Evaluation of the Teacher Incentive Fund: Implementation and Impacts of Pay-for-Performance After Three Years

August 2016

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: Implementation and Impacts of Pay-for-Performance After Three Years

August 2016

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

Elizabeth Warner
Project Officer
Institute of Education Sciences

NCEE 2016-4004
U.S. DEPARTMENT OF EDUCATION

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

U.S. Department of Education
John B. King, Jr.
Secretary

Institute of Education Sciences
Ruth Neild
Deputy Director for Policy and Research
Delegated Duties of the Director

National Center for Education Evaluation and Regional Assistance
Joy Lesnick
Acting Commissioner

August 2016

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.

IES evaluation reports present objective information on the conditions of implementation and impacts of the programs being evaluated. IES evaluation reports do not include conclusions or recommendations or views with regard to actions policymakers or practitioners should take in light of the findings in the reports.

Wellington, Alison, Hanley Chiang, Kristin Hallgren, Cecilia Speroni, Mariesa Herrmann, Paul Burkander. (2016). *Evaluation of the Teacher Incentive Fund: Implementation and Impacts of Pay-for-Performance After Three Years* (NCEE 2016-4004). 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, Dylan Ellis, Kai Filipczak, Chris Jones, William Leith, Serge Lukashanets, Mickey McCauley, Jeremy Page, 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. Tora Davis helped with a variety of critical tasks ranging from tracking data collection activities to 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, Dylan Ellis, Kai Filipczak, and John Hotchkiss. Leah Hackleman-Good edited the report, and Jill Miller carefully and patiently prepared the report for publication.

THIS PAGE IS INTENTIONALLY BLANK

CONTENTS

EXECUTIVE SUMMARY xxi

I INTRODUCTION 1

II STUDY SAMPLE, DESIGN, DATA, AND METHODS 9

III PROGRAMS AND EXPERIENCES OF ALL 2010 TIF DISTRICTS 25

IV TIF IMPLEMENTATION IN EVALUATION DISTRICTS 33

V IMPACTS OF PAY-FOR-PERFORMANCE ON EDUCATORS' ATTITUDES AND BEHAVIORS 67

VI IMPACTS OF PAY-FOR-PERFORMANCE ON EDUCATOR EFFECTIVENESS AND STUDENT ACHIEVEMENT 79

REFERENCES 95

APPENDIX A: SUPPLEMENTAL INFORMATION ON STUDY SAMPLE, DESIGN, DATA, AND METHODS FOR CHAPTER II A.1

APPENDIX B: SUPPLEMENTAL INFORMATION ON ANALYTIC METHODS FOR CHAPTER II B.1

APPENDIX C: SUPPLEMENTAL FINDINGS ON PROGRAMS AND EXPERIENCES OF ALL TIF DISTRICTS FOR CHAPTER III C.1

APPENDIX D: SUPPLEMENTAL FINDINGS ON TIF IMPLEMENTATION IN EVALUATION DISTRICTS FOR CHAPTER IV D.1

APPENDIX E: SUPPLEMENTAL FINDINGS ON IMPACTS OF PAY-FOR-PERFORMANCE ON EDUCATORS' ATTITUDES AND BEHAVIORS FOR CHAPTER V E.1

APPENDIX F: SUPPLEMENTAL FINDINGS ON IMPACTS OF PAY-FOR-PERFORMANCE ON EDUCATOR EFFECTIVENESS AND STUDENT ACHIEVEMENT FOR CHAPTER VI F.1

APPENDIX G: SUPPLEMENTAL FINDINGS ON FACTORS ASSOCIATED WITH DIFFERENCES IN IMPACTS FOR CHAPTER VI G.1

THIS PAGE IS INTENTIONALLY BLANK

TABLES

ES.1 Districts' Reported Implementation of TIF Required Components for Teachers in Year 3 (Percentages) xxvii

II.1 Number of Districts Implementing TIF, by Year 9

II.2 Comparison of TIF Evaluation Districts and Non-Evaluation Districts (Percentages Unless Otherwise Noted) 11

II.3 Number of Schools in the Evaluation, by Cohort and Treatment Status 14

II.4 Characteristics of Students Enrolled in Treatment and Control Schools in the Pre-Implementation School Year (2010–2011) (Percentages Unless Otherwise Indicated) 15

II.5 Characteristics of Educators in Treatment and Control Schools in Year 1 (Percentages Unless Otherwise Noted) 16

II.6 Data Sources for This Report 17

III.1 TIF Districts' Reported Implementation of TIF Required Components for Teachers and Principals (Percentages) 26

III.2 Staff Eligibility for Pay-for-Performance Bonus, Year 3 (Percentages) 29

III.3 Additional Pay Opportunities for Teachers and Principals, Year 3 29

III.4 Planned Professional Development Activities for Teachers, Year 3 (Percentages) 30

IV.1 Evaluation Districts' Reported Implementation of TIF Required Components for Teachers and Principals (Percentages) 35

IV.2 Measures of Student Achievement and Observations of Practices Used to Evaluate Teachers and Principals, as Reported by Evaluation Districts, Year 3 (Percentages) 36

IV.3 Key Features of Evaluation Districts' Teacher Pay-for-Performance Bonus Programs in Year 3 41

IV.4 Evaluation Districts Meeting TIF Grant Goals for Pay-for-Performance Bonuses for Teachers (Percentages) 42

IV.5 Evaluation Districts Meeting TIF Grant Goals for Pay-for-Performance Bonuses for Principals, (Percentages) 48

IV.6 Additional Pay Opportunities, as Reported by Evaluation Districts, Year 3 50

IV.7 Professional Development Activities for Teachers Planned Under TIF, as Reported by Evaluation Districts, Year 3 (Percentages) 51

IV.8 Districts' Communication Activities in Year 3 (Percentages Unless Otherwise Noted) 52

IV.9 Information Districts Provided to Teachers About Actual Pay-for-Performance Bonuses from Year 2 (Percentages) 53

IV.10 Teachers' Reports of the Measures Used to Evaluate Teachers (Percentages) 56

IV.11 Principals' Reports of the Measures Used to Evaluate Principals (Percentages) 56

IV.12 Actual and Reported Receipt of Pay-for-Performance Bonus from Year 2 for Teachers in Treatment Schools in Year 3 61

IV.13 Treatment Teachers' Reported Eligibility for Pay-for-Performance Bonuses and Reported Maximum Bonuses in Year 3, by Selected District and Teacher Characteristics 64

IV.14 Eligibility for Additional Pay Opportunities, as Reported by Teachers and Principals (Percentages) 65

V.1 Teachers' Satisfaction with Professional Opportunities, Evaluation System, and School Environment (Percentages Who Are "Somewhat" or "Very" Satisfied) 69

V.2 Principals' Satisfaction with Professional Opportunities, Evaluation System, and School Environment (Percentages Who Are "Somewhat" or "Very" Satisfied) 70

V.3 Teachers' Attitudes Toward TIF Program (Percentages Who "Agree" or "Strongly Agree") 72

V.4 Principals' Attitudes Toward TIF Program (Percentage Who "Agree" or "Strongly Agree") 73

V.5 Treatment Teachers' Attitudes Toward TIF Program by Bonus Receipt and Report of Bonus Receipt, Year 3 (Percentages Who "Agree" or "Strongly Agree") 75

V.6 Incentives Used to Recruit Teachers (Percentages Who Reported They Were "Always" or "Often" Used) 76

V.7 Teaching Vacancies and Hiring Experiences (Averages Unless Otherwise Noted) 77

VI.1 Student Achievement Growth Ratings (Points on 1-to-4 Scale) 82

VI.2 Observation Ratings for Teachers and Principals (Points on 1-to-4 Scale) 83

VI.3 Impacts of Pay-for-Performance on the Performance Ratings of Returning and Newly Hired Teachers (Points on 1-to-4 Scale) 85

VI.4 Student Achievement in Math and Reading (Student *z*-Score Units) 87

A.1 School Attrition, Cohorts 1 and 2 (Percentages Unless Otherwise Noted) A.4

A.2 Characteristics of Students Enrolled in Treatment and Control Schools in the Pre-Implementation Year, Cohorts 1 and 2 (Percentages Unless Otherwise Noted) A.5

A.3 Characteristics of Educators in Treatment and Control Schools in Year 1, Cohorts 1 and 2 (Percentages Unless Otherwise Noted) A.6

A.4 Characteristics of Educators in Treatment and Control Schools in the Pre-Implementation Year, Cohort 1 (Percentages Unless Otherwise Noted) A.7

A.5 District Survey Response Rates Overall and by Evaluation Status, 2013–2014 School Year, Cohorts 1 and 2 ..... A.9

A.6 District Characteristics by Districts' Response Status, 2013–2014 School Year, Cohorts 1 and 2 (Percentages Unless Otherwise Noted) A.10

A.7 Teacher and Principal Response Rates for the Final Analyses Samples, Cohort 1 ..... A.11

A.8 Teacher and Principal Response Rates for the Final Analyses Samples, Cohort 2 ..... A.12

A.9 Teacher Respondents, by Teaching Assignment and Treatment Status, Cohort 1 ..... A.13

A.10 Characteristics of Teacher Survey Respondents and Nonrespondents, Cohort 1 (Percentages Unless Otherwise Noted) A.14

A.11 Characteristics of Teacher Survey Respondents by Treatment Status, Cohort 1 (Percentages Unless Otherwise Noted) A.15

A.12 Characteristics of Principal Survey Respondents by Treatment Status, Cohort 1 (Percentages Unless Otherwise Noted) A.15

A.13 Number of Full-Time Principals Listed in the Administrative Data and the Number of Schools in Which They Worked, Cohort 1 ..... A.16

A.14 Teachers Who Had Performance Ratings, Cohort 1 (Percentages)..... A.17

A.15 Principals Who Had Observation Ratings, Cohort 1 (Percentages) A.18

A.16 Characteristics of Teachers with and Without Classroom Observation Ratings, Cohort 1 (Percentages Unless Otherwise Noted) A.19

A.17 Characteristics of Teachers with and Without Classroom Achievement Growth Ratings, Cohort 1 (Percentages Unless Otherwise Noted) A.20

A.18 Characteristics of Principals with and Without Observation Ratings, Cohort 1 (Percentages Unless Otherwise Noted) A.21

A.19 Characteristics of Teachers with Classroom Observation Ratings, Cohort 1 (Percentages Unless Otherwise Noted) A.22

A.20 Characteristics of Teachers with Classroom Achievement Growth Ratings, Cohort 1 (Percentages Unless Otherwise Noted) A.23

A.21 Characteristics of Principals with Observation Ratings, Cohort 1 (Percentages Unless Otherwise Noted) A.24

A.22 Students Who Had Test Scores, Cohort 1 (Percentages)..... A.25

A.23 Characteristics of Students Who Did and Did Not Have Math Test Scores, Cohort 1 (Percentages Unless Otherwise Noted) A.26

A.24 Characteristics of Students Who Did and Did Not Have Reading Test Scores, Cohort 1 (Percentages Unless Otherwise Noted) A.27

A.25 Characteristics of Students in the Math Analysis Sample, Cohort 1 (Percentages Unless Otherwise Noted) A.28

A.26 Characteristics of Students in the Reading Analysis Sample, Cohort 1 (Percentages Unless Otherwise Noted) A.29

B.1 Test Scores That Were Dropped or Recoded, Cohort 1 (Percentages)..... B.5

B.2 Students in the Math Analysis Sample with Missing Covariate Data (Percentages)..... B.9

B.3 Students in the Reading Analysis Sample with Missing Covariate Data (Percentages) B.10

B.4 Teachers in the Educator Retention Analysis Sample in Year 1 with Missing Covariate Data, Cohort 1 (Percentages) B.11

B.5 Principals in the Educator Retention Analysis Sample in Year 1 with Missing Covariate Data, Cohort 1 (Percentages) B.11

B.6 Realized Values of Minimum Detectable Impacts B.18

C.1 Observations of Classroom or School Practices to Evaluate Teachers and Principals, Year 3 (Percentages Unless Otherwise Noted) C.3

C.2 Additional Pay Opportunities for Teachers and Principals for Additional Factors, Year 3 ..... C.4

C.3 Challenges Implementing TIF in Year 2 and Year 3 (Percentages) C.5

D.1 Classroom Observations to Evaluate Teachers in Year 3, Cohort 1 (Percentages Unless Otherwise Noted) D.4

D.2 Comparison of Principals' Ratings on Observations and School Achievement Growth in Year 3, Cohort 1 (Percentages) D.4

D.3 Comparison of Teachers' Ratings on Classroom Observations and Classroom Achievement Growth in Year 3, Cohort 1 (Percentages) D.4

D.4 Comparison of Teachers' School Achievement Growth Ratings in Years 2 and 3, Cohort 1 (Percentages) D.5

D.5 Comparison of Teachers' Classroom Achievement Growth Ratings in Years 2 and 3, Cohort 1 (Percentages) D.5

