale	0.25	0.09
Student Math Achievement, Year 1	Student *z*-score units	0.06	0.02
Student Math Achievement, Year 2	Student *z*-score units	0.06	0.02
Student Math Achievement, Year 3	Student *z*-score units	0.07	0.02
Student Math Achievement, Year 4	Student *z*-score units	0.07	0.02
Student Reading Achievement, Year 1	Student *z*-score units	0.05	0.02
Student Reading Achievement, Year 2	Student *z*-score units	0.04	0.01
Student Reading Achievement, Year 3	Student *z*-score units	0.04	0.02
Student Reading Achievement, Year 4	Student *z*-score units	0.06	0.02

Sources: Educator and student administrative data.

APPENDIX C

SUPPLEMENTAL FINDINGS ON PROGRAMS AND EXPERIENCES OF ALL 2010 TIF DISTRICTS FOR CHAPTER III

THIS PAGE IS INTENTIONALLY BLANK

This appendix supplements the findings presented in Chapter III and includes additional analyses on TIF districts' programs and challenges implementing TIF. As explained in Chapter II, the findings presented in this chapter are from districts that were included in a 2010 TIF grant and responded to the district survey. TIF districts completed surveys in the middle of the 2011–2012 school year (Year 1) and the end of the 2012–2013 (Year 2), 2013–2014 (Year 3), and 2014–2015 (Year 4) school years.

TIF Districts and Their Programs

In this section, we provide more details on the measures of educator effectiveness and additional pay opportunities for teachers and principals among all TIF districts. Table C.1 shows additional information on classroom observations for teachers and observations of school practices for principals, as reported by TIF district staff. Table C.2 presents additional pay opportunities for extra work or responsibilities (such as working in a hard-to-staff school) that were not discussed in detail in Chapter III.

Table C.1. Observations of Classroom or School Practices to Evaluate Teachers and Principals (Percentages Unless Otherwise Noted)

	Year 1 (2011–2012)	Year 2 (2012–2013)	Year 3 (2013–2014)	Year 4 (2014–2015)
Teachers				
Average Number of Classroom Observations per School Year	5	4	3	4
Average Length of Classroom Observations (in minutes)	43	45	45	44
Conducting Observations by a Trained Observer [a]	98	99	96	99
Classroom Observations are Conducted by:				
Principal or other administrators at the teacher's school	95	97	93	97
Teacher leaders or peer observers[b]	53	57	52	55
District administrative staff	8	50	49	50
Externally hired observers (non-district employees)	4	9	7	12
Number of Districts—Range[c]	**146-153**	**150-151**	**135-138**	**128-133**
Principals				
Average Number of Observations per School Year	—	3	3	4
Average Length of Observations (in minutes)	—	54	47	45
Observations are Conducted By:				
Superintendent	—	50	49	55
Other central office administrator from the same district	—	58	51	53
Administrator from another district	—	1	4	2
Number of Districts—Range[c]	—	**147-151**	**138-141**	**133-137**

Sources: District surveys (2012, 2013, 2014, and 2015).

Note: The 2012 district survey asked whether observations were used to evaluate principal effectiveness but did not ask about the number or length of observations or who conducted them.

[a]The 2012 district survey asked districts to report on whether classroom observations were conducted. Subsequent surveys asked districts whether observations were conducted by a trained observer.

[b]Department heads, coaches, other senior teachers (at or outside school).

[c]Sample sizes are presented as a range based on the data available for each row in the table.

— is not available.

Table C.2. Additional Pay Opportunities for Teachers and Principals for Additional Factors, Year 4

	Percentage of TIF Districts That Offered Additional Pay	Average Maximum Amount of Additional Pay in Districts Offering it
Teachers		
Additional Factors		
Teaching in a hard-to-staff school or high-need subject area	27	$3,696
Attending professional development activities or enrolling in graduate-level courses	29	$1,128
Number of Districts—Range[a]	**136**	**34-37**
Principals		
Additional Factors		
Working in a hard-to-staff school	13	$8,021
Attending professional development activities or enrolling in graduate-level courses	20	$1,199
Number of Districts—Range[a]	**136**	**14-24**

Source: District survey, 2015.

Note: Table reports on activities funded by TIF.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

Challenges in Implementing TIF

This section provides additional detail on the findings presented in Chapter III on challenges TIF districts faced implementing TIF. Table C.3 presents the percentage of districts that indicated an activity was a "major challenge," "minor challenge," or "not a challenge" in Years 2, 3, and 4. The sample was restricted to the 128 districts that responded to the 2013, 2014, and 2015 district surveys.

Table C.3. Challenges Implementing TIF in Year 2, Year 3, and Year 4 (Percentages)

	In Year 2, Percentage of All TIF Districts Reporting Activity Was:			In Year 3, Percentage of All TIF Districts Reporting Activity Was:			In Year 4, Percentage of All TIF Districts Reporting Activity Was:		
Activity	Major Challenge	Minor Challenge	Not a Challenge	Major Challenge	Minor Challenge	Not a Challenge	Major Challenge	Minor Challenge	Not a Challenge
Incorporating Student Achievement Growth into Teacher Evaluations									
Calculating student achievement growth	27	24	50	20	33	47	15	32	53
Attributing student achievement growth to individual teachers	29	28	43	20+	35	44	12+	34	53
Explaining student achievement measures to educators	27	45	27	20	58+	22	19	53	28
Providing useful and timely feedback on student achievement measures to educators	31	40	29	20+	47	33	26	37	37
Collecting and storing data linking teachers to student achievement data	20	38	41	18	39	43	11+	35	54+
Teacher Classroom Observations									
Choosing a classroom observation tool	8	22	70	2+	8+	90+	2	11	87
Finding a tool that is ready for implementation	9	18	73	2+	7+	91+	0	10	90
Hiring observers	2	19	79	3	10+	87+	3	10	87
Training observers to use the tool	12	46	42	5+	39	56+	2	39	59
Scheduling and/or conducting observations	24	53	24	17	55	28	11	55	34
Providing useful and/or timely feedback from observations	27	46	27	16+	50	34	9+	53	38
Collecting and storing observation data	14	35	51	3+	37	60	3	26+	71+
Principal Observations									
Choosing a principal observation tool	15	33	53	2+	18+	80+	2	17	80
Finding a tool that is ready for implementation	16	26	58	2+	17	80+	5	13	82
Hiring observers	2	15	83	3	9	87	4	6	90
Training observers to use the tool	5	41	54	4	31	65+	2	28	70
Scheduling and/or conducting observations	14	50	37	12	41	47+	6	43	51
Providing useful and/or timely feedback from observations	14	41	45	4+	48	48	5	37+	58+

Table C.3 *(continued)*

Activity	In Year 2, Percentage of All TIF Districts Reporting Activity Was:			In Year 3, Percentage of All TIF Districts Reporting Activity Was:			In Year 4, Percentage of All TIF Districts Reporting Activity Was:		
	Major Challenge	Minor Challenge	Not a Challenge	Major Challenge	Minor Challenge	Not a Challenge	Major Challenge	Minor Challenge	Not a Challenge
Pay-for-Performance Bonuses									
Defining the criteria for earning a pay-for-performance bonus or the amount of the bonus	22	45	34	10+	30+	59+	9	27	63
Calculating pay-for-performance bonuses	20	30	51	6+	30	64+	8	28	64
Distributing pay-for-performance bonuses	9	34	56	4+	28	68+	2	29	69
Communicating the TIF Program to Educators or Other Stakeholders									
Communicating the TIF program to educators	14	47	39	5+	45	50+	6	39	55
Communicating bonus payouts to educators	13	43	45	5+	40	55	6	34	59
Communicating with other stakeholders	13	51	36	9	48	44	7	44	49
Obtaining or Maintaining Support for the TIF Program									
Teachers or teachers' union or association	12	29	59	4+	31	65	7	24	69
Principals or principals' union or association	2	18	80	2	16	82	2	13	84
Superintendent	2	13	85	1	13	86	4	12	84
School board	2	30	68	4	20+	76	6	18	76
Parents or broader community	2	26	72	3	25	72	2	17	80
Other TIF Issues									
Choosing educators for additional roles and responsibilities	8	42	50	4	34	62+	2	41	58
Sustainability of the TIF program	63	30	8	48+	33	19+	58+	31	11+
Number of Districts—Range[a]	**120-128**	**120-128**	**120-128**	**120-128**	**120-128**	**120-128**	**120-128**	**120-128**	**120-128**

Sources: District surveys (2013, 2014, and 2015).

Note: The sample was restricted to the 128 TIF districts that responded to the 2013, 2014, and 2015 district surveys.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

+Difference between prior year is significant at the .05 level, two-tailed test.

THIS PAGE IS INTENTIONALLY BLANK

APPENDIX D

SUPPLEMENTAL FINDINGS ON TIF IMPLEMENTATION IN EVALUATION DISTRICTS FOR CHAPTER IV

THIS PAGE IS INTENTIONALLY BLANK

This appendix supplements the findings presented in Chapter IV on TIF implementation in the evaluation districts. We provide additional details on the four required components, districts' communication activities about the TIF program, and educators' reports about their understanding of the TIF program.

As discussed in Chapter II, evaluation districts were classified into two cohorts—Cohort 1 and Cohort 2—according to the year in which we randomly assigned their schools to a treatment group or a control group. The 10 districts whose schools were randomly assigned in spring and summer 2011 were classified as Cohort 1. Three additional districts, whose schools were randomly assigned in spring and summer 2012, were classified as Cohort 2. Cohort 1 districts completed four years of implementation. Year 1 represents the first year of implementation (2011–2012), Year 2 the second (2012–2013), Year 3 the third year of implementation (2013–2014), and Year 4 the fourth (2014–2015). Cohort 2 districts completed only three years of implementation, 2012–2013, 2013–2014, and 2014–2015, referred to as Years 1, 2, and 3 for this cohort.

The analyses in Chapter IV were based on Cohort 1 only and, in general, focused on findings over four years of implementation. This appendix supplements the findings in Chapter IV in several ways: (1) we present findings for Cohort 1 that were noted but not included in the chapter; (2) we provide findings based on both Cohorts 1 and 2; (3) we show findings when we weight data on pay-for-performance bonuses by the number of schools in a district, rather than giving each district equal weight; and (4) we present findings from subgroup analyses to examine factors that might explain differences in teachers' understanding of their bonus eligibility.

Implementation of the Required Components of TIF

In this section, we show results presented in Chapter IV about the components of TIF programs that the evaluation districts designed and implemented, focusing on the four required components under the TIF grant: (1) measures of educator effectiveness, (2) pay-for-performance bonuses, (3) additional pay opportunities, and (4) professional development.

Requirement 1: Measures of Educator Effectiveness

TIF grantees were required to measure educator effectiveness based on student achievement growth and multiple observations by trained observers. Chapter IV focused on Cohort 1 districts' implementation of this requirement. Table D.1 shows additional details on teacher classroom observations as reported by the Cohort 1 districts in Year 4.

Figure IV.1 in Chapter IV illustrates that school achievement growth and classroom observations sometimes identified the same teachers as high-performing in Year 4, but many had higher ratings from observations of their classroom practices than from school achievement growth. Table D.2 compares principals' ratings based on observations of their school practices and school achievement growth for Cohort 1 in Year 4. Nearly 60 percent of principals received a higher rating based on observations than on the achievement growth of students in their schools. Table D.3 compares teachers' ratings on classroom achievement growth and classroom observations. In Year 4, about one-quarter (26 percent) of teachers received similar ratings based on classroom observations and classroom achievement growth, and nearly 50 percent received a higher classroom observation rating than classroom achievement growth rating.

Table D.1. Classroom Observations to Evaluate Teachers in Year 4, Cohort 1 (Percentages Unless Otherwise Noted)

	Evaluation Districts
Average Number of Observations per School Year	4
Average Length of Observations (in minutes)	39
Observations are Conducted by a Trained Observer	100
Observations are Conducted by	
Principal or other administrators at the teacher's school	89
Teacher leaders or peer observers[a]	33
District administrative staff	56
Externally hired observers (nondistrict employees)	11
Number of Districts	**9-10**

Source: District survey, 2015.

[a]Includes department heads, coaches, or other senior teachers (at or outside school).

Table D.2. Comparison of Principals' Ratings on Observations and School Achievement Growth in Year 4, Cohort 1 (Percentages)

	Principal Observation Rating			
School Achievement Growth Rating	"Ineffective"	"Somewhat Effective"	"Effective"	"Highly Effective"
"Ineffective"	5	6	8	15
"Somewhat Effective"	<3[a]	12	5	20
"Effective"	0	3	9	3
"Highly Effective"	0	<3[a]	5	7
Number of Principals - 117				

Source: Educator administrative data.

Notes: Categories are study-constructed labels to represent quarters of a 1-to-4 rating scale. "Ineffective" = bottom quarter (1.00 to 1.75); "Somewhat Effective" = second quarter (1.75 to 2.50); "Effective" = third quarter (2.50 to 3.25); "Highly Effective" = top quarter (3.25 to 4.00). The table is based on principals with ratings on both observations and school achievement growth in Year 4. One district did not provide school achievement growth ratings for control schools in Year 4, so control schools from this district were excluded from this analysis.

[a]To reduce the risk of disclosing information on an individual, the exact percentage is not shown.

Table D.3. Comparison of Teachers' Ratings on Classroom Observations and Classroom Achievement Growth in Year 4, Cohort 1 (Percentages)

	Classroom Observation Rating				
Classroom Achievement Growth Rating	"Ineffective"	"Somewhat Effective"	"Effective"	"Highly Effective"	**Number of Teachers**
"Ineffective"	1	14	16	3	**576**
"Somewhat Effective"	0	10	12	2	**302**
"Effective"	0	4	9	2	**283**
"Highly Effective"	0	6	15	6	**591**
Number of Teachers	**21**	**518**	**926**	**287**	**1,752**

Source: Educator administrative data.

Notes: Categories are study-constructed labels to represent quarters of a 1-to-4 rating scale. "Ineffective" = bottom quarter (1.00 to 1.75); "Somewhat Effective" = second quarter (1.75 to 2.50); "Effective" = third quarter (2.50 to 3.25); "Highly Effective" = top quarter (3.25 to 4.00). The table is based on teachers with ratings on both classroom observations and classroom achievement growth in Year 4.

Figure IV.2 in Chapter IV compares teachers' classroom observation ratings in Years 3 and 4. Tables D.4 and D.5 compare teachers' school achievement growth ratings and classroom achievement growth ratings in Years 3 and 4, respectively. About half (51 percent) of teachers received similar ratings on these measures in both years.

Table D.4. Comparison of Teachers' School Achievement Growth Ratings in Years 3 and 4, Cohort 1 (Percentages)

	Teacher's School Achievement Growth Rating in Year 4				
Teacher's School Achievement Growth Rating in Year 3	"Ineffective"	"Somewhat Effective"	"Effective"	"Highly Effective"	**Number of Teachers**
"Ineffective"	14	6	2	4	**780**
"Somewhat Effective"	14	26	4	5	**1,256**
"Effective"	1	2	7	3	**390**
"Highly Effective"	6	2	2	4	**631**
Number of Teachers	**994**	**1,039**	**421**	**603**	**3,057**

Source: Educator administrative data.

Notes: Categories are study-constructed labels to represent quarters of a 1-to-4 rating scale. "Ineffective" = bottom quarter (1.00 to 1.75); "Somewhat Effective" = second quarter (1.75 to 2.50); "Effective" = third quarter (2.50 to 3.25); "Highly Effective" = top quarter (3.25 to 4.00). The table is based on teachers with school achievement growth ratings in both Years 3 and 4. One district did not provide school achievement growth ratings for control schools in Year 4, so control schools from this district were excluded from this analysis.

Table D.5. Comparison of Teachers' Classroom Achievement Growth Ratings in Years 3 and 4, Cohort 1 (Percentages)

	Classroom Achievement Growth Rating in Year 4				
Classroom Achievement Growth Rating in Year 3	"Ineffective"	"Somewhat Effective"	"Effective"	"Highly Effective"	**Number of Teachers**
"Ineffective"	21	6	2	3	**432**
"Somewhat Effective"	9	10	4	5	**384**
"Effective"	1	4	4	5	**182**
"Highly Effective"	2	2	4	16	**463**
Number of Teachers	**441**	**238**	**231**	**551**	**1,461**

Source: Educator administrative data.

Notes: Categories are study-constructed labels to represent quarters of a 1-to-4 rating scale. "Ineffective" = bottom quarter (1.00 to 1.75); "Somewhat Effective" = second quarter (1.75 to 2.50); "Effective" = third quarter (2.50 to 3.25); "Highly Effective" = top quarter (3.25 to 4.00). The table is based on teachers with classroom achievement growth ratings in both Years 3 and 4.

Requirement 2: Pay-for-Performance Bonuses

This section presents additional information on districts' pay-for-performance programs and analyses on pay-for-performance bonuses. The additional analyses examine whether the findings change if we base findings on both Cohorts 1 and 2 or weight districts by the number of schools (rather than weight each district equally). We also provide information that supports statements in Chapter IV (such as the distribution of bonuses by district) and provide findings for Cohort 1 for years that were not provided in Chapter IV. We provide additional information first for teachers, then for principals.

Table IV.3 in Chapter IV shows the key features of Cohort 1 pay-for-performance bonus programs in Year 4. Tables D.6 and D.7 provide additional information on Cohorts 1 and 2 pay-for-performance programs. Table D.6 provides summary information on key features of districts' programs, whereas Table D.7 provides more detailed information on their programs. To ensure districts' confidentiality, the numbering of the districts in these tables does not mirror the lettering of districts in other parts of the report.

Table D.6. Key Features of Evaluation Districts' Teacher Pay-for-Performance Bonus Programs in 2014–2015 (Year 4 for Cohort 1 and Year 3 for Cohort 2)

	Cohort 1 Districts										Cohort 2 Districts		
Key Program Feature	1	2	3	4	5	6	7	8	9	10	11	12	13
Teachers could receive a bonus for multiple performance measures	X	X	X	X	X	X	X	X	X	X	X		
Teachers could receive a bonus for a single overall performance rating												X	X
Teachers could receive a bonus for school achievement growth	X	X	X	X	X	X	X	X	X	X	X		
Teachers could receive a bonus for achievement growth of their students			X	X	X		X	X	X	X	X		
Teachers could receive a bonus for the achievement growth of a student subgroup					X	X			X				
Student achievement growth was measured by a value-added model	X	X	X	X	X			X	X	X	X	X	X
Teachers could receive a bonus for classroom observations	X	X	X	X		X	X			X	X		
A maximum bonus was specified for each performance measure or for overall rating		X			X	X	X	X	X			X	X
Maximum bonus possible depended on the number of bonus recipients	X		X	X						X	X		
Bonus amount for a performance measure could be affected by a factor besides the teacher's rating on the measure[a]			X	X	X	X		X	X	X	X		
District changed some aspect of its program from the 2013–2014 to the 2014–2015 school years			X						X	X		X	X

Sources: District interviews (2012, 2013, 2014, and 2015), grantees' Annual Performance Report (APR) documents, and technical assistance documents.

Note: Grantees submit an APR to the U.S. Department of Education that describes how educators are evaluated.

[a]For example, a district could have required teachers to meet a minimum classroom observation rating to receive a bonus based on classroom achievement growth or taken teachers' attendance into account when determining the size of bonuses.

Table D.7. Detailed Information on Measures and Criteria Used for Evaluation Districts' Teacher Pay-for-Performance Bonus Programs in 2014–2015 (Year 4 for Cohort 1 and Year 3 for Cohort 2)

Cohort 1

District 1

Key program features

- Teachers could receive bonuses for school achievement growth and classroom observations
- Maximum bonus possible for classroom observations depended on number of bonus recipients; maximum bonus possible for other measures was fixed

Specific information on performance measures and bonus criteria

1. Bonuses based on school achievement growth
   - Based on school value-added score
   - School's 2012–2013 value-added ranking was compared to school's 2011–2012 value-added ranking
   - Maximum bonus received if school met Target 1, defined as the value-added score the school was estimated to have 25 percent probability of achieving based on 2011–2012 performance
   - Smaller bonus received if school met Target 2, defined as the value-added score the school was estimated to have 50 percent probability of achieving based on 2011–2012 performance

2. Bonuses based on classroom observations
   - Teachers were observed 6 times during the year
   - Pool of money set aside for observation bonuses
   - Could receive up to 4 points for each standard on the rubric
   - Awards were based on the total number of points a teacher received
   - The total possible point count was partitioned into 4 tiers
   - Tiers were determined at the end of the school year
   - Teachers received a bonus if their total score fell within the top 3 tiers and received the maximum bonus if their total score fell in the top tier

District 2

Key program features

- Teachers could receive bonuses for school achievement growth in math, school achievement growth in ELA, and classroom observations
- Set an absolute maximum bonus possible for each criterion

Specific information on performance measures and bonus criteria

1. Bonuses based on school achievement growth in math
   - Based on school math value-added score
   - School achievement growth was partitioned into 4 tiers: (a) Tier 1: 90–100th percentile, (b) Tier 2: 80–89th percentile, (c) Tier 3: 65–79th percentile; (d) Tier 4: below the 65th percentile
   - Teachers in Tier 4 schools did not receive a bonus
   - The maximum bonus went to teachers in the Tier 1 schools

2. Bonuses based on school achievement growth in ELA
   - Based on school ELA value-added score
   - School achievement growth in ELA was partitioned into 4 tiers: (a) Tier 1: 90–100th percentile, (b) Tier 2: 80–89th percentile, (c) Tier 3: 65–79th percentile; (d) Tier 4: below the 65th percentile
   - Teachers in Tier 4 schools did not received a bonus
   - The maximum bonus went to teachers in the Tier 1 schools

3. Bonuses based on classroom observations
   - Teachers were observed 6 times during the year
   - Scores ranged from 1 to 4
   - Teachers received the maximum bonus if their average score was 3.7 or above and they earned at least a 3 on each evaluation
   - Teachers received the second highest bonus if their average score was between 3.4 and 3.69 and they earned at least a 2 on each evaluation
   - Teachers received the smallest bonus if their average score was between 3.0 and 3.39

Districts 3 and 4

Key program features

- Teachers could receive bonuses for school achievement growth, classroom achievement growth (if teaching tested grades and subjects), and classroom observations
- For each performance measure, teachers' ratings were translated into "shares" that determined their bonus amounts
- Maximum bonus possible for each measure depended on the number of bonus recipients
- Bonus based on observations depended on a factor besides the observation score
- District 3 revised some aspect of program between the 2012–2013 and 2013–2014 school years

Specific information on performance measures and bonus criteria

1. Bonuses based on school achievement growth
   - For teachers in tested grades and subjects, 20 percent of their potential bonus was based on school achievement growth
   - For teachers in nontested grades and subjects, 50 percent of their potential bonus was based on school achievement growth
   - Based on school value-added score placed onto a 1–5 scale
   - Teachers in schools rated 1 or 2 earned 0 shares; teachers in schools rated 3 earned 50 shares; teachers in schools rated 4 earned 75 shares; teachers in schools rated 5 earned 100 shares.

2. Bonuses based on classroom achievement growth
   - For teachers in tested grades and subjects, 30 percent of their potential bonus was based on classroom achievement growth
   - Based on classroom value-added score placed onto a 1–5 scale
   - Teachers rated 1 or 2 earned 0 shares; teachers rated 3 earned 1 share; teachers rated 4 earned 6 shares, teachers rated 5 earned 10 shares

3. Bonuses based on classroom observations
   - For all teachers, 50 percent of their potential bonus was based on classroom observations
   - In District 3, teachers were observed 3 times during the year
   - In District 4, teachers were observed 4 times during the year
   - Teachers were classified into 1 of 4 possible categories: (1) career teacher, (2) teacher in a hard-to-fill position, (3) mentor teacher, or (4) master teacher
   - The number of shares earned depended on the teacher's observation rating and position
   - Teachers earned more shares the higher their observation score, but had to be rated above a minimum score to receive any shares
   - The minimum observation score required to receive shares varied depending on their position
   - For a given observation rating, career teachers and teachers in a hard-to-fill position earned more shares than mentor or master teachers

District 5

Key program features

- Teachers could receive bonuses for school achievement growth, grade-level achievement growth, and classroom achievement growth (if teaching tested grades and subjects)
- Set an absolute maximum bonus possible for each criterion
- Teachers could not receive a bonus for classroom observations; however, a teacher's total bonus (based on other measures) was reduced by 25 percent if the teacher's observation score did not meet a minimum threshold
- Bonus based on grade-level achievement growth depended on a factor besides the student subgroups' score

Specific information on performance measures and bonus criteria

1. Bonuses based on school achievement growth
   - Based on school value-added score
   - Bonuses were awarded to teachers in schools whose school value-added score was at least 1 standard error (SE) above the state average

2. Bonuses based on grade-level achievement growth
   - Based on grade-level value-added score
   - All teachers joined a grade-level team
   - Bonus were awarded to teachers in grades whose grade-level value-added score was at least 1 SE above the state average
   - Bonus depended on the percentage of time teacher spent working with that grade

3. Bonuses based on classroom achievement growth
   - Based on classroom value-added score
   - Awards of increasing value were given to teachers whose value added score was at least (1) 0.5 SE above the state average, (2) 1.0 SE above the state average, (3) 1.5 SE above the state average, and (4) 2.0 SE above the state average

District 6

Key program features
- Teachers could receive bonuses for school achievement growth, achievement growth of student subgroups, and classroom observations
- Set an absolute maximum bonus possible for each criterion
- Bonus based on classroom observations depended on factors besides the observation score

Specific information on performance measures and bonus criteria

1. Bonuses based on school achievement growth
   - Based on Colorado Growth Model
   - Each school set a goal for its Colorado Growth Model score
   - Bonuses were awarded if the school met its goal

2. Bonuses based on achievement growth of student subgroups
   - All teachers were assigned to a team
   - Teams of teachers set goals for the achievement growth of their students
   - Bonuses were awarded if the team met its goal

3. Bonuses based on classroom observations
   - Teachers were observed an average of 3 times per year
   - The size of the bonus depended on the teacher's years of education, highest degree earned, and score on the rubric

District 7

Key program features
- Teachers could receive bonuses for school achievement growth, classroom achievement growth, classroom observations, and school achievement levels
- Set an absolute maximum bonus possible for each criterion
- Revised some aspect of program between the 2012–2013 and 2013–2014 school years

Specific information on performance measures and bonus criteria

1. Bonuses based on school achievement growth
   - Fall-to-spring growth targets were set for each student based on the student's fall achievement
   - Schools were rated on a 1–4 scale based on how their students' growth compared with the targets
   - Teachers in schools rated 4 earned a bonus worth 2 percent of average teacher salary; teachers in schools rated 3 earned a bonus worth 1.5 percent of average teacher salary; teachers in schools rated 2 earned a bonus worth 1 percent of average teacher salary; teachers in schools rated 1 did not earn a bonus for this measure

2. Bonuses based on classroom achievement growth
   - Bonus for the measure was available for math, science, and ELA teachers only
   - Fall-to-spring growth targets were set for each student based on the student's fall achievement
   - Teachers were rated on a 1–4 scale based on how their students' growth compared with the targets Teachers rated 4 earned a bonus worth 5 percent of average teacher salary; teachers rated 3 earned a bonus worth 3.5 percent of average teacher salary; teachers rated 2 earned a bonus worth 1 percent of average teacher salary; teachers rated 1 did not earn a bonus for this measure

3. Bonuses based on classroom observations
   - Bonus awarded for score on third party rating of video of a classroom lesson
   - Teachers were rated on a 1–4 scale
   - For math, science, and ELA teachers, those rated 4 earned a bonus worth 4 percent of average teacher salary; those rated 3 earned a bonus worth 3 percent of average teacher salary; those rated 2 earned a bonus worth 1 percent of average teacher salary; those rated 1 did not earn a bonus for this measure
   - For other teachers, those rated 4 earned a bonus worth 6 percent of average teacher salary; those rated 3 earned a bonus worth 4 percent of average teacher salary; those rated 2 earned a bonus worth 1 percent of average teacher salary; those rated 1 did not earn a bonus for this measure

4. Bonuses based on school's achievement level
    - Bonus awarded for school's performance score on the state test
    - Ratings were put on a 1–4 scale
    - Teachers in schools rated 4 earned bonus worth 2 percent of average teacher salary; teachers in schools rated 3 earned bonus worth 1.5 percent of average teacher salary; teachers in schools rated 2 earned bonus worth 1 percent of average teacher salary; teachers in schools rated 1 did not earn a bonus for this measure

District 8

Key program features

- All teachers could receive a bonus for school achievement growth; teachers in tested grades and subjects (except grade 4 teachers) could also receive a bonus for classroom achievement growth
- Set an absolute maximum bonus possible for each criterion
- Teachers could not receive a bonus for classroom observations; however, a teacher had to be rated at least proficient on the summative observation score to earn a bonus for school or classroom achievement growth

Specific information on performance measures and bonus criteria

1. Bonuses based on school achievement growth
    - Based on school value-added score
    - School must receive a rating of "exceeds expected growth" to receive bonus
    - Schools were rated as "exceeds expected growth" if their value-added score was at least 1 standard deviation above the state mean

2. Bonuses based on classroom achievement growth
    - Bonus available to teachers in tested grades and subjects
    - Based on classroom value-added score
    - Teachers with scores between 1 and 1.9 standard deviations above the mean received a rating of 4; teachers with scores at least 2 standard deviations above the mean received a rating of 5
    - Bonuses awarded to teachers with ratings of 4 or 5
    - Math teachers received larger bonuses than non-math teachers

District 9

Key program features

- Teachers could receive bonuses for school achievement growth, achievement growth attributable to teacher teams, achievement growth for subgroups of students, classroom achievement growth, and school achievement levels
- Set an absolute maximum bonus possible for each criterion
- Teachers could not receive a bonus for classroom observations; however, a teacher had to be rated 3 or above on the summative observation measure to receive bonuses based on other measures
- Teachers had their bonuses prorated if they were in attendance for less than 95 percent of the school year, and could not receive any bonus if they were in attendance for less than 80 percent of the school year
- Revised some aspect of program between the 2012–2013 and 2013–2014 school years

Specific information on performance measures and bonus criteria

1. Bonuses based on school achievement growth and achievement growth for student subgroups
    - Could receive a bonus for four measures of school value-added—based on all students in the school, students with disabilities, gifted students, and for low performing students (in bottom 20 percent)
    - Teachers in schools whose value-added score on any of the school value-added measures was rated above expected growth earned a bonus

2. Bonuses based on achievement growth attributable to teacher teams
    - All teachers joined one of four subject-matter teams: math, ELA, science, or social studies
    - Teachers in a subject-matter team received a bonus if their school's value-added score for the specified subject was rated above expected growth

3. Bonuses based on classroom achievement growth
    - Teachers could receive bonuses based on student learning objectives (SLO)

4. Bonuses based on school achievement levels
    - Teachers in schools whose performance index increased by a minimum required amount earned a bonus
    - The minimum required gain in the performance index depended on the school's performance index in the prior year

5. Bonuses based on achievement levels attributable to teacher teams
   - All teachers joined one of four subject-matter teams: math, ELA, science, or social studies
   - Teams set goals for student achievement in their subject; Teachers in teams that met their goals received a bonus

District 10

Key program features
- Teachers could receive bonuses for school achievement growth, classroom achievement growth, and classroom observations
- Maximum bonus possible for each measure depended on the number of bonus recipients
- Bonus based on classroom observations depended on a factor besides the observation score
- Revised some aspect of program between the 2012–2013 and 2013–2014 school years

Specific information on performance measures and bonus criteria
1. Bonuses based on school achievement growth
   - Based on school value-added scores placed on a 1–5 scale
   - 20 percent of their potential bonus was based on school achievement growth

2. Bonuses based on classroom achievement growth
   - Based on student learning targets (SLTsvz)
   - 30 percent of their potential bonus was based on classroom achievement growth

3. Bonuses based on classroom observations
   - For all teachers, 50 percent of their potential bonus was based on classroom observations
   - Teachers were observed 4 times during the year
   - Teachers were classified into 1 of 4 possible positions: (1) career teacher, (2) teacher in a hard-to-fill position, (3) mentor teacher, or (4) master teacher
   - Observation scores were put on a 1–5 scale
   - The size of the bonus earned depended on the teacher's observation rating and position
   - Teachers earned larger bonuses the higher their observation rating, but had to be rated at or above a minimum rating to receive a bonus, which depended on their position

Cohort 2

District 11

Key program features
- Teachers could receive bonuses for school achievement growth, classroom achievement growth (if teaching tested grades and subjects), and classroom observations
- For each performance measure, teachers' ratings were translated into "shares" that determined their bonus amounts
- Maximum bonus possible for each measure depended on the number of bonus recipients
- Bonus for classroom observations depended on a factor besides the observation score

Specific information on performance measures and bonus criteria
1. Bonuses based on school achievement growth
   - For teachers in tested grades and subjects, 20 percent of their potential bonus was based on school achievement growth
   - For teachers in nontested grades and subjects, 50 percent of their potential bonus was based on school achievement growth
   - Based on school value-added score placed on a 1–5 scale
   - Teachers in schools rated 1 or 2 earned 0 shares; teachers in schools rated 3 earned 50 shares; teachers in schools rated 4 earned 75 shares; teachers in schools rated 5 earned 100 shares

2. Bonuses based on classroom achievement growth
   - For teachers in tested grades and subjects, 30 percent of their potential bonus was based on classroom achievement growth
   - Based on classroom value-added score placed on a 1–5 scale
   - Teachers rated 1 or 2 earned 0 shares; teachers rated 3 earned 1 share; teachers rated 4 earned 6 shares, teachers rated 5 earned 10 shares

3. Bonuses based on classroom observations
   - For all teachers, 50 percent of their potential bonus was based on classroom observations
   - Teachers were observed 4 times during the year
   - Teachers were classified into 1 of 4 possible categories: (1) career teacher, (2) teacher in a hard-to-fill position, (3) mentor teacher, or (4) master teacher
   - The number of shares earned depended on the teacher's observation rating and position
   - Teachers earned more shares the higher their observation score, but had to be rated above a minimum score to receive any shares
   - The minimum observation score required to receive shares varied depending on their position
   - For a given observation rating, career teachers and teachers in a hard-to-fill position earned more shares than mentor or master teachers

District 12

Key program features
- Teachers could receive a bonus for 1 overall performance measure that combined ratings based on school achievement growth, classroom achievement growth, and classroom observations
- Set an absolute maximum bonus
- Teachers receiving a score of 4 on a 1–4 scale received a bonus
- Revised some aspect of program between the 2012-2013 and 2013-2014 school years

Specific information on performance measures
1. Rating based on school achievement growth
   - For teachers in tested grades and subjects, based on value-added score on state assessment
   - For teachers in nontested grades and subjects, based on student learning objectives.
   - 20 percent of overall evaluation score based school achievement growth

2. Rating based on classroom achievement growth
   - Based on student learning objectives
   - 20 percent of overall evaluation score based on classroom achievement growth

3. Rating based on classroom observations
   - Teachers were observed 3 times per year
   - 60 percent of overall evaluation score based on classroom observations

District 13

Key program features
- Teachers could receive a bonus for 1 overall performance measure that combined ratings based on school achievement growth, classroom achievement growth, and classroom observations
- Set an absolute maximum bonus
- Teachers receiving a score of 4 on a 1–4 scale received a bonus
- Revised some aspect of program between the 2012-2013 and 2013-2014 school years

Specific information on performance measures
1. Rating based on school achievement growth
   - Based on student learning objectives
   - 20 percent of overall evaluation score based on school achievement growth

2. Rating based on classroom achievement growth
   - For teachers in tested grades and subjects, based on value-added score on state assessment
   - For teachers in nontested grades and subjects, based on student growth on student learning objectives
   - 20 percent of overall evaluation score based on classroom achievement growth

3. Rating based on classroom observations
   - Teachers were observed 3 times per year
   - 60 percent of overall evaluation score based on classroom observations

Sources: District interviews (2012, 2013, 2014, and 2015), grantees' Annual Performance Report (APR) documents, and technical assistance documents.

Note: Grantees submit an APR to the U.S. Department of Education that describes how educators are evaluated.

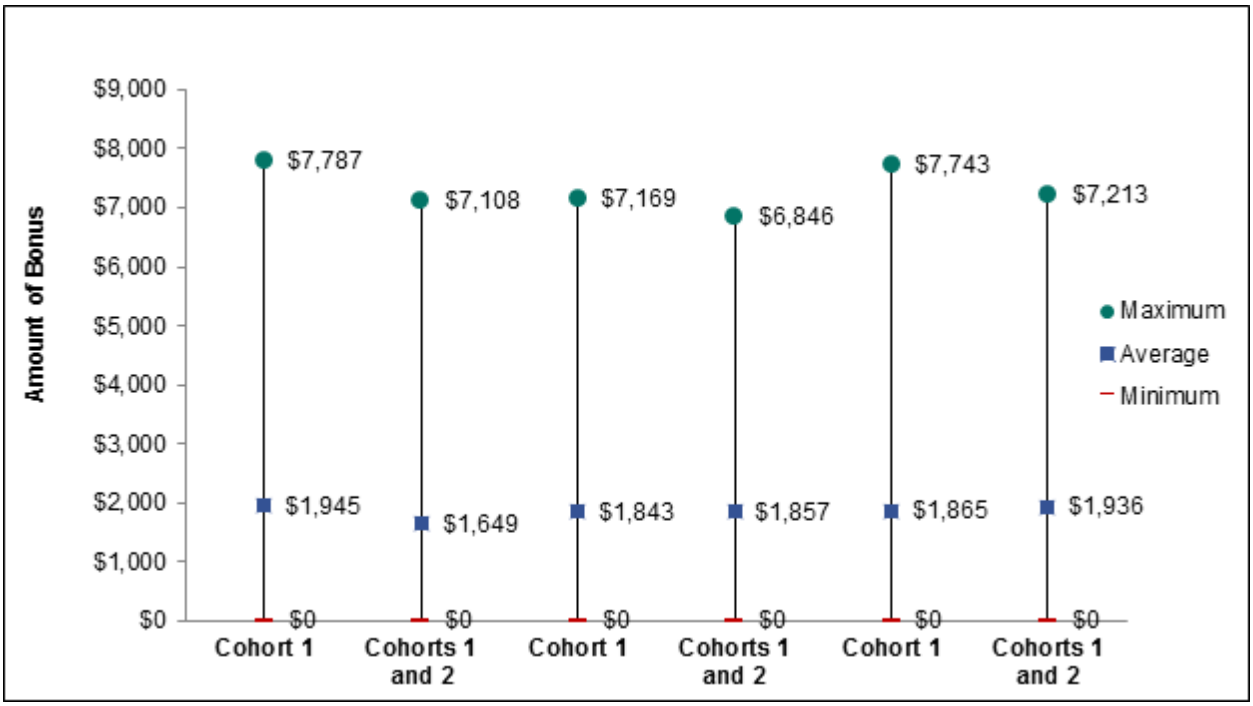
ELA is English language arts.

Teachers

Figure IV.3 in Chapter IV shows the minimum, average, and maximum pay-for-performance bonuses in Years 1 through 4 for teachers in Cohort 1, with each district equally weighted. Figure D.1 compares the minimum, average, and maximum pay-for-performance bonuses in Years 1 through 3 for teachers in Cohort 1 to those in Cohorts 1 and 2. Like Figure IV.3, Figure D.1 weights each district equally. By weighting each district equally, our findings in Chapter IV describe these bonuses for the average Cohort 1 district. Because our findings on educators' understanding and impact findings weight schools equally, Figure D.2 presents the minimum, average, and maximum pay-for-performance bonuses in Years 1 through 4 for teachers in Cohort 1, with districts weighted by the number of schools.