D.6 Key Features of Evaluation Districts' Teacher Pay-for-Performance Bonus Programs in Year 3, Cohorts 1 and 2 ..... D.6

D.7 Detailed Information on Measures and Criteria Used for Evaluation Districts' Teacher Pay-for-Performance Bonus Programs in Year 3, Cohorts 1 and 2 ..... D.7

D.8 Evaluation Districts Meeting TIF Grant Goals for Pay-for-Performance Bonuses for Teachers in Years 1 and 2, Cohorts 1 and 2 (Percentages) D.13

D.9 Comparison of Teachers' Performance Bonus Amounts in Years 2 and 3, Cohort 1 ..... D.17

D.10 Average and Maximum Amounts of Additional Pay Opportunities for Teachers, Cohort 1 ..... D.21

D.11 Actual Amounts of Teachers' Additional Pay, Cohort 1 ..... D.21

D.12 Professional Development Based on Observation and Student Achievement Growth Ratings, As Reported by Evaluation Districts in Year 3, Cohort 1 (Percentages Unless Otherwise Noted) D.22

D.13 Districts' Activities to Communicate to Teachers Their Classroom Observation and Student Achievement Growth Ratings for Year 2, Cohort 1 (Percentages) D.22

D.14 Communication Methods Used to Inform Teachers About Individual Pay-for-Performance Bonuses Based on the Second Year of TIF Implementation (Percentages) D.22

D.15 Bonus Eligibility as Reported by Teachers and Principals, Cohort 1 (Percentages) D.25

D.16 Educators' Reports of the Maximum Possible Bonus Amount: Imputed and Non-Imputed Bonus Amounts, Cohort 1 ..... D.28

D.17 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 ..... D.31

D.18 Treatment Teachers' Reported Eligibility for Pay-for-Performance Bonuses and Reported Maximum Bonuses in Year 3, by Districts' Characteristics, Cohort 1 (Percentages) D.32

D.19 Treatment Teachers' Reported Eligibility for Pay-for-Performance Bonuses and Reported Maximum Bonuses in Year 3, by Principal Understanding and Teacher Characteristics, Cohort 1 (Percentages) D.33

D.20 Professional Development Teachers Reported Receiving or Expecting to Receive During Year 3, Cohort 1 (Percentages) D.34

D.21 Hours of Expected Professional Development in Year 3, as Reported by Teachers, Cohort 1 (Averages) D.34

E.1 Teachers' Satisfaction with Professional Opportunities, Evaluation System, and School Environment, Cohorts 1 and 2 (Percentages Who Are "Somewhat" or "Very" Satisfied) E.4

E.2 Principals' Satisfaction with Professional Opportunities, Evaluation System, and School Environment, Cohorts 1 and 2 (Percentages Who Are "Somewhat" or "Very" Satisfied) E.4

E.3 Teachers' Attitudes Toward TIF Program, Cohorts 1 and 2 (Percentages Who "Agree" or "Strongly Agree") E.5

E.4 Principals' Attitudes Toward TIF Program, Cohorts 1 and 2 (Percentage Who "Agree" or "Strongly Agree") E.6

E.5 Impacts of Pay-for-Performance on Teacher Satisfaction Measures for Teacher Subgroups, Year 3, Cohort 1 (Percentage Points) E.8

E.6 Treatment Teachers' Satisfaction by Bonus Receipt and Report of Bonus Receipt, Year 3, Cohort 1 (Percentages who "Agree" or "Strongly Agree") E.9

E.7 Impacts of Pay-for-Performance on Teacher Attitude Measures for Teacher Subgroups, Year 3, Cohort 1 (Percentage Points) E.10

E.8 Principals' Autonomy in Hiring Teachers, Cohort 1 (Percentages) E.11

E.9 Criteria Used for Teacher Assignments to Grade Levels or Subject Areas, Cohort 1 (Percentages Who Report They Are "Always" or "Often" Used) E.12

E.10 Nonmonetary Benefits Used to Recognize Teachers' Performance or Responsibilities, Cohort 1 (Percentages) E.13

E.11 Teachers' Time Spent on School-Related Activities in the Most Recent Full Week (Average Hours) E.14

F.1 Cluster and School Attrition in the Analysis of the Impacts of Pay-for-Performance on Educator Effectiveness and Student Achievement, Cohort 1 ..... F.4

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) F.6

F.3 Impacts of Pay-for-Performance on School Achievement Growth Ratings in Year 3 Using Alternative Specifications, Cohort 1 (Points on 1-to-4 Scale) F.7

F.4 Impacts of Pay-for-Performance on Teachers' Classroom Observation Ratings in Year 3 Using Alternative Specifications, Cohort 1 (Points on 1-to-4 scale) F.7

F.5 Student Achievement Growth Ratings in Years 1 and 2, Cohorts 1 and 2 (Points on 1-to-4 Scale) F.8

F.6 Observation Ratings for Teachers and Principals in Years 1 and 2, Cohorts 1 and 2 (Points on 1-to-4 Scale) F.9

F.7 Student Achievement Growth Ratings, Consistent Sample of Schools, Cohort 1 (Points on 1-to-4 Scale) F.10

F.8 Teachers Who Continued Teaching in the Same School Across Multiple Years, Cohort 1 (Percentages) F.11

F.9 Principals Who Continued Leading the Same School Across Multiple Years, Cohort 1 (Percentages) F.11

F.10 Characteristics of Teachers and Principals in Year 3, Cohort 1 (Percentages Unless Otherwise Noted) F.12

F.11 Performance Ratings in Year 3 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) F.13

F.12 Observation and School Achievement Growth Ratings of Returning and Newly Hired Principals, Cohort 1 (Points on 1-to-4 Scale) F.13

F.13 Performance Ratings in Year 3 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) F.14

F.14 Impacts of Pay-for-Performance on Student Achievement in Math Using Alternative Specifications in Year 3, Cohort 1 (Student *z*-Score Units) F.15

F.15 Impacts of Pay-for-Performance on Student Achievement in Reading Using Alternative Specifications in Year 3, Cohort 1 (Student *z*-Score Units) F.16

F.16 Student Achievement in Math and Reading, Cohorts 1 and 2 (Student *z*-Score Units) F.17

F.17 Student Achievement in Math and Reading in Elementary and Middle Grades, Cohort 1 (Student *z*-Score Units) F.20

F.18 Impacts of Pay-for-Performance on District-Constructed and Study-Constructed Measures of School Achievement Growth, Cohort 1 ..... F.21

G.1 Program and Implementation Characteristics Used for Subgroup Analyses G.4

G.2 Differences in the Impacts of Pay-for-Performance on Student Achievement in Year 3 Between Subgroups Based on Districts' Program Characteristics (Student *z*-score Units) G.9

G.3 Association Between Continuous Measures of Program Characteristics and the Impacts of Pay-for-Performance on Student Achievement in Year 3 ..... G.10

G.4 Measures of Educator Behaviors for Explaining Differences in Impacts on Student Achievement G.11

G.5 Association Between Impacts on Educator Behaviors and Impacts on Student Achievement in Year 3 ..... G.12

THIS PAGE IS INTENTIONALLY BLANK

FIGURES

ES.1 Random Assignment Evaluation Design xxv

ES.2 Average Student Achievement in Treatment and Control Schools (Percentiles) xxix

ES.3 Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers and Principals xxxi

ES.4 Teachers and Principals in Treatment Schools Who Reported Being Eligible for Pay-for-Performance Bonuses (Percentages) xxxiii

ES.5 Reported and Actual Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools xxxiii

I.1 Logic Model 6

II.1 Two Cohorts of Evaluation TIF Districts 12

II.2 Random Assignment Design 13

III.1 Measures of Student Achievement and Observations Used to Evaluate Teachers and Principals, All TIF Districts, Year 3 (Percentages) 28

III.2 Major Challenges in Implementing TIF (Percentages) 31

IV.1 Comparison of Teachers' Ratings on Classroom Observations and School Achievement Growth in Year 3 (Percentages) 38

IV.2 Comparison of Teachers' Classroom Observation Ratings in Years 2 and 3 (Percentages) 39

IV.3 Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools 43

IV.4 Distribution of Pay-for-Performance Bonuses for Teachers in Treatment Schools 44

IV.5 Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools in Year 3, by District 45

IV.6 Minimum, Average, and Maximum Performance Bonus for Each Performance Measure, in Districts Using Classroom Achievement Growth Measures, Year 3 46

IV.7 Minimum, Average, and Maximum Performance Bonus for Each Performance Measure, in Districts Not Using Classroom Achievement Growth Measures, Year 3 47

IV.8 Minimum, Average, and Maximum Pay-for-Performance Bonuses for Principals in Treatment Schools 49

IV.9 Teachers' Bonus Eligibility, as Reported by Teachers (Percentages) 57

IV.10 Principals' Bonus Eligibility, as Reported by Principals (Percentages) 58

IV.11 Reported and Actual Maximum Pay-for-Performance Bonus for Teachers in Treatment Schools 59

IV.12 Reported and Actual Maximum Pay-for-Performance Bonus for Principals in Treatment Schools 60

IV.13 Treatment Teachers' Reported Pay-for-Performance Bonus Eligibility by School and by District, Year 3 (Percentages) 62

VI.1 Average Student Achievement in Treatment and Control Schools After Years 1, 2, and 3 (Percentiles) 88

VI.2 Impact of Pay-for-Performance on Student Achievement in Math After Year 3, by District (Student *z*-Score Units) 89

VI.3 Impact of Pay-for-Performance on Student Achievement in Reading After Year 3, by District (Student *z*-Score Units) 90

VI.4 Impact of Pay-for-Performance on Student Math Achievement After Year 3, by Treatment School and by District (Student *z*-Score Units) 91

VI.5 Impact of Pay-for-Performance on Student Reading Achievement After Year 3, by Treatment School and by District (Student *z*-Score Units) 92

D.1 Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools in Years 1 and 2, Cohorts 1 and 2 .....D.14

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 .....D.15

D.3 Percentage of Treatment Teachers Earning a Pay-for-Performance Bonus in Year 3, By District, Cohort 1 .....D.16

D.4 Percentage of Treatment Teachers Earning a Pay-for-Performance Bonus in Year 2, by District, Cohorts 1 and 2.....D.16

D.5 Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools in Year 2, by District, Cohorts 1 and 2 ..... D.17

D.6 Minimum, Average, and Maximum Pay-for-Performance Bonuses for Principals for Years 1 and 2, Cohorts 1 and 2 ..... D.18

D.7 Minimum, Average, and Maximum Pay-for-Performance Bonuses for Principals in Treatment Schools, with Districts Weighted by the Number of Schools, Cohort 1 ..... D.19

D.8 Distribution of Pay-for-Performance Bonuses for Principals in Treatment Schools, Cohort 1 ..... D.19

D.9 Minimum, Average, and Maximum Automatic 1 Percent Bonuses for Teachers and Principals in Control Schools, Cohort 1 ..... D.20

D.10 Teachers' Pay-for-Performance Bonus Eligibility in Years 1 and 2, as Reported by Teachers, Cohorts 1 and 2 (Percentages) D.23

D.11 Principals' Pay-for-Performance Bonus Eligibility in Years 1 and 2, as Reported by Principals, Cohorts 1 and 2 (Percentages) D.24

D.12 Reported and Actual Maximum Pay-for-Performance Bonus for Teachers in Treatment Schools, with Districts Weighted by the Number of Schools, Cohort 1 ..... D.26

D.13 Reported and Actual Maximum Pay-for-Performance Bonus for Principals in Treatment Schools, with Districts Weighted by the Number of Schools, Cohort 1 ..... D.27

D.14 Reported and Actual Maximum Pay-for-Performance Bonus for Teachers in Treatment Schools, Cohorts 1 and 2 ..... D.29

D.15 Reported and Actual Maximum Pay-for-Performance Bonus for Principals in Treatment Schools, Cohorts 1 and 2 ..... D.30

F.1 Impact of Pay-for-Performance on Student Math Achievement After Years 1 and 2, by District, Cohorts 1 and 2 (Student *z*-Score Units) F.18

F.2 Impact of Pay-for-Performance on Student Reading Achievement After Years 1 and 2, by District, Cohorts 1 and 2 (Student *z*-Score Units) F.19

G.1 Average Performance Bonus Earned by Teachers in Treatment Schools in Year 2 as a Percentage of Average Teacher Salary G.5

G.2 Standard Deviation of Pay-for-Performance Bonuses Earned by Teachers in Treatment Schools in Year 2 as a Percentage of Average Teacher Salary G.6

G.3 Percentage of Teachers in Treatment Schools Earning a Pay-for-Performance Bonus in Year 2 ..... G.7

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 3 ..... G.8

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, do 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 study focuses on performance-based compensation systems that were established under TIF grants awarded in 2010. It examines grantees' programs and implementation experiences and the impacts of pay-for-performance bonuses on educator effectiveness and student achievement.

This report, the third from the study, describes the programs and implementation experiences of all 2010 TIF grantees in the 2013–2014 school year, the third of four years of implementation for nearly all grantees. The main findings for all districts that received 2010 TIF grants include the following:

- **Overall implementation of TIF requirements among all 2010 TIF districts was very similar in the third year of implementation as in previous years.** Similar to the previous two years, half of TIF districts in the third year reported implementing all four required components for teachers. Nevertheless, most districts (88 percent) reported implementing at least 3 of the 4 required components for teachers.
- **Few TIF districts in the third year reported that key activities related to implementation of their program were a major challenge, and districts were less likely to report major challenges in the third year than in the second year.** No aspect of TIF implementation was a major challenge to more than one-fifth of TIF districts in the third year. Furthermore, fewer districts in the third year than in the second year reported major challenges with program implementation, such as providing feedback on student achievement growth measures or teacher observations, and calculating performance bonuses. Half of the districts in the third year reported that sustainability of the TIF program was a major challenge, a decline from almost two-thirds (64 percent) of the districts in the second year.

This report also provides detailed findings from a subset of 2010 TIF grantees, the evaluation districts, that participated in a random assignment study of the pay-for-performance component of TIF. For the ten evaluation districts that completed three years of TIF implementation, the report provides an in-depth analysis of TIF implementation and the impacts of pay-for-performance bonuses on educator and student outcomes after the first (2011–2012), second (2012–2013), and third (2013-2014) years. The main findings for the ten evaluation districts include the following:

- **Pay-for-performance had small, positive impacts on students' reading and math achievement.** After three years of TIF implementation, average student achievement was 1 to 2 percentile points higher in schools that offered pay-for-performance bonuses than in schools that did not. This difference was equivalent to a gain of about four additional weeks of learning.

- **Few evaluation districts structured pay-for-performance bonuses to align well with TIF grant guidance.** The grant notice provided guidance about how to structure pay-for-performance bonuses to be substantial, differentiated, and challenging to earn. At least half of the evaluation districts each year met the guidance for awarding differentiated performance bonuses for teachers. However, in each year, no more than 20 percent of districts awarded bonuses for teachers that were substantial or challenging to earn.
- **Teachers' understanding of performance measures continued to improve between the second and third year of implementation, but many teachers still did not understand that they were eligible for a bonus or underestimated how much they could earn.** A higher percentage of teachers in the third year reported being evaluated on student achievement growth than in the second year, and a higher percentage of teachers in the second year reported being evaluated on at least two classroom observations than in the first year. In schools that offered performance bonuses, about 60 percent of teachers (62 percent in Year 2 and 57 percent in Year 3) correctly reported being eligible for a performance bonus—implying that about 40 percent were unaware they were eligible. Similar to previous years, teachers believed that the maximum bonus they could earn was no more than two-fifths the size of the actual maximum bonus that districts awarded.

TIF Grants and Requirements

From 2006 to 2012, the U.S. Department of Education awarded about $1.8 billion to support 131 TIF grants. Sixteen grants were awarded in 2006, 18 in 2007, 62 in 2010, and 35 in 2012.[1]

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 performance-based compensation systems implemented in districts to include four components.

Required Program Components of the Performance-Based Compensation Systems

The four required TIF components are:

1. **Measures of educator effectiveness.** Grantees were required to measure the effectiveness of teachers and principals using students' achievement growth and at least two observations of classroom or school practices. They had discretion to include additional measures.
2. **Pay-for-performance bonuses.** Grantees had to offer bonuses to educators based on how they performed on the effectiveness measures. The bonuses aimed to incentivize educators and reward them for being effective in their classrooms and schools. Bonuses had to be substantial in size, differentiated, challenging to earn, and based solely on educators' effectiveness.
3. **Additional pay opportunities.** The performance-based compensation system had to include pay opportunities for educators to take on additional roles or responsibilities.

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

These roles might include becoming a master or mentor teacher who directly counsels other teachers or develops or leads professional development sessions for teachers.

4. **Professional development.** TIF grantees were required to provide professional development to help educators understand the measures being used to evaluate their performance as well as provide feedback based on their actual performance ratings to help improve their instructional practices.

The 2010 TIF grant notice differed from the other rounds in that it included a main and an evaluation competition (Max et al. 2014). By holding two separate competitions, the U.S. Department of Education identified a group 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.

A key difference between the non-evaluation and evaluation grantees is that applicants for the evaluation grants 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). Although applicants had discretion over the proposed structure of the pay-for-performance bonus, these examples provided additional guidance to evaluation grant applicants and might have influenced how they designed their performance-based compensation systems.

Applicants for evaluation grants had to meet the same requirements for the performance-based compensation system as non-evaluation grantees and some additional requirements. One important requirement was that evaluation grant applicants 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, including pay-for-performance bonuses (the treatment group), or all components *except* pay-for-performance bonuses (the control group).

The TIF Study

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

The study is addressing four research questions:

1. What are the characteristics of all TIF districts and their performance-based compensation systems? What implementation experiences and challenges did TIF districts encounter?
2. How do teachers and principals in schools that did or 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 do pay-for-performance bonuses affect educator effectiveness and the retention and recruitment of high-performing educators?
4. What is the impact of pay-for-performance bonuses on students' achievement on state assessments in math and reading?

This report is the third of four planned reports from the study. The first report (Max et al. 2014) addressed the first two research questions based on information from the 2011–2012 school year. The second report used information from the first (2011–2012) and second (2012–2013) years of TIF implementation to describe the ways in which evaluation districts structured the components of their programs and communicated information about those components (question 1). The report also captured the views, attitudes, and behaviors of educators as they evolved over two years of implementation (question 2) and presented initial impacts of pay-for-performance on educator effectiveness and student achievement after the first and second years (questions 3 and 4). This third report also focuses on implementation of TIF and the effect of pay-for-performance (questions 1 through 4), but includes information after an additional year (2013–2014) of program implementation. It captures educators' views and attitudes that, by the end of the third year, were shaped by two years of pay-for-performance bonuses. The report also presents impacts of pay-for-performance on educator effectiveness and student achievement after three years of program implementation.

Districts in the Study

Although this report provides the greatest amount of information on the evaluation districts, it also provides a broad overview of TIF implementation by all 2010 grantees in the 2013–2014 school year. This analysis was based on 144 districts that participated in TIF in 2013–2014.

This report's in-depth analyses of TIF implementation and the effects of pay-for-performance on educator and student outcomes were based on information from the evaluation districts. Of the 13 evaluation districts, 10 completed three years of TIF implementation—2011–2012, 2012–2013, and 2013-2014—during the period covered by the report. The remaining 3 evaluation districts completed two years of TIF implementation—2012–2013 and 2013-2014. This report focuses primarily on the 10 districts for which data were available on three years of TIF implementation. Focusing on districts that completed three years of TIF implementation enabled us to examine changes in educators' perceptions and practices from the first to the third year and assess whether impacts on educator and student outcomes also evolved during that time.

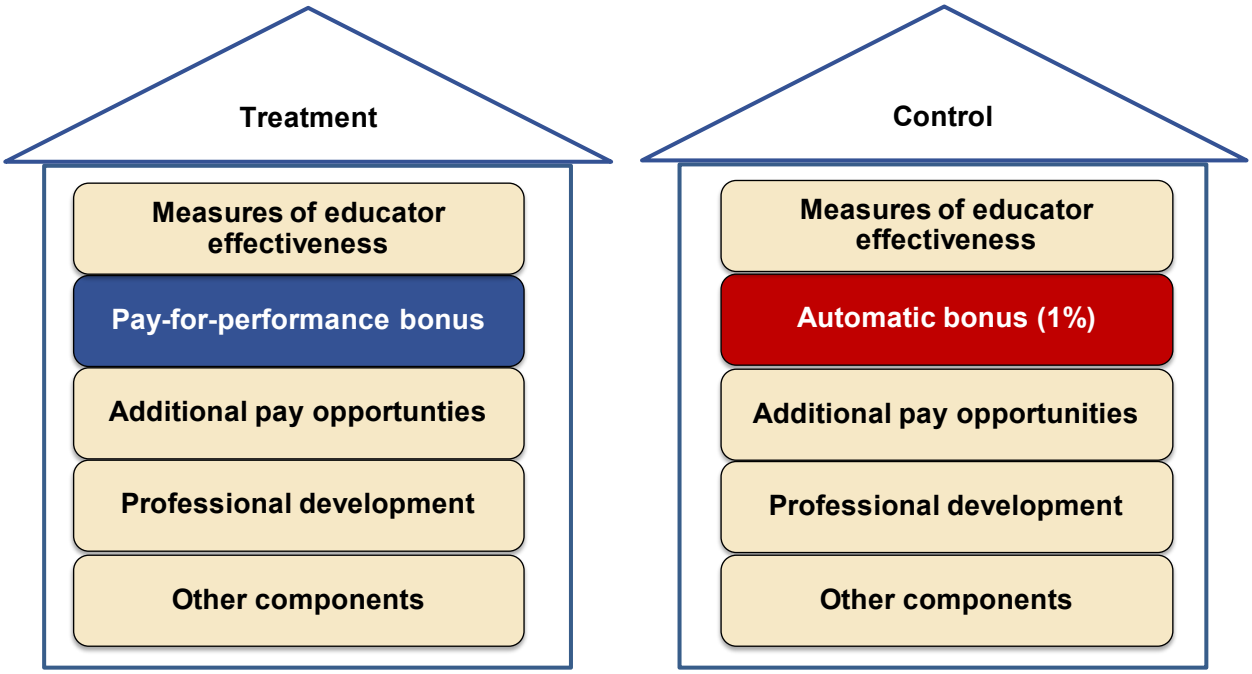
Experimental Study Design

The study used an experimental study design to assess the impacts of pay-for-performance on educator and student outcomes. Elementary and middle schools within the evaluation districts were assigned randomly—that is, completely by chance—to treatment and control groups. As shown in Figure ES.1, treatment and control schools were expected to implement the same required components of the district's performance-based compensation system, except for the pay-for-performance bonus component. As a result, the study measured the impact of pay-for-performance bonuses implemented within the context of broader performance-based compensation systems. The study was not designed to measure the impact of implementing a TIF grant or the multiple components of a performance-based compensation system.

Teachers and principals in treatment schools were eligible to earn a pay-for-performance bonus; teachers and principals in control schools received an automatic bonus worth approximately 1 percent

of their annual salary. 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 reflected 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 ES.1. Random Assignment Evaluation Design



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.

Schools in the Study

Analyses of educator and student outcomes were based on 132 schools—66 treatment schools and 66 control schools—that implemented the TIF program for three years. Before random assignment, evaluation districts chose which schools to include in the evaluation. Because a primary objective of the study was to measure the impact of pay-for-performance on student achievement on state assessments in high-need schools, every participating school had to have at least half of its students receiving free or reduced-price lunch and at least one grade level tested by state assessments (3rd to 8th grade).

Data Sources

Data for this report came from multiple sources. 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.

Data on all 2010 TIF districts. The study team collected data on all TIF districts from two sources. First, to compare characteristics of evaluation and non-evaluation districts, the study team used information from the Common Core of Data. Second, to describe broadly the TIF program features that districts reported implementing and the challenges they encountered in implementation, the study team administered a survey to all TIF district administrators in 2011–2012, 2012–2013, and 2013-2014.

Additional data on evaluation districts. We obtained more detail on TIF programs and implementation experiences from interviews with district staff and technical assistance documents. To examine educators' attitudes toward their job and the TIF program, the study team administered surveys to all principals and a sample of teachers in treatment and control schools in spring of 2012, 2013, and 2014. We collected districts' administrative records on teachers and principals to describe their performance ratings, bonuses, and additional pay, as well as to examine the impact of pay-for-performance on educator effectiveness. Finally, to assess the impact of pay-for-performance on student achievement, the study team collected districts' administrative records on students enrolled in treatment and control schools.

Methods

The study team used several different methods to describe the implementation of TIF and measure the impact of pay-for-performance on educators' and students' outcomes.

Describing TIF implementation in all 2010 TIF districts. To describe broadly the program characteristics and implementation challenges reported by all 2010 TIF districts, we summarized their responses to the district survey with means or percentages, as appropriate.

Describing TIF implementation in evaluation districts. We conducted a variety of analyses to provide an in-depth description of TIF implementation in the evaluation districts. First, as in the analysis of all 2010 TIF districts, we summarized evaluation districts' survey responses about program characteristics and implementation challenges, but we also supplemented these data with information from telephone interviews and technical assistance documents. Second, to describe educators' actual bonus amounts and performance ratings, we summarized administrative data with means, maximum levels, or percentages of educators receiving particular bonus amounts or ratings. Third, to describe educators' understanding of and experiences with the required TIF components, we summarized educators' survey data, making comparisons between treatment and control schools and across years.