In Chapter IV, we noted that the maximum and average bonus amounts for teachers varied substantially across districts. Figure IV.5 shows the distribution of pay-for-performance bonuses for teachers by district for Cohort 1 in Year 4. For comparison, we also show these distributions for Cohort 1 in Years 1, 2, and 3 (Figures D.3, D.4, and D.5). In some districts, fewer than three teachers received the maximum bonus. To reduce the risk of disclosing information on individual teachers, we averaged the top three performance bonuses awarded to teachers in the district and used that value to represent the maximum bonus. Figures IV.5, D.3, D.4, and D.5 present bonuses in dollar amounts for ease of interpretation. However, because average teacher salaries also vary across districts, Figure D.6 presents bonuses for Cohort 1 in Year 4 as a percent of the district's average teacher salary.

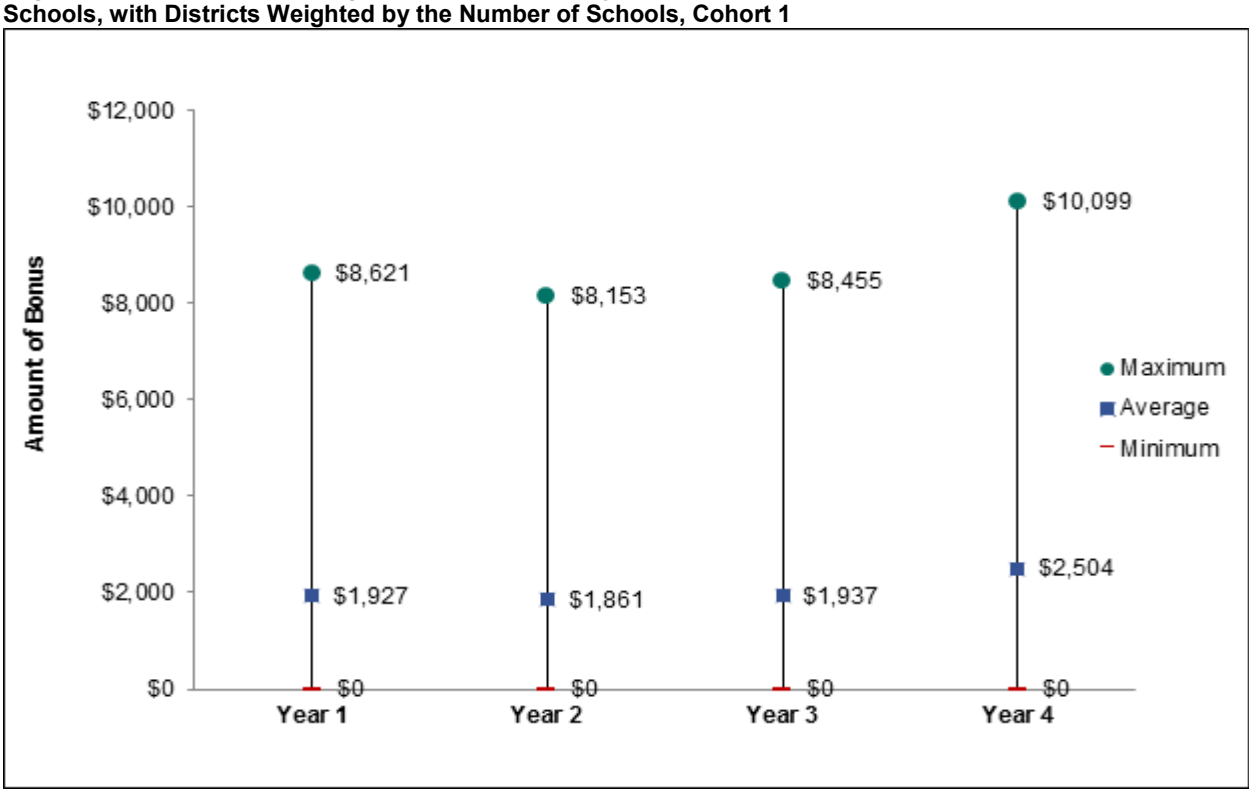
Figure D.1. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools in Years 1 through 3, Cohorts 1 and 2



Source: Educator administrative data (N = 2,151 teachers for Year 1, Cohort 1 and N = 2,949 teachers for Year 1, Cohorts 1 and 2; N = 2,160 teachers for Year 2, Cohort 1 and N = 3,047 teachers for Year 2, Cohorts 1 and 2; N = 2,236 teachers for Year 3, Cohort 1 and N = 3,129 teachers for Year 3, Cohorts 1 and 2).

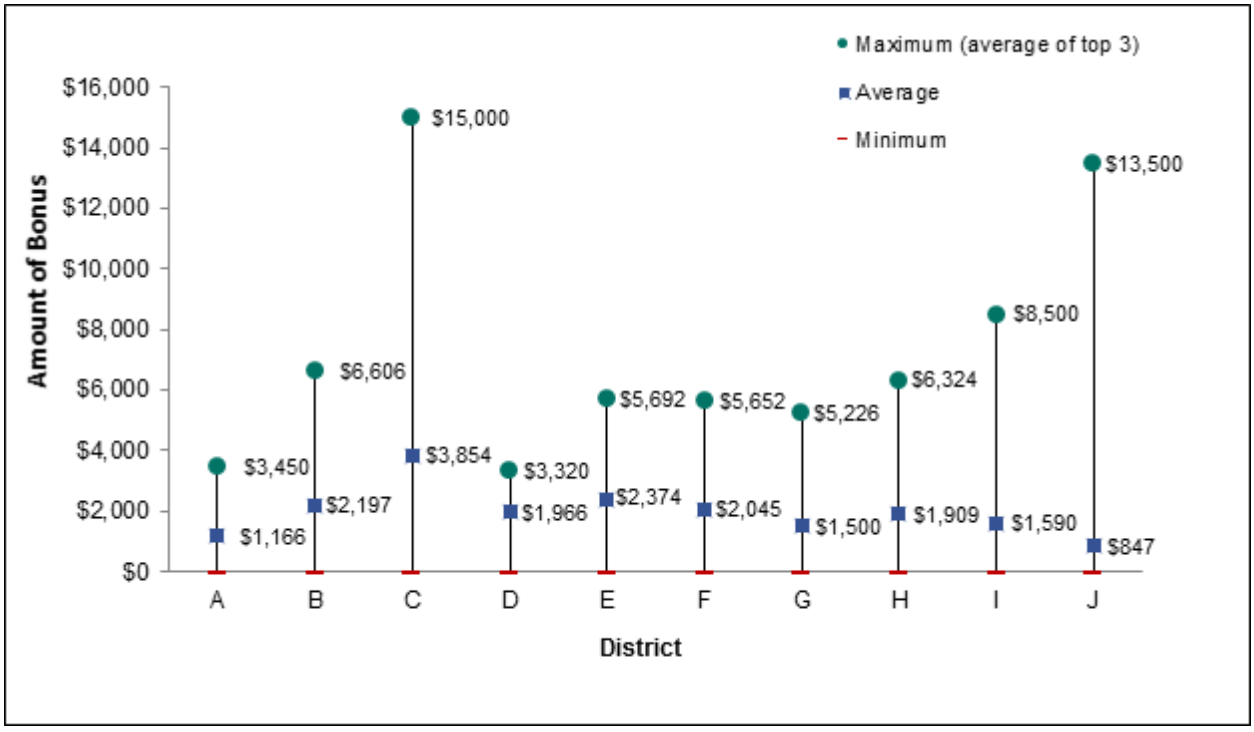
Note: The statistics shown in the figure represent an equal-weighted average of the statistics from the 10 evaluation districts in Cohort 1 or the 13 evaluation districts in Cohorts 1 and 2.

Figure D.2. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools, with Districts Weighted by the Number of Schools, Cohort 1



Source: Educator administrative data (N = 2,151 teachers in Year 1; N = 2,160 teachers in Year 2; N = 2,236 teachers in Year 3; and N = 2,083 teachers in Year 4).

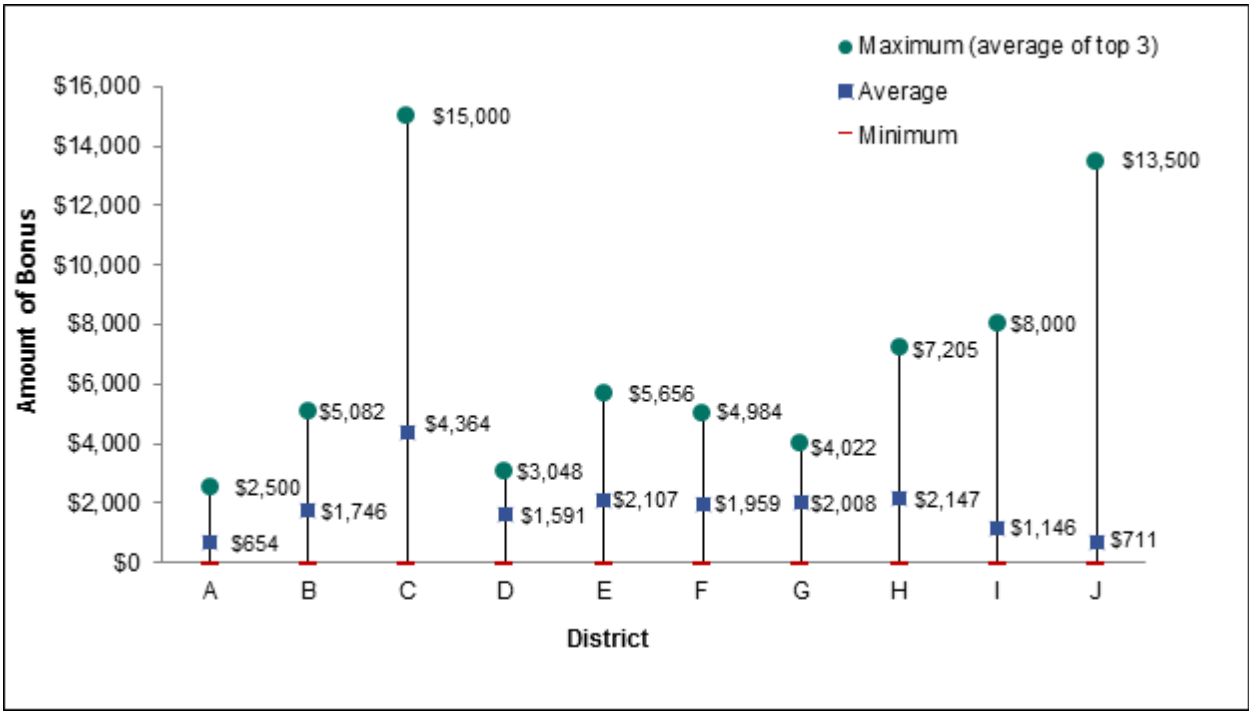
Figure D.3. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools in Year 1, by District



Source: Educator administrative data (N ranges from 68 teachers in District D to 432 in District J).

Note: To reduce the risk of disclosing information on an individual, the maximum bonus in this figure represents the average of the top three performance bonuses awarded to teachers in the district.

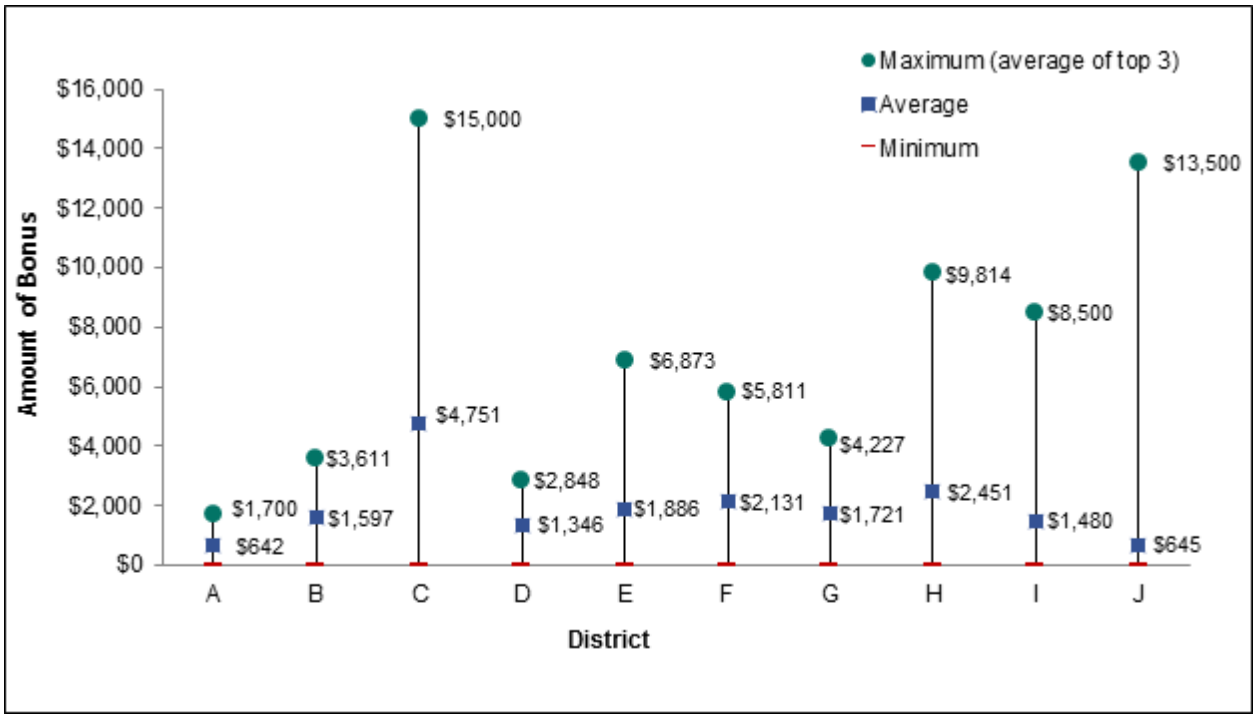
Figure D.4. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools in Year 2, by District



Source: Educator administrative data (N ranges from 78 teachers in District D to 394 in District J).

Note: To reduce the risk of disclosing information on an individual, the maximum bonus in this figure represents the average of the top three performance bonuses awarded to teachers in the district.

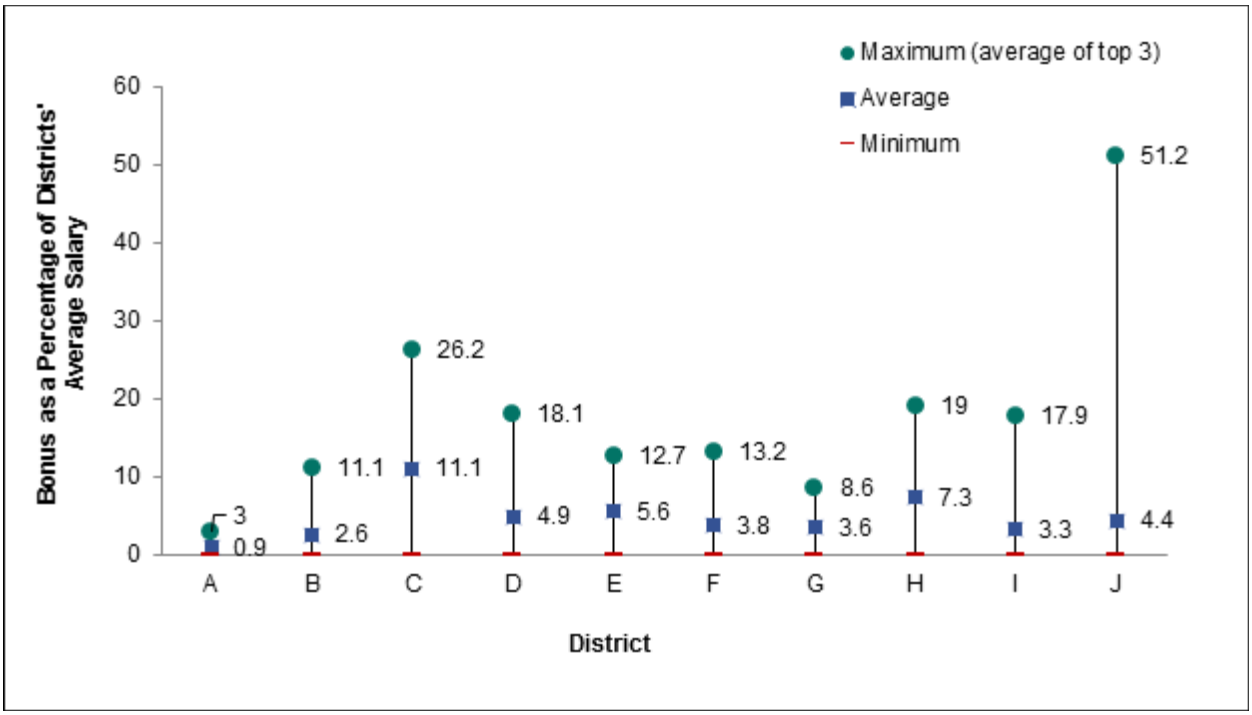
Figure D.5. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools in Year 3, by District



Source: Educator administrative data (N ranges from 81 teachers in District E to 394 in District J).

Note: To reduce the risk of disclosing information on an individual, the maximum bonus in this figure represents the average of the top three performance bonuses awarded to teachers in the district.

Figure D.6. Minimum, Average, and Maximum Pay-for-Performance Bonuses as a Percentage of Districts' Average Salary for Teachers in Treatment Schools in Year 4, by District

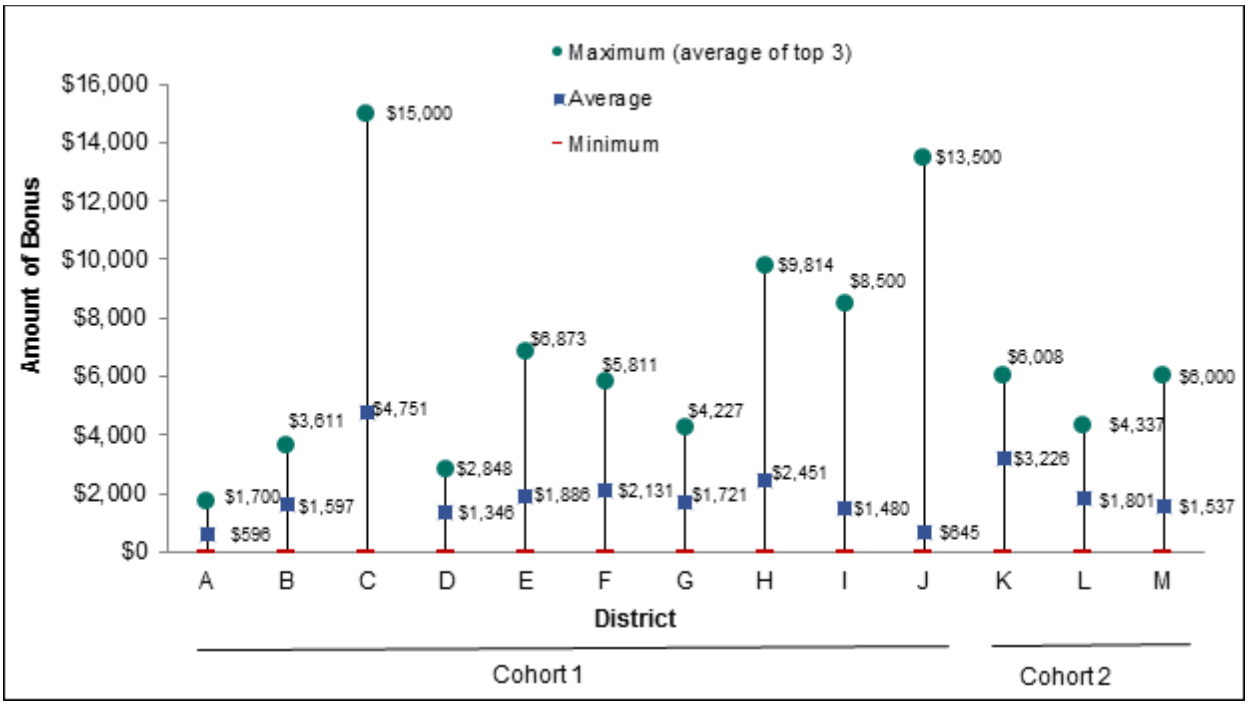


Source: Educator administrative data (N ranges from 68 teachers in District D to 374 in District J).

Note: To reduce the risk of disclosing information on an individual, the maximum bonus in this figure represents the average of the top three performance bonuses awarded to teachers in the district.

In Chapter IV, we noted that the maximum bonus amounts for teachers varied substantially across districts. Figure IV.5 shows the distribution of pay-for-performance bonuses for teachers by district for Cohort 1 in Year 4. Figure D.7 shows the information for Cohorts 1 and 2 in Year 3.

Figure D.7. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools in Year 3, by District, Cohorts 1 and 2

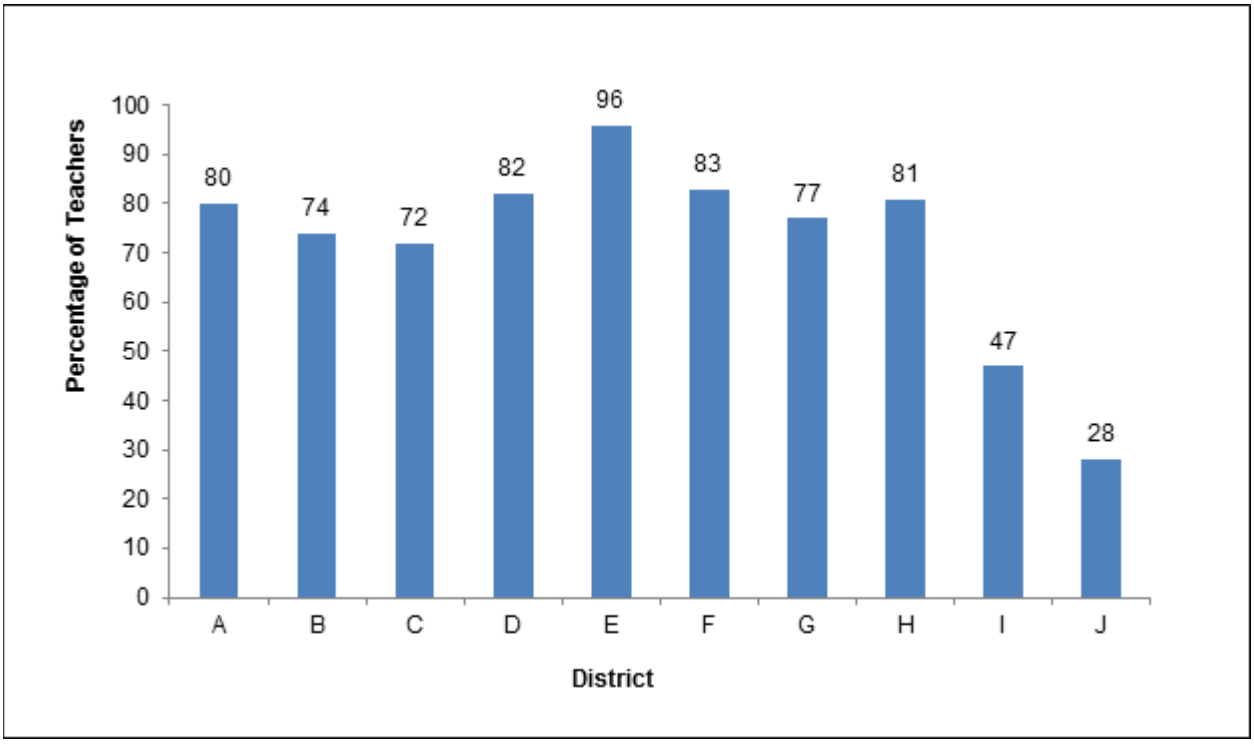


Source: Educator administrative data (N ranges from 26 teachers in District L to 483 in District J).

Note: To reduce the risk of disclosing information on an individual, the maximum bonus in this figure represents the average of the top three performance bonuses awarded to teachers in the district.

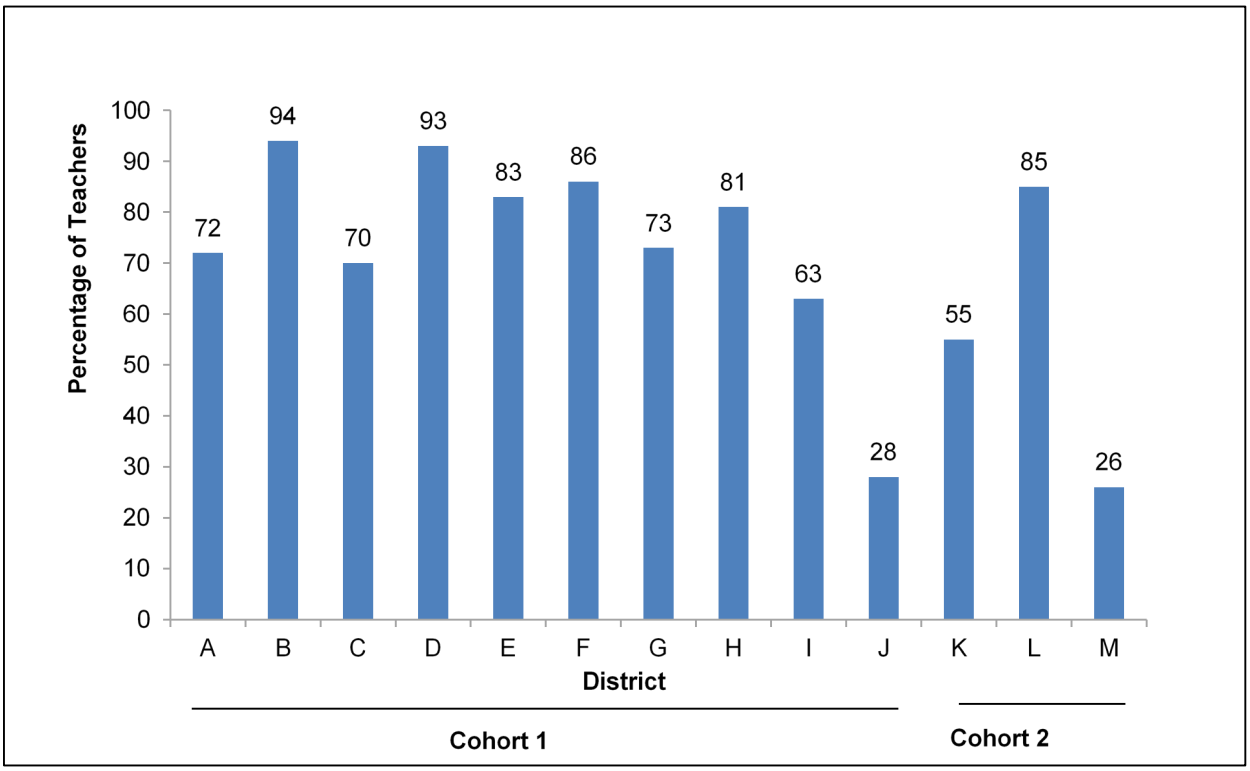
Figure IV.4 shows that across districts, on average, more than 70 percent of treatment teachers received a bonus each year. Figure D.8 shows the percentage of teachers earning pay-for-performance bonuses in Year 4, by district, for Cohort 1. Figure D.9 shows the percentage of teachers earning pay-for-performance bonuses in Year 3, by district, for Cohorts 1 and 2.

Figure D.8. Percentage of Treatment Teachers Earning a Pay-for-Performance Bonus in Year 4, By District, Cohort 1



Source: Educator administrative data (N ranges from 68 teachers in District D to 374 in District J).

Figure D.9. Percentage of Treatment Teachers Earning a Pay-for-Performance Bonus in Year 3, by District, Cohorts 1 and 2



Source: Educator administrative data (N ranges from 26 teachers in District L to 483 in District M).

Applicants for the evaluation grants received guidance on the structure of their pay-for-performance bonus, including the examples of bonuses that were *substantial* (the average bonus was at least 5 percent of teacher salary), *differentiated* (the highest bonus was at least three times the average), and *challenging* to earn, in which only those performing significantly better than the average (therefore, fewer than 50 percent) would receive a bonus. Table D.8 provides information about the percentage of districts that structured teachers' bonuses to meet the grant guidance in each year for Cohorts 1 and 2.

Table D.8. Evaluation Districts Meeting TIF Grant Goals for Pay-for-Performance Bonuses for Teachers, Cohorts 1 and 2 (Percentages)

TIF Grant Goal	Year 1	Year 2	Year 3	Year 4
Cohort 1				
Substantial: Average Bonus Was at Least 5 Percent of Average Salary	20	30	20	30
Differentiated: Highest Bonus Was at Least Three Times the Average Bonus	70	60	50	60
Challenging: Less Than 50 Percent Of Teachers Received a Pay-for-Performance Bonus	20	20	10	20
Number of Districts	**10**	**10**	**10**	**10**
Cohorts 1 and 2				
Substantial: Average Bonus Was at Least 5 Percent of Average Salary	23	31	23	NA
Differentiated: Highest Bonus Was at Least Three Times the Average Bonus	77	62	46	NA
Challenging: Less Than 50 Percent of Teachers Received a Pay-for-Performance Bonus	31	23	15	NA
Number of Districts	**13**	**13**	**13**	NA

Source: Educator administrative data.

NA is not applicable.

Table D.9 compares the amount teachers in treatment schools received in performance bonuses in Years 3 and 4 for Cohort 1. We partitioned bonuses into four categories: (1) $0, or those who did not receive a bonus; (2) $1 to $1,500; (3) $1,501 to $3,000; and (4) above $3,000. Most teachers (51 percent) received similar bonus award amounts in Years 3 and 4.

Table D.9. Comparison of Teachers' Performance Bonus Amounts in Years 3 and 4, Cohort 1 (Percentages)

	Performance Bonus, Year 4				
Performance Bonus, Year 3	$0	$1-1,500	$1,501-3,000	Above $3,000	**Number of Teachers**
$0	20	3	1	2	**769**
$1-1,500	12	11	3	3	**578**
$1,501-3,000	6	5	12	5	**451**
Above $3,000	5	1	2	8	**438**
Number of Teachers	**1,119**	**385**	**322**	**410**	**2,236**

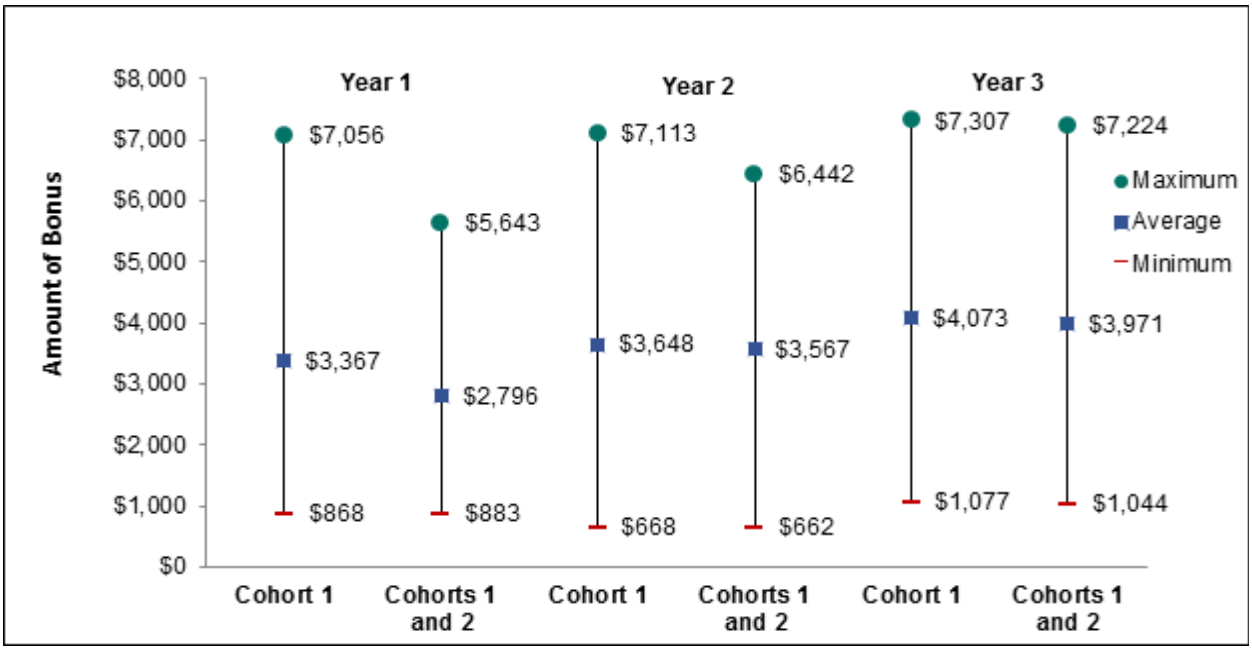
Source: Educator administrative data.

Note: Table is based on teachers who worked in treatment schools in both Years 3 and 4.

Principals

This section provides supplemental information on principals' pay-for-performance bonuses, similar to the previous section on teachers. In Chapter IV, Figure IV.8 shows the minimum, average, and maximum pay-for-performance bonuses in Years 1 through 4 for principals in Cohort 1, with each district equally weighted. Figure D.10 presents the minimum, average, and maximum pay-for-performance bonuses in Years 1 through 3 for principals in Cohorts 1 and 2, also with districts weighted equally. Figure D.11 shows the minimum, average, and maximum pay-for-performance bonuses for principals in Cohort 1 with districts weighted by the number of schools.

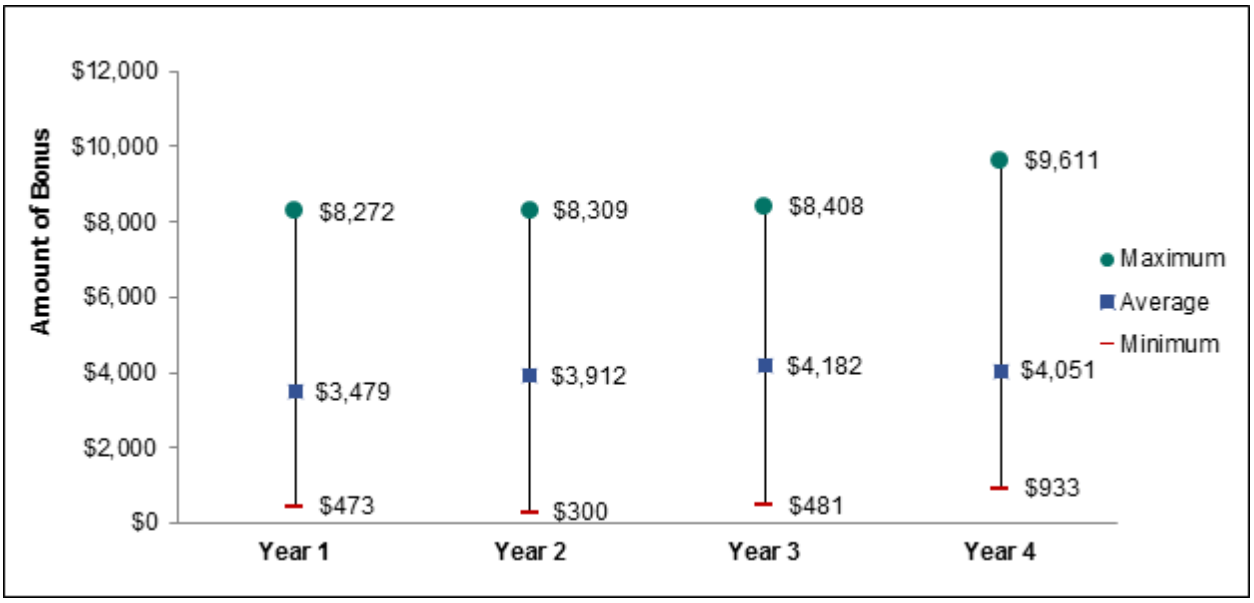
Figure D.10. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Principals in Treatment Schools in Years 1 through 3, Cohorts 1 and 2



Source: Educator administrative data (N = 64 principals in Year 1, Cohort 1 and N = 84 principals in Year 1, Cohorts 1 and 2; N = 67 principals in Year 2, Cohort 1 and N = 87 principals in Year 2, Cohorts 1 and 2; N = 63 principals in Year 3, Cohort 1 and N = 83 principals in Year 3, Cohorts 1 and 2).

Note: The statistics shown in the figure represent an equal-weighted average of the statistics from the 10 evaluation districts in Cohort 1 or the 13 districts in Cohorts 1 and 2.

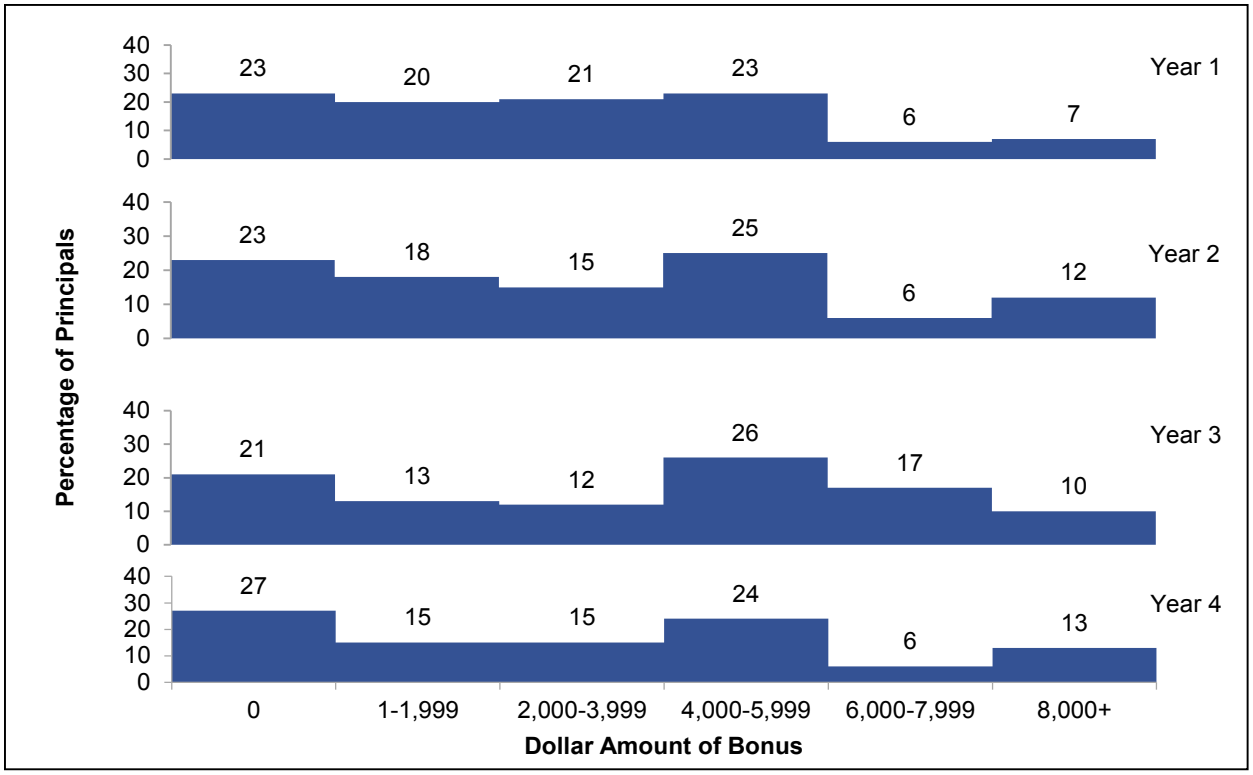
Figure D.11. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Principals in Treatment Schools, with Districts Weighted by the Number of Schools, Cohort 1



Source: Educator administrative data (N = 64 principals in Year 1; N = 67 principals in Year 2; N = 63 principals in Year 3; and N = 64 principals in Year 4). In Chapter IV, we noted that over 70 percent of principals in each year received a pay-for-performance bonus.

Figure D.12 illustrates the distribution of principals' pay-for-performance bonuses in Years 1 through 4 for Cohort 1.

Figure D.12. Distribution of Pay-for-Performance Bonuses for Principals in Treatment Schools, Cohort 1



Source: Educator administrative data (N = 64 principals in Year 1; N = 67 principals in Year 2; N = 63 principals in Year 3; and N = 64 principals in Year 4).