Measuring the impacts of pay-for-performance on educator and student outcomes. Within the evaluation districts, we assessed the impacts of pay-for-performance on several educator and student outcomes, including educators' attitudes and behaviors (measured by survey responses), educator effectiveness (measured by performance ratings that educators received from their districts), and student achievement (measured by scores on state assessments in math and reading). For each outcome, we compared the outcomes of educators and students in treatment schools to those of educators and students in control schools. Because the study used random assignment, any differences in educator or student outcomes between the treatment and control groups could be attributed to pay-for-performance and not some other characteristic of the districts or schools.

Detailed Summary of Findings—All 2010 TIF Districts

As a comprehensive program for reforming educator compensation and improving educator effectiveness, TIF programs were designed to have multiple, interrelated components. Our analysis of implementation in all 144 TIF districts sought to determine whether they could put into place such a comprehensive system, and whether they faced particular challenges doing so.

Most districts implemented each of the four individual required components of TIF, but were least likely to report offering targeted professional development and evaluating principals using both student achievement growth and at least two observations of school practices. In the third year of implementation (2013–2014), nearly all the districts (over 95 percent) reported offering teachers and principals bonuses based on their performance and offering educators opportunities to earn additional pay (88 percent; Table ES.1). Fewer districts reported offering the required professional development to their teachers (70 percent), using both student achievement growth and classroom observations to measure teacher effectiveness (81 percent), and using both student achievement growth and observations of school practices to measure principal effectiveness (69 percent).

Overall implementation of TIF requirements among all 2010 TIF districts was very similar in the third year of implementation as in previous years. Similar to the previous two years, in Year 3 half of TIF districts reported implementing all four required components for teachers (Table ES.1). Nevertheless, most districts (88 percent) reported implementing at least 3 of the 4 required components for teachers. Likewise, more than half of the districts implemented all required components for principals aside from professional development, a component for which data were not available. Districts' reported implementation of each required component and of all components combined was similar across all three years.

Table ES.1. Districts' Reported Implementation of TIF Required Components for Teachers in Year 3 (Percentages)

	All 2010 TIF Districts	Evaluation Districts
Requirements		
Requirement 1: Measures of educator effectiveness[a]	81	100
Requirement 2: Pay-for-performance bonus	100	100
Requirement 3: Additional pay opportunities	88	100
Requirement 4: Professional development	70	60
Implemented all requirements	50	60
Number of Districts—Range[b]	**134-144**	**10**

Source: District surveys and district interviews, 2014.

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

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

Few TIF districts reported that key activities related to implementation of their program were a major challenge, and districts were less likely to report major challenges in the third year of implementation than in the second year. No aspect of TIF implementation was a major challenge to more than one-fifth of TIF districts in Year 3. For example, about 20 percent of the districts reported that explaining student achievement growth to teachers or attributing student achievement growth to individual teachers was a major challenge. In addition, compared to Year 2, fewer districts reported major challenges in Year 3. For example, fewer districts reported major challenges with providing feedback on student achievement growth measures (19 versus 30 percent), teacher observations (14 versus 25 percent), or principal observations (4 versus 15 percent). Although concerns about sustainability stand out among the potential challenges, fewer districts in Year 3 than in Year 2 (50 versus 64 percent) reported sustainability to be a major challenge.

Detailed Summary of Findings—2010 TIF Evaluation Districts

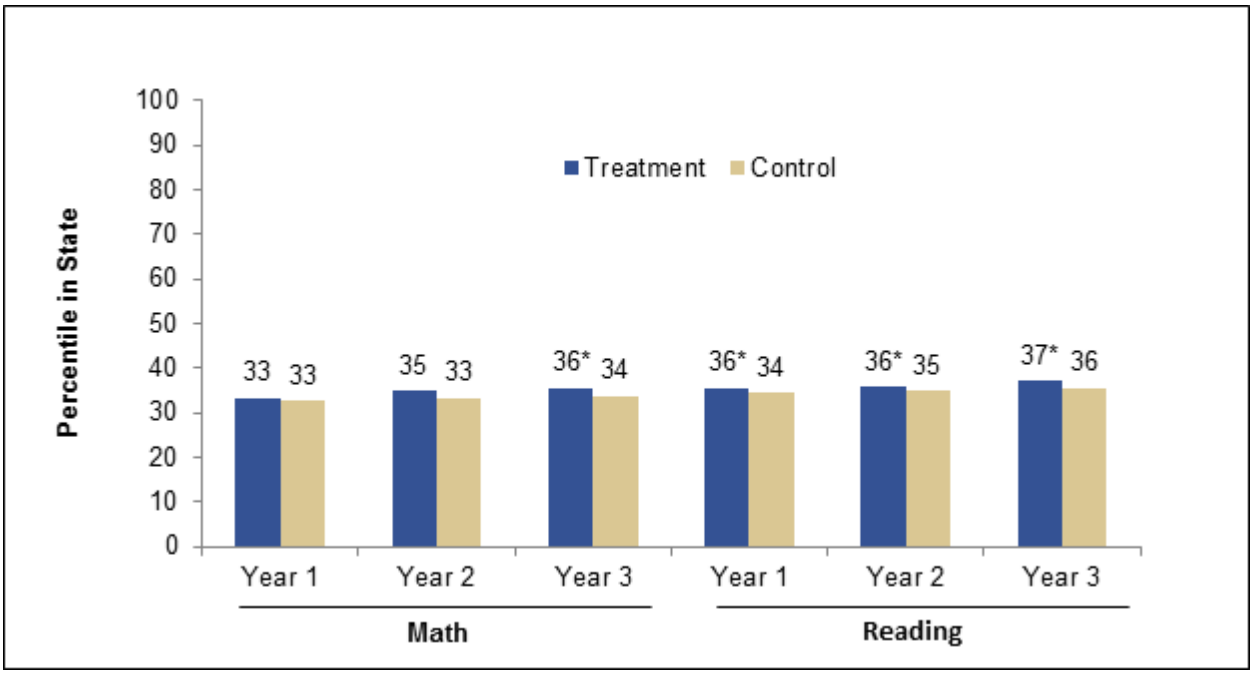
Additional information from the evaluation districts enabled the study team to examine the implementation of pay-for-performance in much greater detail, and measure the impacts of pay-for-performance on educator and student outcomes.[2] Ultimately, the goal of the TIF grants was to improve student achievement in high-need schools. We first present findings on the impacts of pay-for-performance on student achievement. To put those findings in context, we then present in-depth information on evaluation districts' TIF programs, teachers' and principals' understanding of and experiences with key components of their programs, and impacts of pay-for-performance on educators' satisfaction and effectiveness. Given that districts differed in the design and implementation of their programs, we also present findings on whether those differences were associated with differences in student achievement impacts.

Impacts of Pay-for-Performance on Student Achievement

Pay-for-performance had small, positive impacts on students' math and reading achievement. After three years of implementation, the average student in a control school earned a math score at approximately the 34th percentile of student achievement statewide (Figure ES.2). The average student in a treatment school earned a math score at approximately the 36th percentile—a gain of 2 percentile points. Similarly, the impact on reading achievement after Year 3 lifted the average student in these schools from the 36th to the 37th percentile. These differences translated to a gain of about 4 weeks of additional learning in a typical 36-week school year. These impacts, which represent the cumulative effect of schools' exposure to pay-for-performance for three years, were similar in size to the impacts achieved after two years of implementation.

[2] This study examined the impacts of pay-for-performance bonuses on the average outcomes of schools that offered those bonuses, but for simplicity we describe these findings as impacts on educators' or students' outcomes. Student achievement was measured using students' reading and math scores on state assessments.

Figure ES.2. Average Student Achievement in Treatment and Control Schools (Percentiles)



Source: Student administrative data (N = 40,847 students for Year 1 math; N = 40,708 students for Year 2 math; N = 40,037 for Year 3 math; N = 40,571 students for Year 1 reading; N = 40,390 students for Year 2 reading; N = 39,807 for Year 3 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 between treatment and control schools is statistically significant at the .05 level, two-tailed test.

TIF Implementation in Evaluation Districts

To understand the impacts of pay-for-performance on student achievement, the study team collected in-depth information about TIF implementation in the evaluation districts. Using this information, we examined the components of their programs to help assess whether they provided incentives and supports for educators to improve their effectiveness. Finally, we examined whether educators understood those components.

Program Implementation

As a first step, the study team examined the extent to which evaluation districts implemented the four required components. These analyses also examined the types of measures that districts used to evaluate educators' effectiveness and described educators' actual performance on those measures, focusing on whether educators received similar ratings from different measures and whether performance ratings for the same measure were similar across years.

Most evaluation districts reported implementing all required components for teachers. The only component not consistently implemented continued to be professional development. In Year 3, all evaluation districts reported using measures of effectiveness for teachers and principals that included student achievement growth and at least two observations of classroom or school practices, offering bonuses based on how educators performed on effectiveness measures,

and offering additional pay to take on extra roles or responsibilities. Six of 10 evaluation districts reported providing the required professional development for teachers (Table ES.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.

All evaluation districts reported using school achievement growth to evaluate teachers, and some also chose to evaluate teachers based on classroom achievement growth. More than half (70 percent) of evaluation districts reported evaluating teachers based on classroom achievement growth. Within these districts, more than half (about 60 percent) of teachers received classroom achievement growth ratings.

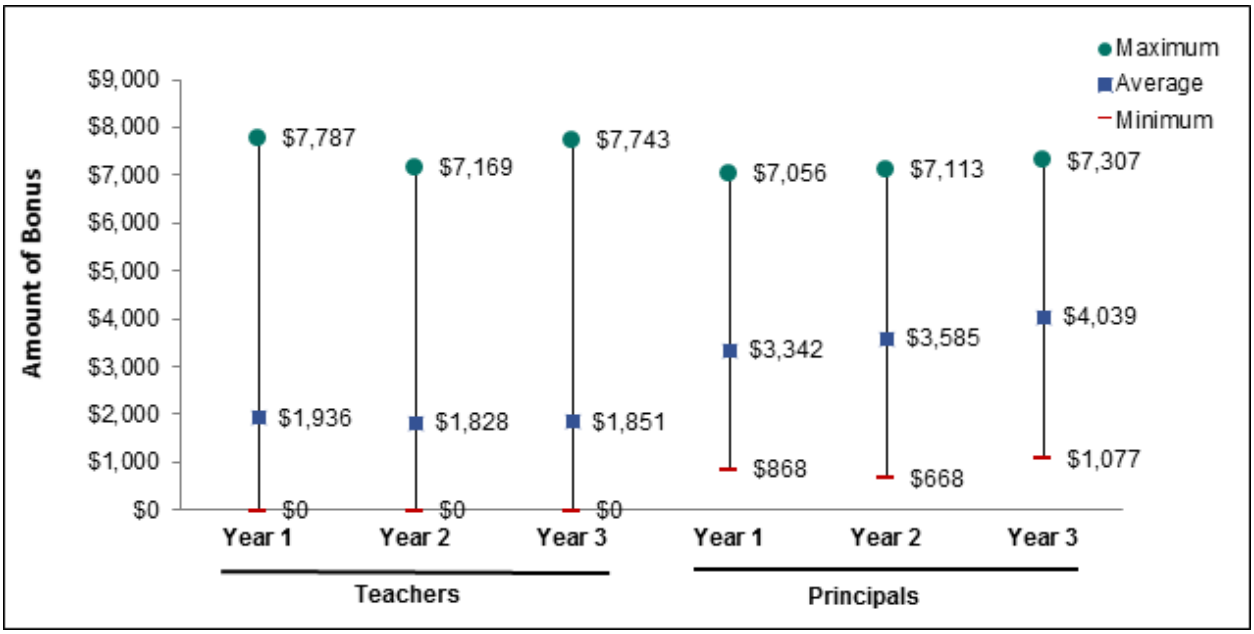
Most teachers received similar performance ratings in the third year of implementation as they did in the second year, with many teachers receiving higher ratings on classroom observations than on student achievement growth. More than half of teachers received similar ratings, based on a 1-to-4 rating scale, in Years 2 and 3. For example, 58 percent of teachers received a similar rating based on classroom observations, and 56 percent received a similar rating based on student achievement growth in their schools. However, in both years, teachers often earned higher ratings on classroom observations than on student achievement growth. For example, in Year 3, slightly more than half (53 percent) of teachers received a higher rating on classroom observations than on student achievement growth in their schools.

Pay-for-Performance Bonuses

The purpose of offering performance bonuses to teachers and principals was to motivate them to improve and reward educators for being effective in their classrooms and schools. To achieve this objective, the TIF notice required that the bonuses had to be substantial in size, differentiated, and challenging to earn.

The highest-performing teachers earned a pay-for-performance bonus about four times the average bonus. Yet, most teachers received a bonus, which, on average, was smaller than suggested by the TIF grant guidance. On average across evaluation districts, the maximum performance bonus for teachers ($7,743 in Year 3) was about four times the average bonus ($1,851 in Year 3), consistent with the example of a differentiated bonus provided in the TIF grant notice (Figure ES.3). However, more than 70 percent of teachers received a performance bonus, suggesting that bonuses were not challenging to earn. Moreover, the average bonus for teachers was about 4 percent of the average teacher salary—less than the 5 percent guidance for substantial bonuses specified in the TIF grant notice. For principals, bonuses were closer to the grant notice's example of a substantial bonus but were not very differentiated or challenging to earn. The average performance bonus in Year 3 ($4,039) was slightly less than 5 percent of the average principal salary, the maximum bonus ($7,307 in Year 3) was less than twice the average bonus, and at least three-fourths of principals received a bonus.

Figure ES.3. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers and Principals



Source: Educator administrative data (N = 2,183 teachers in Year 1, N = 2,193 teachers in Year 2, and N = 2,260 teachers in Year 3; N = 65 principals in Year 1, N = 68 principals in Year 2, and N = 65 principals in Year 3).

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 $1,936, and the maximum pay-for-performance bonus was $7,787.

Teachers' and Principals' Understanding of and Experiences with Key Components

In addition to determining how to implement the required components of TIF, districts had to effectively communicate information about those components to educators, and educators needed to know how to improve their performance. Educators' understanding of the components and how to improve their practices determines how the program can influence educators' behaviors and, ultimately, student achievement.

Most teachers understood that they were evaluated based on student achievement growth and classroom observations, and teachers' awareness of the use of these performance measures continued to improve between the second and third year of implementation. More than 75 percent of teachers in the third year reported being evaluated on student achievement growth, and over 85 percent reported being evaluated on at least two classroom observations. Furthermore, the percentage of teachers who reported being evaluated on these measures continued to increase. For example, a higher percentage of teachers in Year 3 reported being evaluated on student achievement growth (84 percent of treatment teachers and 78 percent of control teachers) than in Year 2 (78 percent of treatment teachers and 72 percent of control teachers), and a higher percentage of teachers in the second year reported being evaluated on at least two classroom observations (87 percent of treatment teachers and 83 percent of control teachers) than in the first year (74 percent of treatment teachers and 76 percent of control teachers).

Many teachers and some principals in schools that offered pay-for-performance bonuses still did not understand that they were eligible for a bonus or underestimated how much they could earn from performance bonuses. By the third year of TIF implementation, about 40 percent

of treatment teachers were still unaware that they could potentially earn a performance bonus (57 percent of treatment teachers reported being eligible for a bonus in Year 3, implying that 43 percent of treatment teachers did not report being eligible for one; Figure ES.4). Although understanding of eligibility was better among principals than teachers, about 20 percent of principals in Year 3 still did not know they were eligible to earn a bonus based on their performance; in fact, fewer principals were aware of their eligibility in the third year of implementation than in the second year (Figure ES.4). Similar to previous years, teachers in treatment schools believed that the maximum bonus they could earn was no more than two-fifths the size of the actual maximum bonus that districts awarded (Figure ES.5).

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 it over the school year. In Year 3, approximately two-thirds of teachers reported that they received or expected to receive professional development focused on understanding performance measures used in TIF, somewhat fewer (about 58 percent) 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 was three hours over the school year.

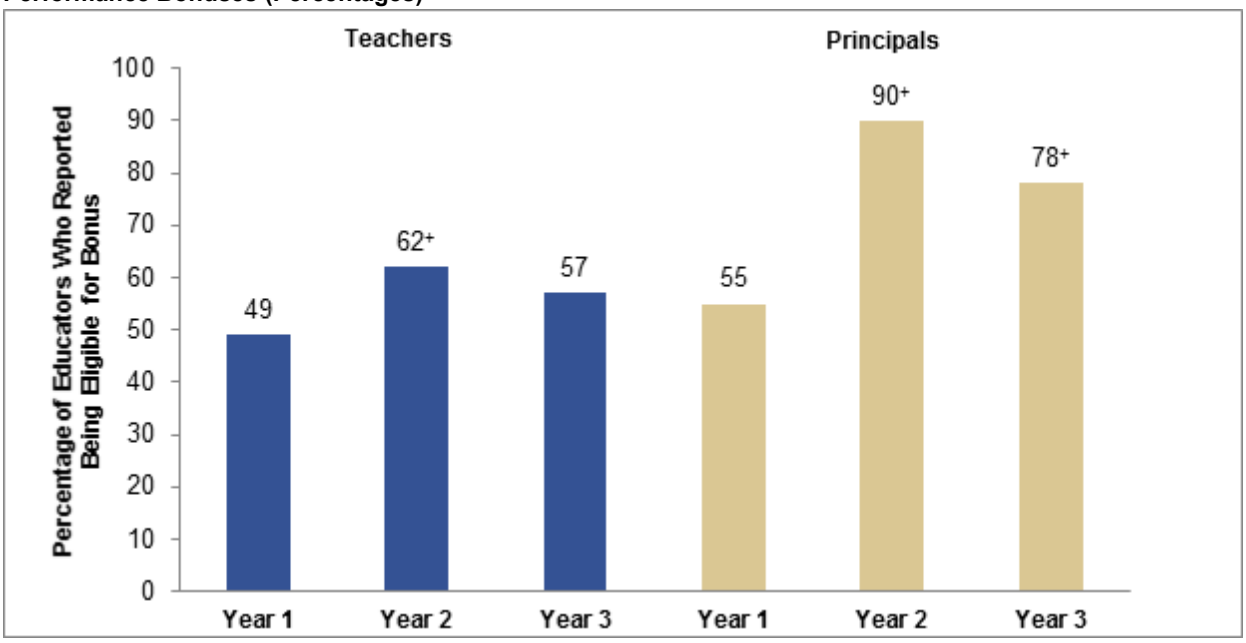
Impacts of Pay-for-Performance on Educators' Attitudes and Behaviors

The ways in which pay-for-performance programs affect educators' attitudes (such as job satisfaction) and behaviors (such as allocation of time) can shape how pay-for-performance affects student outcomes. For example, pay-for-performance could motivate educators to improve their effectiveness if it makes them more satisfied with pay opportunities and the feedback they receive on performance evaluations. 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, ultimately, on student achievement.

Most teachers and principals reported being satisfied with their professional opportunities, how they were evaluated, and their school environment. For example, in Year 3, about 80 percent of teachers reported being satisfied with their opportunities to enhance their skills, the feedback on their performance, the quality of interaction with colleagues, and colleagues' efforts. The percentage of principals satisfied with aspects of their professional opportunities, evaluation system, and school environment ranged from 54 to 96 percent in Year 3.

In contrast to prior years, teachers in treatment schools in the third year of implementation were at least as satisfied as teachers in control schools with their professional opportunities, how they were evaluated, and their school environment. 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, teachers in treatment schools reported being less satisfied than control teachers with recognition of their accomplishments and factors associated with how they were evaluated. Treatment teachers in Year 2 only reported being more satisfied than control teachers with their opportunity to earn extra pay. But in Year 3, treatment teachers reported being more satisfied than control teachers with school morale (62 versus 53 percent), the quality of their interaction with colleagues (83 versus 79 percent), and their opportunities to earn extra pay (61 versus 50 percent). They had similar levels of satisfaction as control teachers did with other aspects of their jobs.

Figure ES.4. Teachers and Principals in Treatment Schools Who Reported Being Eligible for Pay-for-Performance Bonuses (Percentages)

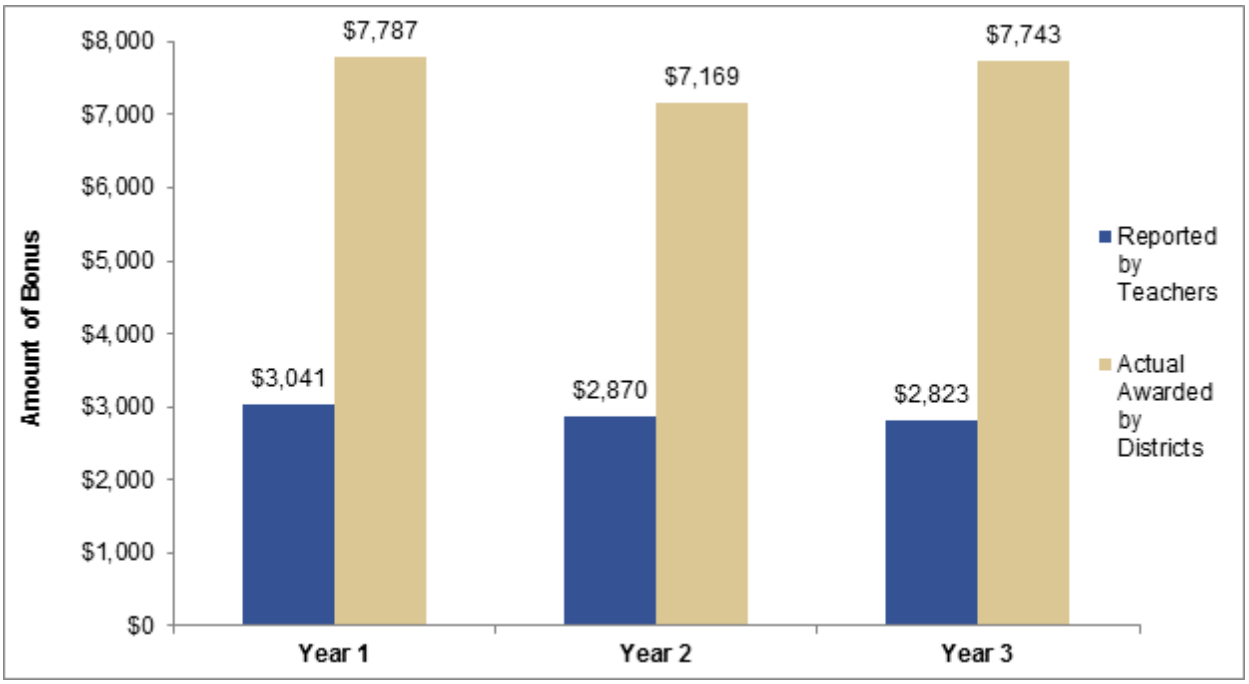


Source: Teacher and principal survey, 2012, 2013, and 2014 (N = 377 teachers in Year 1; N = 444 teachers in Year 2; N = 424 teachers in Year 3; N = 64 principals in Year 1; N = 63 principals in Year 2; and N = 58 principals in Year 3).

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 within treatment status is statistically significant at the .05 level, two-tailed test.

Figure ES.5. Reported and Actual Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools



Source: Teacher survey (2012, 2013, and 2014) and educator administrative data (N = 223 teachers in Year 1; N = 232 teachers in Year 2; N = 232 teachers in Year 3; N = 10 districts).

Figure reads: In Year 1, on average, the maximum pay-for-performance bonus that teachers reported they could earn was $3,041, and the actual maximum pay-for-performance bonus that evaluation districts awarded to teachers was $7,787.

Most teachers had positive attitudes toward their TIF program, and by the third year of implementation, teachers in treatment schools felt at least as positively toward TIF as teachers in control schools did. In Year 3, as in prior years, most teachers were glad to be participating in TIF. However, in contrast to Year 2, treatment teachers in Year 3 no longer felt less favorably than control teachers about the effect of TIF on teacher collaboration, their freedom to teach the way they like, and the use of student test scores to measure student learning. And for the first time, treatment teachers were more likely than control teachers to report that their job satisfaction increased due to the TIF program (39 versus 33 percent in Year 3). However, pay-for-performance continued to cause a higher percentage of treatment teachers than control teachers to feel increased pressure to perform (by 14 and 10 percentage points in Years 2 and 3, respectively).

Impacts of Pay-for-Performance on Educator Effectiveness

The ways in which pay-for-performance programs are implemented and their effects on educators' attitudes could lead to changes in educator effectiveness. In fact, a central objective of the TIF grants is to improve student achievement in high-need schools by increasing educator effectiveness—in particular, by enabling schools to attract and retain more effective educators and motivating educators to improve their effectiveness. This study measured educator effectiveness using the performance ratings that educators received from their districts.

Pay-for-performance had a positive impact on teachers' and principals' performance ratings based on student achievement growth in the first year of implementation, but by the third year educators in treatment and control schools received similar ratings. On measures of school achievement growth, educators in treatment schools earned ratings in Year 1 that were 0.34 points higher on a 1-to-4 rating scale than those of educators in control schools. Likewise, 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 Year 1. However, the impacts of pay-for-performance on both school and classroom achievement growth ratings diminished over the three years of TIF implementation. For example, by Year 3, educators in treatment and control schools earned similar school achievement growth ratings, and by Year 2, teachers in treatment and control schools earned similar classroom achievement growth ratings.

Pay-for-performance led to slightly, but not statistically significantly, higher classroom observation ratings for teachers in each year. Although differences between the classroom observation ratings of teachers in treatment schools and those in control schools were not statistically significant, they were positive and similar in all three years and almost significant by Year 3 (p-value = 0.07 in Year 3). In all three years, there were no statistically significant differences between observation ratings for principals in treatment and control schools.