Table D.10 provides additional information about the structure of principals' bonuses. It shows the percentage of districts that structured principals' bonuses to meet the grant guidance for *substantial*, *differentiated*, and *challenging* to earn—in each year.

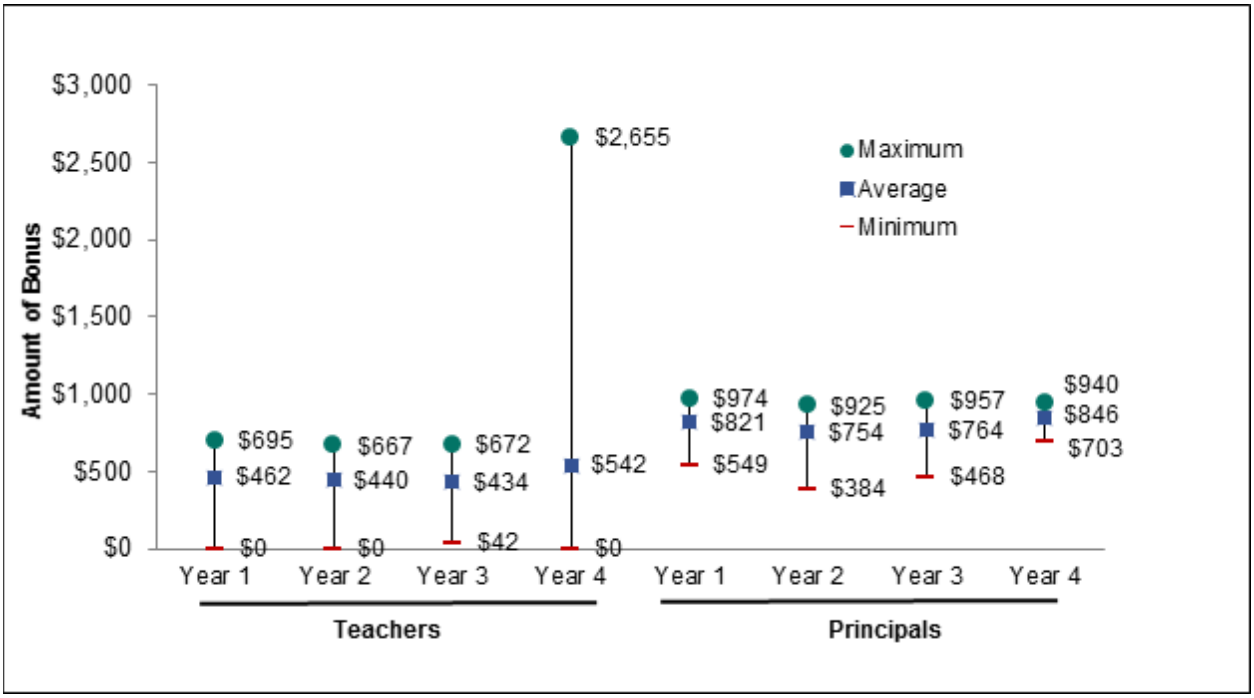
Table D.10. Evaluation Districts Meeting TIF Grant Goals for Pay-for-Performance Bonuses for Principals, Cohort 1 (Percentages)

TIF Grant Goal	Year 1	Year 2	Year 3	Year 4
Substantial: Average bonus was at least 5 percent of average salary	30	30	60	30
Differentiated: Highest bonus was at least three times the average bonus	10	10	10	20
Challenging: Fewer than 50 percent of principals received a pay-for-performance bonus	20	20	10	10
Number of Districts	**10**	**10**	**10**	**10**

Source: Educator administrative data.

Teachers and principals in control schools were expected to receive an automatic 1 percent bonus (see Chapter II). The 1 percent bonus ensured that all educators in evaluation schools received some benefit from participating in the study: either the opportunity to earn a pay-for-performance bonus or the automatic bonus. Figure D.13 presents the minimum, average, and maximum automatic 1 percent bonuses for Cohort 1 teachers and principals. As intended by the study design, the automatic 1 percent bonus provided to teachers and principals in control schools was small and generally did not vary substantially. After Year 4, one district awarded performance bonuses to teachers in control schools who had received high ratings on classroom achievement growth. The control teachers in this district did not know they could receive a performance bonus until after the school year, so the possibility of receiving one should not have affected their behavior. We included both this performance pay and the automatic bonuses to accurately describe the bonuses that control teachers received. The higher average maximum bonus awarded to teachers in Year 4 was driven by the performance bonuses the control teachers in this district received.

Figure D.13. Minimum, Average, and Maximum Automatic 1 Percent Bonuses for Teachers and Principals in Control Schools, Cohort 1



Source: Educator administrative data (N = 2,152 teachers and N = 69 principals in Year 1; N = 2,242 teachers and N = 70 principals in Year 2; N = 2,285 teachers and N = 69 principals in Year 3; N = 2,132 teachers and N = 63 principals in Year 4).

Notes: At the end of Year 4, one district gave performance pay to teachers in control schools who had received high ratings on classroom achievement growth. To accurately describe the bonuses that control teachers received, we included both their performance pay and their automatic bonuses.

Requirement 3: Additional Pay Opportunities

According to the study design, the only difference between treatment and control schools was the pay-for-performance bonus component of the TIF program. Educators in some schools (the treatment schools) were eligible for pay-for-performance, and educators in others (control schools) were not. As explained above, educators in control schools were expected to receive an automatic 1 percent bonus. All other aspects of the districts' TIF program (such as additional pay opportunities) should have been implemented the same in treatment and control schools.

Table D.11 shows the average and maximum payouts for additional pay and the percentage of teachers receiving additional pay for taking on extra roles across treatment and control schools for Cohort 1 in Years 1 through 4. Few teachers (less than 20 percent) received additional pay. Because most teachers received $0 in additional pay, the average amount teachers received (including those who received nothing) was notably less than the average pay-for-performance bonus that treatment teachers received (around $2,000 in each year; Figure IV.3)

Table D.12 compares the amount of additional pay received by Cohort 1 teachers in treatment and control schools in Years 1 through 4. As intended by the study design, the average amount of additional pay for extra roles or any other additional pay earned by teachers in treatment schools and control schools did not differ.

Table D.11. Average and Maximum Amounts of Additional Pay Opportunities for Teachers, Cohort 1

Additional Pay Opportunities	Year 1	Year 2	Year 3	Year 4
Average Amount for Additional Pay Opportunities (dollars)	448	491	497	507
Maximum Amount for Additional Pay Opportunities (dollars)	4,766	5,594	5,275	5,675
Percentage of Districts Offering Additional Pay Opportunities	100	100	100	100
Percentage of Teachers Receiving Additional Pay Opportunities	12	17	17	17
Number of Teachers	**4,303**	**4,402**	**4,521**	**4,215**

Source: Educator administrative data.

Table D.12. Actual Amounts of Teachers' Additional Pay, Cohort 1

	Year 1		Year 2		Year 3		Year 4	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Roles and Responsibilities								
Average additional pay (dollars)	499	504	509	532	516	500	481	457
Received pay (percentage)	12*	14	15*	18	17	17	16	15
Other Additional Pay[a]								
Average additional pay (dollars)	339	329	345	388	364	392	337	323
Received pay (percentage)	22	22	13*	18	14	19	13*	18
Number of Teachers	**2,151**	**2,152**	**2,160**	**2,242**	**2,236**	**2,285**	**2,083**	**2,132**

Source: Educator administrative data.

[a]Other additional pay includes pay for factors such as working in a hard-to-staff school or subject area or professional development and excludes pay-for-performance bonuses.

*Difference between treatment and control group is statistically significant at the .05 level, two-tailed test.

Requirement 4: Professional Development

The TIF grant required that districts provide professional development linked to the measures of educator effectiveness. This support included professional development to help educators understand the measures being used to evaluate their performance as well as feedback based on their actual performance ratings to help improve their instructional practices. Table D.13 shows that all or almost all districts reported providing professional development on how to improve their performance on classroom observations and achievement growth in Years 3 and 4. The table also provides additional details about the average hours districts reported spending on this kind of professional development, the frequency with which districts reported it occurred during the school year, and whether districts reported requiring teachers to participate in it.

Table D.13. Professional Development Based on Observation and Student Achievement Growth Ratings that Teachers Earned in Years 2 and 3, Cohort 1 (Percentages Unless Otherwise Noted)

	Professional Development Based on Ratings That Teachers Earned in Year 2 on		Professional Development Based on Ratings That Teachers Earned in Year 3 on	
	Classroom Observations	Student Achievement Growth	Classroom Observations	Student Achievement Growth
Teachers Received Professional Development on How to Improve their Performance on the Measure	100	90	100	80
Teachers Received Professional Development Based on their Individual Performance on the Measure	80	30	90	60
Average Hours of Professional Development	28	38	25	28
Frequency				
Throughout year	60	80	80	80
1–4 times per year	30	20	20	10
Varies	10	0	0	0
Don't know	0	0	0	10
Teachers Were Required to Participate in Professional Development on How to Improve Their Performance on the Measure	80	60	70	80
Number of Districts—Range[a]	**10**	**10**	**10**	**9-10**

Sources: District interviews (2014 and 2015).

[a]Sample sizes are presented as a range based on the data available for each row in the table.

Communication of TIF Program

We asked district administrators for more detailed information on their communication activities during the district interviews. Table D.14 shows districts' activities to communicate information about how teachers would be evaluated and the potential amounts of TIF bonuses. Table D.15 shows districts' activities to communicate information about ratings based on classroom observations and student achievement growth to teachers. Table D.16 provides information on what districts told Cohort 1 teachers about their individual bonuses for Years 1 through 3.

Table D.14. Districts' Communication Activities in Years 3 and 4 (Percentages Unless Otherwise Noted)

	Year 3	Year 4
Responsible for Majority of Communication about TIF		
District or grantee official	60	50
School-level staff (such as a principal or lead teacher)	30	50
TIF coach	10	0
Communication about		
How teachers would be evaluated	100	100
Potential amounts of TIF bonuses	100	100
Communication Methods on How Teachers Would be Evaluated		
Written materials (including letters, email, brochures, program manuals, newsletters)	90	90
Group presentation	90	80
Individual meetings	10	20
District website	90	70
Communication Methods on Potential Amounts of TIF Bonuses		
Written materials (including letters, email, brochures, program manuals, newsletters)	70	70
Group presentation	90	80
Individual meetings	10	20
District website	80	80
Average Number of Communication Methods Used by Districts About How Teachers Would be Evaluated	2.9	2.6
Average Number of Communication Methods Used by Districts About Potential Amounts of TIF Bonuses	2.6	2.5
Used Survey or Focus Group to Assess if Teachers Understood their Eligibility for a Bonus	60	20
Number of Districts	**10**	**10**

Sources: District interviews (2014 and 2015).

Table D.15. Districts' Activities Used to Communicate to Teachers Their Classroom Observation and Student Achievement Growth Ratings from Years 2 and 3, Cohort 1 (Percentages)

	Ratings from Year 2	Ratings from Year 3
Activities to Communicate Classroom Observation Rating		
In-person meeting	90	90
Online system	80	40
Letter to individual	40	20
E-mail to individual	20	20
Activities to Communicate Student Achievement Growth Rating		
In-person meeting	90	80
Online system	60	40
Letter to individual	50	30
E-mail to individual	20	10
Number of Districts	**10**	**10**

Sources: District interviews (2014 and 2015).

Table D.16. Communication Methods Used to Inform Teachers About Individual Pay-for-Performance Bonuses from Years 1 through 3 (Percentages)

	Bonuses from Year 1	Bonuses from Year 2	Bonuses from Year 3
Letter or Email to Each Teacher with Individual Bonus Amount	80	80	75
Individual Meeting with Each Teacher to Discuss Bonus Amount	30	30	37
Number of Districts	**10**	**10**	**8**

Sources: District surveys (2013, 2014, and 2015).

Educators' Understanding of and Experiences with TIF

This section of the appendix provides additional details and supplemental analyses about educators' reported understanding of the TIF program.

Educators' Understanding of their Eligibility for Pay-for-Performance Bonuses

Findings in Chapter IV on educators' understanding of their eligibility for pay-for-performance bonuses were based on their responses to survey questions. In what follows, we provide the wording of those survey questions.

On the teacher survey, the question that asked teachers to report their eligibility for pay-for-performance bonuses changed between spring 2012 (Year 1 for Cohort 1) and spring 2013 (Year 2 for Cohort 1). Teachers answered the following questions:

- **In Year 1:** For which of the following factors are teachers at your school eligible to receive additional pay?
    - Job performance, as measured by student achievement growth and/or classroom observations
- **In Years 2, 3, and 4:** Is it possible for you to earn a bonus based *solely* on your performance for the [current] school year?

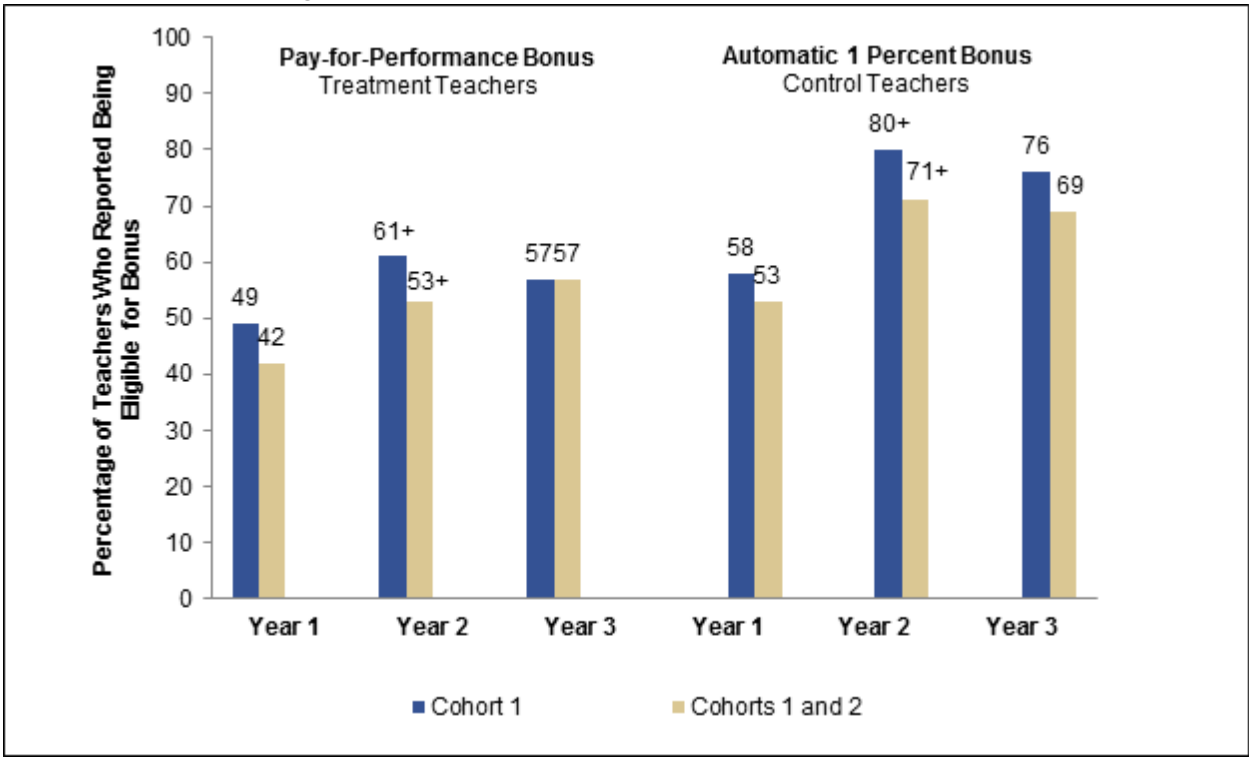
On the principal survey, the question that asked principals to report their eligibility for pay-for-performance bonuses also changed between spring 2012 (Year 1 for Cohort 1) and spring 2013 (Year 2 for Cohort 1). Principals answered the following questions:

- **In Year 1:** For which of the following factors are you eligible to receive additional pay?
    - Job performance, as measured by student achievement growth and/or formal observations
- **In Years 2, 3, and 4:** Could you receive a bonus if your students' achievement improved substantially this year?

On both the teacher and principal surveys, the purpose of the change in wording was to remove the word "eligible." We were concerned that respondents would misinterpret "eligible" to mean "eligible and likely to earn" a bonus. Therefore, the subsequent versions of those questions were rephrased to avoid using the word "eligible."

Our main findings on educators' understanding of their bonus eligibility, reported in Figures IV.9 and IV.10, pertain to Cohort 1 only. Figure D.14 shows the percentage of teachers in treatment schools who reported they were eligible for a pay-for-performance bonus and the percentage of control teachers who reported they were eligible for an automatic 1 percent bonus in Years 1 through 3 for Cohort 1 compared to Cohorts 1 and 2 combined. Figure D.15 shows the same information for principals. When analyses for Years 1 through 3 were based on Cohorts 1 and 2, similar but generally smaller percentages of teachers and principals reported being eligible for the correct type of bonus than the respective estimates based only on Cohort 1.

Figure D.14. Teachers' Pay-for-Performance Bonus Eligibility in Years 1 through 3, as Reported by Teachers, Cohorts 1 and 2 (Percentages)

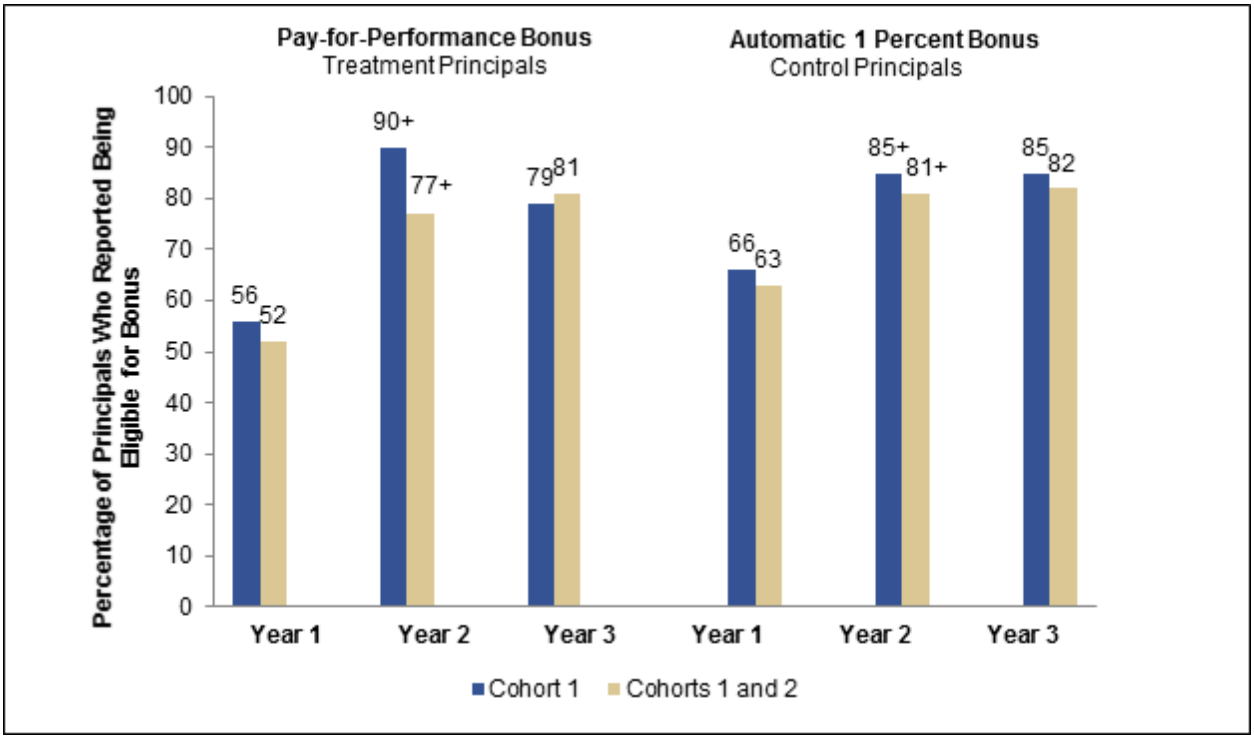


Sources: Teacher surveys (2012, 2013, 2014, and 2015).

Notes: A total of 368 treatment teachers in Cohort 1 and 481 in Cohorts 1 and 2 responded to the question about eligibility for a pay-for-performance bonus in Year 1. A total of 439 treatment teachers in Cohort 1 and 572 in Cohorts 1 and 2 responded to the question about eligibility for a pay-for-performance bonus in Year 2. A total of 420 treatment teachers in Cohort 1 and 538 in Cohorts 1 and 2 responded to the question about eligibility for a pay-for-performance bonus in Year 3. A total of 381 control teachers in Cohort 1 and 471 in Cohorts 1 and 2 responded to the question about eligibility for an automatic 1 percent bonus in Year 1. A total of 445 control teachers in Cohort 1 and 560 in Cohorts 1 and 2 responded to the question about eligibility for an automatic 1 percent bonus in Year 2. A total of 448 control teachers in Cohort 1 and 566 in Cohorts 1 and 2 responded to the question about eligibility for an automatic 1 percent bonus in Year 3.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Figure D.15. Principals' Pay-for-Performance Bonus Eligibility in Years 1 through 3, as Reported by Principals, Cohorts 1 and 2 (Percentages)



Sources: Principal surveys (2012, 2013, 2014, and 2015).

Notes: A total of 63 treatment principals in Cohort 1 and 79 in Cohorts 1 and 2 responded to the question about eligibility for a pay-for-performance bonus in Year 1. A total of 62 treatment principals in Cohort 1 and 81 in Cohorts 1 and 2 responded to the question about eligibility for a pay-for-performance bonus in Year 2. A total of 57 treatment principals in Cohort 1 and 75 in Cohorts 1 and 2 responded to the question about eligibility for a pay-for-performance bonus in Year 3. A total of 64 control principals in Cohort 1 and 80 in Cohorts 1 and 2 responded to the question about eligibility for an automatic 1 percent bonus in Year 1. A total of 61 control principals in Cohort 1 and 77 in Cohorts 1 and 2 responded to the question about eligibility for an automatic 1 percent bonus in Year 2. A total of 61 control principals in Cohort 1 and 78 in Cohorts 1 and 2 responded to the question about eligibility for an automatic 1 percent bonus in Year 3.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Table D.17 shows the percentage of Cohort 1 educators who correctly reported their bonus eligibility as intended by the study design (also shown in Figures D.14 and D.15), but it also shows the percentage that misreported their eligibility. Specifically, it shows the percentage of educators in treatment schools who reported they were eligible for an automatic 1 percent bonus and the percentage of educators in control schools who reported they were eligible for a pay-for-performance bonus. Although more Cohort 1 educators correctly reported their eligibility in Year 2 than Year 1, there were no further improvements after Year 2 (Figures IV.9 and IV.10). Furthermore, even in the fourth year of TIF implementation, many educators continued to misreport their eligibility. In Year 4, 42 percent of treatment teachers did not report being eligible for a pay-for-performance bonus, 39 percent of treatment teachers believed they were eligible for an automatic 1 percent bonus, and 22 percent of control teachers believed they were eligible for a pay-for-performance bonus.

Table D.17. Bonus Eligibility as Reported by Teachers and Principals, Cohort 1 (Percentages)

	Year 1		Year 2		Year 3		Year 4	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Teachers								
Pay-for-Performance Bonus	49*	17	61*+	17	57*	18	58*	22
Automatic 1 Percent Bonus	38*	58	41*	80+	40*	76	39*	72
Number of Teachers—Range[a]	**368-369**	**379-381**	**430-439**	**445-449**	**408-420**	**448-455**	**380-391**	**384-396**
Principals								
Pay-for-Performance Bonus	56*	13	90*+	15	79*	13	81*	20
Automatic 1 Percent Bonus	26*	66	30*	85+	22*	85	28*	84
Number of Principals—Range[a]	**62-63**	**63-64**	**62-63**	**61**	**57-58**	**61**	**59-60**	**61-62**

Sources: Teacher and principal surveys (2012, 2013, 2014, and 2015).

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference between treatment and control group is statistically significant at the .05 level, two-tailed test.

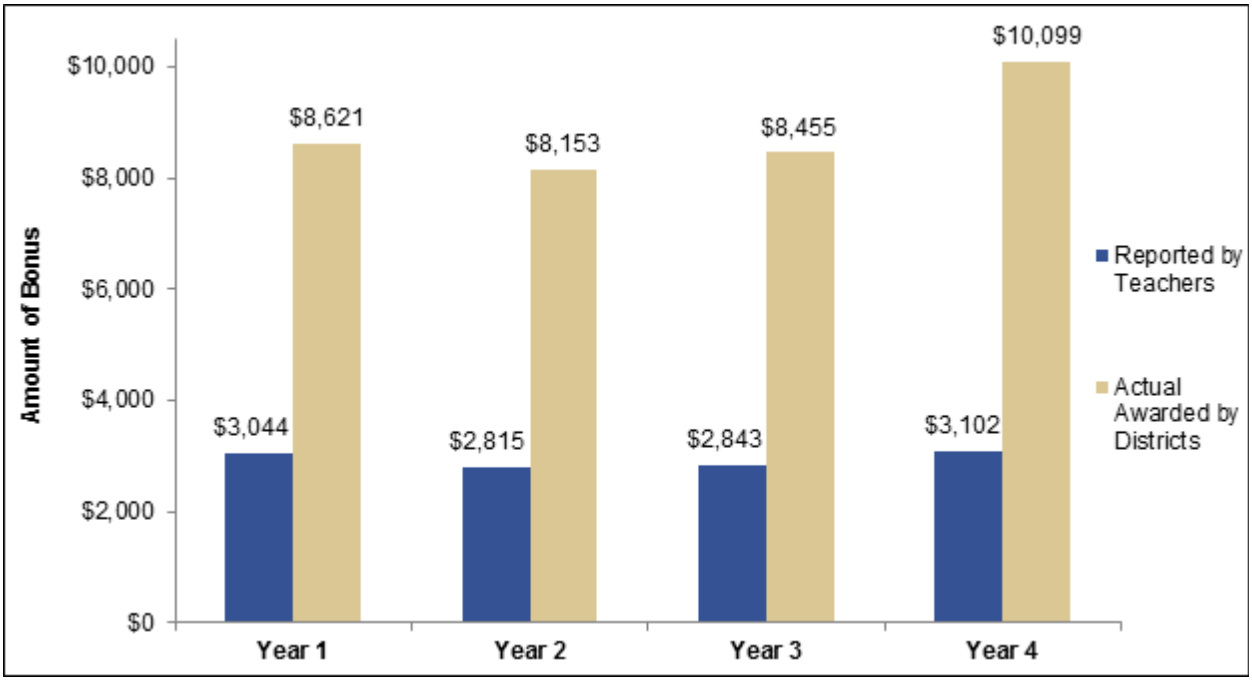
+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Educators' Understanding of the Potential Amounts of Pay-for-Performance Bonuses

Figures IV.11 and IV.12 show the reported and actual maximum pay-for-performance bonuses for teachers and for principals, respectively, for Cohort 1 in Years 1 through 4. For teachers and principals who reported being eligible for the bonus but left the amount missing, bonus amounts were imputed through multiple imputation methods (see Appendix B). Teachers' and principals' amounts are based on survey responses, with each school receiving an equal weight. Districts' actual maximum bonus amounts are based on administrative data, with each district receiving an equal weight. This section shows analyses that do not use imputed values for missing data, analyses that calculate districts' actual maximum bonus amounts weighting each school equally, and estimates for Years 1 through 3 for Cohorts 1 and 2.

Figures D.16 and D.17 show the actual and reported maximum pay-for-performance bonuses for teachers and for principals with the districts weighted by the number of schools. Unlike Figures IV.11 and IV.12, Figures D.16 and D.17 compare districts' amounts to educators' reported amounts using the same weighting approach. These figures show that our results are similar if we only use school weights.

Figure D.16. Reported and Actual Maximum Pay-for-Performance Bonus for Teachers in Treatment Schools, with Districts Weighted by the Number of Schools, Cohort 1

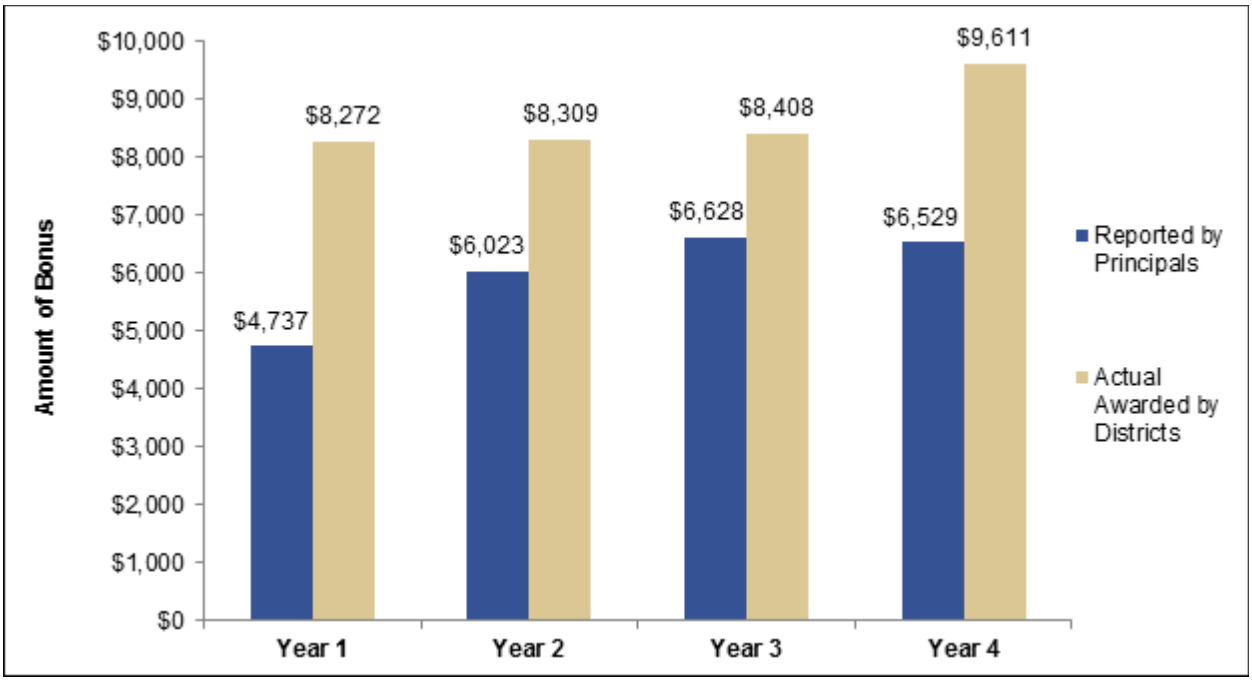


Sources: Teacher surveys (2012, 2013, 2014, and 2015), district interviews, and educator administrative data.

Notes: Teachers' reports are based on data for teachers in tested grades and subjects, with each school weighted equally. Districts' payouts are based on data for all teachers, with districts weighted by the number of schools.

A total of 194 treatment teachers in tested grades and subjects responded to this survey question in Year 1, a total of 215 in Year 2, a total of 201 in Year 3, and a total of 192 in Year 4. The maximum bonus amount was set to zero for all respondents who indicated they were ineligible for a bonus. For teachers who reported being eligible for the bonus but left the amount missing, bonus amounts were imputed through multiple imputation methods. This led to 24 additional responses for treatment teachers in Year 1, 14 additional responses in Year 2, 14 additional responses in Year 3, and 9 additional responses in Year 4. See Appendix B for additional discussion on the imputation methods. Appendix D, Table D.18 shows that our results are similar if we do not impute the missing bonus amounts.

Figure D.17. Reported and Actual Maximum Pay-for-Performance Bonus for Principals in Treatment Schools, with Districts Weighted by the Number of Schools, Cohort 1



Sources: Principal surveys (2012, 2013, 2014, and 2015), district interviews, and educator administrative data.

Notes: Principals' reports are based on weighting each school equally. Districts' payouts are based on weighting districts by the number of schools.

A total of 55 treatment principals responded to this survey question in Year 1, a total of 60 in Year 2, a total of 57 in Year 3 and a total of 58 in Year 4. The maximum bonus amount was set to zero for all respondents who indicated they were ineligible for a bonus. For educators who reported being eligible for the bonus but left the amount missing, bonus amounts were imputed through multiple imputation methods. This led to 8 additional responses for treatment principals in Year 1, 2 in Year 2, 0 in Year 3, and 2 in Year 4. See Appendix B for additional discussion on the imputation methods. Appendix D, Table D.18 shows that our results are similar if we do not impute the missing bonus amounts.

Table D.18 shows the maximum possible bonus amounts as reported by educators with missing values imputed (as shown in Figures IV.11 and IV.12) and non-imputed bonus amounts. Table D.18 shows that our results are similar if we do not impute the missing bonus amounts.

Table D.18. Educators' Reports of the Maximum Possible Bonus Amount: Imputed and Non-Imputed Bonus Amounts, Cohort 1

	Year 1		Year 2		Year 3		Year 4	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Teachers								
Pay-for-Performance								
With imputed amounts	$3,044*	$395	$2,815*	$506	$2,843*	$135	$3,102*	$538
Only non-imputed amounts	$2,834*	$293	$2,766*	$460	$2,808*	$122	$3,048*	$449
Automatic 1 Percent Bonus								
With imputed amounts	$765	$1,086	$1,020	$872	$1,271	$1,932	$1,898	$1,932
Only non-imputed amounts	$495	$970	$771	$764	$1,622	$2,602	$1,302	$1,145
Number of Teachers—Range[a]	**194-219**	**190-222**	**191-229**	**185-252**	**175-215**	**186-234**	**176-201**	**155-194**
Principals								
Pay-for-Performance								
With imputed amounts	$4,737*	$652	$6,023*	NA	NA	NA	$6,529*	NA
Only non-imputed amounts	$4,505*	$207	$5,962*	$321	$6,628*	$374	$6,431*	$511
Automatic 1 Percent Bonus								
With imputed amounts	$1,776	$1,060	$1,032	$1,214	$803	$1,295	$1,187	$1,071
Only non-imputed amounts	$1,663	$979	$801	$992	$855	$1,286	$1,022	$944
Number of Principals—Range[a]	**55-63**	**58-64**	**59-63**	**46-61**	**57-58**	**53-61**	**55-60**	**54-62**

Sources: Teacher and principal surveys (2012, 2013, 2014, and 2015).

Notes: All treatment principals who reported being eligible for pay-for-performance bonuses in Year 3 and all control principals who reported being eligible for pay-for-performance bonuses in Years 2 through 4 responded to the survey question about maximum possible bonus amount. Therefore, no multiple imputation was needed for these principals' maximum possible pay-for-performance bonus amount in these years.

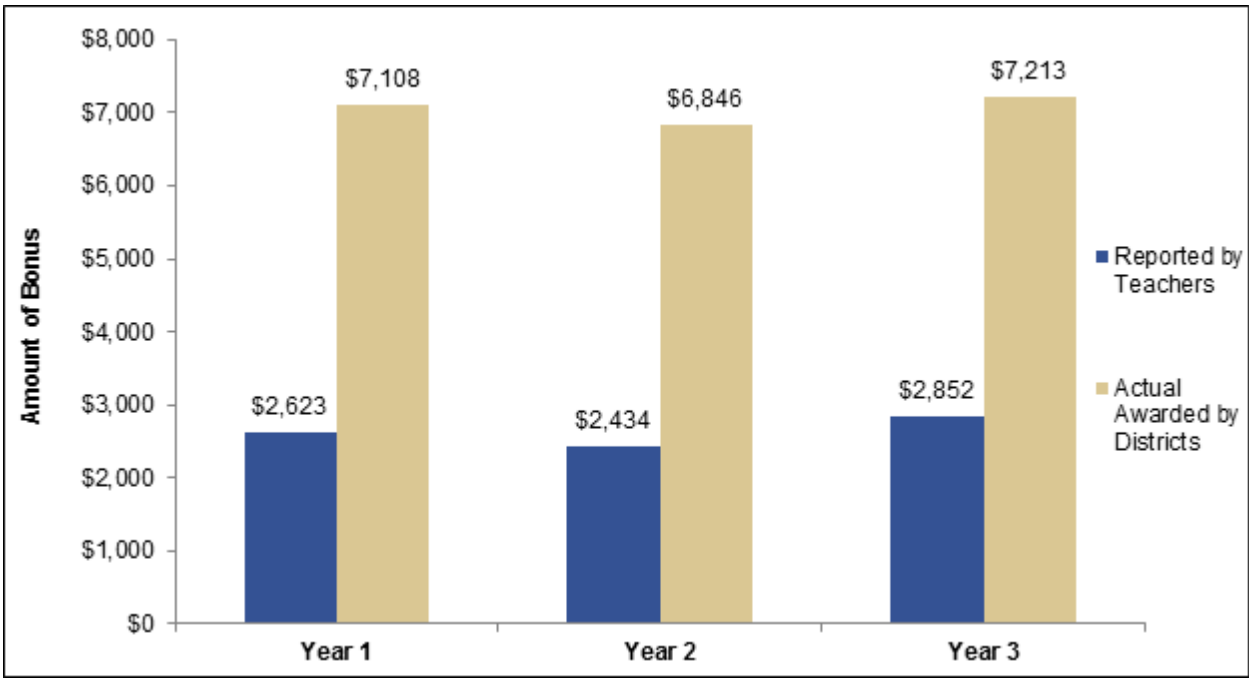
[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference between treatment and control group is statistically significant at the .05 level, two-tailed test.

NA is not applicable.

Figures D.18 and D.19 show the actual and reported maximum pay-for-performance bonuses for teachers and for principals for Years 1 through 3 for Cohorts 1 and 2. Similar to findings based on Cohort 1 only, teachers underestimated the potential amount they could earn in a bonus. Principals also underestimated the maximum bonus they could earn, although their expectations aligned more closely with the actual bonuses awarded.

Figure D.18. Reported and Actual Maximum Pay-for-Performance Bonus for Teachers in Treatment Schools, Cohorts 1 and 2

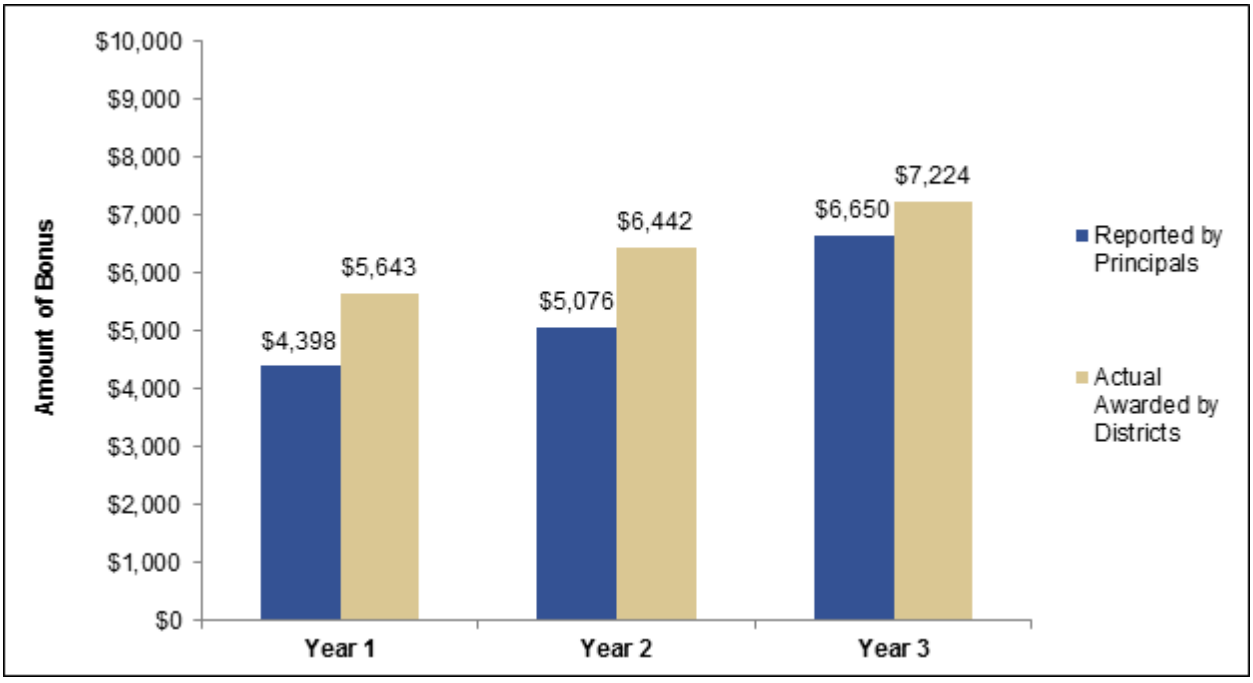


Sources: Teacher surveys (2012, 2013, 2014, and 2015) and educator administrative data.

Notes: Teachers' reports are based on data for teachers in tested grades and subjects, with each school receiving an equal weight. Districts' payouts are based on data for all teachers, with each district receiving an equal weight.

A total of 258 treatment teachers in tested grades and subjects in Cohort 1 and 2 schools responded to this survey question in Year 1, a total of 287 in Year 2, and a total of 264 in Year 3. The maximum bonus amount was set to zero for all respondents who indicated they were ineligible for a bonus. For teachers who reported being eligible for the bonus but left the amount missing, bonus amounts were imputed through multiple imputation methods. This led to 26 additional responses for treatment teachers in Year 1, 21 additional responses in Year 2, and 17 in Year 3.

Figure D.19. Reported and Actual Maximum Pay-for-Performance Bonus for Principals in Treatment Schools, Cohorts 1 and 2



Sources: Principal surveys (2012, 2013, 2014, and 2015) and educator administrative data.

Notes: Principals' reported values are calculated giving each school an equal weight. Actual payouts are calculated giving each district an equal weight.

A total of 70 treatment principals in Cohorts 1 and 2 responded to this survey question in Year 1, a total of 76 in Year 2 and a total of 75 in Year 3. The maximum bonus amount was set to zero for all respondents who indicated they were ineligible for a bonus. For educators who reported being eligible for the bonus but left the amount missing, bonus amounts were imputed through multiple imputation methods. This led to 9 additional responses for treatment principals in Year 1, 5 additional responses in Year 2, and 0 additional responses in Year 3.

Examining Why Teacher Understanding Varies

As explained in Chapter IV, because understanding about eligibility for a bonus and the potential size of the bonus are critical for changing behavior, we explored how teacher understanding varied across districts, across schools within the same district, and within the same school. Table D.19 shows the percentage of the variation in teachers' understanding of their bonus eligibility and maximum possible bonus that can be attributed to variation across districts, variation across schools within the same district, and variation across teachers within the same schools.[67] We found that most of the difference in treatment teachers' understanding (more than 85 percent of the variation in understanding of bonus eligibility and more than 70 percent of the variation in understanding of the maximum bonus amount) occurs among teachers in the same school.

Table D.19. Percentages of Total Variance in Treatment Teachers' Understanding of Their Pay-for-Performance Bonus Eligibility and Maximum Possible Bonus Amount Attributable to Districts, Schools, and Teachers, Cohort 1

	Pay-for-Performance Bonus Eligibility				Maximum Possible Pay-for-Performance Bonus Amount			
	Year 1	Year 2	Year 3	Year 4	Year 1	Year 2	Year 3	Year 4
Variation Across Districts	12	5	11	9	13	15	12	16
Variation Across Schools Within Districts	3	4	1	3	5	13	11	14
Variation Across Teachers Within Schools	85	91	88	87	82	72	78	70
Number of Teachers	**368**	**439**	**420**	**391**	**368**	**439**	**420**	**391**
Number of Schools	**65-368**	**65-439**	**65-420**	**63-391**	**65-368**	**65-439**	**65-420**	**63-391**

Sources: Teacher surveys (2012, 2013, 2014, and 2015).

Notes: Percentages may not add up to 100 because of rounding.

Tables D.20 and D.21 present subgroup results that examine district, principal, and teacher factors that might account for differences in treatment teachers' understanding of their eligibility for a performance bonus and the maximum possible bonus amount.

[67] We disaggregated the variance components by estimating a random effect model of bonus eligibility (or maximum possible bonus amount) on intercepts for schools and districts that account for the nesting of teachers in schools and schools in districts.

Table D.20. Treatment Teachers' Reported Eligibility for Pay-for-Performance Bonuses and Reported Maximum Bonuses in Year 4, by Districts' Characteristics, Cohort 1 (Percentages)

	Percentage of Teachers Reporting They Are Eligible for Pay-for-Performance Bonuses	Teachers' Reported Maximum Pay-for-Performance Bonuses as a Percentage of the Actual Awarded	**Number of Treatment Teachers**
All Teachers (primary analysis)	**58**	**27**	**391**
Subgroup Analyses By District Characteristics			
District Communication Approach			
(1) Centralized—relied primarily on district staff	49	28	**179**
(2) Decentralized—relied primarily on school staff	64	25	**212**
Difference, (1) – (2)	-15*	3	
District Expectations of Teachers' Participation in Professional Development for Current Year			
(1) At least 75 percent of teachers will participate	66	17	**169**
(2) Fewer than 75 percent of teachers will participate	56	26	**222**
Difference, (1) – (2)	10	-9	
Districts' Use of Classroom Achievement Growth to Determine Pay-for-Performance Bonuses			
(1) Used classroom achievement growth	62	23	**295**
(2) Did not use classroom achievement growth	66	23	**96**
Difference, (1) – (2)	-4	1	
Districts' Average Prior Year Pay-for-Performance Bonuses			
(1) High—at least 4.5 percent of average salary	60	27	**141**
(2) Low—less than 4.5 percent of average salary	52	25	**250**
Difference, (1) – (2)	8	2	
Districts' Prior Year Pay-for-Performance Bonus Distribution Method			
(1) Pay-for-performance bonus paid in separate check	57	28	**175**
(2) Pay-for-performance bonus paid in regular paycheck	59	23	**216**
Difference, (1) – (2)	-2	6	
Extent to which Districts Communicated Prior Year Actual Bonuses			
(1) Told all treatment teachers the total bonus amount that they earned (including $0 for nonrecipients)	54	24	**235**
(2) Did not tell all treatment teachers the total bonus amount that they earned	65	30	**156**
Difference, (1) – (2)	-11	-7	
Method by which Districts Communicated Prior Year Actual Bonuses			
(1) Held individual meetings with teachers to discuss their bonus amounts	57	24	**132**
(2) Did not hold individual meetings with teachers to discuss their bonus amounts	66	29	**183**
Difference, (1) – (2)	-8	-5	

Sources: Teacher and district surveys (2015), district interviews (2015), and administrative data.

Notes: For teachers who reported being eligible for the bonus but left the amount missing, bonus amounts were imputed through multiple imputation methods. See Appendix B for additional discussion on the imputation methods.

*Difference between subgroups is statistically significant at the .05 level, two-tailed test.

Table D.21. Treatment Teachers' Reported Eligibility for Pay-for-Performance Bonuses and Reported Maximum Bonuses in Year 4, by Principal Understanding and Teacher Characteristics, Cohort 1 (Percentages)

	Percentage of Teachers Reporting They Are Eligible for Pay-for-Performance Bonuses	Teachers' Reported Maximum Pay-for-Performance Bonuses as a Percentage of the Actual Awarded	**Number of Treatment Teachers**
All Teachers (primary analysis)	**58**	**27**	**391**
Subgroup Analysis by Principal Understanding			
Principal Understanding of Teachers' Eligibility			
(1) Principal correctly reported teachers' eligibility	64	22	**159**
(2) Principal incorrectly reported teachers' eligibility	58	28	**186**
Difference, (1) – (2)	6	-6	
Subgroup Analyses by Teacher Characteristics			
Teacher Experience in Current School			
(1) More than one year in school	64	29	**332**
(2) First year in school	50	19	**59**
Difference, (1) – (2)	14	10	
Teaching Assignment			
(1) Tested grade and subject	59	29	**201**
(2) Nontested grade and subject	51	23	**190**
Difference, (1) – (2)	8	7	
Report About Receiving a Pay-for-Performance Bonus Based on Prior Year's Performance			
(1) Reported receiving a pay-for-performance bonus	87	43	**153**
(2) Reported not receiving a pay-for-performance bonus	37	16	**234**
Difference, (1) – (2)	50*	27*	
Actual Receipt of a Pay-for-Performance Bonus Based on Prior Year's Performance			
(1) Received a pay-for-performance bonus	70	34	**254**
(2) Did not receive a pay-for-performance bonus	48	19	**133**
Difference, (1) – (2)	22*	16*	
Participation in Professional Development About TIF Performance Measures			
(1) Teacher participated in professional development	53	26	**205**
(2) Teacher did not participate in professional development	52	25	**175**
Difference, (1) – (2)	1	1	
Mentoring Role			
(1) Teacher had a mentor teacher	57	29	**174**
(2) Teacher did not have a mentor teacher	51	22	**215**
Difference, (1) – (2)	7	7	
(1) Teacher mentored other teachers	70	35	**106**
(2) Teacher did not mentor other teachers	52	24	**283**
Difference, (1) – (2)	18*	11*	
(1) Teacher mentored other teachers as part of TIF	73	33	**54**
(2) Teacher did not mentor other teachers as part of TIF	53	25	**335**
Difference, (1) – (2)	20	8	

Sources: Teacher and principal surveys (2015) and administrative data.

Notes: For teachers who reported being eligible for the bonus but left the amount missing, bonus amounts were imputed through multiple imputation methods. See Appendix B for additional discussion on the imputation methods.

*Difference between subgroups is statistically significant at the .05 level, two-tailed test.

Educators' Understanding of and Experiences with Professional Development

The TIF grant required that teachers receive professional development focused on understanding performance measures used in TIF and feedback based on their performance ratings. This requirement applied equally to teachers in treatment and control schools. Tables D.22 and D.23 show that teachers in treatment and control schools generally reported similar professional development experiences. These tables also support the finding discussed in Chapter IV that at least half of teachers reported they received the professional development required under the TIF grant but indicated they received no more than six hours of professional development over the school year.

Table D.22. Professional Development Teachers Reported Receiving or Expecting to Receive, Cohort 1 (Percentages)

	Year 1		Year 2		Year 3		Year 4	
Professional Development Topics	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Understanding Components of TIF	81*	76	66 +	70	62	57+	48 +	48+
Understanding Performance Measures of TIF	77	73	63*+	68	68	62	53 +	57
Feedback Based on TIF Performance Ratings	53	55	52*	57	60 +	57	51*+	58
Differentiated Instructional Strategies Based on Student Assessments	69	73	76 +	76	80	78	79	73
Instructional Techniques and Strategies	88*	82	87	90+	89*	92	90	90
Aligning Curricula to State or District Standards	80	81	81	85	80	79+	77	76
Number of Teachers—Range[a]	**374-376**	**387-391**	**430-432**	**438-440**	**405-409**	**442-448**	**384-388**	**390-392**

Sources: Teacher surveys (2012, 2013, 2014, and 2015).

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Table D.23. Hours of Expected Professional Development, as Reported by Teachers, Cohort 1 (Averages)

	Expected Hours Among Teachers Who Expected to Receive Any Professional Development in the Specified Topic							
	Year 1		Year 2		Year 3		Year 4	
Professional Development Topics	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Understanding Components of TIF	9	5	4 +	5	4	4+	2*+	4
Understanding Performance Measures of TIF	6	4	3*+	4	3	3+	2*+	5
Feedback Based on TIF Performance Ratings	5	4	3 +	4	3	3	2*+	5
Differentiated Instructional Strategies Based on Student Assessments	9	8	8	7	9	8+	9	8
Instructional Techniques and Strategies	14	12	15*	12	12	13	12	13
Aligning Curricula to State or District Standards	11	10	9*	8	9	9	9	9
Number of Teachers—Range[a]	**186-315**	**200-314**	**217-361**	**245-380**	**228-359**	**237-391**	**185-335**	**190-335**

Sources: Teacher surveys (2012, 2013, 2014, and 2015).

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

APPENDIX E

SUPPLEMENTAL FINDINGS ON IMPACTS OF PAY-FOR-PERFORMANCE ON EDUCATORS' ATTITUDES AND BEHAVIORS FOR CHAPTER V

THIS PAGE IS INTENTIONALLY BLANK

This appendix supplements the findings presented in Chapter V. As discussed in Chapter II, evaluation districts were classified into two cohorts—Cohort 1 and Cohort 2—according to the year in which we randomly assigned their schools to a treatment group or a control group. The 10 districts whose schools were randomly assigned in spring and summer 2011 were classified as Cohort 1. Three additional districts, whose schools were randomly assigned in spring and summer 2012, were classified as Cohort 2. Cohort 1 districts completed four years of implementation. Year 1 represents the first year of implementation (2011–2012), Year 2 the second (2012–2013), Year 3 the third year of implementation (2013–2014), and Year 4 the fourth (2014–2015). Cohort 2 districts completed only three years of implementation, 2012–2013, 2013–2014, and 2014–2015, referred to as Years 1, 2, and 3 for this cohort.

Tables E.1 through E.7 present impact estimates for three years of TIF implementation using all evaluation schools (Cohorts 1 and 2) and additional findings based on teachers' subgroups for Cohort 1 only. Tables E.8 through E.10 provide evidence on the impact of pay-for-performance on additional measures for Cohort 1: principals' hiring autonomy, staffing, and compensation decisions. Although these factors are not the main drivers of teachers' productivity or mobility captured in our logic model, they may still contribute to teachers' school environment and job satisfaction. Table E.11 shows the impact of pay-for-performance on teachers' time on school-related activities.

Impacts for Cohort 1 Schools Compared to Cohorts 1 and 2

In Chapter V, we presented impact estimates based on Cohort 1 schools that have implemented the program for four full years. Here, we show estimates for the first, second, and third years of implementation (Years 1, 2, and 3) for all study schools that have implemented the program, combining Cohorts 1 and 2.

Table E.1. Teachers' Satisfaction with Professional Opportunities, Evaluation System, and School Environment, Cohorts 1 and 2 (Percentages Who Are Somewhat or Very Satisfied)

	Year 1			Year 2			Year 3		
Satisfaction Dimension	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
Opportunities for Pay and Development									
Opportunities for professional advancement	68	74	-6*	73	73	-1	73	71	2
Opportunities to enhance skills	76	77	-2	81	80	1	79	80	-1
Opportunities to earn extra pay	65	62	3	66	59	7*	64	55	10*
Key Features of Evaluation System									
Use of student achievement scores to assess performance	58	63	-5	56	65	-9*	64	62	2
Feedback on my performance	—	—	—	76	78	-2	80	79	0
School Environment									
Recognition of accomplishments	53	60	-7*	61	63	-1	65	61	4
Quality of interaction with colleagues	75	82	-6*	80	81	0	84	81	3*
Colleagues' efforts	82	84	-2	82	83	-1	84	83	0
School morale	46	52	-6	55	56	-1	61	53	8*
Job Satisfaction									
Overall job satisfaction	66	71	-5	70	71	-1	76	72	4
Number of Teachers—Range[a]	**497-502**	**489-498**		**571-575**	**562-566**		**540-545**	**577-581**	

Sources: Teacher surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

— is not available.

*Impact is statistically significant at the .05 level, two-tailed test.

Table E.2. Principals' Satisfaction with Professional Opportunities, Evaluation System, and School Environment, Cohorts 1 and 2 (Percentages Who Are Somewhat or Very Satisfied)

	Year 1			Year 2			Year 3		
Satisfaction Dimension	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
Opportunities for Pay and Development									
Opportunities for professional advancement	—	—	—	88	86	2	90	86	4
Opportunities to enhance skills	94	95	-1	87	83	4	97	85	12*
Opportunities to earn extra pay	69	62	8	58	60	-2	73	51	22*
Key Features of Evaluation System									
Use of observations to assess skills	—	—	—	67	83	-16*	92	84	9
Use of student achievement scores to assess performance	—	—	—	61	71	-10	82	71	11
Feedback on my performance	79	86	-6	71	75	-5	81	81	0
School Environment									
Recognition of accomplishments	73	77	-4	60	69	-9	75	67	8
Quality of interaction with colleagues	92	95	-3	86	88	-2	94	87	7
Colleagues' efforts	92	97	-5	88	87	1	94	91	3
School morale	74	83	-9	75	79	-4	92	82	9
Number of Principals—Range[a]	**79-80**	**73-77**		**81-82**	**76-77**		**76-77**	**78-79**	

Sources: Principal surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]Sample sizes are presented as a range, based on the data available for each row in the table.

— is not available.

*Impact is statistically significant at the .05 level, two-tailed test.

Table E.3. Teachers' Attitudes Toward TIF Program, Cohorts 1 and 2 (Percentages Who Agree or Strongly Agree)

	Year 1			Year 2			Year 3		
Statement	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
Overall Attitudes Toward TIF									
I am glad that I am participating in the TIF program	67	64	3	65	71	-6	70	70	0
The TIF program is fair	53	58	-5	53	57	-4	57	58	-1
My job satisfaction has increased due to the TIF program	30	34	-4	36	37	-1	41	33	8*
Attitudes Toward Evaluation and Compensation Approaches Used by TIF									
My principal is a good judge of teacher talent	66	75	-8*	72	73	-1	79	77	2
Standardized student test scores in my district measure what students have learned	32	28	4	30	34	-4	26	29	-3
Teachers who do the same job should receive the same pay	58	60	-1	63	66	-3	67	68	-1
The process used to determine how bonuses are determined was adequately explained to me	60	53	6*	60	54	5	64	56	8*
Attitudes Toward Effects of TIF on Teachers' Effort and Practices									
I feel increased pressure to perform due to the TIF program	63	51	12*	62	47	15*	57	48	8*
The TIF program has harmed the collaborative nature of teaching	27	25	1	32	21	11*	24	25	-1
The TIF program has caused teachers to work more effectively	46	46	0	48	52	-4	54	50	4
I have less freedom to teach the way I would like to teach due to the TIF program	37	34	3	38	29	9*	33	34	-1
Number of Teachers—Range[a]	**455-496**	**450-495**		**480-563**	**467-555**		**479-537**	**475-568**	

Sources: Teacher surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]Sample sizes are presented as a range, based on the data available for each row in the table.

*Impact is statistically significant at the .05 level, two-tailed test.

Table E.4. Principals' Attitudes Toward TIF Program, Cohorts 1 and 2 (Percentage Who Agree or Strongly Agree)

	Year 1			Year 2			Year 3		
Statement	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
Attitudes Toward Effects of TIF on Teachers' Behavior									
The TIF program contributes to greater collegiality and professionalism among the staff at this school	43	53	-9	52	63	-11	59	50	9
Teachers at this school are more comfortable with frequent formal observations of their teaching because of the TIF program	52	58	-6	57	62	-5	64	56	8
Attitudes Toward Fairness of TIF									
The evaluation system omits important aspects of school administration that should be considered	41	38	3	58	49	9	53	51	2
This school has less chance of earning a bonus because of the characteristics of our student population	23	24	-1	39	26	14	33	32	1
Attitudes Toward Stakeholders' Buy-in and Sustainability of TIF									
I played an important role in implementing the TIF program at my school	76	79	-3	79	78	0	83	66	17*
The TIF program has been clearly communicated to me	81	84	-3	85	91	-5	88	82	6
Parents and the school community believe the TIF program is important	38	43	-5	45	36	10	35	43	-8
The TIF program is likely to continue for the foreseeable future	80	83	-3	70	68	2	53	45	8
Number of Principals—Range[a]	**78-81**	**74-80**		**76-80**	**73-76**		**76-77**	**74-78**	

Sources: Principal surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]Sample sizes are presented as a range, based on the data available for each row in the table.

*Impact is statistically significant at the .05 level, two-tailed test.

Additional Findings on Teachers' Attitudes by Subgroup

Tables E.5 through E.7 show supplementary analyses of teachers' satisfaction and attitudes in the fourth year of TIF implementation. Table E.5 shows the impacts of pay-for-performance on teachers' satisfaction with their professional opportunities, evaluation system, and school environment by subgroups based on teaching assignment and teaching experience. Table E.6 examines treatment teachers' satisfaction on these dimensions by whether the teacher received (or reported receiving) a bonus based on their Year 3 performance. Table E.7 shows the impacts of pay-for-performance on teachers' attitudes toward their job and the TIF program by subgroups based on teaching assignment and teaching experience.

Table E.5. Impacts of Pay-for-Performance on Teacher Satisfaction Measures for Teacher Subgroups, Year 4, Cohort 1 (Percentage Points)

	Impacts on Whether Teachers Were Somewhat or Very Satisfied with…											
Subgroup	Opportunities for Professional Advancement	Opportunities to Enhance My Skills	Opportunities to Earn Extra Pay	Use of Student Achievement Scores to Assess Teacher Effectiveness	Use of Classroom Observations to Assess Skills	Feedback on My Performance	Recognition of Accomplish-ments	Quality of Interactions with Colleagues	Colleagues' Efforts	School Morale	Overall Job Satisfaction	**Number of Teachers[a]**
All Teachers (primary analysis)	-6	0	7*	-1	-2	10*	4	0	0	5	7	**782–789**
Teaching Assignment												
(1) Tested grades and subjects	-3	1	8	3	-3	9*	2	2	2	7	3	**391–396**
(2) Nontested grades and subjects	-8	0	6	-4	0	11*	6	-2	-1	4	11*	**391–393**
Difference, (1) - (2)	**5**	**1**	**2**	**8**	**-3**	**-2**	**-4**	**4**	**3**	**3**	**-8**	
Teacher Experience												
(1) Fewer than 5 years	-8	0	8	4	-1	5	18*	-3	2	6	3	**214–215**
(2) 5 to 15 years	5	3	11*	0	2	10*	-3	-1	-3	10	14*	**354–358**
(3) Greater than 15 years	-20*	-3	0	-10	-10	14*	3	2	4	-2	0	**211–213**
Difference, (1) - (2)	**-13**	**-3**	**-4**	**4**	**-3**	**-6**	**21***	**-2**	**4**	**-4**	**-11**	
Difference, (3) - (2)	**-25***	**-6**	**-11**	**-11**	**-12**	**4**	**6**	**3**	**7**	**-13**	**-14**	

Source: Teacher survey, 2015.

[a]Sample sizes are presented as a range, based on the data available for each row in the table.

*Impact is statistically significant at the .05 level, two-tailed test.

Table E.6. Treatment Teachers' Satisfaction by Bonus Receipt and Report of Bonus Receipt, Year 4, Cohort 1 (Percentages Who Agree or Strongly Agree)

	Actual Year 3 Bonus Receipt			Report of Year 3 Bonus Receipt		
Statement	Received a Bonus	Did Not Receive a Bonus	Difference	Reported Receiving Bonus	Reported Not Receiving a Bonus	Difference
Opportunities for Pay and Development						
Opportunities for professional advancement	65	74	-9	64	69	-6
Opportunities to enhance skills	79	82	-3	80	78	1
Opportunities to earn extra pay	55	63	-8	64	52	12
Key Features of Evaluation System						
Use of classroom observations to assess skills	72	73	-1	73	68	5
Use of student achievement scores to assess teacher effectiveness	66	61	5	63	57	7
Feedback on teacher performance	80	82	-2	83	83	0
School Environment						
Recognition of accomplishments	60	64	-4	64	61	3
Quality of interaction with colleagues	81	77	3	74	73	1
Colleagues' efforts	78	80	-2	79	77	1
School morale	50	57	-7	57	53	4
Job Satisfaction						
Overall job satisfaction	67	71	-4	71	72	-1
Number of Teachers—Range[a]	**250-252**	**136-137**		**150-152**	**231-232**	

Sources: Teacher survey (2015) and educator administrative data.

Notes: Pay-for-performance bonus receipt information comes from Year 3 educator administrative data. The difference between those who received (or reported receiving) a bonus and those who did not may not equal the difference shown in the table because of rounding.

[a]Sample sizes are presented as a range, based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

Table E.7. Impacts of Pay-for-Performance on Teacher Attitude Measures for Teacher Subgroups, Year 4, Cohort 1 (Percentage Points)

	Impacts on Whether Teachers Responded They Agreed or Strongly Agreed with…											
Subgroup	I Am Glad I Am Participating in the TIF Program	The TIF Program Is Fair	My Job Satisfaction Has Increased due to the TIF Program	My Principal Is a Good Judge of Teacher Talent	Standardized Student Test Scores in My District Measure What Students Have Learned	Teachers Who Do the Same Job Should Receive the Same Pay	The Process Used to Determine How Bonuses Are Determined Was Adequately Explained to Me	I Feel Increased Pressure to Perform due to the TIF Program	The TIF Program Has Harmed the Collaborative Nature of Teaching	The TIF Program Has Caused Teachers to Work More Effectively	I Have Less Freedom to Teach the Way I Would Like to Teach due to the TIF Program	**Number of Teachers**
All Teachers (primary analysis)	0	3	4	2	-2	2	2	11*	5	8*	5	**695–770**
Teaching Assignment												
(1) Tested grades and subjects	3	0	-1	7	1	8	2	12*	2	0	4	**346–390**
(2) Nontested grades and subjects	-2	6	10	-2	-5	-3	2	10	7	15*	5	**347–380**
Difference, (1) - (2)	**5**	**-6**	**-11**	**9**	**6**	**11**	**0**	**2**	**-5**	**-15***	**-1**	
Teacher Experience												
(1) Fewer than 5 years	5	12	13	-6	10	6	18*	-1	-5	6	0	**165–207**
(2) 5 to 15 years	3	0	1	3	-9	1	-1	17*	10*	13*	9	**323–353**
(3) Greater than 15 years	-8	-1	3	9	-2	1	-11	9	6	2	2	**196–208**
Difference, (1) - (2)	**2**	**12**	**12**	**-9**	**19***	**5**	**19**	**-18**	**-14**	**-6**	**-9**	
Difference, (3) - (2)	**-10**	**-1**	**2**	**6**	**7**	**0**	**-11**	**-8**	**-3**	**-11**	**-7**	

Source: Teacher survey, 2015.

[a]Sample sizes are presented as a range, based on the data available for each row in the table.

*Impact is statistically significant at the .05 level, two-tailed test.

Impacts on Principals' Hiring Autonomy, Staffing, and Compensation Decisions

In this section, we report findings on principals' hiring autonomy, staffing, and compensation decisions. Principals' autonomy in hiring is a necessary, though not sufficient, condition for pay-for-performance to have an effect on principal recruitment strategies. Most principals in both treatment and control schools reported having input in hiring decisions (over 95 percent reported either they had complete authority over hiring, were part of a team responsible for hiring teachers, or could chose whom to hire from an approved list), although less than one quarter reported having complete autonomy over teacher hiring (Table E.8). In addition, the introduction of pay-for-performance in treatment schools may generate incentives for principals to strategically assign teachers to classrooms or use nonmonetary compensation. Because pay-for-performance bonuses depend on students' achievement growth on standardized tests, principals in schools eligible for such bonuses may use different criteria to assign teachers to tested grades and subjects. For example, if school staff can earn a pay-for-performance bonus based on student achievement growth measured at the school level, a principal may decide to assign teachers to tested grades and subjects based on a belief in a teacher's ability to raise student achievement scores. Control schools could also compensate for the lack of pay-for-performance bonuses in their schools by making more extensive use of nonmonetary benefits to reward performance, such as giving effective teachers more time for leadership activities or priority in teaching assignments.

We found no consistent evidence that principals determined teacher assignments or compensated teachers differently in response to pay-for-performance. Across all years of the study, pay-for-performance had no significant impact on most measures of principals' staffing decisions (Table E.9). For example, treatment and control principals were equally likely to report that they used teacher's experience in a grade level of subject area or ability to produce high test scores when making decisions. The notable exception is that treatment and control principals often differed (in Years 1, 2, and 4) in reporting that they made assignments based on teachers' seniority. However, there was no consistent pattern in terms of which principals were more likely to consider teachers' seniority.

Principals in control schools were not more likely than principals in treatment schools to offer teachers nonmonetary benefits to compensate their teachers for not being eligible to earn a performance bonus. About 40 percent of principals (37 to 46 percent of treatment principals and 35 to 38 percent of control principals) offered nonmonetary benefits, such as release from classroom teaching, increased decision-making authority, or priority in student assignments (Table E.10). In general, treatment and control principals were equally likely to use any particular nonmonetary benefit.

Table E.8. Principals' Autonomy in Hiring Teachers, Cohort 1 (Percentages)

	Year 1			Year 2			Year 3			Year 4		
	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
Principal has complete autonomy over teacher hiring	12	5	7*	23	15	8	20	13	8	26	8	18*
Principal is part of a school-level team responsible for teacher hiring	50	52	-1	47	57	-11	52	60	-7	43	62	-20*
Principal receives a set of prescreened candidates from the district office as the pool from which he or she can interview and hire	36	41	-4	26	25+	1	24	24	0	26	26	0
Number of Principals	**64**	**64**		**63**	**61**		**58**	**62**		**61**	**61**	

Sources: Principal surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

*Impact is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test

Table E.9. Criteria Used for Teacher Assignments to Grade Levels or Subject Areas, Cohort 1 (Percentages Who Report They Are Always or Often Used)

	Year 1			Year 2			Year 3			Year 4		
	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
The teacher's experience in a grade level or subject area	84	89	-5	88	90	-1	88	82	6	91	82	9
The teacher's seniority	<10[a]	<15[a]	-12*	<15[a]	<10[a]+	9*	14	13+	1	11	24	-14*
The teacher's content knowledge	91	97	-6	93	93	-1	96	93	3	97	89	8*
The teacher's ability to produce high test scores in grades/classes in which state or federal assessments are administered	72	74	-3	66	66	0	61	64	-3	66	61	5
The teacher's ability to work with certain student populations	85	80	5	86	81	5	91	85	6	88	79	9
To balance teacher experience and expertise in a grade level or subject	72	71	1	70	73	-3	71	75	-3	71	67	4
Number of Principals—Range[b]	**63-64**	**62-64**		**61-62**	**58-59**		**57-58**	**59-61**		**60-61**	**60-62**	

Sources: Principal surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]To reduce the risk of disclosing information on an individual, the exact percentage is not shown.

[b]Sample sizes are presented as a range, based on the data available for each row in the table.

*Impact is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Table E.10. Nonmonetary Benefits Used to Recognize Teachers' Performance or Responsibilities, Cohort 1 (Percentages)

	Year 1			Year 2			Year 3			Year 4		
	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
Use of Nonmonetary Benefits	37	38	0	39	37	3	46	35	11	42	35	7
Type of Nonmonetary Benefits:												
Release from classroom teaching for mentoring or other leadership activities	34	27	8	30	32	-2	32	32	1	34	34	0
Decision-making authority on issues such as hiring staff or adopting curriculum	31	28	3	31	30	1	35	30	5	35	34	1
Priority in teaching assignments	5	11	-6	10	18	-9	9	18	-9*	12	18	-5
Priority in student assignments	<5[a]	<5[a]	-1	<5[a]	<10[a]	-4	8	5	3	6	8	-2
Number of Principals—Range[b]	**63**	**64**		**62-63**	**60**		**58**	**60**		**61**	**62**	

Sources: Principal surveys (2012, 2013, 2014, and 2015).

Note: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]To reduce the risk of disclosing information on an individual, the exact percentage is not shown.

[b]Sample sizes are presented as a range, based on the data available for each row in the table.

*Impact is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

Teachers' Use of Time Throughout the School Day

We asked teachers to report how they spent their time in the most recent full week of teaching. In theory, pay-for-performance could motivate teachers to allocate more time to activities aimed at improving their performance ratings. For example, if efforts to improve performance ratings entail revamping lessons to better align with state assessments, treatment teachers may decide to spend more time than control teachers on class preparation.

Pay-for-performance did not affect teachers' time on school-related activities. On average, across all years of the study teachers reported working approximately 45 hours during school hours in the most recent full week of work (Table E.11). Treatment and control teachers reported spending a similar amount of time on specific activities both during and outside school hours. In cases where there was a significant difference in their reported hours on an activity, the difference was almost always an hour.

Table E.11. Teachers' Time Spent on School-Related Activities in the Most Recent Full Week (Average Hours)

	Year 1			Year 2			Year 3			Year 4		
	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact	Treatment	Control	Impact
Time Spent During School Hours on												
Teaching students in the classroom, small groups, or individually	27	26	1	27	28	0	28	29	-1	28	28	0
Supervising students in other activities	4	4	0	4	4	0	3	3	0	3	3	0
Preparation on your own (for example, lessons, grading, assignment)	6	7	0	8+	7	1	8	8	0	9	7	1*
Preparation and professional development with colleagues (for example, common lesson planning, workshops, staff meetings, mentoring)	3	3	0	4	4	0	4	4	-1	5	3+	1*
Other activities	2	2	1	2	2	0	2	2	0	2	2	0
Total hours during school hours (calculated)	42	41	2	45	44+	0	45	46	0	46	43	3
Time Spent During Nonschool Hours on												
Academic-related activities with students	2	2	-1*	3	4+	-1*	2	3+	0	3	3	0
Other activities with students	1	1	0	1	1	0	1	1	0	1	1	0
Preparation on your own	9	8	0	10	8	2*	9	9	0	10	9	1*
Preparation and professional development with colleagues	2	3	-1*	3+	3	0	3	3	0	3	2+	1*
Other school-related activities	1	1	0	1	1	0	1	1	0	1	1	0
Total hours during nonschool hours (calculated)	14	16	-2	18	17	1	16	16	-1	18	15	2
Number of Teachers—Range[a]	**312-366**	**315-379**		**429-442**	**443-448**		**370-426**	**398-460**		**340-390**	**338-397**	

Sources: Teacher surveys (2012, 2013, 2014, and 2015).

Notes: The categories in the table are identical to the language used in the survey. The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Impact is statistically significant at the .05 level, two-tailed test.

+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

THIS PAGE IS INTENTIONALLY BLANK

APPENDIX F

SUPPLEMENTAL FINDINGS ON IMPACTS OF PAY-FOR-PERFORMANCE ON EDUCATOR EFFECTIVENESS AND STUDENT ACHIEVEMENT FOR CHAPTER VI

THIS PAGE IS INTENTIONALLY BLANK

This appendix supplements the findings presented in Chapter VI that examined the impacts of pay-for-performance on educator effectiveness and student achievement.

As discussed in Chapter II, evaluation districts were classified into two cohorts—Cohort 1 and Cohort 2—according to the year in which we randomly assigned their schools to a treatment group or a control group. The 10 districts whose schools were randomly assigned in spring and summer 2011 were classified as Cohort 1. Three additional districts, whose schools were randomly assigned in spring and summer 2012, were classified as Cohort 2. Cohort 1 districts completed four years of implementation. Year 1 represents the first year of implementation (2011–2012), Year 2 the second (2012–2013), Year 3 the third (2013–2014), and Year 4 the fourth (2014–2015). Cohort 2 districts completed three years of implementation, 2012–2013, 2013–2014, and 2014–2015, referred to as Years 1, 2, and 3 for this cohort.

This appendix includes supplemental findings for Cohort 1 (for example, supplemental information for systematic reviews and subgroup findings), findings for Cohorts 1 and 2, and sensitivity analyses that assess the robustness of the main impact estimates reported in Chapter VI.

Supplemental Information for Systematic Reviews

Systematic reviews of evidence on the impacts of educational interventions often require specific types of information to evaluate the quality of a study. This section provides supplemental information that a systematic review would potentially need to assess the quality of the main impact findings reported in Chapter VI—specifically, findings about the impacts of pay-for-performance on educator effectiveness and student achievement in Cohort 1 schools.

Cluster and School Attrition

Because this study was a randomized controlled trial, the extent of attrition from the original randomly assigned sample is the key factor determining the internal validity of the impact findings. As discussed in Appendix A, we randomly assigned clusters—either schools or groups of schools—to the treatment or control groups. We then made conclusions (or "inferences") about the impacts of pay-for-performance on schools, a subcluster unit. Therefore, the attrition rates of both clusters and schools are central to evaluating the evidence in Chapter VI.

Table F.1 shows the original number of clusters that we randomly assigned and the final number of clusters included in the analysis of each outcome. Among the original ("baseline") sample of clusters relevant to most outcomes, we assigned 48 clusters to the treatment group and 48 clusters to the control group. Some educator effectiveness outcomes were not applicable to particular districts because either the districts did not use those types of effectiveness measures or those measures were not based on a rating scale with a defined minimum and maximum value. Whenever an outcome was not applicable to a particular district, we excluded the treatment and control clusters in that district from the definition of the original, randomly assigned sample. For each outcome, the number of clusters in the final analysis sample differed from the original number of randomly assigned clusters because of cases in which (1) all schools in a cluster closed or dropped out of the study; (2) the study team dropped clusters that, for random assignment, had been paired with clusters that closed or dropped out; or (3) all schools in a cluster had missing data on the specified outcome.

Table F.1. Cluster and School Attrition in the Analysis of the Impacts of Pay-for-Performance on Educator Effectiveness and Student Achievement, Cohort 1

	Original Number of Clusters that were Randomly Assigned		Final Number of Clusters that Remained in the Analysis Sample		Original Number of Schools in the Remaining Clusters		Final Number of Schools that Remained in the Analysis Sample	
Outcome	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
School Achievement Growth Ratings								
Year 1	44[a]	44[a]	41	41	62	62	61	62
Year 2	48	48	45	44	66	65	65	65
Year 3	48	48	45	45	66	66	65	66
Year 4	48	48	44	43	61	61	60	61
Classroom Achievement Growth Ratings								
Year 1	23[b]	23[b]	21	21	37	36	36	36
Year 2	23[b]	23[b]	21	21	37	36	36	36
Year 3	33[c]	33[c]	30	30	46	45	45	45
Year 4	33[c]	33[c]	30	30	46	45	45	45
Teachers' Classroom Observation Ratings								
Year 1	48	48	45	45	66	66	65	66
Year 2	48	48	45	45	66	66	65	66
Year 3	48	48	45	45	66	66	65	66
Year 4	48	48	45	45	66	66	65	66
Observation Ratings for Principals								
Year 1	48	48	37	37	55	55	53	52
Year 2	48	48	43	40	64	61	61	56
Year 3	48	48	43	43	64	64	60	58
Year 4	48	48	44	43	65	63	62	60
Student Math Achievement								
Year 1	48	48	45	45	66	66	65	66
Year 2	48	48	45	45	66	66	65	66
Year 3	48	48	45	45	66	66	65	66
Year 4	48	48	45	45	66	66	65	66
Student Reading Achievement								
Year 1	48	48	45	45	66	66	65	66
Year 2	48	48	45	45	66	66	65	66
Year 3	48	48	45	45	66	66	65	66
Year 4	48	48	45	45	66	66	65	66

Sources: Educator and student administrative data.

[a]Count excludes one district in which school achievement growth ratings did not place educators into performance categories or onto a numeric scale. Neither treatment nor control schools from this district are included in the count.

[b]Count excludes four districts that did not use classroom achievement growth to evaluate teachers in Years 1 and 2. Neither treatment nor control schools from those four districts are included in the count.

[c]Count excludes three districts that did not use classroom achievement growth to evaluate teachers in Years 3 and 4. Neither treatment nor control schools from those three districts are included in the count.

School attrition (within clusters that remained in the study) also determines the internal validity of the impact findings because for every outcome examined in Chapter VI, we sought to make conclusions about impacts on schools. As explained in Chapters I, II, and VI, pay-for-performance could affect the average educator effectiveness of schools in the study by either enabling schools to retain and recruit more effective educators or motivating educators to improve their performance. Impacts on average educator effectiveness in the study schools, reported in Tables VI.1 and VI.2, could reflect a combination of these influences. Likewise, as stated in Chapter VI, the study's main findings on student achievement captured the impacts of implementing pay-for-performance for one, two, three, and four years on schools' average student achievement. In Chapter II and Appendix B, we explained that these impacts on student achievement potentially reflected changes in individual students' achievement and changes in the schools' student composition resulting from pay-for-performance. Therefore, for the outcomes examined in Chapter VI, the units for which we made inferences (schools) were not the same as the ultimate units of analysis (educators or students).