Differences in Student Achievement Impacts Across Districts

The study's main findings on the impact of pay-for-performance on student achievement represent an average impact of pay-for-performance across the 10 evaluation districts. However, these districts differed in many ways, including the design and implementation of their pay-for-performance programs. 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 student achievement differed across districts, but differences in impacts were not related to differences in key program characteristics measured by this study. The impacts of pay-for-performance on reading and math achievement varied substantially, but were not related to a variety of program and implementation characteristics, 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 level of differentiation of bonuses, (4) the degree to which earning a bonus was challenging, (5) the timing of awarding bonuses based on the prior year, and (6) teachers' understanding of their pay-for-performance eligibility.

Concluding Thoughts

Overall, the 2010 TIF districts were able to implement most required components of a comprehensive performance-based compensation system without major, widespread challenges. In fact, fewer districts in the third year of implementation reported major challenges to implementing their TIF program than in the second year. However, many districts still did not put into place all the required components by the end of the third year of implementation.

A primary objective of TIF grants is to raise student achievement in high-need schools. Based on the experiences of ten districts that participated in the national evaluation and completed three years of program implementation, the pay-for-performance component of TIF made a small contribution toward achieving this objective. Pay-for-performance bonuses generated slightly higher student achievement in reading and math. Most of the impact emerged in the first two years, and did not significantly grow in the third year.

The theory underlying the belief that pay-for-performance bonuses can lead to large impacts on student achievement depends on many factors. First, educators must understand their eligibility for a performance bonus. Yet, near the end of the third year of implementation, many educators continued to misreport their eligibility, and their understanding was no better than it was in the previous year.

Second, pay-for-performance needs to provide educators with the motivation to improve and cause effective educators to want to work in schools offering pay-for-performance bonuses. However, bonuses continued to be small on average and generally not challenging to earn, which may have dampened the motivation for teachers to improve. Furthermore, teachers still underestimated how much they could earn from the bonuses, so they may not have perceived a compelling monetary incentive to become a high performer. On the other hand, in contrast to previous years, by the third year teachers who were eligible for pay-for-performance were at least as satisfied with their jobs as those who were not eligible. However, this improvement in satisfaction was not accompanied by a larger impact on student achievement in Year 3. The improvement in satisfaction may not have been large enough to trigger changes in educator effectiveness, or it may take time for more favorable attitudes to translate into better classroom and school practices.

Third, educators need to know how to change their practices in ways that improve student achievement. We found that pay-for-performance did have small (although insignificant), positive impacts on teachers' classroom observation ratings each year. This suggests that teachers may have changed their practices slightly in response to pay-for-performance. Yet, teachers reported receiving few hours of professional development aimed at helping them improve their practices based on their performance. From this evidence, it is unclear whether teachers could really identify the changes to their practices that would most effectively improve their performance and raise student achievement.

Although the overall impact of pay-for-performance on student achievement was small, impacts were larger in some districts than in others. This raises the question of whether particular ways of designing or implementing their TIF programs could lead to larger impacts. However, none of the characteristics we examined could help explain observed differences in student achievement impacts across districts.

Evidence from the fourth and final year of implementation may provide more clarity on whether an additional year of implementation enhances educators' understanding of and experience with this program, and how impacts of pay-for-performance may evolve.

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, do 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.

This report is the third of four planned reports from a multiyear study focusing on the TIF grants awarded in 2010.[2] 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 report (Chiang et al. 2015) examined grantees' implementation experiences and educators' understanding of, and attitudes toward, the program near the end of the second year of program implementation, as well as changes in educators' understanding and attitudes. It also examined the impacts of pay-for-performance bonuses on educator effectiveness and student achievement after one and two years of TIF implementation.

This study has two main goals. First, it will inform program development and improvement by describing how grantees implemented their performance-based compensation systems and the implementation challenges they faced. Second, it will test whether pay-for-performance bonuses as part of a comprehensive reform system lead to increases in educator effectiveness and student achievement.

Previous Research on Pay-for-Performance Programs for Educators

Research on the effectiveness of pay-for-performance initiatives in U.S. public schools is inconclusive, and few studies of U.S. pay-for-performance programs have found consistent impacts on student achievement.[3]

[1] The 2015 reauthorization of the Elementary and Secondary Education Act renamed TIF the Teacher and School Leader Incentive Grants program. This program will provide 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. For this report, all references to TIF are for the 2010 awardees.

[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

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 used nonexperimental designs that leave 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 experimental 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 a dearth of 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 will be 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, the U.S. Department of Education (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 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. 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

(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); and Springer and Taylor (2016).

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

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 bonus. 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 bonus, 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.

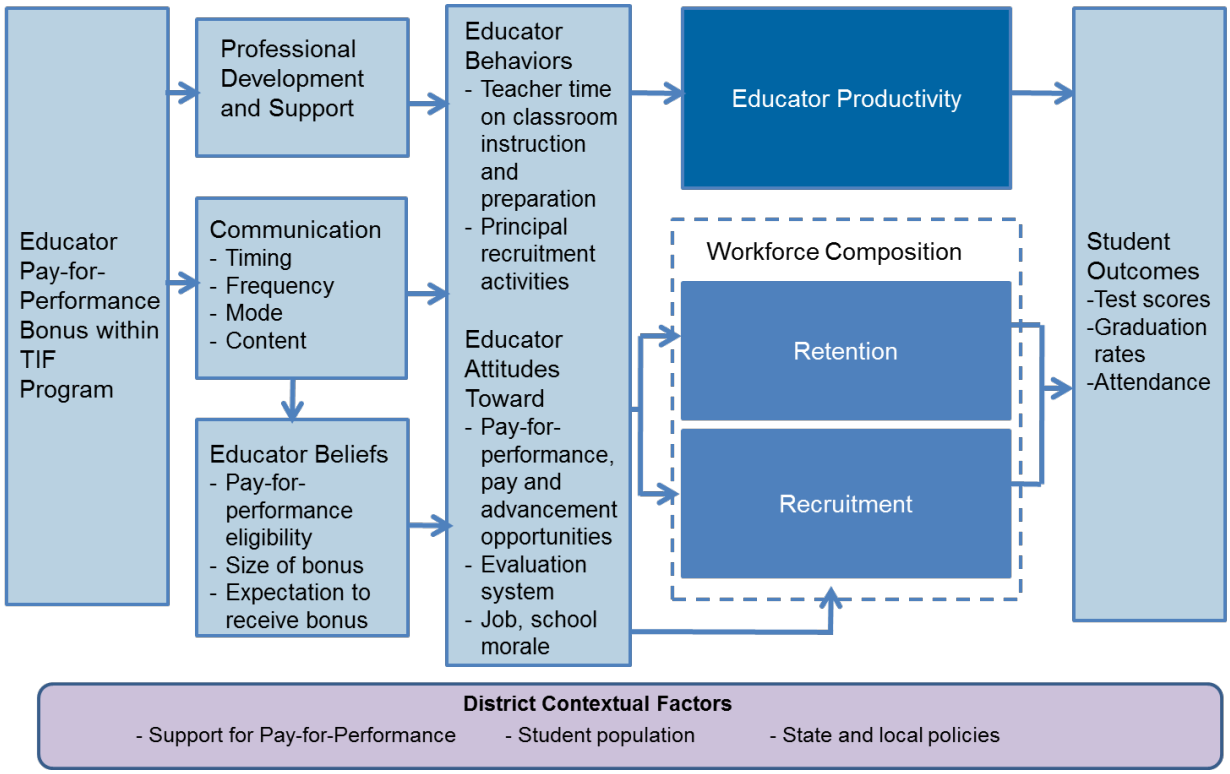
Logic Model: How Pay-for-Performance Could Influence Student Outcomes

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.

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. The ability to earn a pay-for-performance bonus, as well as the fact that the criteria to earn a bonus depend on student achievement gains, could affect teachers' attitudes toward their school choice, alter their teaching practices, and increase their productivity. For example, pay-for-performance bonuses may serve as incentives for effective teachers to remain in a school that provides bonuses and may attract other effective teachers to the school. In addition, pay-for-performance bonuses based on schoolwide student achievement gains may encourage teacher collaboration, which may increase educator productivity. Educators rewarded for student achievement gains on standardized tests may allocate more time to instructional practices intended to improve test scores.

However, whether and how pay-for-performance bonuses actually lead to changes in educator productivity and the composition of the teaching workforce depend on many factors. 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 highlights the emphasis on these activities required by the grant.

Figure I.1. Logic Model



Educators' understanding of their TIF program will depend on districts' communication activities, timing of communication, and educators' receiving the information. Educators' awareness and understanding of the program can depend on the frequency, content, and types of district communication. 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. 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.

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 is to describe the program characteristics and implementation experiences of 2010 TIF grantees and estimate 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 plans to follow the grantees for all four years of TIF implementation.

The study addresses four research questions:

1. What are the characteristics of all TIF districts and their performance-based compensation systems? What implementation experiences and challenges did TIF districts encounter?
2. How do teachers and principals in schools that did or 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 do pay-for-performance bonuses affect educator effectiveness and the retention and recruitment of high-performing educators?
4. What is 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 report (Chiang et al. 2015) focused on implementation of TIF and the effect of pay-for-performance bonuses in the 10 evaluation districts after one and two years of program implementation. This report found that after one and two years of implementation, pay-for-performance had small, positive impacts on students' reading achievement; impacts on students' math achievement were not statistically significant but were similar in magnitude. The report 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 from the first to the second year, but many teachers still misunderstood whether they were eligible for performance bonuses or the amount they could earn.

This third report also focuses on implementation of TIF and the effect of pay-for-performance in the 10 evaluation districts, but includes information after an additional year of program implementation. It captures educators' views and attitudes that, by the end of the third year, were shaped by two years of pay-for-performance bonuses. The report also presents impacts of pay-for-performance on educator effectiveness and student achievement after three years of program implementation. 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. Finally, in Chapter VI, we present findings on the impact of pay-for-performance on educator effectiveness and student achievement.

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 in the 2010 grants participated in TIF in subsequent years. A total of 171 districts implemented TIF—that is, had a performance-based compensation system supported by TIF funds—in 2011–2012, 164 districts implemented TIF in 2012–2013, and 158 districts implemented TIF in 2013–2014 (Table II.1).[6] Among the districts that implemented TIF in 2013–2014, 13 were evaluation districts.

Table II.1. Number of Districts Implementing TIF, by Year

	Implemented TIF in 2011–2012	Implemented TIF in 2012–2013	Implemented TIF in 2013–2014	Responded to 2014 District Survey
Non-Evaluation Districts	159	151	145	131
Evaluation Districts	12	13	13	13
Total	**171**	**164**	**158**	**144**

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

Note: 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.

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 to 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/ethnic minorities and eligible for free or reduced-price lunch.

[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] Between 2011–2012 and 2012–2013, eight non-evaluation districts withdrew from their grants, and one evaluation grantee added a district to its TIF grant. Between 2012–2013 and 2013–2014, six non-evaluation districts withdrew from their grants.

This report provides an overview of TIF implementation in all TIF districts in 2013–2014 and, within the evaluation districts, an in-depth analysis of implementation and the impacts of pay-for-performance on educator and student outcomes after three years. 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) in the 2013–2014 school year—the third year of implementation for nearly all those districts. We describe the districts' reported compliance with implementing the four required components of TIF and the challenges they encountered in implementing TIF. As discussed later, this analysis relied on districts' responses to a survey we administered in 2014. Therefore, the final sample for this analysis consisted of 144 districts in 2014—13 evaluation and 131 non-evaluation districts—that participated in TIF in 2013–2014 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. This information—obtained from surveys, interviews, technical assistance documents, and administrative data—allowed 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 educator 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. Evaluation districts were larger, on average, than non-evaluation districts. Evaluation districts were also more likely than non-evaluation districts to be in urban areas (69 versus 30 percent) and the West (46 versus 14 percent), and less likely to be in towns (8 versus 22 percent), rural areas (0 versus 28 percent), the Midwest (15 versus 29 percent), the South (23 versus 48 percent), and states with collective bargaining agreements (54 versus 69 percent). Evaluation and non-evaluation districts had similar proportions of students who were black or Hispanic or that 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/Ethnic Distribution		
White, non-Hispanic	39	49
Black, non-Hispanic	32	26
Hispanic	22	19
Student Socioeconomic Status		
Eligible for free/reduced-price lunch	65	65
Title 1 eligible schools (schoolwide)	70	79
Enrollment (Average)		
Number of students	32,317	20,298
District Location		
Urban	69	30*
Suburban	23	20
Town	8	22
Rural	0	28*
Geographic Region		
Northeast	15	8
Midwest	15	29
South	23	48
West	46	14*
Collective Bargaining[a]		
In state with collective bargaining	54	69
Number of States	**8**	**24**
Number of Districts	**13**	**137-145**

Source: Common Core of Data for 2012–2013 school year.

Notes: The table is based on all 158 districts that implemented TIF in 2013–2014. Seven non-evaluation districts were not included in the 2012–2013 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.