The final four columns of Table F.1 show the original number of schools at the time of random assignment and the final number of schools included in the analysis of each outcome. Both types of school counts are based only on the clusters that remained in the analysis for the specified outcome.

Effect Sizes

Table F.2 provides complete information needed for computing effect sizes. The adjusted mean outcomes, impacts, and *p*-values are identical to those reported in Chapter VI. The additional information in this table consists of the unadjusted standard deviations of the outcomes in the treatment and control groups.

Educator Performance Ratings

This section presents six types of additional analyses of the impact of pay-for-performance on educator performance ratings: (1) sensitivity analyses that assess the robustness of the main impact estimates, (2) findings that include both Cohorts 1 and 2, (3) findings that use a consistent sample of districts and schools across years, (4) impacts of pay-for-performance on educator retention, (5) impacts of pay-for-performance on educator demographic and professional characteristics, and (6) subgroup analyses that assess impacts for returning and newly hired educators separately.

Sensitivity Analyses

Tables F.3 and F.4 explore the sensitivity of the main impact estimates for school achievement growth ratings and teacher observation ratings to several changes to the regression model or estimation sample, described below.

Table F.2. Detailed Statistics about the Impacts of Pay-for-Performance on Educator Effectiveness and Student Achievement, Cohort 1 (Points on 1-to-4 Scale Unless Otherwise Noted)

	Treatment Schools		Control Schools			
Outcome	Adjusted Mean	Unadjusted Standard Deviation	Adjusted Mean	Unadjusted Standard Deviation	Impact	*p*-value
School Achievement Growth Ratings						
Year 1	2.62	0.99	2.25	0.99	0.36*	0.04
Year 2	2.55	1.05	2.27	0.95	0.27	0.07
Year 3	2.41	1.02	2.37	0.93	0.04	0.75
Year 4	2.59	1.18	2.20	1.02	0.39*	0.03
Classroom Achievement Growth Ratings						
Year 1	2.26	0.97	2.08	0.95	0.18*	0.04
Year 2	2.23	1.00	2.17	1.04	0.06	0.25
Year 3	2.53	1.11	2.53	1.13	0.00	0.97
Year 4	2.71	1.15	2.53	1.20	0.18*	0.00
Teachers' Classroom Observation Ratings						
Year 1	2.94	0.50	2.91	0.55	0.03	0.21
Year 2	2.98	0.48	2.93	0.53	0.05	0.08
Year 3	2.97	0.70	2.91	0.69	0.05*	0.02
Year 4	2.88	0.76	2.80	0.77	0.09*	0.02
Observation Ratings For Principals						
Year 1	3.08	0.60	3.18	0.60	-0.10	0.20
Year 2	3.14	0.68	3.01	0.72	0.13	0.19
Year 3	3.37	0.64	3.32	0.58	0.05	0.52
Year 4	3.25	0.89	3.21	0.85	0.04	0.63
Student Math Achievement (student *z*-score units)						
Year 1	-0.43	0.93	-0.45	0.93	0.02	0.33
Year 2	-0.38	0.92	-0.43	0.92	0.04	0.07
Year 3	-0.36	0.93	-0.42	0.93	0.06*	0.02
Year 4	-0.34	0.94	-0.37	0.95	0.04	0.13
Student Reading Achievement (student z-score units)						
Year 1	-0.37	0.95	-0.40	0.96	0.03*	0.04
Year 2	-0.36	0.95	-0.39	0.95	0.03*	0.02
Year 3	-0.33	0.95	-0.37	0.95	0.04*	0.02
Year 4	-0.34	0.96	-0.38	0.97	0.04	0.08

Sources: Educator and student administrative data.

Notes: Means were adjusted by the regression model described in Appendix B. Unadjusted standard deviations were the standard deviations across schools for school achievement growth outcomes, across teachers for teachers' performance rating outcomes, across principals for principals' performance rating outcomes, and across students for student achievement outcomes.

Using alternative weighting approaches. In our main specification, we normalized the analysis weights so that each school received the same weight in the final analysis sample. Therefore, in the main impact estimates, districts with more schools received more weight than those with fewer schools. In addition, teachers in large schools received less weight than those in small schools. We explored three alternative approaches to normalizing sample weights. In the first alternative approach (for analyses of school achievement growth ratings and teacher observation ratings), each district received the same weight. This approach produced estimates of the impact of pay-for-

performance in the average Cohort 1 district, which could be of interest because each district designed its TIF program in a different way. In the second alternative approach (for analyses of school achievement growth ratings), each school received a weight in proportion to the number of students enrolled at the school. This approach produced estimates of the impact of pay-for-performance on the educator effectiveness experienced by the average student, which could be of interest because pay-for-performance was intended to affect students' achievement growth. In the third alternative approach (for analyses of teacher observation ratings), each teacher received the same weight. This approach produced estimates of the impact of pay-for-performance on the average teacher, which could be of interest because pay-for-performance was intended to change teachers' behavior.

In Year 4, impact estimates from the main model and the models that used alternative weighting approaches were statistically significant for school achievement growth ratings (Table F.3, models 1 and 2) and teacher observation ratings (Table F.4, models 1 and 2).

Excluding covariates. Our main estimation model controlled for randomization block indicators and the school-level pre-implementation means of student achievement and student race/ethnicity. Controlling for schools' pre-implementation characteristics accounted for treatment schools having slightly lower student math achievement and slightly different student racial/ethnic composition than control schools at the beginning of the study. Failure to account for these preexisting differences could generate an inaccurate estimate of the effects of pay-for-performance. Nevertheless, because some researchers have expressed methodological concerns about the use of covariates in analyzing experimental data (Freedman 2008), we also estimated a model that included no other covariates aside from the randomization block indicators. When covariates were excluded, the estimated impacts of pay-for-performance on school achievement growth ratings in Year 4 were smaller than the main estimates and were no longer statistically significant (Table F.3, model 3). In contrast, for teacher observation ratings, both the main model and the specification excluding covariates found impacts of pay-for-performance in Year 4 that were statistically significant and similar in magnitude (Table F.4, model 3).

Table F.3. Impacts of Pay-for-Performance on School Achievement Growth Ratings in Year 4 Using Alternative Specifications, Cohort 1 (Points on 1-to-4 Scale)

Model	Treatment Schools	Control Schools	Impact	*p*-value	**Number of Schools**
Main Model	2.59	2.20	0.39*	0.03	**121**
Alternative Specifications					
Weights					
(1) Districts are weighted equally	2.50	2.15	0.35*	0.04	**121**
(2) Schools are weighted by the number of enrolled students	2.51	2.17	0.34*	0.04	**121**
Covariates					
(3) No covariates except randomization block indicators	2.45	2.20	0.25	0.17	**121**

Source: Educator administrative data.

Note: The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

*Impact is statistically significant at the .05 level, two-tailed test.

Table F.4. Impacts of Pay-for-Performance on Teachers' Classroom Observation Ratings in Year 4 Using Alternative Specifications, Cohort 1 (Points on 1-to-4 Scale)

Model	Teachers in Treatment Schools	Teachers in Control Schools	Impact	*p*-value	**Number of Teachers**	**Number of Schools**
Main Model	2.88	2.80	0.09*	0.02	**3,544**	**131**
Alternative Specifications						
Weights						
(1) Teachers are weighted equally	2.86	2.77	0.10*	0.01	**3,544**	**131**
(2) Districts are weighted equally	2.79	2.70	0.08*	0.03	**3,544**	**131**
Covariates						
(3) No covariates except randomization block indicators	2.88	2.80	0.08*	0.03	**3,544**	**131**

Source: Educator administrative data.

Note: The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

*Impact is statistically significant at the .05 level, two-tailed test.

Findings for Cohorts 1 and 2

In Tables F.5 and F.6, we present the impact of pay-for-performance on performance ratings of educators in schools in Cohorts 1 and 2 in the first three years of implementation, as well as the main impact estimates from Chapter VI, which only included educators in Cohort 1 schools. Unlike estimates based on Cohort 1 only, the estimated impacts of pay-for-performance on school achievement growth ratings and classroom achievement growth ratings in Year 1 were no longer statistically significant (*p*-values = 0.10 and 0.09) when both cohorts were included in the analysis (Table F.5). Impact estimates for Cohort 1 and for Cohorts 1 and 2 were also not significant for school achievement growth ratings in Years 2 and 3 and classroom achievement growth ratings in Year 2. For classroom achievement growth ratings in Year 3, pay-for-performance had no impact on these ratings when based on Cohort 1 only, but it had a *negative* impact when based on both cohorts. For teachers' and principals' observation ratings, none of the estimated impacts based on either Cohort 1 only or Cohorts 1 and 2 were statistically significant, with one exception. In Year 3, pay-for-performance had a statistically significant impact on teachers' observation ratings in Cohort 1 only, but not when both cohorts were included in the analysis (Table F.6).

Findings for a Consistent Sample of Schools

The number of districts and schools included in the main impact estimates varied across years for two of the educator performance ratings: school achievement growth and classroom achievement growth. The analyses of school achievement growth ratings excluded one district in Year 1 and one district in Year 4, but they included these districts in the other years. The district that was excluded in Year 1 did not place school achievement growth ratings into performance categories or onto a numeric scale, and the district that was excluded in Year 4 did not provide school achievement growth ratings for control schools. The main impact estimates for classroom achievement growth ratings included the six districts that evaluated teachers on this measure in Years 1 and 2 and the seven districts (the original six plus one additional district) that evaluated teachers on this measure in Years 3 and 4. Moreover, one of the six districts that evaluated teachers

on classroom achievement growth in all four years changed its measure over the course of the grant: it expanded the number of teachers who were evaluated on classroom achievement growth in Year 2 and then, in Year 3, changed the type of measure it used (from a value-added model to a student learning objective). Thus, differences in the main impact estimates across years could reflect the combination of changes in the impacts over time, changes in the composition of districts and schools included in the analyses, and changes in the particular measures used to evaluate educators.

Table F.5. Student Achievement Growth Ratings in Years 1 through 3, Cohorts 1 and 2 (Points on 1-to-4 Scale)

	Treatment	Control	Impact	*p*-value	**Number of Teachers**	**Number of Schools**
Year 1, Cohort 1						
School Achievement Growth Ratings	2.62	2.25	0.36*	0.04	NA	**123**
Classroom Achievement Growth Ratings	2.26	2.08	0.18*	0.04	**1,076**	**72**
Year 1, Cohorts 1 and 2						
School Achievement Growth Ratings	2.47	2.25	0.22	0.10	NA	**162**
Classroom Achievement Growth Ratings	2.37	2.27	0.10	0.09	**2,254**	**109**
Year 2, Cohort 1						
School Achievement Growth Ratings	2.55	2.27	0.27	0.07	NA	**130**
Classroom Achievement Growth Ratings	2.23	2.17	0.06	0.25	**1,332**	**72**
Year 2, Cohorts 1 and 2						
School Achievement Growth Ratings	2.76	2.60	0.16	0.21	NA	**169**
Classroom Achievement Growth Ratings	2.24	2.30	-0.05	0.29	**2,619**	**111**
Year 3, Cohort 1						
School Achievement Growth Ratings	2.41	2.37	0.04	0.75	NA	**131**
Classroom Achievement Growth Ratings	2.53	2.53	0.00	0.97	**2,040**	**90**
Year 3, Cohorts 1 and 2						
School Achievement Growth Ratings	2.65	2.67	-0.02	0.89	NA	**170**
Classroom Achievement Growth Ratings	2.50	2.59	-0.09*	0.04	**3,359**	**129**

Source: Educator administrative data.

Notes: School achievement growth ratings for one district in Year 1 are omitted because they did not place educators into performance categories or onto a numeric scale. Classroom achievement growth ratings are available in Years 1 and 2 only for the six districts in Cohort 1 and three districts in Cohort 2 that evaluated teachers based on classroom achievement growth in these years. Classroom achievement growth ratings are available in Year 3 only for the seven districts in Cohort 1 and three districts in Cohort 2 that evaluated teachers based on classroom achievement growth in this year. The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

NA is not applicable.

*Impact is statistically significant at the .05 level, two-tailed test.

Table F.6. Observation Ratings for Teachers and Principals in Years 1 through 3, Cohorts 1 and 2 (Points on 1-to-4 Scale)

	Treatment	Control	Impact	*p*-value	**Number of Educators**	**Number of Schools**
Year 1, Cohort 1						
Teachers' Classroom Observation Ratings	2.94	2.91	0.03	0.21	**3,599**	**131**
Observation Ratings for Principals	3.08	3.18	-0.10	0.20	**105**	**105**
Year 1, Cohorts 1 and 2						
Teachers' Classroom Observation Ratings	3.00	2.98	0.02	0.42	**4,900**	**170**
Observation Ratings for Principals	3.05	3.13	-0.08	0.24	**143**	**143**
Year 2, Cohort 1						
Teachers' Classroom Observation Ratings	2.98	2.93	0.05	0.08	**3,598**	**131**
Observation Ratings for Principals	3.14	3.01	0.13	0.19	**118**	**117**
Year 2, Cohorts 1 and 2						
Teachers' Classroom Observation Ratings	3.09	3.09	0.00	0.99	**4,939**	**170**
Observation Ratings for Principals	3.29	3.19	0.11	0.17	**155**	**154**
Year 3, Cohort 1						
Teachers' Classroom Observation Ratings	2.97	2.91	0.05*	0.02	**3,622**	**131**
Observation Ratings for Principals	3.37	3.32	0.05	0.52	**119**	**118**
Year 3, Cohorts 1 and 2						
Teachers' Classroom Observation Ratings	3.09	3.06	0.03	0.15	**5,000**	**170**
Observation Ratings for Principals	3.40	3.36	0.04	0.47	**158**	**157**

Source: Educator administrative data.

Notes: One district did not provide observation ratings for principals in Year 1. The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

*Impact is statistically significant at the .05 level, two-tailed test.

In Table F.7, we present the impacts of pay-for-performance on school and classroom achievement growth ratings based on consistent samples of districts and schools across years. For school achievement growth ratings, the consistent sample consisted of districts and schools for which ratings were available in all four years. For classroom achievement growth ratings, the consistent sample consisted of districts and schools that evaluated teachers on this measure in all four years and did not change the type of measure used. Therefore, in this table, differences in the impact estimates across years reflect only changes in impacts over time. Findings based on this consistent sample of schools and measures were generally similar to the main impact findings. However, in the consistent sample, the impacts on school achievement growth ratings and classroom achievement growth ratings in Year 1

were no longer significant (*p*-values = 0.08 and 0.16). The similarity of the main estimates to those based on a consistent sample of schools suggests that changes in the samples of districts and schools and changes in the measures included in the analyses were not primarily responsible for the evolution of the impact findings over time.

Table F.7. Student Achievement Growth Ratings, Consistent Sample of Schools, Cohort 1 (Points on 1-to-4 Scale)

Performance Measure and Year	Treatment	Control	Impact	*p*-value	Number of Teachers	Number of Schools
School Achievement Growth (Based on 8 districts)[a]						
Ratings in Year 1	2.62	2.29	0.33	0.08	**NA**	**114**
Ratings in Year 2	2.47	2.26	0.21	0.16	**NA**	**114**
Ratings in Year 3	2.41	2.40	0.01	0.93	**NA**	**114**
Ratings in Year 4	2.60	2.24	0.36*	0.05	**NA**	**114**
Classroom Achievement Growth (Based on 5 districts)[b]						
Ratings in Year 1	2.22	2.10	0.12	0.16	**974**	**63**
Ratings in Year 2	2.08	2.06	0.02	0.65	**1,066**	**63**
Ratings in Year 3	2.24	2.24	0.00	0.94	**1,320**	**63**
Ratings in Year 4	2.37	2.08	0.29*	0.00	**1,251**	**63**

Source: Educator administrative data.

Note: The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]For all four years, these analyses exclude one district in which school achievement growth ratings did not place educators into performance categories or onto a numeric scale in Year 1, and one other district that did not provide school achievement growth ratings for control schools in Year 4.

[b]For all four years, these analyses include only the five districts that evaluated teachers based on the same measure of classroom achievement growth in all years.

NA is not applicable.

*Impact is statistically significant at the .05 level, two-tailed test.

Educator Retention Rates

Educator retention rates—the percentages of educators who stayed in their schools between years—provide important context for analyzing whether pay-for-performance had different impacts on the effectiveness of returning or newly hired educators. The extent of educator turnover at a school determines how much scope there is for impacts on each group to shape the overall effectiveness of the school's staff. For example, if at the end of each year few teachers departed and were replaced by newly hired teachers, then even large differences between treatment and control schools in the effectiveness of newly hired teachers would make little contribution to the overall impacts of pay-for-performance on educator effectiveness.

We measured retention for all full-time teachers and principals working in study schools in Year 1. Educators were considered retained if they returned to the same school and position (teacher or

principal) in the fall of Year 2 (one-year retention), fall of Year 3 (two-year retention), or fall of Year 4 (three-year retention). We also measured one-year retention for all full-time educators working in study schools in Years 2 and 3, and we measured two-year retention for all full-time educators working in study schools in Year 3. Differences in retention rates between treatment and control schools measured the impact of pay-for-performance on educator retention.

In the study schools, about 20 to 30 percent of teachers departed between consecutive years, and 30 to 40 percent of teachers departed over a two-year period (Table F.8). After a three-year period, about half of the teachers working in study schools had departed. Likewise, about 20 to 30 percent of principals departed between consecutive years, and 40 percent of principals departed over a two-year period (Table F.9). After a three-year period, about 60 percent of principals working in study schools had departed. Therefore, although many educators were retained, there was also plenty of turnover. This means that treatment-control differences in the effectiveness of both returning and newly hired educators could contribute to shaping the impacts of pay-for-performance on overall educator effectiveness.

We also found that pay-for-performance had a small, positive impact on the overall retention rates of teachers, but not principals. Among teachers working in study schools in Years 1, 2, or 3, those in treatment schools were three percentage points more likely to return to their schools in Year 4 than those in control schools (Table F.8).

Table F.8. Teachers Who Continued Teaching in the Same School Across Multiple Years, Cohort 1 (Percentages)

Period	Treatment	Control	Impact	*p*-value	**Number of Teachers**	**Number of Schools**
One-Year Period						
Between Years 1 and 2	83	81	2	0.17	**4,303**	**131**
Between Years 2 and 3	78	77	1	0.39	**4,402**	**131**
Between Years 3 and 4	74	70	3*	0.03	**4,521**	**131**
Two-Year Period						
Between Years 1 and 3	66	64	2	0.12	**4,303**	**131**
Between Years 2 and 4	61	57	3*	0.01	**4,402**	**131**
Three-Year Period						
Between Years 1 and 4	51	49	3*	0.03	**4,303**	**131**

Source: Educator administrative data.

Note: The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

*Impact is statistically significant at the .05 level, two-tailed test.

Impacts of Pay-for-Performance on Other Characteristics of Schools' Staff

Given that pay-for-performance was intended to help schools retain and attract more effective educators, any staffing changes resulting from pay-for-performance could have also altered other characteristics of the schools' staff, including the demographic and professional characteristics of teachers and principals. However, we found little evidence that pay-for-performance led to changes in those staff characteristics. In Year 4, educators working in treatment and control schools had similar demographic characteristics and professional backgrounds, with one exception: teachers in treatment schools were less likely to be Hispanic or other race than those in control schools (Table F.10). Likewise, in Year 2 (Chiang et al. 2015) and Year 3 (Wellington et al. 2016), educators working in

treatment and control schools also had similar demographic characteristics and professional backgrounds.

Table F.9. Principals Who Continued Leading the Same School Across Multiple Years, Cohort 1 (Percentages)

Period	Treatment	Control	Impact	*p*-value	**Number of Principals**	**Number of Schools**
One-Year Period						
Between Years 1 and 2	80	74	6	0.42	**133**	**127**
Between Years 2 and 3	78	79	-2	0.84	**137**	**128**
Between Years 3 and 4	71	75	-4	0.49	**132**	**127**
Two-Year Period						
Between Years 1 and 3	64	59	5	0.63	**133**	**127**
Between Years 2 and 4	57	58	-1	0.92	**137**	**128**
Three-Year Period						
Between Years 1 and 4	41	44	-3	0.80	**133**	**127**

Source: Educator administrative data.

Notes: The difference between treatment and control estimates may not equal the impact shown in the table because of rounding. None of the impacts are statistically significant at the .05 level, two-tailed test.

Table F.10. Characteristics of Teachers and Principals in Year 4, Cohort 1 (Percentages Unless Otherwise Noted)

	Teachers			Principals		
	Treatment	Control	Difference	Treatment	Control	Difference
Demographic Characteristics						
Female	86	85	0	61	68	-7
Race or ethnicity						
White, non-Hispanic	74	72	2	62	51	11
Black, non-Hispanic	20	20	-1	32	38	-6
Hispanic or other	6	8	-2*	6	11	-5
Age (average years)	41	42	0	47	49	-2
Education						
Master's degree or higher	47	49	-1	93	93	0
Experience in K–12 Education						
Total experience (average years)	10	10	0	18	16	2
Fewer than 5 years	32	36	-4	12	15	-3
5–15 years	43	41	2	32	40	-7
More than 15 years	25	24	1	56	45	11
Number of Educators—Range[a]	**1,648-2,081**	**1,713-2,129**		**48-64**	**46-63**	
Number of Schools—Range[a]	**52-65**	**53-66**		**48-64**	**46-63**	

Source: Educator administrative data.

Note: The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

[a]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference is statistically significant at the .05 level, two-tailed test.

Impacts of Pay-for-Performance on the Effectiveness of Returning and Newly Hired Teachers and Principals

In Chapter VI, we examined the impacts of pay-for-performance on the performance ratings of returning and newly hired teachers (Table VI.3). In those main analyses, we classified returning teachers as those who had stayed at their school since the previous year and newly hired teachers as those who were new to their school in the current year. Findings were similar in analyses that classified returning teachers as those who had stayed at their school since Year 1 and newly hired teachers as those who entered the school in any year after Year 1 (Table F.11).

Table F.11. Performance Ratings in Years 3 and 4 for Teachers Who Stayed at Their School from Year 1 and Teachers Who Were Hired at Their School After Year 1, Cohort 1 (Points on 1-to-4 Scale)

	Teachers Who Stayed from Year 1		Teachers Who Were Hired After Year 1				
Performance Measure	Impact	*p*-value	Impact	*p*-value	**Number of Returning Teachers**	**Number of Newly Hired Teachers**	**Number of Schools**
Year 3							
Classroom Observation Rating	0.07*	0.03	0.01	0.76	**2,304**	**1,318**	**131**
Classroom Achievement Growth Rating	0.05	0.42	-0.09	0.24	**1,272**	**768**	**90**
Year 4							
Classroom Observation Rating	0.09	0.08	0.07	0.13	**1,746**	**1,798**	**131**
Classroom Achievement Growth Rating	0.20*	0.01	0.14	0.06	**971**	**986**	**90**

Source: Educator administrative data.

*Impact is statistically significant at the .05 level, two-tailed test.

We also examined the impacts of pay-for-performance on the performance ratings of returning and newly hired principals. We found no impacts of pay-for-performance on the observation ratings of returning principals, regardless of whether they were defined as those who stayed at their school since the previous year (Table F.12) or since Year 1 (Table F.13). Likewise, pay-for-performance did not affect the school achievement growth ratings of returning principals in Year 3. In Year 4, pay-for-performance led to higher school achievement growth ratings earned by returning principals who had stayed at their school since the previous year (Table F.12); the impact on returning principals who had stayed at their school since Year 1 was similar in magnitude but not statistically significant (*p*-value = 0.11, Table F.13). We found no impacts of pay-for-performance on the performance ratings of newly hired principals, regardless of whether they were defined as those who entered their school in the current year (Table F.12) or any year after Year 1 (Table F.13).

Table F.12. Observation and School Achievement Growth Ratings of Returning and Newly Hired Principals, Cohort 1 (Points on 1-to-4 Scale)

	Returning Principals		Newly Hired Principals				
Performance Measure and Year	Impact	*p*-value	Impact	*p*-value	**Number of Returning Principals**	**Number of Newly Hired Principals**	**Number of Schools**
Year 2							
Observation Ratings	0.08	0.53	0.42	0.22	**99**	**19**	**117**
School Achievement Growth Ratings	0.24	0.21	0.42	0.46	**110**	**26**	**127**
Year 3							
Observation Ratings	0.06	0.49	-0.09	0.69	**96**	**23**	**118**
School Achievement Growth Ratings	0.15	0.35	-0.56	0.21	**108**	**24**	**127**
Year 4							
Observation Ratings	0.08	0.36	-0.11	0.73	**100**	**22**	**122**
School Achievement Growth Ratings[a]	0.41*	0.04	0.26	0.39	**96**	**22**	**118**

Source: Educator administrative data.

Notes: Returning principals were those who had stayed at their school since the previous school year, and newly hired principals were those who were new to their school in the current year. For example, in Year 3, returning principals were those who had stayed at their school between Years 2 and 3, and newly hired principals were those who were new to their school in Year 3.

[a]One district did not provide school achievement growth ratings for control schools in Year 4, so all schools in the district were excluded from this analysis.

*Impact is statistically significant at the .05 level, two-tailed test.

Student Achievement

This section presents four types of additional analyses of the impacts of pay-for-performance on student achievement: (1) sensitivity analyses that assess the robustness of the main impact estimates; (2) findings that include both Cohorts 1 and 2; (3) subgroup analyses that assess impacts within elementary and middle grades separately; and (4) analyses of the impacts of students' exposure to pay-for-performance, based on students who were in a study school at the time of random assignment.

Sensitivity Analyses

We explored the sensitivity of the main impact estimates after four years of implementation to several alternative ways of specifying the regression model or estimation sample (Tables F.14 and F.15). Findings from these specifications were generally similar to the main impact estimates, with some exceptions described below.

Table F.13. Performance Ratings in Years 3 and 4 for Principals Who Stayed at Their School from Year 1 and Principals Who Were Hired at Their School After Year 1, Cohort 1 (Points on 1-to-4 Scale)

	Principals Who Stayed from Year 1		Principals Who Were Hired After Year 1				
Performance Measure	Impact	*p*-value	Impact	*p*-value	**Number of Returning Principals**	**Number of New Principals**	**Number of Schools**
Year 3							
Observation Rating	0.05	0.66	0.03	0.81	**76**	**43**	**118**
School Achievement Growth Rating	0.17	0.39	-0.21	0.39	**83**	**49**	**127**
Year 4							
Observation Rating	0.18	0.13	-0.08	0.56	**54**	**68**	**122**
School Achievement Growth Rating[a]	0.54	0.11	0.23	0.39	**55**	**63**	**118**

Source: Educator administrative data.

Notes: None of the impacts are statistically significant at the .05 level, two-tailed test.

[a]One district did not provide school achievement growth ratings for control schools in Year 4, so all schools in the district were excluded from this analysis.

Standardizing test scores. For the main analysis, we standardized outcome and baseline test scores into z-scores based on grade-specific means and standard deviations of test scores in each statewide population (for more details, see Appendix B). We explored an alternative method of standardizing test scores into z-scores based on the grade-specific means and standard deviations of test scores for students in control schools in the same state. Findings from these specifications were similar to the main impact estimates (Tables F.14 and F.15, model 1).

Using alternative weighting approaches. In our main specification, we normalized the analysis weights so that each school received the same weight in the final analysis sample. Therefore, in the main impact estimates, students in large schools received less weight than those in small schools, and districts with more schools received more weight than those with fewer schools. We explored two alternative approaches to normalizing sample weights. In the first alternative approach, each district received the same weight. This approach produced estimates of the impact of pay-for-performance in the average Cohort 1 district, which could be of interest because each district designed its TIF program in a different way. In the second alternative approach, each student received the same weight. This approach produced estimates of the impact of pay-for-performance on the average student, which could be of interest because pay-for-performance was ultimately intended to improve student outcomes. Findings from these models were similar in magnitude to the main impact estimates, though when students were weighted equally the impacts of pay-for-performance on student achievement in math and reading were statistically significant (Tables F.14 and F.15, models 2 and 3).

Table F.14. Impacts of Pay-for-Performance on Student Math Achievement After Four Years of Implementation, Alternative Specifications, Cohort 1 (Student *z*-Score Units)

	Impact	*p*-value	Number of Students	Number of Schools
Main Model	0.04	0.13	**38,939**	**131**
Alternative Specifications				
Standardizing Test Scores				
(1) Compute *z*-scores using sample means and standard deviations	0.04	0.11	**38,939**	**131**
Weights				
(2) Students weighted equally	0.04*	0.05	**38,939**	**131**
(3) Districts weighted equally	0.03	0.30	**38,939**	**131**
Covariates				
(4) No covariates except randomization block indicators	0.01	0.69	**38,939**	**131**
(5) Only covariates are school-level pre-implementation means of student achievement, student race or ethnicity, and randomization block indicators	0.06*	0.04	**38,939**	**131**
(6) All covariates interacted with state indicators	0.03	0.13	**38,939**	**131**
(7) Include student pretests interacted with grade indicators	0.04	0.13	**38,939**	**131**
(8) Include student pretests, squared and cubed	0.04	0.12	**38,939**	**131**

Source: Student administrative data.

*Impact is statistically significant at the .05 level, two-tailed test.

Table F.15. Impacts of Pay-for-Performance on Student Reading Achievement After Four Years of Implementation, Alternative Specifications, Cohort 1 (Student *z*-Score Units)

	Impact	*p*-value	Number of Students	Number of Schools
Main Model	0.04	0.08	**38,929**	**131**
Alternative Specifications				
Standardizing Test Scores				
(1) Compute *z*-scores using sample means and standard deviations	0.03	0.12	**38,929**	**131**
Weights				
(2) Students weighted equally	0.04*	0.03	**38,929**	**131**
(3) Districts weighted equally	0.03	0.32	**38,929**	**131**
Covariates				
(4) No covariates except randomization block indicators	0.00	0.88	**38,929**	**131**
(5) Only covariates are school-level pre-implementation means of student achievement, student race or ethnicity, and randomization block indicators	0.04	0.07	**38,929**	**131**
(6) All covariates interacted with state indicators	0.02	0.36	**38,929**	**131**
(7) Include student pretests interacted with grade indicators	0.04	0.08	**38,929**	**131**
(8) Include student pretests, squared and cubed	0.04	0.08	**38,929**	**131**

Source: Student administrative data.

*Impact is statistically significant at the .05 level, two-tailed test.

Changing covariates. Our main estimation model controlled for randomization block indicators and the student- and school-level covariates described in Appendix B. To assess the sensitivity of the estimates to the choice of covariates or the method of controlling for pretest scores, we estimated several alternative models.

First, we omitted all covariates except the randomization block indicators (Tables F.14 and F.15, model 4). Unlike this alternative model, the main model controlled for schools' pre-implementation characteristics (along with student-level covariates) to account for treatment schools having slightly lower student math achievement and slightly different student racial/ethnic composition than control schools at the beginning of the study. Failure to account for these preexisting differences could generate an inaccurate estimate of the effects of pay-for-performance. Nevertheless, because some researchers have expressed methodological concerns about the use of covariates in analyzing experimental data (Freedman 2008), we estimated this alternative model, dropping all covariates aside from the randomization block indicators. As expected, when we did not account for preexisting differences between treatment and control schools, the alternative estimates differed from our main findings. In both our main model and this alternative model we found that the impacts of four years of pay-for-performance on student achievement in math and reading were not statistically significant. However, the estimated impact was smaller in this alternative model than in the main model in both math (0.01 versus 0.04 standard deviations) and reading (0.00 versus 0.04 standard deviations).

Second, we omitted student-level covariates—those measuring the individual characteristics of students in the analysis sample—but included randomization block indicators and school-level pre-implementation means of student achievement and student race/ethnicity (Tables F.14 and F.15, model 5). Because pay-for-performance could have affected families' decisions on where to enroll their children and, thus, the characteristics of a school's student population, omitting student-level covariates could avoid biases from controlling for factors that might have been influenced by pay-for-performance. In math, this model produced an impact estimate that was slightly larger than that produced by our main model (0.06 versus 0.04) and was statistically significant, unlike our main estimate. The results for reading were similar between our main model and this alternative model.

We also explored models that permitted more flexible functional forms for the covariates. These models differed from the main model in that they (1) added interactions between all covariates in the main estimation model and state indicators, (2) added interactions between the student pretest scores and grade indicators, or (3) included a cubic polynomial of student pretests. The findings from these models were, in general, similar to the main impact estimates (Tables F.14 and F.15, models 6 through 8).

Findings for Cohorts 1 and 2

In Table F.16, we present the impact of pay-for-performance on math and reading achievement in Years 1, 2, and 3 for Cohorts 1 and 2, as well as the main impact estimates from Chapter VI, which included Cohort 1 only. After one year of pay-for-performance, the impact on student reading achievement for both cohorts combined was similar in magnitude to the impact for Cohort 1 only, but was no longer statistically significant. The impact in math after one year was not significant in either sample. After two years of pay-for-performance, results that included both cohorts were similar to those that included Cohort 1 only, though the impact in math was significant only when both cohorts were included. After three years, impacts in math and reading based on both cohorts were smaller than those based on Cohort 1 only, and were no longer significant.

Table F.16. Student Achievement in Math and Reading, Cohorts 1 and 2 (Student *z*-Score Units)

Year, Cohort and Subject	Treatment	Control	Impact	*p*-value	**Number of Students**	**Number of Schools**
Year 1, Cohort 1						
Math	-0.43	-0.45	0.02	0.33	**40,535**	**131**
Reading	-0.37	-0.40	0.03*	0.04	**40,256**	**131**
Year 1, Cohorts 1 and 2						
Math	-0.53	-0.55	0.01	0.41	**52,958**	**170**
Reading	-0.48	-0.50	0.02	0.14	**52,475**	**170**
Year 2, Cohort 1						
Math	-0.38	-0.43	0.04	0.07	**40,454**	**131**
Reading	-0.36	-0.39	0.03*	0.02	**40,122**	**131**
Year 2, Cohorts 1 and 2						
Math	-0.49	-0.53	0.04*	0.04	**51,860**	**170**
Reading	-0.46	-0.49	0.03*	0.02	**51,645**	**170**
Year 3, Cohort 1						
Math	-0.36	-0.42	0.06*	0.02	**39,770**	**131**
Reading	-0.33	-0.37	0.04*	0.02	**39,538**	**131**
Year 3, Cohorts 1 and 2						
Math	-0.49	-0.52	0.03	0.10	**49,609**	**170**
Reading	-0.47	-0.49	0.02	0.18	**49,952**	**170**

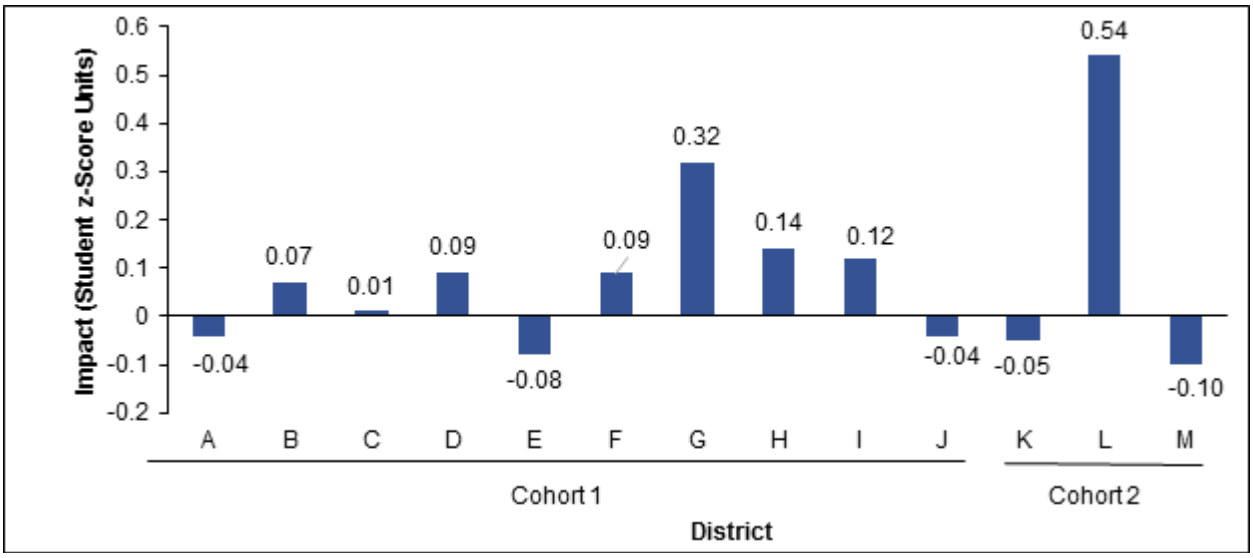
Source: Student administrative data.

Note: The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

*Impact is statistically significant at the .05 level, two-tailed test.

Figures F.1 and F.2 show the impacts of three years of pay-for-performance on student achievement in math and reading, respectively, in each of the 13 districts in Cohorts 1 and 2. Similar to the findings after four years of pay-for-performance for Cohort 1 (Figures VI.2 and VI.3), these figures illustrate that impacts after three years also varied across all 13 districts.

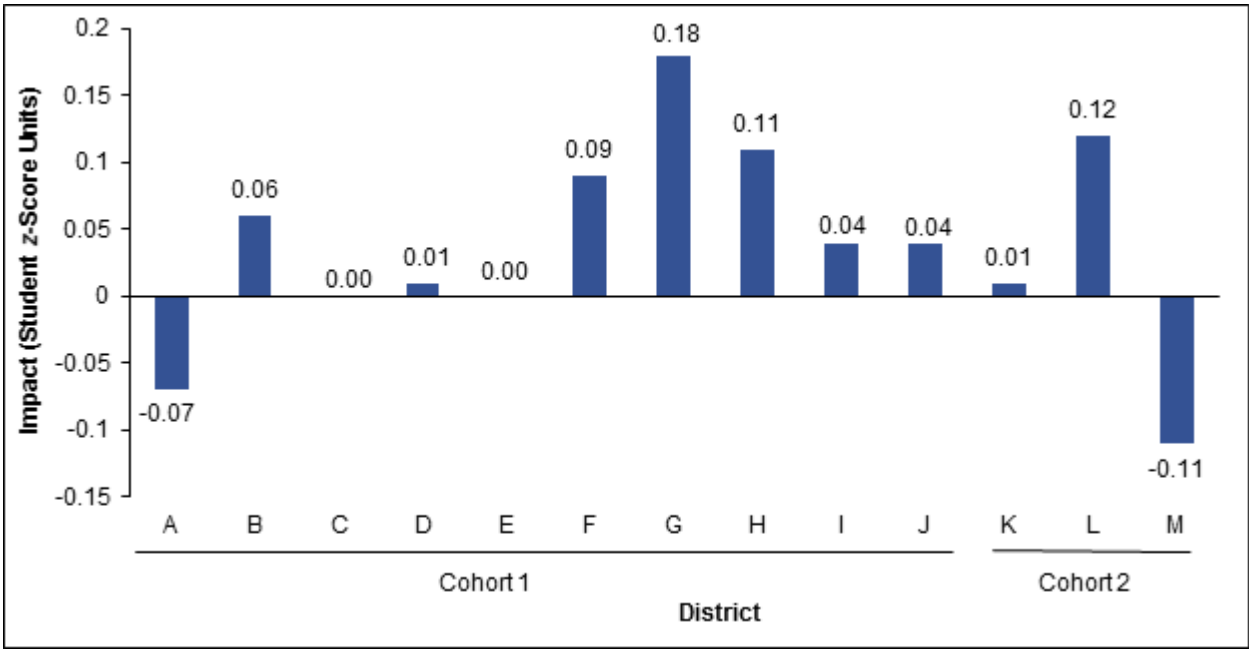
Figure F.1. Impact of Pay-for-Performance on Student Math Achievement After Three Years of Implementation, by District, Cohorts 1 and 2 (Student *z*-Score Units)



Source: Student administrative data (N = 49,609).

Note: An F-test of the null hypothesis that impacts are equal across districts has a *p*-value of less than 0.01.

Figure F.2. Impact of Pay-for-Performance on Student Reading Achievement After Three Years of Implementation, by District, Cohorts 1 and 2 (Student *z*-Score Units)



Source: Student administrative data (N = 49,952).

Note: An F-test of the null hypothesis that impacts are equal across districts has a *p*-value of less than 0.01.

Subgroup Findings for Elementary and Middle School Grades

In Table F.17, we present the impacts of pay-for-performance on student achievement separately within elementary school grades (grades 3 through 5) and middle school grades (grades 6 through 8). Across most years and grade spans, impacts on math achievement were small and positive, but not statistically significant. The one exception is that the impact of three years of pay-for-performance on math achievement was statistically significant. In reading, impacts were likewise small and positive across years and grade spans, and were also statistically significant for middle school grades after one, three, and four years of pay-for-performance. However, impacts on reading achievement in the elementary school grades were statistically significant in only one out of the four years.

In most years and subjects, impacts in the middle school grades were larger than in the elementary school grades, but those differences in impacts were generally not statistically significant. The one exception is that pay-for-performance had a significantly larger impact on reading achievement in the middle school grades than in the elementary school grades after four years.

Findings on the Impacts of Student Exposure to Pay-for-Performance

The main impacts on student achievement that we reported in Chapter VI measured the impacts of pay-for-performance on schools' average test scores after one to four years of implementation. We focused on those impacts because they were the only types of impacts that we could measure based on the available data for all four years of the study. However, there are two complications to interpreting the impacts on schools' average test scores. First, pay-for-performance could have affected schools' average test scores either by changing how much individual students learned, or by changing the types of students who enrolled in treatment schools compared to control schools. Second, impacts on schools' average test scores after two, three, and four years of implementation reflected a mix of student exposure to pay-for-performance because students who were tested at the end of each year included those who had been enrolled at the same school since the beginning of the study and those who had joined in the intervening years. With each successive year, fewer students were exposed to pay-for-performance for the full duration of its implementation (see Appendix B for details on the distribution of student exposure to pay-for-performance).

To explore an alternative type of impact whose interpretation did not have these complications, we conducted an additional analysis examining impacts on the achievement of students who had the same duration of exposure to pay-for-performance and for whom any impacts would reflect only changes in learning. To do so, our analysis used only students who were enrolled in a study school in the pre-implementation year—the year in which schools were randomly assigned to the treatment and control groups. We refer to this sample as the randomization sample, divided between treatment students (those who were enrolled in treatment schools in the pre-implementation year) and control students (those who were enrolled in control schools in the pre-implementation year). By excluding students outside the randomization sample, we removed any differences in outcomes between students in treatment and control schools resulting from different types of students having joined the two groups of schools after random assignment.

Table F.17. Student Achievement in Math and Reading in Elementary and Middle Grades, Cohort 1 (Student z-Score Units)

	Math				Reading			
Year and Grades	Treatment	Control	Impact	*p*-value	Treatment	Control	Impact	*p*-value
Year 1								
(1) Grades 3–5	-0.45	-0.45	0.00	0.92	-0.40	-0.42	0.02	0.29
(2) Grades 6–8	-0.40	-0.45	0.05	0.08	-0.31	-0.37	0.06*	0.02
Difference, (1) - (2)			-0.05	0.08			-0.04	0.17
Number of Students	**20,213**	**20,322**			**20,028**	**20,228**		
Number of Schools	**65**	**66**			**65**	**66**		
Year 2								
(1) Grades 3–5	-0.41	-0.45	0.04	0.13	-0.38	-0.42	0.03*	0.04
(2) Grades 6–8	-0.34	-0.39	0.05	0.11	-0.31	-0.34	0.03	0.22
Difference, (1) - (2)			-0.01	0.77			0.01	0.86
Number of Students	**19,997**	**20,457**			**19,763**	**20,359**		
Number of Schools	**65**	**66**			**65**	**66**		
Year 3								
(1) Grades 3–5	-0.39	-0.43	0.05	0.11	-0.37	-0.39	0.03	0.21
(2) Grades 6–8	-0.32	-0.40	0.08*	0.02	-0.26	-0.32	0.06*	0.02
Difference, (1) - (2)			-0.04	0.38			-0.03	0.32
Number of Students	**19,759**	**20,011**			**19,611**	**19,927**		
Number of Schools	**65**	**66**			**65**	**66**		
Year 4								
(1) Grades 3–5	-0.35	-0.37	0.02	0.38	-0.38	-0.39	0.01	0.57
(2) Grades 6–8	-0.31	-0.37	0.06	0.06	-0.26	-0.35	0.08*	0.00
Difference, (1) - (2)			-0.04	0.31			-0.07*	0.04
Number of Students	**19,160**	**19,779**			**19,159**	**19,770**		
Number of Schools	**65**	**66**			**65**	**66**		

Source: Student administrative data.

Note: The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

*Impact is statistically significant at the .05 level, two-tailed test.

Furthermore, when examining the impact of each duration of exposure to pay-for-performance, we limited the randomization sample to students who were expected, under normal grade progression, to stay at the same school for that duration based on the school's grade span. For example, students who were enrolled as fourth graders within a K–5 treatment school in the pre-implementation year were expected to stay in the same school in Year 1, but fifth graders were not. Because of this restriction, all treatment students had one year of expected exposure to pay-for-performance at the end of Year 1, and two years of expected exposure at the end of Year 2. This restriction led to different randomization samples for the Year 1 analysis (students who were expected to stay at the same school for at least one year after random assignment) and the Year 2 analysis (students who were expected to stay at the same school for at least two years after random assignment). As discussed in Appendix B, we had insufficient data to identify many students who were expected to stay for three or four years, so we could not estimate the impacts of those durations of exposure.

Among students in each randomization sample—those who were expected to stay at their schools through Year 1 or Year 2—our final analysis sample consisted of students who were actually tested in that year within any study school. Defining the analysis sample required dealing with two issues. First, some students in each randomization sample were not included in the final analysis sample because they exited the study schools—for example, they moved to a nonstudy school or out of the district—even though they had been expected to stay based on their grade level. This type of departure from the study is known as attrition. Second, some students stayed in the study schools—and were therefore included in the analysis sample—but switched from a treatment school to a control school or vice versa. This type of movement is known as crossover. In what follows, we provide more details on the extent of attrition and crossover and our approaches to dealing with these issues.

Attrition. Attrition could pose a potential threat to interpreting differences in outcomes between treatment and control students as reflecting the impacts of pay-for-performance on student learning. For example, if the students who left treatment schools differed from those who left control schools in unmeasured ways, then the remaining treatment and control students—those in the final analysis sample—would also differ on unmeasured characteristics, giving rise to differences in outcomes that did not reflect pay-for-performance having changed student learning.

However, we found that attrition rates were similar among treatment and control students, suggesting that attrition may not have led to differences in unmeasured characteristics between the remaining students in the two groups. For example, among students in the randomization sample who were expected to stay at their schools through Year 1, 22 percent of treatment students and 24 percent of control students did not have math test scores in Year 1 (Table F.18). Among students in the randomization sample who were expected to stay at their schools through Year 2, 35 percent of treatment students and 34 percent of control students did not have math test scores in Year 2. Attrition rates were generally similar for the analysis of impacts on reading achievement. Although in some districts we could conceivably have reduced attrition further by including the test scores of students who left the study schools but remained in the same district, not all districts provided such data. In light of the similar attrition rates among treatment and control students, we adopted a consistent approach across all districts, defining the final analysis sample based on students who remained in the study schools.

Table F.18. Attrition Among Students in Study Schools at Random Assignment, Cohort 1

	Original Number of Students who Were Randomly Assigned[a]		Number of Students who Left the Study		Percentage of Students Who Left the Study (attrition rate)	
Outcome Year and Subject	Treatment	Control	Treatment	Control	Treatment	Control
Year 1						
Math	14,877	14,930	3,316	3,515	22	24
Reading	14,825	14,911	3,376	3,569	23	24
Year 2						
Math	8,747	8,932	3,026	3,057	35	34
Reading	8,705	8,918	3,079	3,085	35	35

Source: Student administrative data.

Note: Among clusters in the randomization sample (schools or groups of schools that were randomly assigned to the treatment or control group), every cluster had at least one student who was included in the analysis sample. In other words, there was no cluster attrition.

[a]Numbers of students are based on students who were enrolled in the pre-implementation year and who were expected, under normal grade progression, to stay in the same school through the end of the specified outcome year.

Crossover. Among students who were originally in control schools at the time of random assignment, those who moved to treatment schools could have done so because of pay-for-performance—for example, because their families were aware of changes in teacher satisfaction due to pay-for-performance. Similarly, student movement from treatment schools to control schools could also be influenced by pay-for-performance. This type of movement could lead to differences in the characteristics of students enrolled at treatment and control schools—and, therefore, differences in outcomes—for reasons that do not reflect changes in student learning. For this reason, crossover also poses a threat to interpreting test-score differences between treatment and control schools as differences in learning.

To deal with crossover, we always classified students into the treatment or control group based on the school in which they were enrolled at the time of random assignment, regardless of any subsequent movement into other study schools. Because this classification remained fixed across years, students' decisions to move between study schools could not generate differences in outcomes between the treatment and control groups. By comparing the outcomes of students who were originally in treatment and control schools at the time of random assignment—an approach known as intent-to-treat estimation—the impacts we measured should only have reflected differences in student learning.

In practice, crossover was rare. In the analysis of Year 1 impacts, no more than 4 percent of students transferred to a school with the opposite status (treatment or control) to the one in which they had been enrolled at the time of random assignment (Table F.19). In the analysis of Year 2 impacts, no more than 10 percent of students spent any time in a school with the opposite status.

Although intent-to-treat estimates are faithful to the original random assignment design, they must be interpreted appropriately. As discussed earlier, Year 1 and Year 2 impacts reflected the effects of one and two years of expected exposure to pay-for-performance. Because of crossover,

students' expected exposure differed slightly from their actual exposure. Treatment students who transferred to a control school did not have full exposure to pay-for-performance; control students who transferred to a treatment school had some exposure to pay-for-performance. However, given that crossover was rare, expected and actual exposure were similar. In the results that follow, we present the impacts of expected exposure to pay-for-performance.

Table F.19. Distribution of Time Spent in Treatment and Control Schools After Random Assignment, Among Students Who Were Included in the Analysis of Impacts of Student Exposure to Pay-for-Performance (Percentages)

Outcome Year and Subject	In Treatment School at Random Assignment	In Control School at Random Assignment
Year 1		
Math		
Spent one year in treatment schools	96.3	2.9
Spent one year in control schools	3.7	97.1
Number of students	**11,561**	**11,415**
Reading		
Spent one year in treatment schools	96.3	2.9
Spent one year in control schools	3.7	97.1
Number of students	**11,449**	**11,342**
Year 2		
Math		
Spent two years in treatment schools	90.1	2.6
Spent one year each in treatment and control schools	6.5	5.6
Spent two years in control schools	3.4	91.8
Number of students	**5,721**	**5,875**
Reading		
Spent two years in treatment schools	89.9	2.5
Spent one year each in treatment and control schools	6.7	5.7
Spent two years in control schools	3.4	91.7
Number of students	**5,626**	**5,833**

Source: Student administrative data.

Note: For simplicity, a small number of students who were in a study school in Year 0 and Year 2, but not in a study school in Year 1, are counted as having spent one year each in treatment and control schools.

Findings. Similar to the main impacts of pay-for-performance on schools' average test scores, we found small, positive impacts of exposure to pay-for-performance on students' test scores. One year of exposure to pay-for-performance raised student achievement by 0.03 standard deviations in math and 0.04 standard deviations in reading, though only the impact in reading scores was statistically significant (Table F.20). Two years of exposure to pay-for-performance raised student achievement by 0.06 standard deviations in math and 0.05 standard deviations in reading.

Table F.20. Student Achievement in Math and Reading, by Years of Exposure to Pay-for-Performance, Based on Students Enrolled in Study Schools at Random Assignment, Cohort 1 (Student *z*-Score Units)

Duration of Exposure and Subject	Treatment	Control	Impact	*p*-value	**Number of Students**	**Number of Schools**
One Year						
Math	-0.39	-0.42	0.03	0.12	**22,976**	**129**
Reading	-0.33	-0.37	0.04*	0.05	**22,791**	**129**
Two Years						
Math	-0.31	-0.37	0.06*	0.02	**11,596**	**121**
Reading	-0.30	-0.35	0.05*	0.01	**11,459**	**121**

Source: Student administrative data.

Notes: The Year 1 and Year 2 analyses exclude two schools whose highest grade was 3rd grade at the time of random assignment. The Year 2 analyses exclude an additional eight schools whose lowest grade was 7th grade at the time of random assignment. The difference between treatment and control estimates may not equal the impact shown in the table because of rounding.

*Impact is statistically significant at the .05 level, two-tailed test.

APPENDIX G

SUPPLEMENTAL FINDINGS ON FACTORS ASSOCIATED WITH DIFFERENCES IN IMPACTS FOR CHAPTER VI

THIS PAGE IS INTENTIONALLY BLANK

This appendix supplements the information presented in Chapter VI examining district and school factors that were associated with the impacts of pay-for-performance on student achievement. We examine whether the characteristics of districts' TIF programs and their implementation, and the political and institutional context in which implementation took place, were associated with impacts on student achievement. We also examine whether schools that experienced greater impacts of pay-for-performance on measures of educators' strategic behavior, effort, and practices also experienced greater impacts on student achievement. Such a relationship would suggest that pay-for-performance improved student achievement by affecting those educator behaviors. Finally, we examine whether schools' average student achievement prior to pay-for-performance implementation was associated with the subsequent impacts of pay-for-performance on student achievement.

The information and analyses in this appendix pertain to the 10 evaluation districts, referred to as Cohort 1, whose schools were randomly assigned to the treatment or control group in spring and summer 2011. As discussed in Chapter II, Cohort 1 completed four years of TIF implementation—2011–2012, 2012–2013, 2013–2014, and 2014–2015—referred to as Years 1, 2, 3, and 4, respectively.

Explaining Differences in Impacts Across Districts

This section discusses the relationships between districts' program and contextual characteristics and impacts on student achievement. First, we provide details on each characteristic, including the way in which we measured it on a continuous scale (when appropriate) and the way in which we divided each continuous scale into two subgroups of districts that differed on the characteristic. Second, we compare the impacts of pay-for-performance on student achievement between subgroups based on program characteristics. Third, we report findings from a sensitivity analysis that examined the relationships between the continuous measures of program characteristics and impacts on student achievement. Fourth, we compare the impacts of pay-for-performance on student achievement between subgroups based on district contextual factors and report findings from a sensitivity analysis examining the relationship between one continuous measure of district context and impacts on student achievement.

Program and Contextual Characteristics Examined

We examined five district-level program characteristics and two contextual characteristics (Table G.1). Three of these program characteristics—the use of classroom achievement growth to measure teacher effectiveness and award bonuses, the size of the average bonus, and the amount of differentiation in bonuses—pertain to how the programs were designed. Two of the program characteristics—the timing of awarding bonuses based on the prior year and teachers' understanding of their pay-for-performance eligibility—relate to how the programs were implemented. The two contextual characteristics examined were district size and whether the district was in a state with right-to-work laws.

Table G.1. District-level Characteristics Used for Subgroup Analyses

Characteristic	Reason for Examining This Characteristic	Subgroup Definition	Number of Districts in Subgroup
Program Characteristics (Year 4 Analysis)			
Use of Classroom Achievement Growth to Measure Teacher Effectiveness and Award Bonuses[a]	Measure may give teachers a greater incentive to improve their students' performance on the same outcomes that this study measured.	Districts used classroom achievement growth measures to award performance bonuses.	7
Size of Average Bonus[b]	Teachers may pay more attention to bonuses that are larger on average.	Districts had a larger average bonus if the average bonus in Year 3 was at least 4.5 percent of average teacher salary.	4
Amount of Differentiation in Bonuses[b]	More differentiation implies a larger monetary gain from performing well on the performance ratings.	Districts had a larger amount of differentiation if the standard deviation of bonuses in Year 3 was at least 4 percent of average teacher salary.	5
Timing of Awarding Bonuses[a]	Early awarding of prior-year bonuses allows more time for teachers to revise their teaching practices for the current year.	Districts carried out earlier awarding of bonuses from Year 3 if they awarded at least one component of the bonuses no later than the August after Year 3.	3
Teachers' Understanding of Their Pay-for-Performance Eligibility[c]	Understanding of eligibility is necessary for bonuses to affect behavior.	Districts had higher levels of teacher understanding if there was at least a 50 percentage point difference between treatment and control teachers in the percentage who believed they were eligible for performance bonuses in Year 4.	4
Contextual Characteristics (Years 2, 3, and 4 Analyses)			
District enrollment[d]	In larger districts, teachers may have weaker connections to district officials, leading to less buy-in for new policies.	Districts were considered larger if their total prior-year enrollment was greater than 30,000 students.	5
Right-to-work laws[e]	Districts with right-to-work laws have weaker teachers' unions, and teachers in these districts may be more supportive of pay-for-performance.	Districts had right-to-work laws according to the National Right to Work Legal Defense Foundation.	6

[a]Based on district interviews, 2015.

[b]Based on educator administrative data from Year 3.

[c]Based on teacher survey, 2015.

[d]Based on Common Core of Data, 2011–2012, 2012–2013, and 2013–2014.

[e]Based on data from the National Right to Work Legal Defense Foundation.

For each characteristic, we constructed two measures to capture differences across districts. First, most of the characteristics varied across districts on a continuous scale (a spectrum), and for those characteristics we selected a continuous measure to capture this variation. (One characteristic—whether a state had right-to-work laws—was simply present or absent in each district, so it did not have a continuous measure.) Second, we identified a subgroup of districts that had higher levels of the characteristic (or, in the case of right-to-work laws, those that had the characteristic at all). The final column in Table G.1 indicates the number of districts that met the study's definition for having higher levels of the characteristic. The remaining districts (out of the total of 10 districts) made up its comparison subgroup. To classify districts into two subgroups, we first ranked the 10 districts according to the continuous measure of the specified characteristic. We then grouped districts into "higher" and "lower" categories such that there was a clear decline in the characteristic when moving from the "higher" to "lower" group. As we discuss later in this appendix, our main analysis examined differences in student achievement impacts between subgroups, but we also conducted sensitivity analyses to examine the relationships between the original, continuous measures of the characteristics and impacts.

We used data from the teacher survey, district interview, administrative records, and the Common Core of Data to measure these characteristics and categorize districts into subgroups. For program characteristics, the findings in this section are based on district subgroups in Year 4, and we compare the findings to those for Years 2 and 3 that we reported previously (Chiang et al. 2015; Wellington et al. 2016). Since teachers in each year would likely be responding to information about actual bonus awards from the prior year, we used prior-year bonus data to place districts into subgroups based on characteristics related to bonuses.

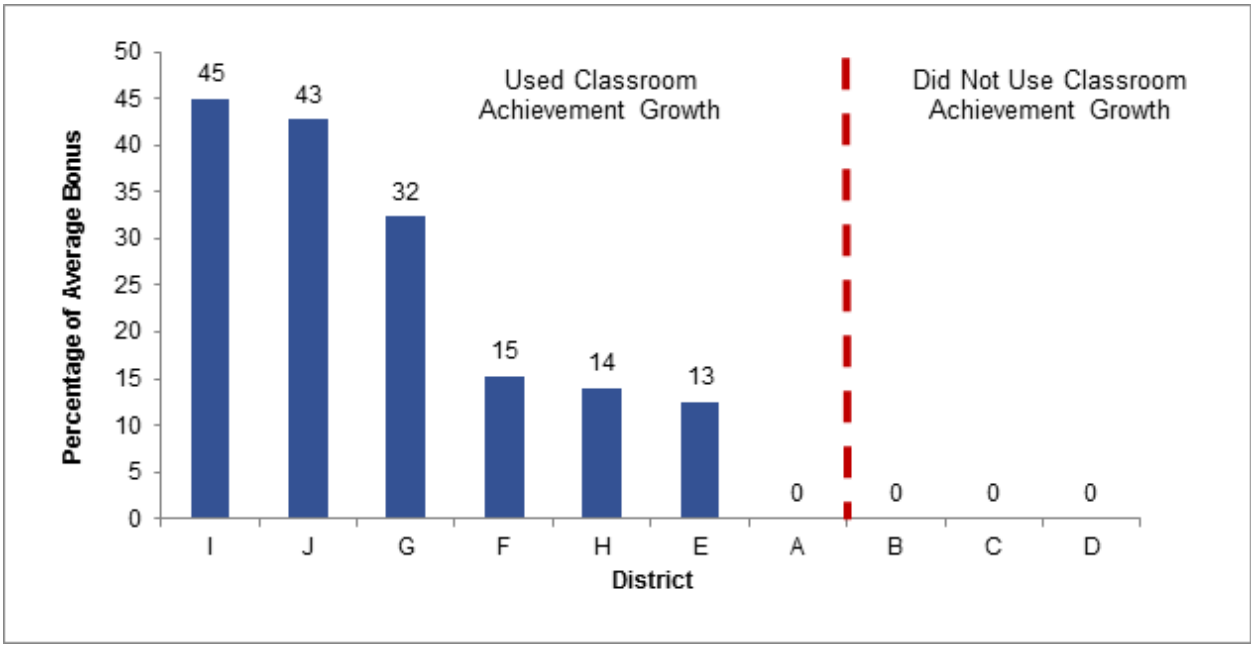
For contextual characteristics, we categorized districts into subgroups in Years 2, 3, and 4, given that these characteristics had not been examined in previous reports. We determined whether a district was in a right-to-work state in each of these years based on information obtained from the National Right to Work Legal Defense Foundation (http://www.nrtw.org/rtws.htm). The number of students enrolled in each district, our measure of district size, came from the Common Core of Data in the prior year. For both of these contextual characteristics, we found that districts' subgroup classification remained consistent across Years 2 through 4.

In what follows, we provide more details on the rationale for examining these characteristics, specify the continuous measure of each characteristic, and discuss how we divided the continuous scale into two subgroups.

Use of classroom achievement growth. In the seven districts that used classroom achievement growth to measure teacher effectiveness and award bonuses, teachers in treatment schools who taught grades and subjects that were tested by state assessments had a clear, direct monetary incentive to raise their students' test scores. In fact, in these districts, the average and maximum bonus amounts that were linked to this measure exceeded the amounts that were linked to each of the other effectiveness measures (Chapter IV, Figure IV.6). In contrast, in the remaining three districts that did not evaluate teachers based on classroom achievement growth, teachers faced a potentially weaker incentive to raise the test scores of their own students because their achievement growth ratings were determined more heavily by the performance of students they did not directly teach (for example, students in the entire school or in the same grade). Given that test scores were the outcome in our impact analysis, our expectation was that the impacts of pay-for-performance on test scores would be larger in districts that gave teachers a direct monetary incentive to raise those scores in their own classrooms.

Although districts either did or did not use classroom achievement growth—directly giving rise to two subgroups of districts—we also created a continuous measure of the extent to which teachers' bonuses were based on this measure. Specifically, among the seven districts that used classroom achievement growth, there was variation in the percentage of treatment teachers' average bonus that was based on this measure (Figure G.1). For example, for bonuses awarded in Year 3 (whose characteristics were used to explain differences in student achievement impacts in Year 4), treatment teachers in three districts received more than 30 percent of their average bonus from classroom achievement growth; in three other districts, this percentage ranged from 13 to 15 percent; and in one district (District A in Figure G.1), no teachers received a bonus based on classroom achievement growth even though they potentially could have.

Figure G.1. Percentage of Average Bonus Based on Classroom Achievement Growth Among Teachers in Treatment Schools in Year 3

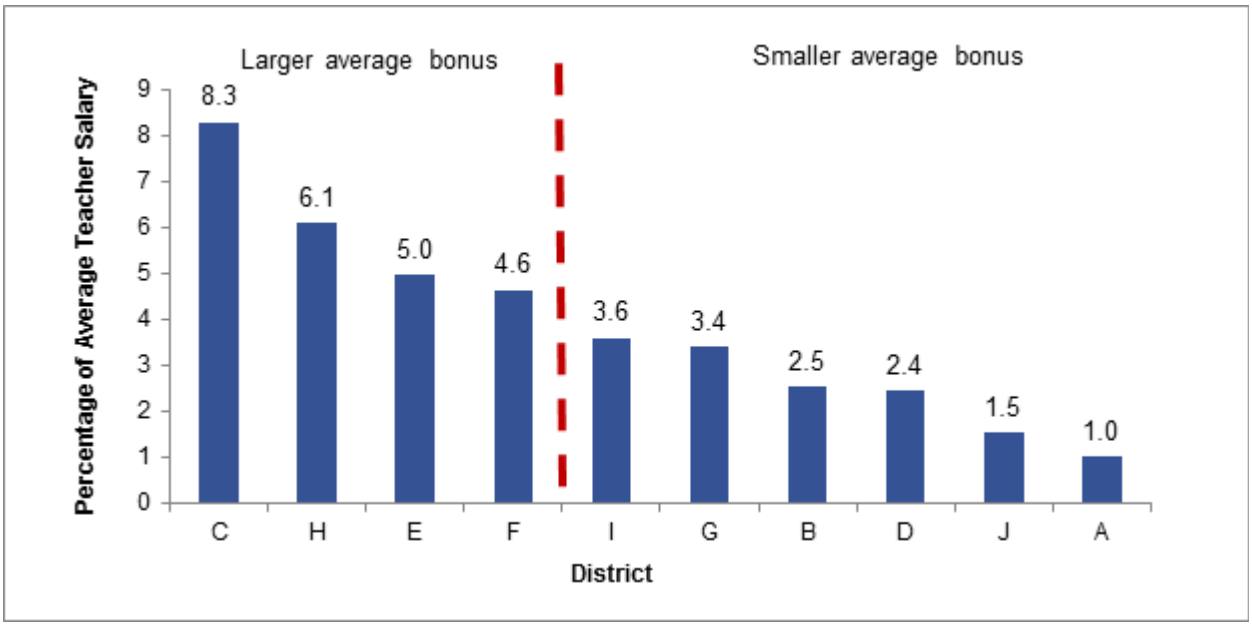


Source: Educator administrative data, Year 3 (N = 2,266 teachers).

Size of average bonus. Teachers may pay more attention to bonuses that are larger on average. To the extent that bonuses represent, on average, a larger portion of teachers' total compensation, teachers may find greater reason to learn about the details of the bonus program. In other findings from this evaluation, the evidence to support this possibility was inconsistent. The average size of bonuses awarded by districts was positively related to teachers' understanding of their pay-for-performance eligibility in Year 3 (Wellington et al. 2016) but not in Year 4 (Chapter IV). Nevertheless, average bonus amounts could have been related to teachers' level of interest in the bonus program in ways that we did not measure, such as their interest in learning about the specific criteria for earning a bonus and the resources available to help increase performance ratings. If larger average bonuses generate greater interest in the bonus program, they may also be associated with larger impacts on student achievement.

For explaining differences in impacts after four years of pay-for-performance, our continuous measure of the size of the average bonus was the average bonus amount for treatment teachers in the prior year (Year 3) as a percentage of the average teacher's salary. Average performance bonuses for treatment teachers ranged from one to eight percent of average salary (Figure G.2). The four districts that awarded an average bonus of at least 4.5 percent of average salary were classified as having a larger average bonus.

Figure G.2. Average Performance Bonus Earned by Teachers in Treatment Schools in Year 3 as a Percentage of Average Teacher Salary

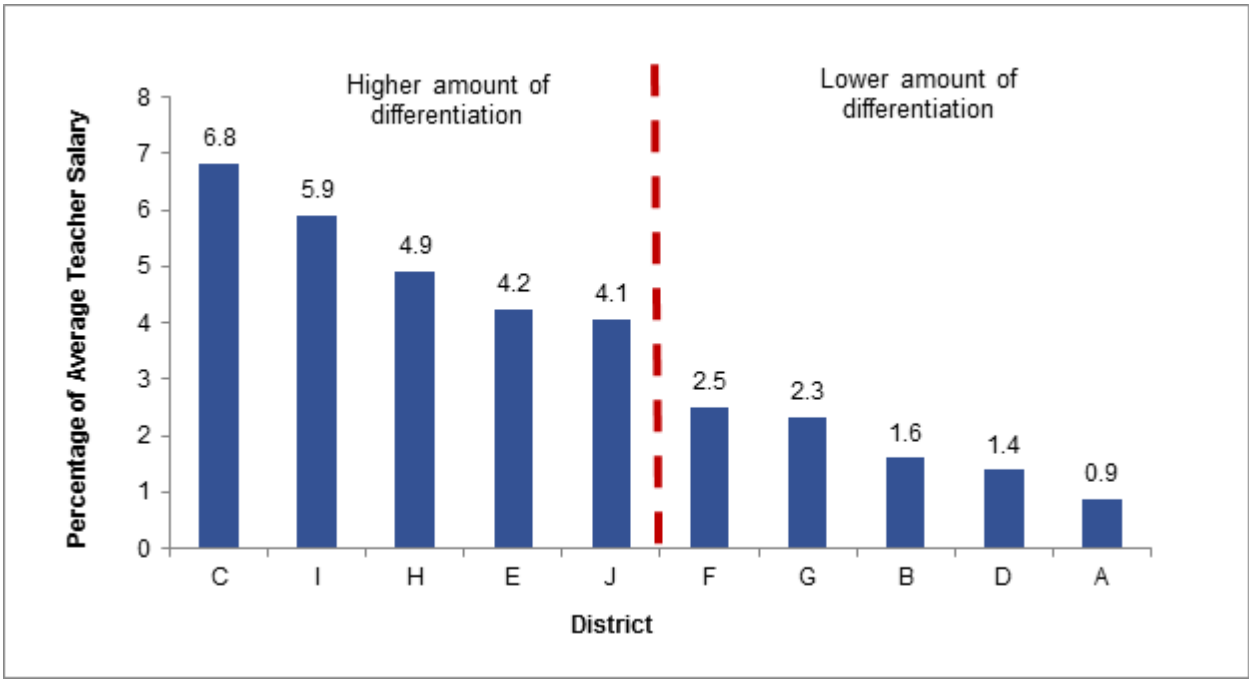


Source: Educator administrative data, Year 3 (N = 2,266 teachers).

Amount of differentiation in bonuses. Districts that award bonuses with larger differences between the amounts earned by teachers with higher and lower performance ratings may provide a greater monetary incentive for teachers to perform well on their performance ratings. On the other hand, for those who believe that teachers should be paid similarly (or based on tenure), pay-for-performance with large differences in payouts among teachers may lower satisfaction and have a negative impact on teachers' effectiveness.

For explaining differences in impacts after four years of pay-for-performance, our continuous measure of the amount of differentiation in each district's bonuses was the standard deviation of treatment teachers' bonuses in Year 3 as a percentage of the average teacher's salary. This measure captured how extensively below- and above-average bonuses differed in value from the average bonus. In five of the ten districts, the standard deviation of treatment teachers' bonuses exceeded four percent of average teacher salary, and we classified these districts as having a higher amount of differentiation (Figure G.3). This measure of differentiation was different from the example given in the grant notice—a bonus in which the maximum amount was at least three times the average amount. Districts that met the grant notice's example of differentiation but had very small bonuses would not be classified as a district with high differentiation of bonuses by our measure, because the dollar value of the differences in bonus amounts between teachers would still be small.

Figure G.3. Standard Deviation of Pay-for-Performance Bonuses Earned by Teachers in Treatment Schools in Year 3 as a Percentage of Average Teacher Salary



Source: Educator administrative data, Year 3 (N = 2,266 teachers).

Timing of awarding bonuses. Bonuses may affect teacher effectiveness by encouraging teachers to change their practices or to change schools. Districts that distribute awards earlier allow their teachers more time to respond in these ways.

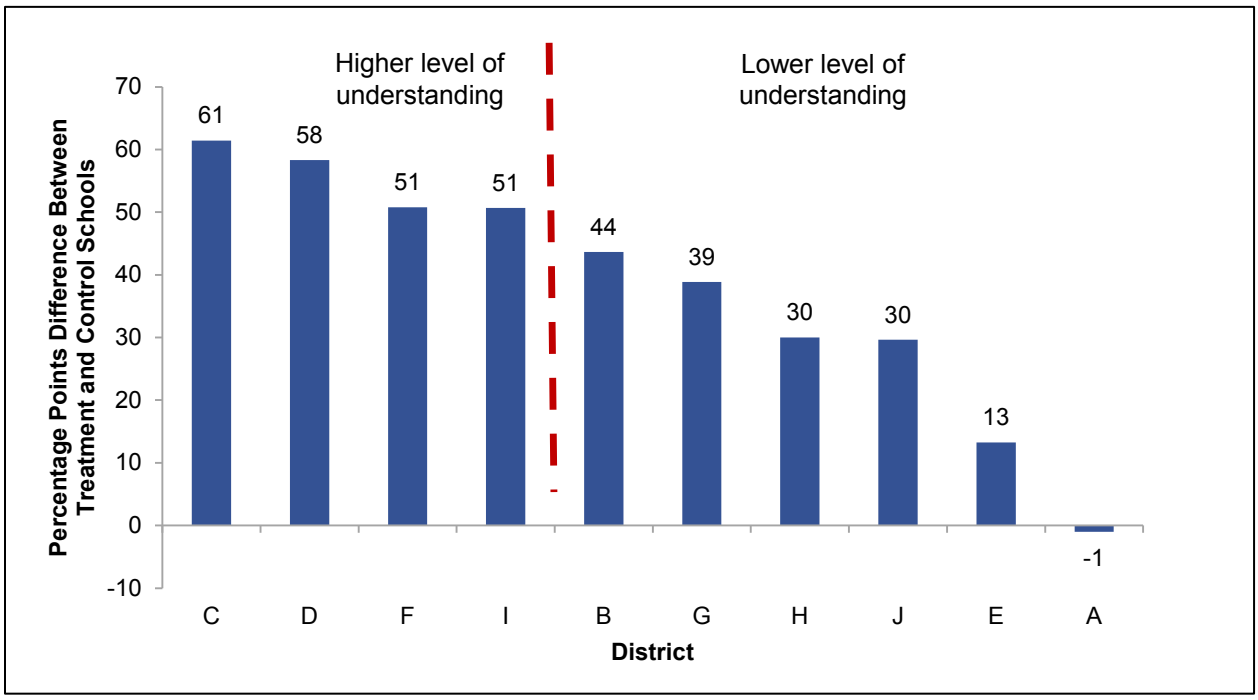
As discussed in Chapter IV, there were differences among evaluation districts in the timing of awarding bonuses from Year 3 (2013–2014). To create a continuous measure of the timing of bonus awards, we measured the number of months that elapsed after May 2014 before each district awarded the first component of its bonuses. Three districts paid out at least some components of the bonuses before the start of the 2014–2015 school year, and we classified those districts as having awarded bonuses earlier. Among the remaining districts, five reported paying teachers between October 2014 and January 2015, one paid teachers in March 2015, and one paid teachers in June 2015. Although none of the districts distributed awards early enough for teachers to respond by changing schools for the next school year, those that notified teachers sooner did provide teachers with more time to revise their teaching practices.

Teachers' understanding of their eligibility for a pay-for-performance bonus. Teachers must understand they are eligible for pay-for-performance bonuses for those bonuses to affect their decisions and behavior. If understanding of pay-for-performance eligibility had been perfect, all teachers in treatment schools would have been aware that they were eligible, and all teachers in control schools would have recognized that they were not.

In each district, our continuous measure of teachers' understanding of their pay-for-performance eligibility was the difference between the percentage of teachers in treatment and control schools who believed they were eligible for a performance bonus. This difference ranged from -1 to 61 percentage points across districts in Year 4 (Figure G.4). In four districts, there was at least a 50 percentage point difference between treatment and control teachers in the percentage who

believed they were eligible for a performance bonus. We classified these districts as having higher levels of teacher understanding of pay-for-performance eligibility.

Figure G.4. Difference Between the Percentages of Teachers in Treatment and Control Schools Who Believed They Were Eligible for Pay-for-Performance Bonuses in Year 4



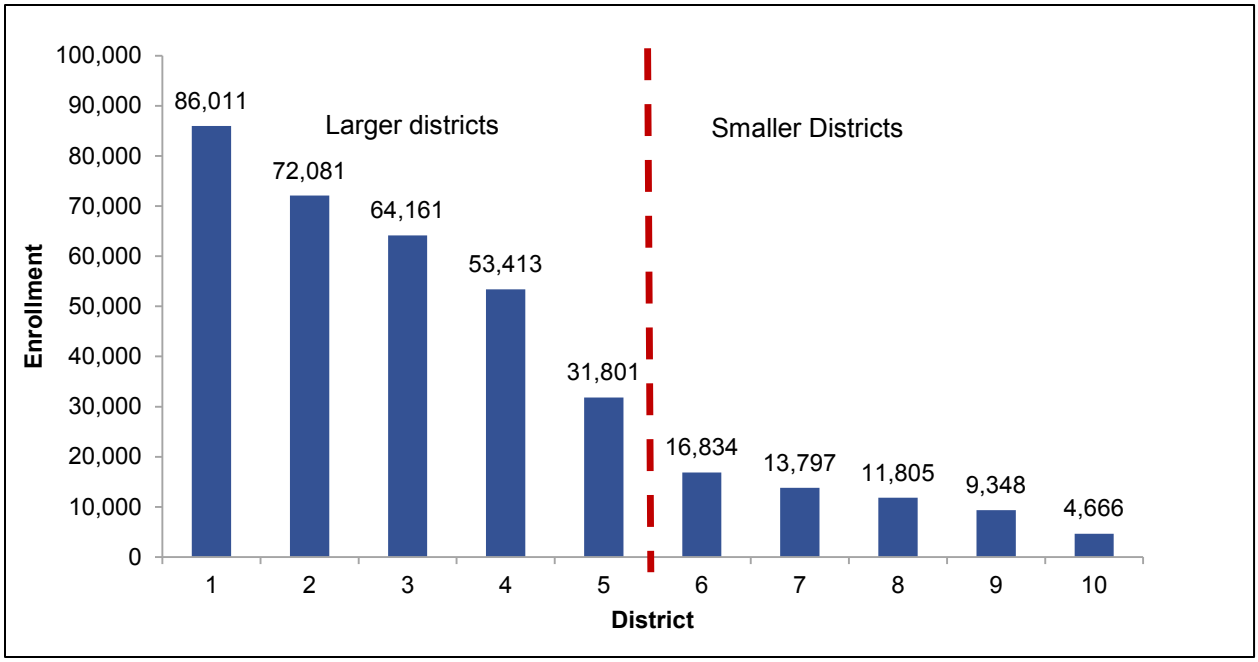
Source: Teacher survey, Year 4 (N = 797 teachers).

District enrollment. Larger districts may be faced with challenges to implementing pay-for-performance not faced by smaller districts. For example, teachers in larger districts may have less input into district policies, and may therefore be less likely to buy into new district-initiated policies such as pay-for-performance.

In each year, we measured the number of students enrolled in the district based on the prior year's Common Core of Data, the most recent information available. Five districts that we classified as larger districts had total annual enrollment consistently above 30,000 students in each year. For example, in the analysis of impacts after four years, the larger districts had Year 3 enrollment ranging from 31,801 to 86,011 students (Figure G.5). The remaining districts had Year 3 enrollment ranging from 4,666 to 16,834 students.