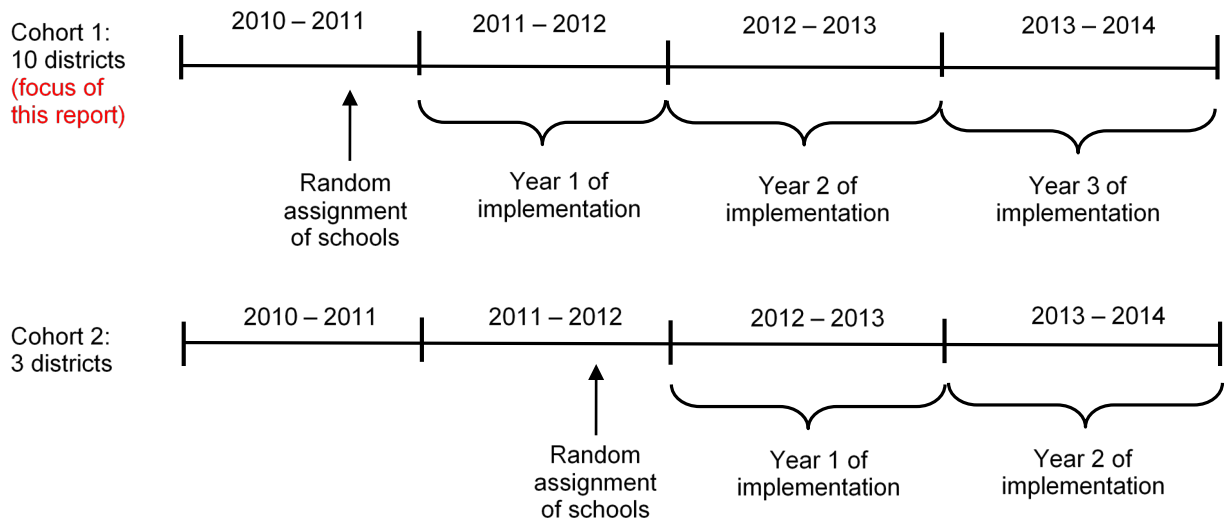
[a]Collective bargaining is a state-level indicator from the National Right to Work Legal Defense Foundation (http://www.nrtw.org/rtws.htm).

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

We classified 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 and a 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 three years of TIF implementation: 2011–2012 (Year 1), 2012–2013 (Year 2), and 2013–2014 (Year 3). Cohort 2 consists of three districts in which we randomly assigned schools in spring and summer 2012 and obtained data on two years of TIF implementation, 2012–2013 and 2013–2014, representing Years 1 and 2 of this cohort's implementation of TIF.[7]

[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 2013–2014 as Year 2 of the districts' implementation of TIF.

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 the U.S. Department of Education. The remaining six Cohort 1 districts were part of multidistrict grants that were 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.

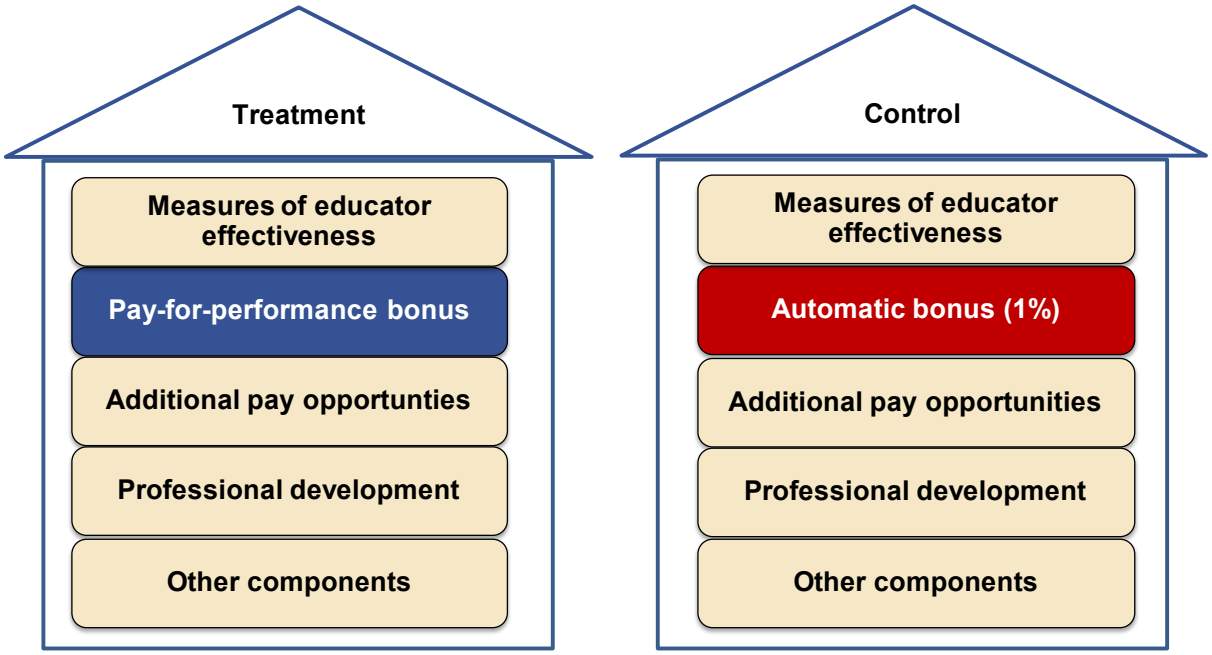
This report primarily focuses on the 10 Cohort 1 evaluation districts—those for which data were available on three years of TIF implementation. As explained in Chapter I, because TIF is a comprehensive program for reforming educator compensation and improving educator effectiveness, it may take time for educators to fully understand the incentives available, the measures on which they are evaluated, and the improvements they need to make to earn bonuses. An earlier report from this study (Max et al. 2014) presented findings for Cohort 1 districts on educators' understanding, attitudes, and behaviors from the first year of TIF implementation—before performance ratings were determined and bonuses were distributed. Educators' perceptions and practices may have changed after they experienced the results of the performance evaluations and bonuses and determined how to respond to this new information. The second report from this study primarily focused on findings for Cohort 1 districts after two years of implementation and examined changes in educators' perceptions and impacts on educator and student outcomes (Chiang et al. 2015). This report, which examines outcomes for educators and students in Cohort 1 districts after three years of implementation, allows us to examine whether understanding of the TIF program and the impact of pay-for-performance bonuses continued to evolve. Focusing on Cohort 1 districts ensures that the same schools were included in the analyses for all three years. Unless otherwise noted, all findings in Chapters IV through VI are based on these 10 Cohort 1 districts.[8]

[8] For key implementation features and outcomes, the appendices of this report provide findings from Year 2 of implementation for Cohorts 1 and 2 together—that is, findings from 2012–2013 for Cohort 1 and from 2013–2014 for Cohort 2.

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 approximately 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



Evaluation districts chose which schools would be included in the evaluation. Because a primary objective of the study was to measure the impact of pay-for-performance on student achievement on state assessments in high-need schools, every participating school needed to have (1) at least half of its students receiving free or reduced-price lunch, and (2) 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 132 schools that implemented the TIF program for three years.[9] This sample excluded schools that closed or dropped out of the study along with the schools with which they were paired—a total of six schools (4 percent of all Cohort 1 schools). 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 3 years)		**66**	**66**	**132**

Source: Study authors' calculations.

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

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 pre-implementation year—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=0.09; 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

[9] Analyses that used administrative data were based on all 132 schools. Analyses that used educator survey data were based on 131 schools in 2011–2012 and 132 schools in 2012–2013 and 2013–2014. When we administered the spring 2012 educator surveys, we did not know that one school was a multicampus school with different administrative structures, and therefore only one of the campuses was surveyed.

[10] Forty-one of the 45 schools in Cohort 2 implemented TIF for two years and were also paired with schools that did so. Therefore, supplemental analyses that include Cohorts 1 and 2 together are based on 173 schools—132 schools from Cohort 1 and 41 schools from Cohort 2.

similar fractions of students who received free or reduced-price lunch, had an Individualized Education Program, were overage 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/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.

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	-0.02
Race/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/reduced-price lunch	77	76	1
Disabled or has an Individualized Education Program	12	12	0
Overage 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			0.09
Number of Students—Range[a]	**12,624-22,141**	**12,540-22,037**	
Number of Schools—Range[a]	**42-66**	**42-66**	

Source: Student administrative data.

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

[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 educator characteristics within treatment and control schools in the pre-implementation year for 9 of 10 districts that provided educator 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.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	60	57	3
Race/ethnicity						
White, non-Hispanic	74	73	1	60	56	3
Black, non-Hispanic	19	21	-2	32	36	-4
Hispanic	3	3	0	3	2	2
Other	4	4	0	5	6	-1
Age (average years)	42	41	0	49	48	1
Education						
Master's degree or higher	50	50	0	94	93	1
Experience in K-12 Education						
Total experience (average years)	12	11	0	16	15	2
Less than 5 years	24	25	-1	18	14	4
5-15 years	45	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			0.54			0.80
Number of Educators—Range[a]	**1,456-2,136**	**1,499-2,136**		**40-65**	**45-68**	
Number of Schools—Range[a]	**49-66**	**49-66**		**38-63**	**43-64**	

Source: Educator administrative data.

Note: None of the differences 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.

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. We used data from the 2012– 2013 school year to compare the characteristics of evaluation and non-evaluation districts.

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. We addressed these surveys to the person identified as overseeing or directing each district's TIF program. Districts' responses allowed us to describe programs in all TIF districts and to determine their compliance with the four required components of the TIF grant.

Table II.6. Data Sources for This Report

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

Note: 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.

NA is not applicable.

We administered the survey in 2012 (in the middle of the 2011–2012 school year), 2013 (near the end of the 2012–2013 school year), and 2014 (near the end of the 2013–2014 school year) to all districts participating in TIF in those years. This report primarily used data from the 2014 survey to describe the programs in 2013–2014; in some cases, however, we used data from the 2012 and 2013 surveys to examine whether compliance with required components changed over time. In 2014, 91 percent of TIF districts responded to the district survey (Appendix A, Table A.5). Districts that responded and did not respond to the survey did not differ by a statistically significant margin on most characteristics—including the districts' student racial composition, student socioeconomic status, and size (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 revisions to the program to overcome those challenges. Information from the interviews allowed us to develop a

comprehensive description of implementation in evaluation districts and, when appropriate, to fill in missing information or supplement survey responses. This report used data from the first, second, and third 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 educator 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, and 2014, 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 and 2014. These groups represent elementary and middle school grades and subjects both with and without annual accountability testing.[13]

Response rates for principals and teachers were over 90 percent in each year. (Appendix A, Table A.7).[14] The response rates of treatment and control educators were generally similar; 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).

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'

[13] In 2013 and 2014, we also surveyed teachers from the prior-year sample even if they left teaching, left the study schools, or switched teaching assignments. These teachers were not included in the final analysis sample. 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.

background characteristics and their scores on state assessments in math and reading, allowing us to examine the impact of pay-for-performance on student achievement. Within Cohort 1 districts—those that completed three years of TIF implementation—the data covered all students in study schools in 2010–2011 to 2013–2014, representing the period from the pre-implementation year to Year 3 of implementation. We obtained similar data from Cohort 2 districts for 2011–2012 to 2013–2014.

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 allowed 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 3 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 2.[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.

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 three 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 the equal-weighted average across all districts.

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

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 done in our impact analyses, described below.

Educators' understanding of program components may change as they gain more exposure to those components. We examined how educators' understanding changed from Year 2 of TIF implementation (when educators had experienced one year of bonuses) to Year 3 (when educators had experienced two years of bonuses). Separately for treatment and control schools, we compared average reports in Year 2 and Year 3 and conducted hypothesis tests to determine whether differences between years 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 teachers' allocation of time and 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/ethnic composition before TIF implementation. Therefore, all regressions in the impact analyses accounted for the baseline differences by controlling for school averages of those student characteristics from the pre-implementation year. In some analyses, we also controlled for the individual characteristics of students or educators in the analysis samples to enhance precision (see Appendix B for a full description of these 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 an average study school after one, two, and three years of TIF implementation.

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

[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.

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. 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 after random assignment.[20] Therefore, this analysis measured the impact of pay-for-performance on schools' average student achievement after one, two, and three years of TIF implementation, potentially reflecting changes in individual students' achievement and changes in the schools' student composition resulting from pay-for-performance.[21] In Chapter VI, for simplicity, we describe the findings as impacts on students'

[18] Appendix B includes an explanation of 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 educators with and without observation ratings, but teachers who received classroom achievement growth ratings in Year 3 were more likely to be female, younger, less educated, and less experienced than those who did not. In Year 3, there were no significant differences between the characteristics of treatment and control educators who received performance ratings, with one exception: among teachers with classroom achievement growth ratings, a larger percentage in treatment schools than control schools had more than 15 years of experience.