Districts with right-to-work laws. Right-to-work laws prohibit unions from collecting agency fees from nonmembers, and states with such laws are likely to have weaker unions. Since teachers' unions may promote skepticism of pay-for-performance, teachers in states with right-to-work laws may be more generally supportive of pay-for-performance and more willing to respond to these incentives. Among the evaluation districts, six were in states with right-to-work laws in all years that we examined.

Figure G.5. Number of Students Enrolled in Year 3



Source: Common Core of Data, 2013–2014.

Note: District labels in this figure do not correspond to district labels in any other table or figure in this report so that the impacts of pay-for-performance in specific districts cannot be identified.

Findings on Differences in Student Achievement Impacts Between District Program Subgroups

For each pair of program subgroups that differed on a particular characteristic, we estimated the impacts of pay-for-performance on student achievement within the two subgroups and examined whether the impacts differed between the subgroups. As discussed in Chapters II and VI, we expressed achievement outcomes as z-scores based on statewide means and standard deviations of scores in each grade. A statistically significant difference in impacts between the two subgroups would represent an association between the characteristic and impacts.

Overall, we found no consistent evidence that any of the program characteristics that we considered explained differences in impacts across districts. After four years of pay-for-performance, impacts in one subject or the other tended to be more positive in districts that used classroom achievement growth, awarded larger average bonuses, and awarded bonuses earlier (Table G.2). However, given that we examined a large number of relationships between program characteristics and student achievement impacts, there was an increased likelihood that these significant relationships occurred just by chance. Moreover, none of the program characteristics that were associated with impacts after four years had previously been associated with impacts after two or three years (Chiang et al. 2015; Wellington et al. 2016). In fact, in previous years, only one other characteristic—higher differentiation of bonuses—was associated with impacts (in a negative manner), and this association appeared in only one year and subject (Chiang et al. 2015).

Table G.2. Differences in the Impacts of Pay-for-Performance on Student Achievement After Four Years of Implementation Between Subgroups Based on Districts' Program Characteristics (Student *z*-Score Units)

	Math		Reading	
Subgroup of Districts	Difference in Impacts Between Specified Subgroup and Remaining Districts	*p*-value	Difference in Impacts Between Specified Subgroup and Remaining Districts	*p*-value
Used Classroom Achievement Growth to Measure Teacher Effectiveness and Award Bonuses	0.11*	0.02	0.07	0.10
Larger Average Bonus	0.08	0.11	0.08*	0.05
Higher Amount of Differentiation in Bonuses	0.05	0.41	0.04	0.31
Earlier Awarding of Bonuses	0.07	0.17	0.09*	0.02
Higher Level of Teacher Understanding of Pay-for-Performance Eligibility	-0.03	0.59	0.02	0.57
Number of Students	**38,939**		**38,929**	
Number of Schools	**131**		**131**	

Source: Student administrative data.

*Difference is statistically significant at the .05 level, two-tailed test.

Sensitivity Analysis for District Program Characteristics

As discussed earlier, most of the program characteristics varied across districts on a continuous scale (a spectrum), even though our main analysis divided that spectrum into two subgroups. Therefore, we also examined whether the continuous measures of those characteristics were associated with the impacts of pay-for-performance on student achievement.

Findings on the relationships between continuous measures of program characteristics and student achievement impacts were generally consistent with the earlier findings based on categorizing districts into two subgroups. For example, after four years of implementation, impacts in at least one subject were positively related to the percentage of the average bonus based on classroom achievement growth and the total average bonus amount as a percentage of teacher salary, and negatively related to the amount of time that districts took to award bonuses (Table G.3). These were the same program characteristics that were associated with achievement impacts after four years in the subgroup analyses presented earlier (Table G.2). However, similar to the findings from the subgroup analyses, the continuous measures of program characteristics that were related to impacts after four years were not related to impacts after two or three years (Chiang et al. 2015; Wellington et al. 2016). In previous years, one other continuous measure of a program characteristic—the standard deviation of bonus amounts as a percentage of average teacher salary—was related (negatively) to impacts, but in only one year and subject (Chiang et al. 2015). Therefore, these sensitivity analyses support the conclusion that the program characteristics we considered did not consistently explain differences in impacts across districts.

Table G.3. Association Between Continuous Measures of Program Characteristics and the Impacts of Pay-for-Performance on Student Achievement After Four Years of Implementation

	Math		Reading	
Program Characteristic	Association	*p*-value	Association	*p*-value
Percent of Average Bonus in Year 3 Based on Classroom Achievement Growth	0.002*	0.05	0.002*	0.03
Average Performance Bonus in Year 3 as a Percentage of Average Teacher Salary	0.015	0.11	0.015*	0.03
Standard Deviation of Performance Bonuses in Year 3 as a Percentage of Average Teacher Salary	0.012	0.31	0.014	0.14
Number of Months Since May 2014 When District Paid Out Year 3 Performance Bonuses	-0.010	0.18	-0.012*	0.02
Difference Between the Percentages of Teachers in Treatment and Control Schools Who Believed They Were Eligible for Performance Bonuses in Year 4	0.001	0.62	0.002	0.10
Number of Students	**38,939**		**38,929**	
Number of Schools	**131**		**131**	

Source: Student administrative data.

Note: The association between each characteristic and student achievement impacts is expressed as the difference in student achievement impacts (in student *z*-score units) associated with a one-unit change in the measure of the characteristic.

*Association is statistically significant at the .05 level, two-tailed test.

Findings on District Contextual Factors

We examined differences in student achievement impacts between districts in states with and without right-to-work laws, and between larger and smaller districts. Although the relationships between program characteristics and student achievement impacts after two and three years of pay-for-performance were examined in previous reports (Chiang et al. 2015; Wellington et al. 2016), the relationships between these district contextual factors and impacts had not been previously examined. We therefore present findings on these relationships after two, three, and four years of implementing pay-for-performance.

We found that by Year 4, the impact of pay-for-performance on student achievement was larger in districts with right-to-work laws than in districts without such laws. The difference in student achievement impacts between districts with and without right-to-work laws grew over the final three years of TIF implementation (Table G.4). Although these differences were positive in all of those years and in both subjects, they were initially small and not statistically significant. However, by the end of four years of implementation, the impact in right-to-work districts was larger by 0.15 standard deviations in math and 0.12 standard deviations in reading than in the other districts.

Table G.4. Association Between District Contextual Factors and the Impacts of Pay-for-Performance on Student Achievement

	Math		Reading	
Subgroup of Districts and Year	Difference in Impacts Between Specified Subgroup and Remaining Districts	*p*-value	Difference in Impacts Between Specified Subgroup and Remaining Districts	*p*-value
Right-to-Work Districts				
Year 2	0.06	0.22	0.03	0.24
Year 3	0.06	0.21	0.08*	0.01
Year 4	0.15*	0.00	0.12*	0.00
Larger Districts				
Year 2	-0.14*	0.00	-0.06*	0.03
Year 3	-0.11*	0.02	-0.07*	0.02
Year 4	-0.08	0.13	-0.04	0.28
Continuous Measure of Program Characteristic	Association	*p*-value	Association	*p*-value
Log Enrollment[a]				
Year 2	-0.08*	0.00	-0.02	0.06
Year 3	-0.07*	0.00	-0.03*	0.03
Year 4	-0.05*	0.04	-0.03	0.14
Number of Students—Range[b]	**38,939-40,454**		**38,929-40,122**	
Number of Schools	**131**		**131**	

Source: Student administrative data.

Note: The association between log enrollment and student achievement impacts is expressed as the difference in student achievement impacts (in student *z*-score units) associated with a one-unit change in log enrollment.

[a]Defined as the natural logarithm of the number of students enrolled in the prior year.

[b]Sample sizes are presented as a range based on the data available for each row in the table.

*Difference or association is statistically significant at the .05 level, two-tailed test.

When comparing impacts in larger districts—those that enrolled more than 30,000 students—with impacts in the remaining districts, we found that impacts were generally smaller in larger districts**.** The impact of pay-for-performance on student math achievement was lower by 0.08 to 0.14 standard deviations in larger districts than in smaller districts after two to four years of implementation, though the difference after four years was not statistically significant (Table G.4). Over the same period, the impact of pay-for-performance on student reading achievement was lower by 0.04 to 0.07 standard deviations in larger districts, though again the difference after four years was not significant.

Because our classification of districts as larger or smaller districts was based on enrollment counts that varied along a spectrum, we also conducted a sensitivity analysis that examined the direct relationship between the number of students in a district and the impacts of pay-for-performance. When examining this relationship, we refined the measure of district enrollment to account for the highly skewed distribution of enrollment across districts. As shown in Figure G.5, the largest evaluation districts had enrollment that exceeded the median by far more than the smallest districts fell short of the median. Because of this skew, the relationship between the raw number of students in a district and students achievement impacts would be dominated by the largest districts. To avoid

this issue, we took the natural logarithm of each district's enrollment (abbreviated as log enrollment), which resulted in a distribution with much less skew. We then estimated the relationship between log enrollment and student achievement impacts. Because an increase of 0.01 in log enrollment is approximately a 1 percentage point increase in enrollment, the relationships reported in Table G.4 can be multiplied by 0.1 to approximate the difference in student achievement impacts associated with a 10 percentage point increase in enrollment.

Findings from the sensitivity analysis generally support the conclusion that larger numbers of students were associated with smaller impacts of pay-for-performance, with somewhat stronger evidence in math than in reading. After two to four years of pay-for-performance, an increase in enrollment of 10 percent was associated with a reduction in impacts on math achievement ranging from 0.005 to 0.008 standard deviations (Table G.4). These relationships were statistically significant in all years. In reading, a 10 percent increase in enrollment was associated with a reduction in impacts of 0.002 to 0.003 standard deviations, with the relationship reaching statistical significance after three years of implementation but not after two years (p-value = 0.06) or four years (p-value = 0.14). To place the magnitudes of these relationships in context, we can also consider the predicted difference in impacts that would be associated with an increase in enrollment from 17,000 to 32,000—approximately the difference in enrollment between districts 6 and 5 in Figure G.5. This increase in enrollment would be associated with a reduction in impacts on math achievement of about 0.03 to 0.05 standard deviations, and a reduction in impacts on reading achievement of about 0.01 to 0.02 standard deviations.

Explaining Differences in Impacts Across Schools

As discussed in Chapter VI, the impacts of pay-for-performance on student achievement also differed across treatment schools, even within the same district. We sought to determine whether those differences were related to differences in two types of school-level factors: (1) impacts on teacher and principal behaviors at each school and (2) each school's baseline student achievement. If schools with larger impacts of pay-for-performance on certain behaviors also had larger impacts on student achievement, this pattern would provide suggestive evidence that pay-for-performance might affect student achievement by influencing those behaviors. If schools with higher or lower baseline student achievement tended to experience larger impacts, this information could suggest ways to target pay-for-performance policies to the types of schools that would benefit the most. This section provides detail on each set of school-level factors that we examined and presents findings on their relationships with the impacts of pay-for-performance.

Associations Between Impacts on Educator Behaviors and Impacts on Student Achievement

We considered the possibility that pay-for-performance may have affected teacher and principal behaviors differently across schools, leading to differences in impacts on student achievement. To assess this possibility, we examined whether treatment schools with larger impacts of pay-for-performance on certain types of educator behaviors also had larger impacts on student achievement. For each behavior and student achievement outcome, we measured the impact of pay-for-performance in each treatment school—the extent to which outcomes in the treatment school differed from those in the control school to which it was paired for random assignment (see Chapter II and Appendix B for details). We then examined the association between impacts on behaviors and impacts on student achievement.

The educator behaviors we examined were based on the theory of change for how pay-for-performance might affect student achievement (see Chapter I). In an effort to earn pay-for-performance bonuses, principals and teachers may act strategically, shifting attention towards activities that improve measures on which those bonuses are based; they may increase their effort on the job; or they may adopt different teaching practices known to be more effective. To measure these behaviors, we used educators' responses to survey questions on topics that could reflect strategic behavior, effort, and teaching practices. In addition, we used observation ratings (from administrative data) as a direct measure of teaching practices. Table G.5 details the specific items used.

Table G.5. Measures of Educator Behaviors for Explaining Differences in Impacts on Student Achievement

Type of Educator Behavior	Data Source	Data Item	Rationale for Use of This Item
Principal and Teacher Strategic Behavior	Principal survey	Principals often or always uses teachers' ability to produce high test scores as a criterion for assigning them to grade levels or subject areas	Principals who report having used this criterion frequently are acting strategically to improve test scores.
Principal and Teacher Strategic Behavior	Teacher survey	Weekly hours spent by teachers during the school day on instructional activities (teaching, preparation, and professional development)	Teachers reporting more time on these activities could be shifting the focus of their school hours towards improving student achievement.
Teacher Effort	Teacher survey	Weekly hours spent by teachers outside the school day on instructional activities (tutoring, preparation, professional development)	Spending more non-school hours on school activities entails greater total effort.
Teacher Effort	Teacher survey	Teachers believe that TIF caused them to work more effectively	Increased effectiveness may be a consequence of increased effort.
Teacher Effort	Teacher survey	Teachers feel increased pressure to perform due to TIF	Increased pressure suggests teachers feel a need to work harder.
Teaching Practices	Teacher survey	Teachers believe that TIF harmed the collaborative nature of teaching	Less agreement with this statement suggests increased collaboration, a type of change in teaching practices.
Teaching Practices	Teacher survey	Teachers feel they have less freedom in teaching due to TIF	Less freedom to choose teaching practices suggests a change in teaching practices.
Teaching Practices	Teacher survey	Teachers believe that students will benefit from the feedback received from classroom observations	Feedback should benefit students through a change in practice.
Teaching Practices	Educator administrative data	Observation ratings	Observation ratings are a direct measure of teaching practices.

Overall, we found no consistent evidence of any relationship between impacts on educator behaviors and impacts on student achievement**.** For example, after four years of pay-for-performance, only one of the eighteen relationships we examined between impacts on educator behaviors and impacts on student achievement (nine measures of behaviors and two subjects) was statistically significant (Table G.6). Given the large number of relationships examined, there was an increased likelihood that the single significant finding could have occurred just by chance. In fact, the single relationship that was significant after four years of implementation was not significant after three years (Wellington et al. 2016). Likewise, only one other relationship was significant after three years, and that relationship was not significant after four years.

Table G.6. Association Between Impacts of Pay-for-Performance on Educator Behaviors and Impacts of Pay-for-Performance on Student Achievement After Four Years of Implementation

	Math		Reading	
Measure of Educator Behavior (and units in which impacts on the behavior are expressed)	Association	*p*-value	Association	*p*-value
Strategic Behavior				
Principals Often or Always Use Teachers' Ability to Produce High Test Scores as a Criterion in Teaching Assignments (percentage points)	-0.035	0.65	-0.042	0.48
Weekly Hours Spent by Teachers During School Day on Instructional Activities	0.001	0.85	-0.001	0.62
Effort				
Weekly Hours Spent by Teachers Outside the School Day on Instructional Activities	-0.012*	0.05	-0.007	0.24
Teachers Believe that TIF Caused Them to Work More Effectively (percentage points)	-0.065	0.56	-0.065	0.49
Teachers Feel Increased Pressure to Perform due to TIF (percentage points)	-0.082	0.62	-0.086	0.43
Teaching Practices				
Teachers Believe that TIF Harmed the Collaborative Nature of Teaching (percentage points)	-0.116	0.36	-0.109	0.23
Teachers Feel They Have Less Freedom in Teaching due to TIF (percentage points)	-0.003	0.98	-0.060	0.59
Teachers Believe that Students Will Benefit from the Feedback Received from Classroom Observations (percentage points)	0.120	0.40	0.035	0.82
Observation Ratings (points on 1-to-4 scale)	-0.054	0.68	-0.033	0.73
Number of Random Assignment Blocks—Range[a]	**42-44**		**42-44**	

Sources: Educator and student administrative data (Year 4) and principal and teacher surveys (2015).

Notes: Random assignment blocks (matched pairs of treatment and control schools or matched groups of treatment and control schools) are the units of analysis. Associations are measured as the difference in the impact of pay-for-performance on student achievement (in student *z*-score units) that is associated with a one-unit difference in the impact of pay-for-performance on the measure of educator behavior. Sample sizes differ across analyses because two random assignment blocks with impacts on weekly hours spent by teachers outside and during the school day on instructional activities exceeded 30 hours, and are omitted from those analyses.

[a]Number of random assignment blocks presented as a range based on the data available for each row in the table.

*Association is statistically significant at the .05 level, two-tailed test.

Associations Between Schools' Baseline Student Achievement and Subsequent Impacts of Pay-for-Performance on Student Achievement

Additionally, we considered the possibility that differences in student achievement across schools prior to the first year of implementation could explain differences in the degree to which pay-for-performance improved student achievement. For example, schools with lower baseline student achievement may benefit more from increases in teacher effectiveness since they have more room to improve. However, having lower baseline student achievement may indicate that the school faces more challenges—such as a dysfunctional school culture, challenging student home environments, or high student mobility—that render student achievement less responsive to policy interventions.

To explore these possibilities, we measured baseline student achievement in each treatment school and the control school to which it had been paired for random assignment by calculating the average of math and reading scores among students in that pair within the pre-implementation year (see Appendix B for details). Across these pairs of schools, we examined the association between baseline student achievement and the subsequent student achievement impacts after four years of pay-for-performance. Because we did not previously conduct similar analyses after three years of pay-for-performance (unlike the analyses of educator behaviors discussed earlier in this section), we also examined the association between baseline student achievement and impacts after three years.

Differences in student achievement prior to implementing pay-for-performance were not associated with the impacts of pay-for-performance on student achievement (Table G.7). Across years (Year 3 and Year 4) and subjects (math and reading), none of the relationships we measured were statistically significant.

Table G.7. Association Between Schools' Baseline Student Achievement and Impacts of Pay-for-Performance on Student Achievement After Three and Four Years of Implementation

	Math		Reading	
Year in Which Impact is Measured	Association	*p*-value	Association	*p*-value
Year 3	-0.114	0.57	0.002	0.99
Year 4	-0.002	0.99	-0.037	0.83
Number of Random Assignment Blocks	**44**		**44**	

Source: Student administrative data.

Notes: Random assignment blocks (matched pairs of treatment and control schools or matched groups of treatment and control schools) are the units of analysis. The measure of baseline student achievement in each random assignment block is the average of math and reading *z*-scores within the block in the pre-implementation year, with scores weighted such that each school and subject within a block contribute equally to the block-level average. Associations are measured as the difference in the impact of pay-for-performance on student achievement (in student *z*-score units) that is associated with an increase of one student *z*-score unit in baseline achievement.

THIS PAGE IS INTENTIONALLY BLANK

APPENDIX H

COST-EFFECTIVENESS METHODS AND DETAILED FINDINGS FOR CHAPTER VI

THIS PAGE IS INTENTIONALLY BLANK

Districts have limited budgets and must choose how to allocate their funds among many different policies. Before implementing a pay-for-performance program, districts would likely want to compare the benefit of this policy—for example, improved student achievement—with its cost and consider whether alternative policies could achieve these benefits in a more cost-effective manner. To explore this question, we conducted a cost-effectiveness analysis to determine how the cost of improving student achievement through the offer of pay-for-performance bonuses compared to two other alternative policies: (1) transfer incentives for high-performing teachers to move to low-performing schools and (2) class-size reduction.

This appendix provides a comprehensive discussion of the methods and findings from the cost-effectiveness analysis presented in Chapter VI. First, we provide an overview of our approach to calculating the cost-effectiveness ratio—the per-student cost needed for a policy to achieve a given impact on student achievement—and to testing whether these costs differ statistically among policies. Second, we present estimates of the impact of each policy as reported by large-scale random assignment studies, and we explain the need to refine those estimates so that they identify the impacts of specific durations of student exposure to the policies. Third, we describe the approach we used to obtain (and to model when needed) the impacts of specific durations of student exposure to each policy. We also discuss how we calculated the per-student cost of the pay-for-performance policy evaluated in this study. Fourth, we present all findings from the cost-effectiveness analysis, including those discussed in Chapter VI and supplemental findings, and we discuss the limitations of the analysis. Last, we provide the technical details for calculating the cost-effectiveness ratios and their standard errors, and present the formulas used to model the impacts of student exposure to each policy and their associated standard errors.

Our evaluation of pay-for-performance classified the evaluation districts into two cohorts—Cohort 1 and Cohort 2—according to the year in which we randomly assigned their schools to a treatment group or a control group (Chapter II). The 10 districts whose schools were randomly assigned in spring and summer 2011 were classified as Cohort 1. Three additional districts, whose schools were randomly assigned in spring and summer 2012, were classified as Cohort 2. Cohort 1 districts completed four years of implementation, whereas Cohort 2 districts completed three years. Unless otherwise noted, all of the cost-effectiveness findings for pay-for-performance in this appendix (as in Chapter VI) are based on costs and impacts in Cohort 1 districts. In some cases, this appendix provides supplemental information on the cost-effectiveness of pay-for-performance in Cohorts 1 and 2 combined.

Overview of the Cost-Effectiveness Approach

This section describes our approach to assessing the cost-effectiveness of pay-for-performance bonuses within the Teacher Incentive Fund (TIF) program. The goal of this analysis was to determine how the cost of improving student achievement through the offer of pay-for-performance bonuses compared to alternative policies. In particular, we describe our approach to calculating the cost-effectiveness ratio—the measure of cost-effectiveness on which we compared different policies—and to testing whether the cost-effectiveness ratios for alternative policies differed statistically.

Cost-Effectiveness Ratio

A cost-effectiveness analysis compares policies by their cost-effectiveness ratio—the policy's cost divided by its impact. For each policy, we used information from a large-scale random assignment evaluation of the policy to calculate its cumulative cost per student and cumulative

impact on student achievement from the start of the policy through the end of each year of implementation. The cumulative cost per student was the sum of the additional cost from each year of exposing a student to the policy; likewise, a cumulative impact was the sum of the additional impact from each year of exposure. For a specified duration of exposure, we divided a policy's cumulative cost per student by its cumulative impact to obtain the policy's cost-effectiveness ratio based on that duration of exposure.

Our main approach to calculating cost-effectiveness ratios had the following elements:

Impacts. We used estimates of the impact of exposing students to specific durations of the policy on the students' math and reading achievement. For example, in the case of class-size reduction, we used estimates that captured the effects of attending small classes for durations of one, two, three, and four years. Because the evaluations of pay-for-performance and transfer incentives assigned schools and teacher teams, respectively, to implement those policies, we needed estimates of the impacts that students experienced from spending specific numbers of years in those schools or teams. Instead, these studies reported the impacts of implementing each policy on the average test scores of these schools or teams, based on students who had not necessarily been enrolled for all years since the policies began. In those cases, we modeled the impacts of students' full exposure to the policy—for example, the impacts of attending a school for all of the years since it started to offer pay-for-performance—from the study-reported impacts so that impacts would be comparable across policies (see the section "Calculating Costs and Impacts of Each Policy" in this appendix).

Costs. We expressed costs on a per-student basis to facilitate comparisons between policies. Because cost information on the various policies came from different sets of years, we adjusted all costs for inflation by expressing costs in May 2016 dollars using the U.S. Consumer Price Index for All Urban Consumers (CPI-U).

Discounting. The cost-effectiveness of a policy implemented over multiple years depends not only on its cumulative cost and cumulative impact, but also on the timing with which those costs and impacts emerge. Costs and impacts that are realized sooner are valued more than those realized later. Even without inflation, $100 today is worth more than $100 in the future because today's money can be invested to earn interest. In a cost-effectiveness analysis, this interest rate is called the discount rate. Future benefits are also discounted (even when benefits are not monetary) under the assumption that people value a current benefit more than they value the same benefit realized in the future (for example, Caulkins et al. [1999]; Harris [2009]). To account for the timing with which costs and impacts emerged, we discounted the portions of the costs and impacts that emerged in the second year and beyond by expressing them in their present value—specifically, their value in the first year of the policy. Following guidelines in research, we used a discount rate of 3 percent for the main analysis and tested the sensitivity of the analysis to alternative discount rates (Levin and McEwan 2001; Cellini and Kee 2010).

Dealing with multiple academic subjects. The evaluation of each policy measured impacts on both math and reading achievement. We used an equal-weighted average of the math and reading impacts for calculating the cost-effectiveness ratios in our main approach, but also conducted supplemental analyses of cost-effectiveness separately by subject.

Scale. Before any rescaling, each cost-effectiveness ratio represented the per-student cost to achieve a one standard deviation increase in test scores. We rescaled the cost-effectiveness ratios to indicate the per-student cost needed to achieve an impact of 0.04 standard deviations—the actual average impact of pay-for-performance—by multiplying the original ratio by 0.04.

Formally, we calculated cost-effectiveness ratios using equation (1) later in this appendix.

Hypothesis Testing

Because the impact estimate in the denominator of each cost-effectiveness ratio had imprecision, so too did the ratio as a whole. To measure the imprecision in each cost-effectiveness ratio, we calculated its standard error, using information on the standard errors of the impact estimates (see equation [4] later in this appendix). Because evaluation studies included in the analysis did not report standard errors for their cost estimates, we assumed that costs were measured with no error. We tested whether the cost-effectiveness ratios for different policies were statistically different using a two-tailed *t*-test.

Policies and Impact Estimates Included in the Cost-Effectiveness Analysis

We compared pay-for-performance to two alternative policies: (1) transfer incentives for high-performing teachers to move to low-performing schools and (2) class-size reduction. In Chapter VI we explain why we selected these alternative policies for the cost-effectiveness analysis. We obtained impact and cost information for transfer incentives from the U.S. Department of Education's national random assignment evaluation of a program that gave bonuses to high-performing teachers to teach in low-performing schools for two years (Glazerman et al. 2013). We obtained impact and cost information for class-size reduction from Tennessee's Student Teacher Achievement Ratio Project (Project STAR), a large random assignment evaluation that followed students who were assigned to small and regular-sized classes for four consecutive years (from kindergarten to grade 3). Specifically, we used impact information from Schanzenbach (2006) and Nye et al. (2000) and cost information from Harris (2009).

Table H.1 presents impacts, standard errors, and costs as reported by the evaluation studies—all of which were inputs in our cost-effectiveness calculations. The impacts of pay-for-performance are those reported in Tables VI.4 (Chapter VI) and F.16 (Appendix F).

A limitation of directly using the experimental impact estimates in the cost-effectiveness calculations is that students' exposure to the three policies differed and, therefore, the impacts were not comparable. The reason is that although each study measured the impact of implementing a policy for a specific number of years, the types of study participants assigned to those policies differed. These differences in study designs led to the following difference in students' exposure to the policies:

- The evaluation of **class-size reduction** assigned students to small or regular-sized classes. Nye et al. (2000) followed students who stayed in their assigned class sizes for one, two, three, and four years from the beginning of the study. Therefore, in each of those periods of time, Nye et al. (2000) measured impacts on students who had been exposed to the full duration of the policy.
- The evaluation of **transfer incentives** assigned teacher teams (defined based on grade and subject in a school) to be eligible to hire high-performing teachers who had received bonuses to transfer to low-performing schools or to hire from a normal pool of applicants (Glazerman et al. 2013). After each of the two years of the policy, the study measured the impacts of transfer incentives based on students who were taught by these teacher teams. Each team taught an entirely new cohort of students in the second year than the first. Therefore, experimental impacts at the end of the second year were based on students who had not been exposed to the policy in the first year.

Table H.1. Study-Reported Impact Estimates and Cost Information for the Policies Examined in the Cost-Effectiveness Analysis

Policy and Year of Implementation	Cost per Student	Cumulative Cost per Student	Math		Reading		Math and Reading Combined	
			Impact	Standard Error	Impact	Standard Error	Impact	Standard Error
Pay-for-Performance								
Cohort 1								
Year 1	$102	$102	0.02	0.02	0.03*	0.02		
Year 2	$102	$204	0.04	0.02	0.03*	0.01		
Year 3	$106	$310	0.06*	0.02	0.04*	0.02		
Year 4	$127	$424	0.04	0.02	0.04	0.02		
Cohorts 1 and 2								
Year 1	$74	$74	0.01	0.02	0.02	0.01		
Year 2	$108	$182	0.04*	0.02	0.03*	0.01		
Year 3	$116	$298	0.03	0.02	0.02	0.01		
Transfer Incentives								
Year 1	$202	$202	0.00	0.03	0.03	0.02		
Year 2	$107	$309	0.05	0.03	0.05*	0.02		
Class-Size Reduction								
Nye et al. (2000)[a]								
Year 1			0.22*	—	0.20*	—		
Year 2			0.33*	—	0.25*	—		
Year 3			0.30*	—	0.34*	—		
Year 4			0.35*	—	0.39*	—		
Schanzenbach (2006)[a]								
Year 1							0.19*	0.04
Year 2							0.19*	0.04
Year 3							0.14*	0.03
Year 4							0.15*	0.03
Harris (2009)								
Year 1	$1,812	$1,812						
Year 2	$1,812	$3,624						
Year 3	$1,812	$5,436						
Year 4	$1,812	$7,248						

Sources: Impacts and costs of pay-for-performance are based on educator and student administrative data and the Common Core of Data; impacts and costs of transfer incentives are based on Glazerman et al. (2013); impacts of class-size reduction are based on Nye et al. (2000) and Schanzenbach (2006); costs of class-size reduction are based on Harris (2009).

Notes: Costs are adjusted for inflation and expressed in May 2016 dollars. Impacts are expressed in student *z*-score units. Standard errors for class-size reduction impacts from Nye et al. (2000) are not available.

[a]Schanzenbach (2006) measured the impact of being assigned to small classes on the test scores of all study students regardless of the grade at which they entered the study. Nye et al. (2000) measured the impact of staying in small classes on the test scores of students who were part of the study since the beginning of kindergarten.

— is not available.

*Impact is statistically significant at the .05 level, two-tailed test.

- The evaluation of **pay-for-performance**, the current study, assigned schools to offer pay-for-performance bonuses or an automatic 1 percent bonus (Chapter II). After each of the four years of the policy, the study measured the impacts of pay-for-performance based on students who were tested in the study schools. From the second year onward, some students had been at their school since the beginning of the policy whereas other had not. For example, in the second year, about 30 percent of students had not attended the same school in the previous year (Appendix B, Table B.6).

For impacts to be comparable across studies, students needed to have been exposed to the same number of years of each policy. Therefore, for each duration of policy implementation, we converted the experimental impacts of pay-for-performance (measured on schools' performance) and transfer incentives (measured on teacher teams' performance) into the impacts of exposing students to the full duration of the policy. The next section provides an overview of how we modeled the impacts of student exposure to the policies, with detailed formulas at the end of this appendix. Table H.2 presents the student-level impacts (which, for pay-for-performance and transfer incentives, were based on modeling) used in the cost-effectiveness analysis.

Table H.2. Impacts of Student Exposure to Policies Examined in the Cost-Effectiveness Analysis (Student *z*-Score Units)

	Math		Reading	
Policy and Years of Exposure	Impact	Standard Error	Impact	Standard Error
Pay-for-Performance				
Cohort 1				
1 Year	0.02	0.02	0.03*	0.02
2 Years	0.05	0.03	0.03*	0.02
3 Years	0.08*	0.03	0.04	0.02
4 Years	0.03	0.05	0.04	0.04
Cohorts 1 and 2				
1 Year	0.01	0.02	0.02	0.01
2 Years	0.05*	0.02	0.03*	0.01
3 Years	0.04	0.03	0.01	0.02
Transfer Incentives				
1 Year	0.00	0.03	0.03	0.02
2 Years	0.05	0.04	0.08*	0.03
Class-Size Reduction				
1 Year	0.22*	—	0.20*	—
2 Years	0.33*	—	0.25*	—
3 Years	0.30*	—	0.34*	—
4 Years	0.35*	—	0.39*	—

Sources: Student administrative data (pay-for-performance); Glazerman et al. 2013 (transfer incentives); Nye et al. 2000 (class-size reduction).

Note: Standard errors for class-size reduction impacts from Nye et al. (2000) are not available.

— is not available.

*Impact is statistically significant at the .05 level, two-tailed test.

Calculating Costs and Impacts of Each Policy

This section explains how we calculated the cost and impact estimates used as inputs into the cost-effectiveness ratios. It provides an overview of methods used to model the impacts of specific

durations of student exposure to the policies (when those were not reported by studies) and specifies the types of costs that were and were not included in the cost estimates from each study. As discussed in Chapter VI, cost information available for each policy captured the main types of expenses required for implementation, but not all possible costs.

Costs and Impacts of Pay-for-Performance

Costs. The policy contrast examined by this study consisted of pay-for-performance compared with an automatic 1 percent bonus. Therefore, the cost that we used in the cost-effectiveness formulas was the cost of this contrast—the per-student cost of pay-for-performance in treatment schools minus the per-student cost of the automatic 1 percent bonus in control schools.

We calculated the cost of this policy contrast directly from our study data. Specifically, in each school, we took the total pay-for-performance or automatic bonus amount that teachers and principals received (from educator administrative data) and divided it by the number of students in the school (obtained from the Common Core of Data). We then took the average of the per-student cost across all treatment schools and across all control schools. Consistent with the weighting approach used for estimating impacts, we calculated these two averages weighting schools equally. The per-student cost of pay-for-performance equaled the difference between the per-student cost of bonuses in treatment and control schools. As Table H.1 shows, the annual cost of implementing pay-for-performance within the TIF program was about $100 per student.

Although these cost estimates captured the main cost drivers of the policy, they did not include two types of costs that evaluation districts might have incurred to implement this policy. First, we had no data on the administrative costs associated with calculating, explaining, and distributing bonuses. If it was more time-consuming to calculate and communicate pay-for-performance bonuses than the automatic 1 percent bonuses, we would be underestimating the cost of the policy contrast. Second, we had no data on bonuses given to nonteaching staff. Our analysis included bonuses paid to all principals and teachers (part-time, full-time, and substitutes) in the study schools. However, some districts also offered bonuses to other school staff, such as assistant principals, counselors, librarians, or custodians. Thus, our cost-effectiveness analysis assumed that bonuses to other school staff were not essential for generating the impacts found by our study. A third type of cost commonly associated with pay-for-performance, the cost of measuring educators' effectiveness, was appropriately excluded from our cost estimates. Because the evaluation districts established identical types of performance measures for educators in treatment and control schools, performance measurement did not contribute to the cost of the policy contrast between these schools.

Impacts. Our study's impact analysis estimated the impacts of specific durations of pay-for-performance implementation on schools (Chapter VI). To model the impacts of specific durations of exposure to pay-for-performance on students, we expressed impacts on schools as a weighted average of the impacts on students who had been exposed to the policy for different lengths of time. For example, the experimental impact of two years of pay-for-performance implementation on schools, measured using students at the end of Year 2 of implementation, was a weighted average of the impacts on new students (who were exposed to pay-for-performance for one year) and returning students (who were exposed to pay-for-performance for two years), with the weights being the percentage of students who were new and returning. Assuming the impact on new students was the same as the overall impact observed in Year 1—when all students in study schools were new to pay-for-performance—we solved for the impact on returning students to derive the impact of two years

of exposure to pay-for-performance. We applied a similar approach in subsequent years of implementation to model the impacts of three and four years of student exposure to pay-for-performance (see equations [6] through [13] later in this appendix for technical details).

Costs and Impacts of Transfer Incentives

Costs. The impact study report (Glazerman et al. 2013) provided information on the cost of transfer incentives. The cost estimate for transfer incentives included the costs of recruiting high-performing teachers to transfer to low-performing schools and the costs of the incentive payments (paid in installments over the study's two-year period). It did not include, however, the costs associated with identifying high-performing teachers, which the study team calculated using value-added modeling. To convert these aggregate costs into costs per student, we divided by the total number of students in the study (number of teacher teams multiplied by average number of students in each team). The cost of transfer incentives was higher in the first year ($202 per student) than in the second ($107 per student) owing to the costs incurred in the first year to recruit high-performing teachers (Table H.1).

Impacts. The study measured the impacts of transfer incentives on the average test scores of teacher teams, which taught a new cohort of students each year (Glazerman et al. 2013). The experimental impact in Year 1 of the policy's implementation captured the impact of students' exposure to one year—the first year—of the policy, and we used this impact to calculate the cost-effectiveness of one year of transfer incentives. The experimental impact in Year 2 of the policy's implementation captured the impact of another cohort's exposure to one year—the second year—of the policy. Therefore, no cohorts were actually exposed to two years of the policy. However, we modeled the impact of two years of exposure to the policy by adding the Year 1 and Year 2 impacts—that is, by assuming that those impacts would have been cumulative if the same cohort had been exposed to both the first and second years of the policy.

A key limitation of the impact estimates is that they capture impacts only within the two years in which highly effective teachers received bonuses for working in low-performing schools. Sixty percent of these teachers remained at their schools in the year after the bonuses ended (Glazerman et al. 2013), and some of these teachers may have stayed for even a longer amount of time. If the teachers who stayed continued to improve student achievement, then this policy might have longer-term effects beyond the time in which districts would need to incur costs and, therefore, be more cost-effective than suggested by our short-term analyses.

Costs and Impacts of Class-Size Reduction

Costs. We used Harris's (2009) estimate for the cost of class-size reduction, which includes the cost of paying salaries and fringe benefits to the additional teachers hired because of smaller classes, as well as capital costs associated with additional portable classrooms. This cost estimate does not include, however, the administrative costs associated with hiring the additional teachers.

The cost of class-size reduction was similar across years, reflecting the fact that the main component of the cost (salaries and benefits for additional teachers) had to be paid every year. The annual cost of reducing class size from about 22 to 15 students—the reduction that Project STAR implemented—was about $1,800 per student (Table H.1).

Impacts. We used the estimated impacts of class-size reduction reported by Nye et al. (2000). Their reanalysis of the Project STAR data estimated the effect of being in a small class for one, two, three, and four years, based on a comparison between students who remained in small classes and those who remained in regular-sized classes for the full length of each duration. These impact estimates are not the most internally valid estimates the experiment allowed. Although students' assigned class-size group was randomly determined, the actual size of the class in which they enrolled depended on decisions by principals and families. In fact, a large percentage of the students initially randomly assigned to class-size groups in kindergarten dropped out of the study before grade 3 (because they moved) or changed to a classroom with a different class-size status (because of principals' decisions or parental requests). Therefore, students who remained in small classes could have differed in unmeasured ways from students who remained in regular-sized classes, making estimates such as those reported by Nye et al. (2000) susceptible to self-selection bias. An intent-to-treat approach—which analyzes data based on the initial treatment assignment rather than the treatment received—would be less susceptible to selection bias. Numerous other analyses of the STAR data have reported these intent-to-treat estimates for class-size reduction (for example, Schanzenbach 2006; Krueger 1999). The disadvantage of using those intent-to-treat estimates for a cost-effectiveness analysis is that new students joined the study schools each year and were randomly assigned to small or regular-sized classes. Therefore, intent-to-treat estimates in each year of the policy include students who were assigned at varying times and, therefore, exposed to the policy for varying lengths of time. To facilitate comparisons to other policies, we used impacts from Nye et al. (2000) in our main cost-effectiveness analysis because those impacts measure the effect of being in a small class for a well-defined duration. Nevertheless, we present supplemental analyses using intent-to-treat estimates from Schanzenbach (2006).

Cost-Effectiveness Analysis Findings

In this section, we describe our findings from the cost-effectiveness analysis. First, we repeat the main findings on cost-effectiveness reported in Chapter VI, because those findings serve as a benchmark for the supplemental findings presented subsequently. Second, we report findings from various supplemental analyses, including analyses of cost-effectiveness separately by subject, analyses in which data for pay-for-performance are based on Cohorts 1 and 2 combined, and sensitivity analyses that present findings under alternative approaches to conducting the cost-effectiveness analysis.

Main Findings

The three policies included in our analysis were implemented with the goal of improving student achievement overall. Therefore, as discussed in Chapter VI, our main analysis compared the policies' cost-effectiveness at raising math and reading achievement averaged together.