[19] To ensure that all outcomes were measured in the spring, we used a grantee-administered test for one district located in a state that administered fall state assessments during the period covered by this study.

[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, and 3 (Appendix A, Table A.22). Compared to students without scores, those with scores had higher baseline achievement, were more likely to be female, and were less likely to have an Individualized Education Program or be overage for their grade (Appendix A, Tables A.23 and A.24).

[21] In Years 1, 2, and 3, 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 Year 1, 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 pre-implementation year (Table II.4).

achievement, but these statements are shorthand for impacts on the average student achievement of schools.

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 teacher behaviors, influenced the direction and magnitude of the achievement impacts. In Chapter VI (and Appendix G), we explore whether such factors were associated with the achievement impacts after Year 3.

Differences in student achievement impacts across districts provided an opportunity to examine whether the characteristics of districts' TIF programs and their implementation were associated with the impacts of pay-for-performance. Districts varied substantially in the design and implementation of their TIF programs in ways that could have influenced 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. Knowing whether the impacts of pay-for-performance were systematically larger or smaller in districts with particular program characteristics can suggest best practices for developing and improving these programs.

We assessed whether any of the following characteristics could help explain differences across districts in impacts on student achievement: (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 in bonuses, (4) the degree to which earning a bonus was challenging, (5) the timing of awarding bonuses based on the prior year, and (6) teachers' understanding of their pay-for-performance eligibility. We selected these six characteristics because of their potential to motivate teachers to change their behavior in response to pay-for-performance bonuses, which may, in turn, affect student achievement. For each feature, 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 then compared the impacts of pay-for-performance on student achievement in Year 3 between these two subgroups of districts. A significant difference in impacts between the two subgroups provides only suggestive evidence that the characteristic may have influenced impacts, given that the two groups may differ on other measured and unmeasured characteristics.

Impacts of pay-for-performance bonuses on student achievement could also differ among treatment schools. Such variation could be due to bonuses triggering educators to change their behaviors differently across schools. 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 selected educator survey items on which educators' responses could reflect their effort, strategic behavior, or classroom practices. We also used teachers' classroom observation ratings as a direct measure of their practices. For every treatment school, we estimated the impacts of pay-for-performance on these educator behaviors by comparing educators' behaviors in that school with those in the control school to which it was paired for random assignment. In a similar manner, we estimated the impacts of pay-for-performance on student achievement for every treatment school. Across treatment schools, we

then examined the association between impacts on educator behaviors and impacts on student achievement. Although these associations can suggest which behavioral changes may be responsible for impacts on achievement, they may also reflect the influence of other behaviors that the study did not measure.

THIS PAGE IS INTENTIONALLY BLANK

III. PROGRAMS AND EXPERIENCES OF ALL 2010 TIF DISTRICTS

In this chapter, we broadly describe TIF program implementation in 2013–2014. 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 144 districts that were included in the 2010 TIF grants and implemented a TIF program in the 2013–2014 school year. The information in this chapter is based on surveys completed by TIF districts between April and July 2014, when nearly all TIF districts had completed Year 3 of program implementation. We also draw upon districts' responses to the 2012 and 2013 surveys to compare findings from the third year of implementation to those from the previous two years.

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.

Key Findings on Programs and Experiences of All 2010 TIF Districts

- **Most districts implemented each individual required component of TIF, but were least likely to report offering targeted professional development and evaluating principals using both student achievement growth and at least two observations.**
- **Overall implementation of TIF requirements among all 2010 TIF districts was very similar in the third year of implementation as in previous years.**
- **Few TIF districts reported that key activities related to implementation of their program were a major challenge, and districts were less likely to report major challenges in Year 3 than in Year 2.**

Implementation of TIF Required Components

Most districts implemented each of the four individual required components of TIF, but were least likely to report offering targeted professional development and evaluating principals using both student achievement growth and at least two observations. In the third year of implementation (2013–2014), nearly all the districts (over 95 percent) reported offering teachers and principals bonuses based on their performance, and 88 percent reported offering educators opportunities to earn additional pay (Table III.1). In contrast, 70 percent of districts reported that they offered the required professional development to their teachers, 81 percent reported using both student achievement growth and classroom observations to measure teacher effectiveness, and 69 percent reported using both student achievement growth and observations of school practices to measure principal effectiveness.

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

Source: Max et al. (2014); district survey (2013 and 2014).

[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. The decrease in the number of districts between Years 2 and 3 is due to some districts dropping out of TIF (Table II.1) and a lower response rate in Year 3 than Year 2 (91 versus 95 percent).

Overall implementation of TIF requirements among all 2010 TIF districts was very similar in Year 3 as in previous years. Similar to the previous two years, in Year 3 half of TIF districts reported implementing all four required components for teachers (Table III.1). Nevertheless, most districts (88 percent) reported implementing at least 3 of the 4 required components for teachers. Likewise, more than half of the districts implemented all required components for principals aside from professional development.[22] Districts' reported implementation of each required component and of all components combined was similar across all three years.

Next, we provide an overview of districts' implementation of each individual required component in 2013–2014.[23]

[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 notice also required pay for additional opportunities for educators. Most 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.

[23] Districts' implementation of each required component in 2013–2014 was similar to their implementation of each component in 2011–2012 (Max et al. 2014) and 2012–2013 (Chiang et al. 2015).

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. Eighty-one percent of TIF districts reported using student achievement growth and classroom observations to measure teacher effectiveness, and 69 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 (88 percent), followed by classroom achievement growth (69 percent) and achievement growth of student subgroups (50 percent; Figure III.1).

Most TIF districts reported using at least two formal observations to evaluate teachers. Eighty-three percent of districts reported using at least two formal observations by trained observers to evaluate teachers (Figure III.1). Districts planned to conduct, on average, 3 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 observations were conducted by principals (93 percent).

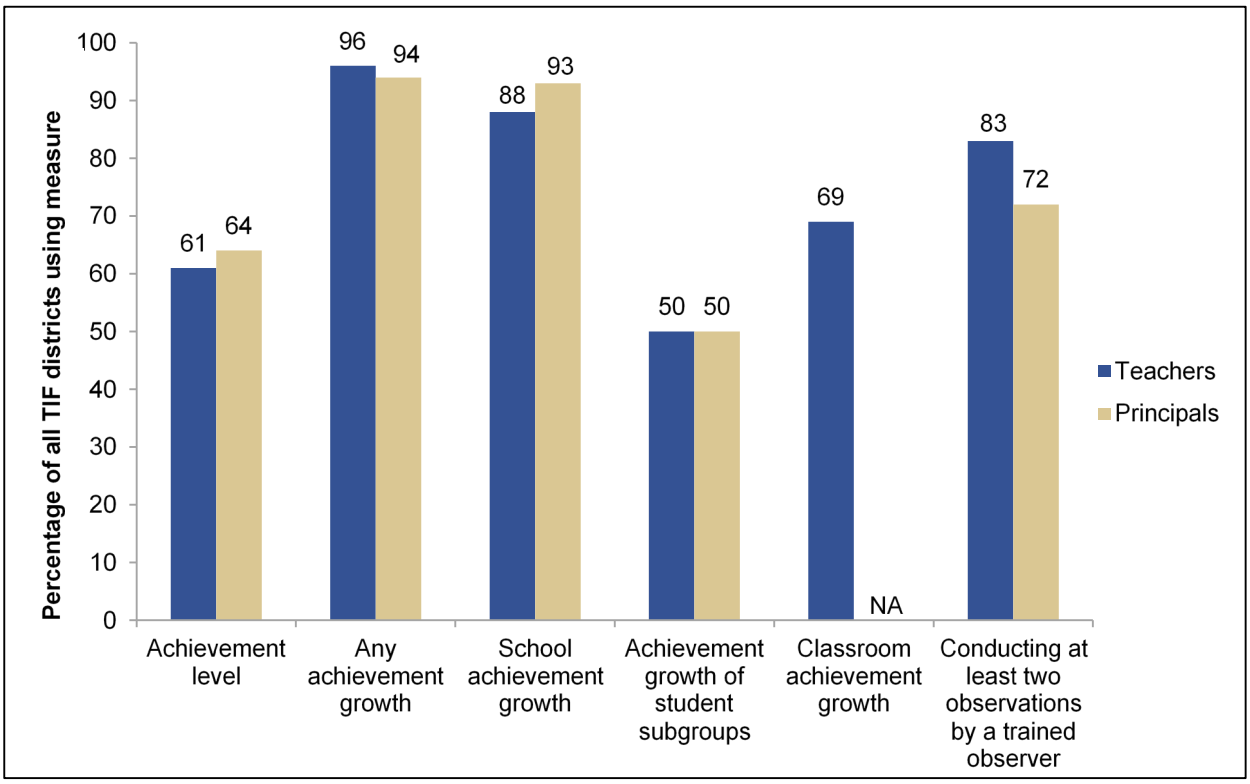
Most TIF districts reported using student achievement growth and observations by trained observers to evaluate principals. Most frequently, districts reported using school achievement growth to evaluate principals (93 percent) (Figure III.1). Most districts (72 percent) also reported conducting observations by trained observers. Districts planned to conduct, on average, about three observations per principal, lasting about 47 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 (51 percent).

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 determination of who is eligible 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.

Figure III.1. Measures of Student Achievement and Observations Used to Evaluate Teachers and Principals, All TIF Districts, Year 3 (Percentages)



Source: District survey, 2014.

Notes: Between 135 and 138 districts responded to the survey questions for teachers, and between 132 and 138 districts responded to the survey questions for principals. Teacher evaluation measures are those for teachers in tested grades and subjects.

Figure reads: In Year 3, 61 percent of all TIF districts reported using achievement level to evaluate teachers, and 64 percent reported using achievement level to evaluate principals.

NA is not applicable.

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. Almost all (99 percent) of TIF districts reported that teachers were eligible for performance bonuses, and 97 percent reported that principals were eligible (Table III.2). Teachers' eligibility for performance bonuses was almost never contingent upon teaching a grade or subject with annual, end-of-year state assessments. In fact, 93 percent of districts reported that teachers in grades or subjects without annual assessments (referred to as "nontested") were eligible for performance bonuses (Table III.2). Moreover, districts tended not to restrict eligibility to teachers and principals. Seventy-nine percent of districts reported that assistant/vice principals were eligible for performance bonuses. Almost half of districts (46 percent) reported making nonteaching staff, such as counselors, librarians, or custodians, eligible for such bonuses.

Nearly all TIF districts offered additional pay for teachers to take on roles and responsibilities, most often to support mentor or master/lead teacher opportunities. Eighty-five percent of TIF districts reported offering teachers additional pay for roles and responsibilities (Table III.3). Most often, districts offered additional pay for mentor (61 percent) and master or lead teachers (53 percent). About one fifth of districts (22 percent) reported offering principals extra pay for assuming additional roles or responsibilities.

Table III.2. Staff Eligibility for Pay-for-Performance Bonus, Year 3 (Percentages)

	All TIF Districts
Teachers	
Teachers in tested grades and subjects	99
Teachers in nontested grades and subjects	93
Principals	97
Other School Staff	
Assistant/vice principal	79
Other school administrators	25
Other teaching staff (e.g., part-time teachers, substitutes, aides)	20
Nonteaching staff (e.g., counselors, librarians, custodians)	46
Number of Districts—Range[a]	**122-144**

Source: District survey, 2014.

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

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

	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	85	NA
Roles and Responsibilities		
Mentor teacher	61	$4,111
Master or lead teacher	53	$7,771
Department chair or head	18	$1,628
Lead curriculum specialist	15	$4,437
Schoolwide committee or task force member	17	$874
Leadership team member	30	$1,635
Number of Districts—Range[a]	**139-142**	**19-75**
Principals		
Principals Could Receive Additional Pay for Taking on Extra Roles or Responsibilities in School or District	22	$3,698
Number of Districts—Range[a]	**143**	**30**

Source: District survey, 2014.

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.

The TIF 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 (33 percent) offered teachers additional pay for doing so (Appendix C, Table C.2). Twelve percent of the districts reported offering principals extra pay for working in a hard-to-staff school.

Requirement 4: Professional Development

The TIF 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 being used to evaluate their performance, as well as to provide feedback based on their actual performance ratings to help improve their instructional practices.

About three-quarters of the TIF districts provided the required professional development to teachers. Although most TIF districts (87 percent) offered professional development to help teachers understand the performance measures used in the program, fewer districts (76 percent) offered the more targeted professional development based on teachers' actual performance (Table III.4).[24]

Table III.4. Planned Professional Development Activities for Teachers, Year 3 (Percentages)

	All TIF Districts
Focus of Professional Development	
Understanding performance measures of TIF program	87
Feedback based on TIF performance ratings	76
Number of Districts	**142**

Source: District survey, 2014.

Challenges in Implementing and Sustaining TIF

The 2013 and 2014 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 like TIF, and to examine if 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.[25]

By Year 3, few TIF districts reported that key activities related to implementation of their program were a major challenge. No aspect of TIF implementation was a major challenge to more