At the end of two years, pay-for-performance and transfer incentives were similar in their cost-effectiveness. To raise student achievement by 0.04 standard deviations after two years, transfer incentives would require $193 per student, nearly identical to the $196 per-student cost of pay-for-performance (Table H.3). Transfer incentives were initially less cost-effective than pay-for-performance—though not by a statistically significant difference—due to the high start-up costs associated with recruiting high-performing teachers to transfer and the small size of the impacts observed in the first year. However, because larger impacts of transfer incentives emerged in the second year but the start-up costs did not recur, transfer incentives became as cost-effective as pay-for-performance at the end of two years.

Pay-for-performance was generally more cost-effective than class-size reduction. In each policy duration that we studied (one to four years), pay-for-performance required about one-third to two-thirds the cost that class-size reduction would have needed to achieve the same impact on student achievement (Table H.3). For example, after four years, raising student achievement by 0.04 standard deviations required spending $499 per student on pay-for-performance, but would have required spending $767 per student on class-size reduction. Differences in cost-effectiveness between pay-for-performance and class-size reduction were statistically significant in two of the four durations that we examined (two years and three years).

Table H.3. Cumulative Cost per Student Needed to Achieve the Actual Impact of Pay-for-Performance, by Years of Student Exposure to the Policy

Policy and Subject	1 Year	2 Years	3 Years	4 Years
Math and Reading Combined[a]				
Pay-for-Performance (Cohort 1)	$153	$196	$199	$499
	($98)	($90)	($80)	($617)
Transfer Incentives	$539	$193	NA	NA
	($841)	($99)		
Class-Size Reduction	$345	$497*	$668*	$767
	($64)	($70)	($96)	($98)
Math				
Pay-for-Performance (Cohort 1)	$207	$161	$150	$574
	($206)	($84)	($62)	($955)
Transfer Incentives	ND	$252	NA	NA
		($217)		
Class-Size Reduction[b]	$329	$437	$707	$806
	(—)	(—)	(—)	(—)
Reading				
Pay-for-Performance (Cohort 1)	$121	$252	$293	$442
	($59)	($120)	($165)	($473)
Transfer Incentives	$269	$156	NA	NA
	($180)	($55)		
Class-Size Reduction[b]	$362	$575	$633	$732
	(—)	(—)	(—)	(—)

Source: Impacts and costs of pay-for-performance are based on educator and student administrative data and the Common Core of Data; impacts and costs of transfer incentives are based on Glazerman et al. (2013); impacts of class-size reduction are based on Nye et al. (2000) with standard errors approximated from Schanzenbach (2006); costs of class-size reduction are based on Harris (2009).

Notes: Table shows the per-student cost of obtaining a 0.04 standard deviation impact on test scores. All costs were adjusted for inflation and expressed in May 2016 dollars. All costs and impacts were discounted at a 3 percent discount rate to their value in the first year of the policy. Standard errors are shown in parenthesis.

[a]Impacts on math and reading scores were averaged, giving equal weight to both subjects.

[b]We could not test whether the cost-effectiveness of class-size reduction differed from that of pay-for-performance separately for each subject because Nye et al. (2000) did not provide standard errors of class-size reduction impacts separately for each subject.

— is not available.

NA is not applicable because the policy was not evaluated beyond two years.

ND is not defined because the Year 1 impact of transfer incentives on math achievement was zero.

*Difference in cost-effectiveness from pay-for-performance is statistically significant at the .05 level, two-tailed test.

Findings from Supplemental Analyses

We conducted four types of supplemental analyses. First, we examined the cost-effectiveness of the three policies separately by subject. Second, we compared the cost-effectiveness of pay-for-performance based on Cohorts 1 and 2 combined with the cost-effectiveness of the alternative policies. Third, we explored the sensitivity of the main findings to using alternative discount rates. Fourth, we assessed the extent to which findings would be similar if we used the experimental impact estimates without adjusting for incomplete student exposure to the policies.

Examining cost-effectiveness separately by subject. The cost-effectiveness of a policy could vary by subject area, and policymakers might be interested in implementing a policy in only one subject area (for example, a hard-to-staff subject or a subject in which students needed improvement). When we examined the cost-effectiveness of the three policies separately for math and reading, the results were generally consistent with the main results. Compared to pay-for-performance, transfer incentives would require a somewhat higher cost ($91 more per student) to raise math achievement by 0.04 standard deviations after two years and a somewhat lower cost ($96 less per student) to raise reading achievement, but neither of those differences was statistically different (Table H.3). Across all subjects and policy durations, class-size reduction required a higher cost (an extra $122 to $557 per student) than pay-for-performance to achieve the same impact, but published information was not available to test whether those differences were statistically significant.

Estimating cost-effectiveness of pay-for-performance using Cohorts 1 and 2 combined. As in our main findings, using data from Cohorts 1 and 2 suggests that pay-for-performance was similar in cost-effectiveness to transfer incentives. To raise student achievement by 0.04 standard deviations after two years, pay-for-performance would require about the same cost per student ($177) as transfer incentives ($193; Table H.4). Pay-for-performance was more cost-effective than class-size reduction for each of the three years for which data were available for Cohorts 1 and 2, though the difference was statistically significant for only one of the years.

Table H.4. Cumulative Cost per Student Needed to Achieve the Actual Impact of Pay-for-Performance, by Years of Student Exposure to the Policy, Cohorts 1 and 2

Policy	1 Year	2 Years	3 Years
Math and Reading Combined[a]			
Pay-for-Performance (Cohorts 1 and 2)	$173 ($141)	$177 ($70)	$447 ($379)
Transfer Incentives	$539 ($841)	$193 ($99)	NA
Class-Size Reduction	$345 ($64)	$497* ($70)	$668 ($96)

Source: Impacts and costs of pay-for-performance are based on educator and student administrative data and the Common Core of Data; impacts and costs of transfer incentives are based on Glazerman et al. (2013); impacts of class-size reduction are based on Nye et al. (2000) with standard errors approximated from Schanzenbach (2006); costs of class-size reduction are based on Harris (2009).

Note: Table shows the per-student cost of obtaining a 0.04 standard deviation impact on test scores. All costs were adjusted for inflation and expressed in May 2016 dollars. All costs and impacts were discounted at a 3 percent discount rate to their value in the first year of the policy. Standard errors are shown in parenthesis.

[a]Impacts on math and reading scores were averaged, giving equal weight to both subjects.

NA is not applicable because the policy was not evaluated beyond two years.

*Difference in cost-effectiveness from pay-for-performance is statistically significant at the .05 level, two-tailed test.

Using alternative discount rates. Our main analyses used a 3 percent discount rate to discount each policy's costs and impacts in the second year and beyond. Because there is no consensus in the cost-effectiveness literature on the appropriate choice of the discount rate (Levin and McEwan 2001), we tested the sensitivity of the findings to using a rate of 0 percent (no discounting) and 10 percent (a higher real interest rate than any observed in the U.S. for the last 30 years). In Table H.5, we apply these alternative discount rates to calculate the cost effectiveness of exposing students to each policy for two years. Given that the evaluations were of relatively short length, discounting over the years studied did not meaningfully affect the results.

Table H.5. Cumulative Cost per Student Needed to Achieve the Actual Impact of Pay-for-Performance at the End of Two Years, Using Alternative Discount Rates

Policy	0 Percent Undiscounted	3 Percent Benchmark	10 Percent High-Interest
Pay-for-Performance (Cohort 1)	$197 ($91)	$196 ($90)	$194 ($89)
Transfer Incentives	$190 ($97)	$193 ($99)	$198 ($104)
Class-Size Reduction	$500* ($70)	$497* ($70)	$489* ($70)

Source: Impacts and costs of pay-for-performance are based on educator and student administrative data and the Common Core of Data; impacts and costs of transfer incentives are based on Glazerman et al. (2013); impacts of class-size reduction are based on Nye et al. (2000) with standard errors approximated from Schanzenbach (2006); costs of class-size reduction are based on Harris (2009).

Notes: Table shows per-student cost of obtaining a 0.04 standard deviation impact on test scores. All costs were adjusted for inflation and expressed in May 2016 dollars. Impacts on math and reading were averaged, giving equal weight to both subjects. Standard errors are shown in parenthesis.

*Difference in cost-effectiveness from pay-for-performance is statistically significant at the .05 level, two-tailed test.

Using experimental impact estimates. To ensure that impact estimates from different studies were comparable for a given policy duration, our main analyses adjusted the estimates so that they measured the impacts of students' exposure to the policy for the full duration. However, these adjustments relied on a number of assumptions, which we discussed earlier in this appendix. Therefore, we also explored using the experimental impacts as reported by the studies (and reproduced in Table H.1) directly in the cost-effectiveness ratios. Following the studies' random assignment designs, these impact estimates captured the effects of implementing each policy for specific numbers of years on the students (for class-size reduction), teacher teams (for transfer incentives), or schools (for pay-for-performance) that were assigned to those policies. As discussed earlier, some of the students who contributed to these experimental impact estimates had incomplete exposure to the policies. (For class-size reduction, our analyses here used the intent-to-treat estimates from Schanzenbach [2006] rather than estimates by Nye et al. [2000] of the impacts of remaining in small classes, because, as discussed earlier in this appendix, the intent-to-treat estimates aligned better with the evaluation's random assignment design.) Although these impacts were not directly comparable across studies, using experimental impacts without adjusting further for incomplete student exposure is common in the cost-effectiveness literature to date and therefore provides a point of comparison to our main approach. Because experimental impacts (on students with potentially incomplete exposure to the policies) tended to be smaller than the adjusted impacts (on students with complete exposure), the per-student cost to achieve a target impact was generally higher. Yet, when using the experimental impact estimates, the conclusions were similar to those from the main approach (Table H.6). At the end of two years, transfer incentives would have required a similar cumulative per-student cost ($252) as pay-for-performance ($217) to achieve the

same impact. However, for all policy durations of one to four years, class-size reduction would have required a higher cost than pay-for-performance to achieve the same impact, with the difference being statistically significant not only for two- and three-year policy durations (as in the main findings), but also for a four-year duration.

Table H.6. Cumulative Cost per Student Needed to Achieve the Actual Impact of Pay-for-Performance, Calculated Using Experimental Impact Estimates, by Year of Policy Implementation

Policy	Year 1	Year 2	Year 3	Year 4
Pay-for-Performance (Cohort 1)	$153 ($98)	$217 ($100)	$256 ($95)	$453 ($254)
Transfer Incentives	$539 ($841)	$252 ($170)		
Class-Size Reduction	$381 ($78)	$752* ($138)	$1,478* ($352)	$1,825* ($360)

Sources: Impacts and costs of pay-for-performance are based on educator and student administrative data and the Common Core of Data; impacts and costs of transfer incentives are based on Glazerman et al. (2013); impacts of class-size reduction are based on Schanzenbach (2006); costs of class-size reduction are based on Harris (2009).

Notes: Table shows the per-student cost of obtaining a 0.04 standard deviation impact on test scores. All costs were adjusted for inflation and expressed in May 2016 dollars. Impacts on math and reading were averaged, giving equal weight to both subjects. All costs and impacts were discounted at a 3 percent discount rate to their value in the first year of the intervention. Standard errors are shown in parenthesis.

*Difference in cost-effectiveness from pay-for-performance is statistically significant at the .05 level, two-tailed test.

Limitations of Cost-Effectiveness Analysis

The goal of a cost-effectiveness analysis is to enable policymakers to compare, and ultimately choose from, alternative policies. When using the results of any cost-effectiveness analysis to help inform the selection of policies, policymakers should take into account a number of limitations that cost-effectiveness analyses commonly face.

First, these analyses can inform policymakers about the cost-effectiveness of policies only for the length of time they were implemented and studied. Accordingly, our cost-effectiveness analysis provided insights on the short-term cost-effectiveness of the three policies (up to durations of two or four years) and could not extrapolate these findings beyond the years for which there was available evidence. As discussed in this appendix, the cost-effectiveness of pay-for-performance and class-size reduction worsened over time as impacts did not grow at the pace that costs did. On the other hand, transfer incentives became more cost-effective between the first and second year. Because the cost of achieving a specified impact did not stabilize by the end of each study, we cannot use those findings to infer how cost-effective each policy would have been had they been evaluated for longer durations. For example, transfer incentives might become more cost-effective than pay-for-performance bonuses in the years after these policies end if the teachers who received transfer incentives stay in the low-performing schools and remain effective at improving student achievement. On the other hand, if both policies remain in place over the long term, pay-for-performance may become more cost-effective than transfer incentives as principals of schools that offer pay-for-performance bonuses find new ways of advertising those bonuses to recruit more effective teachers. Answering questions such as these will require further research to evaluate the impacts of policies that remain in place over longer durations, as well as research on the lingering impacts of policies after they are no longer in place.

Second, a cost-effectiveness analysis provides relevant information to help districts choose among policies only if the costs that are included in the analysis resemble the likely costs that districts would incur when implementing their own versions of these policies. The cost information that we used for all three policies omitted particular expenses that districts in some situations might need to incur. For example, as described earlier in this appendix, we appropriately did not account for the cost of measuring educators' performance when estimating the cost of pay-for-performance because performance measures were not part of the policy contrast between the treatment and control schools, which evaluated educators in identical ways. Cost estimates for transfer incentives also did not include the cost of identifying highly effective teachers. Therefore, the results of our analysis may be most applicable to districts that already have a well-established educator evaluation system, but may be less relevant to districts that would need to revise their evaluation system substantially in order to offer pay-for-performance bonuses or transfer incentives. Likewise, the cost estimates for pay-for-performance and class-size reduction excluded the costs of certain types of human resource tasks, such as the costs of calculating and distributing bonuses (in the case of pay-for-performance) and recruiting additional teachers (in the case of class-size reduction). These tasks may add little burden to districts with well-developed human resource systems but may be a notable burden for other districts. Overall, although the cost estimates used in our analysis captured the main cost drivers of the policies, they may still have understated the costs that many districts would need to absorb in order to implement these policies.

Third, a cost-effectiveness analysis provides accurate information only if it uses valid estimates of policy impacts. Although all of the policies we considered were evaluated with rigorous random assignment designs, the impact estimates were not always directly comparable across studies because the extent of student exposure to those policies differed. To provide proper comparisons, we adjusted these estimates so that they measured impacts of the same duration of student exposure, but we had to rely on assumptions to make those adjustments.

Last, policymakers could value the benefits that policies have beyond those included in typical cost-effectiveness analyses. As with most other cost-effectiveness analyses, ours focused solely on one particular outcome—student achievement—and did not consider other potentially relevant benefits associated with each policy. For example, district officials might value teacher satisfaction and school morale beyond the potential impact those might have on test scores. Future research may have to expand the range of benefits considered in cost-effectiveness analyses and develop methods for combining those benefits in a systematic way.

Cost-Effectiveness Analysis Formulas

This section describes the technical approach for assessing the cost-effectiveness of pay-for-performance bonuses within the TIF program relative to two alternative policies: (1) transfer incentives for high-performing teachers to move to low-performing schools and (2) class-size reduction. First, we provide the formula for the cost-effectiveness ratio—the measure of cost-effectiveness on which policies were compared. Second, we derive the standard error of the cost-effectiveness ratio. Third, we describe the approach for testing whether the cost-effectiveness ratios for different policies differed statistically. Last, we model (when not available from the studies) the impacts of specific durations of students' exposure to each policy and the standard errors of those impacts.

Cost-Effectiveness Ratio

We calculated the discounted cost-effectiveness ratio after T years (CER_T) using the following formula:

$$(1)\quad CER_T = \left[\frac{\sum_{t=1}^{T} d_t AC_t}{\sum_{t=1}^{T} d_t AI_t}\right] * S,$$

where *t* is the year of the implementation (starting in Year 1 and ending in Year T), AC_t is the additional cost incurred in Year *t*; AI_t is the additional impact in Year *t*; $d_t = 1/(1+i)^{t-1}$ is the discount factor using a discount rate of *i*, and S is the scale factor which expresses the cost-effectiveness ratio in terms of the cost needed to increase test scores by S standard deviation units.

Standard Errors for the Cost-Effectiveness Ratio

We derived standard errors for the cost-effectiveness ratio using the delta method. Let the function $f(\hat{C},\hat{\gamma}) = \frac{\hat{C}}{\hat{\gamma}}$ be the estimated cost-effectiveness ratio, where $\hat{C}$ is the estimated (and discounted) cumulative cost and $\hat{\gamma}$ is the estimated (and discounted) impact. Let C and γ be the true values of the cost and impact. A Taylor series approximation of $f(\hat{C},\hat{\gamma})$ around its true value $f(C,\gamma)$ is:

$$(2)\quad f(\hat{C},\hat{\gamma}) \approx f(C,\gamma) + \frac{\partial f\left(\hat{C},\hat{\gamma}\right)}{\partial \hat{C}} \times \left(\hat{C} - C\right) + \frac{\partial f\left(\hat{C},\hat{\gamma}\right)}{\partial \hat{\gamma}} \times \left(\hat{\gamma} - \gamma\right).$$

The first term on the right is a constant with no uncertainty. This means that

$$(3)\quad Var\left[f\left(\hat{C},\hat{\gamma}\right)\right] = \left(\frac{\partial f\left(\hat{C},\hat{\gamma}\right)}{\partial \hat{C}}\right)^2 Var\left(\hat{C}\right) + \left(\frac{\partial f\left(\hat{C},\hat{\gamma}\right)}{\partial \hat{\gamma}}\right)^2 Var\left(\hat{\gamma}\right) + 2\left(\frac{\partial f\left(\hat{C},\hat{\gamma}\right)}{\partial \hat{C}}\right)\left(\frac{\partial f\left(\hat{C},\hat{\gamma}\right)}{\partial \hat{\gamma}}\right) Cov(\hat{C},\hat{\gamma})$$

$$= \left(\frac{1}{\hat{\gamma}}\right)^2 Var\left(\hat{C}\right) + \left(-\frac{\hat{C}}{\hat{\gamma}^2}\right)^2 Var\left(\hat{\gamma}\right) + 2\left(\frac{1}{\hat{\gamma}}\right)\left(-\frac{\hat{C}}{\hat{\gamma}^2}\right) Cov(\hat{C},\hat{\gamma}).$$

Because costs and impacts were based on different data sources, we assumed that $Cov(\hat{C},\hat{\gamma}) = 0$. We further assumed that costs were measured with certainty, as the standard errors of cost estimates were not reported, making $Var(\hat{C}) = 0$.

Finally, because our cost-effectiveness ratio was scaled to a constant S, we multiplied the variance by the square of the constant, yielding the final formula for the variance of the cost-effectiveness ratio:

$$(4)\ Var(CER_T) = (\frac{\hat{C}^2}{\hat{\gamma}^4})Var(\hat{\gamma}) * S^2.$$

Comparing Cost-Effectiveness Ratios Across Policies

To test the statistical significance of the difference between the cost-effectiveness ratios of pay-for-performance ($CER_{P4P,t}$) and another policy ($CER_{comparison,t}$) after *t* years, we used the following t-statistic:

$$(5)\ tstat_{P4P-comparison,t} = (\frac{CER_{P4P,t} - CER_{comparison,t}}{\sqrt{Var(CER_{P4P,t}) + Var(CER_{comparison,t})}}).$$

When the t-statistic was larger than 1.96 in absolute value, the two cost-effectiveness ratios were deemed statistically different at the .05 level.

Impacts of Student Exposure to Pay-for-Performance

The cost-effectiveness analysis required knowing the impacts of each duration of student exposure to pay-for-performance on students' achievement. However, our main impact analysis estimated the impacts of each duration of schools' implementation of pay-for-performance on the schools' average student achievement, and these schools encompassed students who had been exposed for varying lengths of time.

To model the impact of each duration of student exposure, we expressed the impact of each duration of schools' implementation (from Chapter VI) as a weighted average of the impacts of all possible durations of student exposure. We then solved for the impacts of student exposure.[68] Let $\hat{\beta}_t$ be the impact of *t* years of pay-for-performance implementation on schools' average student

[68] The method outlined here enabled us to model the impact of every possible duration of students' exposure to pay-for-performance (from one to four years). In separate analyses reported in Appendix F, Table F.20, we also directly estimated the impact of one and two years of exposure to pay-for-performance on students who were expected to stay at their schools for those durations. As explained in Appendix B, we could not directly estimate the impacts of three or four years of students' exposure to pay-for-performance because our data did not span enough grade levels to observe many students who were expected to stay at their schools for three or four years. We considered using the one- and two-year estimates from Table F.20 directly in the cost-effectiveness ratios, but decided against this approach. Although the estimates from Table F.20 had the advantage of being directly estimated from the data without the need for mathematical modeling, those estimates were based on narrow sets of student cohorts that could be observed in our data for one or two years at the same school. In contrast, the method here enabled us to model impacts on all students who were exposed for specified durations. Applying our method to all possible durations also enabled us to use a consistent approach for obtaining the impacts of one, two, three, and four years of exposure. In any case, for both subjects (reading and math) and durations (one and two years), the impacts modeled using this method (Table H.2) were similar to the direct estimates shown in Table F.20.

achievement (as reported in Chapter VI). The goal was to derive the impact of *r* years of exposure to pay-for-performance on students' achievement, denoted by $\hat{\delta}_r$, for *r* = 1 to 4. To do so, we expressed each $\hat{\beta}_t$ as a weighted average of the $\hat{\delta}_r$ impacts, with weights equal to the fractions of students exposed for those specified durations. We assumed that δ_r did not vary with *t*. For example, we assumed that the effects of a student's first year of exposure to pay-for-performance was the same regardless of whether the student's school was in its first, second, third, or fourth year of implementation.

Based on these assumptions and notation, the following four equations specify the relationships between the experimental impacts of pay-for-performance implementation on schools ($\hat{\beta}_t$) and the impacts of pay-for-performance exposure on students ($\hat{\delta}_r$):

$$(6)\quad \hat{\beta}_1 = \hat{\delta}_1$$

$$(7)\quad \hat{\beta}_2 = a_{21}\hat{\delta}_1 + a_{22}\hat{\delta}_2$$

$$(8)\quad \hat{\beta}_3 = a_{31}\hat{\delta}_1 + a_{32}\hat{\delta}_2 + a_{33}\hat{\delta}_3$$

$$(9)\quad \hat{\beta}_4 = a_{41}\hat{\delta}_1 + a_{42}\hat{\delta}_2 + a_{43}\hat{\delta}_3 + a_{44}\hat{\delta}_4.$$

In equations (6) through (9), among students enrolled in study schools at the end of Year *t* of implementation, a_{tr} is the fraction of students in treatment schools who had been exposed to *r* years of pay-for-performance. We reported estimates of a_{tr} in Appendix B, Table B.6.

By successively substituting earlier equations into later equations, we expressed each $\hat{\delta}_r$ as a function of the $\hat{\beta}_t$ impacts:

$$(10)\quad \hat{\delta}_1 = \hat{\beta}_1$$

$$(11)\quad \hat{\delta}_2 = \left(-\frac{a_{21}}{a_{22}}\right)\hat{\beta}_1 + \left(\frac{1}{a_{22}}\right)\hat{\beta}_2$$

$$(12)\quad \hat{\delta}_3 = \left(\frac{a_{21}a_{32}}{a_{22}a_{33}} - \frac{a_{31}}{a_{33}}\right)\hat{\beta}_1 + \left(-\frac{a_{32}}{a_{22}a_{33}}\right)\hat{\beta}_2 + \left(\frac{1}{a_{33}}\right)\hat{\beta}_3$$

$$(13)\quad \hat{\delta}_4 = \left(-\frac{a_{41}}{a_{44}} + \frac{a_{21}a_{42}}{a_{22}a_{44}} - \frac{a_{21}a_{32}a_{43}}{a_{22}a_{33}a_{44}} + \frac{a_{31}a_{43}}{a_{33}a_{44}}\right)\hat{\beta}_1 + \left(-\frac{a_{42}}{a_{22}a_{44}} + \frac{a_{32}a_{43}}{a_{22}a_{33}a_{44}}\right)\hat{\beta}_2 + \left(-\frac{a_{43}}{a_{33}a_{44}}\right)\hat{\beta}_3 + \left(\frac{1}{a_{44}}\right)\hat{\beta}_4.$$

Although the coefficients that multiply the $\hat{\beta}_t$ impacts in equations (11) through (13) reflect complex formulas, ultimately each coefficient is just a single number. To simplify the notation, we re-expressed those equations by substituting each coefficient formula with a single coefficient symbol c_{rt}, where c_{rt} denotes the coefficient that multiplies $\hat{\beta}_t$ in the equation for $\hat{\delta}_r$:

(14) $\hat{\delta}_1 = \hat{\beta}_1$

(15) $\hat{\delta}_2 = c_{21}\hat{\beta}_1 + c_{22}\hat{\beta}_2$

(16) $\hat{\delta}_3 = c_{31}\hat{\beta}_1 + c_{32}\hat{\beta}_2 + c_{33}\hat{\beta}_3$

(17) $\hat{\delta}_4 = c_{41}\hat{\beta}_1 + c_{42}\hat{\beta}_2 + c_{43}\hat{\beta}_3 + c_{44}\hat{\beta}_4$.

Last, within each duration of exposure, we had to account for the fact that impacts that emerged later in the duration were valued less than those that emerged earlier—a fact not reflected in $\hat{\delta}_r$, which captured the cumulative impact at the end of the duration without regard to the timing with which it emerged. To account for timing, we disaggregated each cumulative impact into the additional impact of each year (that is, the cumulative impact at the end of that year minus the cumulative impact at the end of the prior year), discounted the additional impacts beyond the first year of exposure, and then summed together the discounted additional impacts. The end result was to obtain an estimate of the cumulative, discounted impact of *r* years of exposure to pay-for-performance, which we denote by $\hat{\gamma}_r$. As before, let $d_t = 1/(1+i)^{t-1}$ be the discount factor for the additional impact in year *t* with a discount rate of *i*. This leads to the following final formulas for $\hat{\gamma}_r$:

(18) $\hat{\gamma}_1 = \hat{\delta}_1 = \hat{\beta}_1$

(19) $\hat{\gamma}_2 = \hat{\delta}_1 + d_2(\hat{\delta}_2 - \hat{\delta}_1) = (1-d_2)\hat{\delta}_1 + d_2\hat{\delta}_2 = (1-d_2)\hat{\beta}_1 + d_2(c_{21}\hat{\beta}_1 + c_{22}\hat{\beta}_2)$
$= (1-d_2+d_2c_{21})\hat{\beta}_1 + d_2c_{22}\hat{\beta}_2$

(20) $\hat{\gamma}_3 = \hat{\delta}_1 + d_2(\hat{\delta}_2 - \hat{\delta}_1) + d_3(\hat{\delta}_3 - \hat{\delta}_2) = (1-d_2)\hat{\delta}_1 + (d_2-d_3)\hat{\delta}_2 + d_3\hat{\delta}_3$
$= (1-d_2)\hat{\beta}_1 + (d_2-d_3)(c_{21}\hat{\beta}_1 + c_{22}\hat{\beta}_2) + d_3(c_{31}\hat{\beta}_1 + c_{32}\hat{\beta}_2 + c_{33}\hat{\beta}_3)$
$= (1-d_2+d_2c_{21}-d_3c_{21}+d_3c_{31})\hat{\beta}_1 + (d_2c_{22}-d_3c_{22}+d_3c_{32})\hat{\beta}_2 + d_3c_{33}\hat{\beta}_3$

$$
\begin{aligned}
(21)\ \hat{\gamma}_4 &= \hat{\delta}_1 + d_2(\hat{\delta}_2 - \hat{\delta}_1) + d_3(\hat{\delta}_3 - \hat{\delta}_2) + d_4(\hat{\delta}_4 - \hat{\delta}_3) \\
&= (1-d_2)\hat{\delta}_1 + (d_2 - d_3)\hat{\delta}_2 + (d_3 - d_4)\hat{\delta}_3 + d_4\hat{\delta}_4 \\
&= (1-d_2)\hat{\beta}_1 + (d_2 - d_3)(c_{21}\hat{\beta}_1 + c_{22}\hat{\beta}_2) + (d_3 - d_4)(c_{31}\hat{\beta}_1 + c_{32}\hat{\beta}_2 + c_{33}\hat{\beta}_3) \\
&\quad + d_4(c_{41}\hat{\beta}_1 + c_{42}\hat{\beta}_2 + c_{43}\hat{\beta}_3 + c_{44}\hat{\beta}_4) \\
&= (1 - d_2 + d_2c_{21} - d_3c_{21} + d_3c_{31} - d_4c_{31} + d_4c_{41})\hat{\beta}_1 \\
&\quad + (d_2c_{22} - d_3c_{22} + d_3c_{32} - d_4c_{32} + d_4c_{42})\hat{\beta}_2 \\
&\quad + (d_3c_{33} - d_4c_{33} + d_4c_{43})\hat{\beta}_3 + d_4c_{44}\hat{\beta}_4.
\end{aligned}
$$

Equations (18) through (21) depend on the study-reported experimental impact estimates ($\hat{\beta}_t$), which are subject-specific. We combined impacts on math and reading by replacing each $\hat{\beta}_t$ on the right side of the equations with an equal-weighted average of the math and reading impacts, $(1/2)(\hat{\beta}_t^{math} + \hat{\beta}_t^{read})$.

Standard Errors of the Estimated Impacts of Student Exposure to Pay-for-Performance

As shown in equation (4), the variance (or squared standard error) of the cost-effectiveness ratio depends on the variance of the impact estimate. It was therefore important to estimate the variance of each of the $\hat{\gamma}_r$ estimates from equations (18) through (21). Each $\hat{\gamma}_r$ estimate—regardless of whether it was subject-specific or combined across subjects—is a linear combination of the study-reported impacts. Therefore, the variance of each $\hat{\gamma}_r$ estimate is a linear combination of the variances of the study-reported impacts and the covariances among them, all of which we directly obtained from the impact estimation results. We used Stata's lincom command to compute the appropriate linear combinations of these variances and covariances to estimate $Var(\hat{\gamma}_r)$.

Impacts of Student Exposure to Transfer Incentives

To make the cost-effectiveness findings comparable across policies, we inferred what the impact of each policy would be on students exposed to it for the duration of the study. In the two-year transfer incentives study, each teacher team worked with an entirely new cohort of students in the second year than it had in the first, meaning that each student was exposed to the policy for at most one year. We modeled what the impact of two years of student exposure to transfer incentives would have been by assuming that the Year 1 and Year 2 impacts from Glazerman et al. (2013)—estimated on two separate cohorts of students—would be cumulative if estimated on the same cohort of students. That is, the impact on students of being exposed to two years of transfer incentives is the sum of the Year 1 and Year 2 impacts, with the Year 2 impacts discounted to its value in Year 1. As before, let $\hat{\beta}_t$ be the study-reported impact at the end of Year *t* of implementation; $\hat{\gamma}_r$ be the discounted, cumulative impact of *r* years of exposure to the policy; and d_t be the discount factor in Year *t*. Under the assumptions described above,

$$(22)\ \hat{\gamma}_1 = \hat{\beta}_1$$

$$(23)\ \hat{\gamma}_2 = \hat{\beta}_1 + d_2\hat{\beta}_2.$$

Like the impacts of pay-for-performance reported by our evaluation, the impacts of transfer incentives reported by Glazerman et al. (2013) were subject-specific. To combine impacts on math and reading, we replaced each $\hat{\beta}_t$ on the right side of equations (22) and (23) with an equal-weighted average of the math and reading impacts, $(1/2)(\hat{\beta}_t^{math} + \hat{\beta}_t^{read})$.

Standard Errors of the Estimated Impacts of Student Exposure to Transfer Incentives

Within each subject separately, all of the information needed to estimate the variances of the estimators in equations (22) and (23) was available from Glazerman et al. (2013). The estimated impact of one year of student exposure (equation [22]) was directly reported by the study, so its variance was just the variance of the study-reported Year 1 impact. The estimated impact of two years of student exposure (equation [23]) was a linear combination of statistically independent impact estimates reported by the study ($\hat{\beta}_1$ and $\hat{\beta}_2$, estimated from two separate cohorts of students). Therefore, its variance was a linear combination of the variances of the study-reported impact estimates:

$$(24)\ Var(\hat{\gamma}_2) = Var(\hat{\beta}_1) + d_2^2 Var(\hat{\beta}_2).$$

However, estimating the variance of the $\hat{\gamma}_1$ impacts for math and reading combined required some information that was not available from the study report. Let $\hat{\gamma}_r^{combined} = (1/2)(\hat{\gamma}_r^{math} + \hat{\gamma}_r^{read})$ denote the estimated impact of *r* years of student exposure to transfer incentives on math and reading achievement averaged together. The variance of $\hat{\gamma}_r^{combined}$ is given by:

$$\begin{aligned}(25)\ Var(\hat{\gamma}_r^{combined}) &= (1/4)[Var(\hat{\gamma}_r^{math}) + Var(\hat{\gamma}_r^{read}) + 2 * Cov(\hat{\gamma}_r^{math}, \hat{\gamma}_r^{read})] \\ &= (1/4)[Var(\hat{\gamma}_r^{math}) + Var(\hat{\gamma}_r^{read}) + 2\rho SE(\hat{\gamma}_r^{math}) SE(\hat{\gamma}_r^{read})],\end{aligned}$$

where ρ is the sampling correlation between the math and reading impact estimates. An estimate for ρ could have been obtained if the study had reported information on the covariance between its math and reading impact estimates, but it did not provide this information.

Therefore, we had to impute ρ. To do so, we estimated the sampling correlation between the math and reading pay-for-performance impacts, averaged across the four years of the intervention. In each implementation year *t*, let $\hat{\beta}_t^{pfp,math}$ and $\hat{\beta}_t^{pfp,read}$ be the study-reported impacts of pay-for-performance in math and reading, respectively. We calculated:

$$(26)\ \hat{\rho} = (1/4)\sum_{t=1}^{4} \frac{Cov(\hat{\beta}_t^{pfp,math}, \hat{\beta}_t^{pfp,read})}{SE(\hat{\beta}_t^{pfp,math}) SE(\hat{\beta}_t^{pfp,read})}.$$

Under the assumption that pay-for-performance and transfer incentives gave rise to a similar correlation between math and reading impacts, we used the estimate $\hat{\rho}$ from equation (26) in the variance formula in equation (25).

Impacts of Student Exposure to Class-Size Reduction

Unlike the evaluations of pay-for-performance and transfer incentives, the evaluation of class-size reduction—Project STAR—randomly assigned students to the treatment and control groups and followed those students for the duration of the study. Therefore, we did not have to model the impacts of specific durations of student exposure to the policy, as those were already available in the literature. We used impact estimates from Nye et al. (2000), who measured the effect of being in a small class for one, two, three, and four years on the achievement of students who remained in small classes for the entire duration of the study.

The only adjustment that we had to make to the impacts reported by Nye et al. (2000) was to apply discounting. Our objective was to derive $\hat{\gamma}_r$, the estimate of the cumulative, discounted impact of *r* years of exposure to class-size reduction on students' achievement. As before, let $\hat{\beta}_t$ be the study-reported impacts at the end of Year *t* of implementation and $d_t = 1/(1+i)^{t-1}$ be the discount factor for the additional impact in Year *t* with a discount rate of *i*. This leads to the following final formulas for $\hat{\gamma}_r$:

(27) $\hat{\gamma}_1 = \hat{\beta}_1$

(28) $\hat{\gamma}_2 = \hat{\beta}_1 + d_2(\hat{\beta}_2 - \hat{\beta}_1)$

(29) $\hat{\gamma}_3 = \hat{\beta}_1 + d_2(\hat{\beta}_2 - \hat{\beta}_1) + d_3(\hat{\beta}_3 - \hat{\beta}_2)$

(30) $\hat{\gamma}_4 = \hat{\beta}_1 + d_2(\hat{\beta}_2 - \hat{\beta}_1) + d_3(\hat{\beta}_3 - \hat{\beta}_2) + d_4(\hat{\beta}_4 - \hat{\beta}_3)$.

Like the impacts of pay-for-performance reported by our evaluation, the impacts of class-size reduction reported by Nye et al. (2000) were subject-specific. To combine impacts on math and reading, we replaced each $\hat{\beta}_t$ on the right side of equations (27) through (30) with an equal-weighted average of the math and reading impacts, $(1/2)(\hat{\beta}_t^{math} + \hat{\beta}_t^{read})$.

Standard Errors of the Estimated Impacts of Student Exposure to Class-Size Reduction

Because standard errors for the impacts reported by Nye et al. (2000) were not available, we approximated them with those reported by Schanzenbach (2006). To be able to apply the standard errors from Schanzenbach (2006) to the impacts from Nye et al. (2000), we had to account for two notable differences in the studies' estimation approaches. First, whereas Nye et al. (2000) reported impact estimates separately for math and reading (which we subsequently averaged in our main analyses), Schanzenbach (2006) estimated impacts averaged only across the two subjects. Therefore, we could apply Schanzenbach's (2006) standard errors only to cost-effectiveness analyses for math and reading combined—not to the supplemental analyses that examined math and reading separately. Second, the two studies' estimates pertained to different samples of students. In each year of the study, Nye et al. (2000) compared students who had stayed in small classes since the beginning of the study with students who had stayed in regular-sized classes for the same length of time—a stayer analysis. Instead, Schanzenbach (2006) compared students assigned to small versus

regular-sized classes—an intent-to-treat analysis—which included students who entered study schools after the beginning of the study (and were randomly assigned at the time of entering) and those who changed classrooms non-randomly. Because Nye and colleagues used a smaller sample than Schanzenbach, their estimates were expected to be less precise.

Therefore, applying the standard errors from Schanzenbach (2006) to the impacts from Nye et al. (2000) required accounting for differences in the studies' sample sizes. Although neither Nye et al. (2000) nor Schanzenbach (2006) reported sample sizes for the specific estimates we used, their sample sizes ought to have differed as a result of two main factors that could be quantified: attrition (departure from the study) and crossover (switching between class-size groups). First, sample sizes for Schanzenbach's (2006) intent-to-treat analyses were not expected to differ substantially across years despite the presence of attrition, because new students who were randomly assigned to class-size groups replaced departing students. (In fact, Krueger [1999] also conducted intent-to-treat analyses and reported sample sizes that did not differ by more than 10 percent across years.) However, attrition should lead to year-to-year reductions in the sample sizes used by Nye et al. (2000), whose analysis of students who stayed in their assigned class sizes did not include new students who arrived after the beginning of the study. We derived the year-to-year attrition rate based on information from Krueger (1999), who reported that about 50 percent of students who entered the study in kindergarten remained in a study classroom through the end of 3rd grade. This means that in each grade transition, about 80 percent of students (cube root of 0.5) stayed in a study classroom. Second, students who switched to a different class-size group between years remained in the sample for the intent-to-treat analysis of Schanzenbach (2006) but not in the sample for the stayer analysis of Nye et al. (2000). Of students who did not leave the study between years, Krueger (1999) reported that 94 percent stayed in their assigned class type. On net, when taking into account both attrition and crossover, the sample size for the stayer analysis was expected to be about 0.8*0.94 = 0.75 times (or three-fourths) the sample size in the previous year, whereas the sample size for the intent-to-treat analysis was expected to be similar across years.

Based on these attrition and crossover rates, we inflated the standard errors reported by Schanzenbach (2006) for the second year and beyond to account for the expected smaller sample size used by Nye et al. (2000). Because standard errors are inversely proportional to the square root of the sample size, we multiplied the standard errors from Schanzenbach (2006) by the following adjustment factor:

$$(31)\quad Adj\_factor = \frac{1}{\sqrt{(0.75)^{t-1}}}$$

for each year t=2, 3, 4.

Note that these standard errors do not account for discounting. Although we used discounted impact estimates (see equations [27] through [30]), we could not derive standard errors for class-size reduction that took into account discounting, as that would require knowing the sampling covariance of the study-reported impact estimates across years, for which we had no information. We also could not impute this sampling covariance based on the sampling covariance of pay-for-performance impacts across years, because the sampling covariance of impact estimates across years depends on the degree to which analysis samples remain consistent over time, a factor that differed across studies.

THIS PAGE IS INTENTIONALLY BLANK

THIS PAGE IS INTENTIONALLY BLANK


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA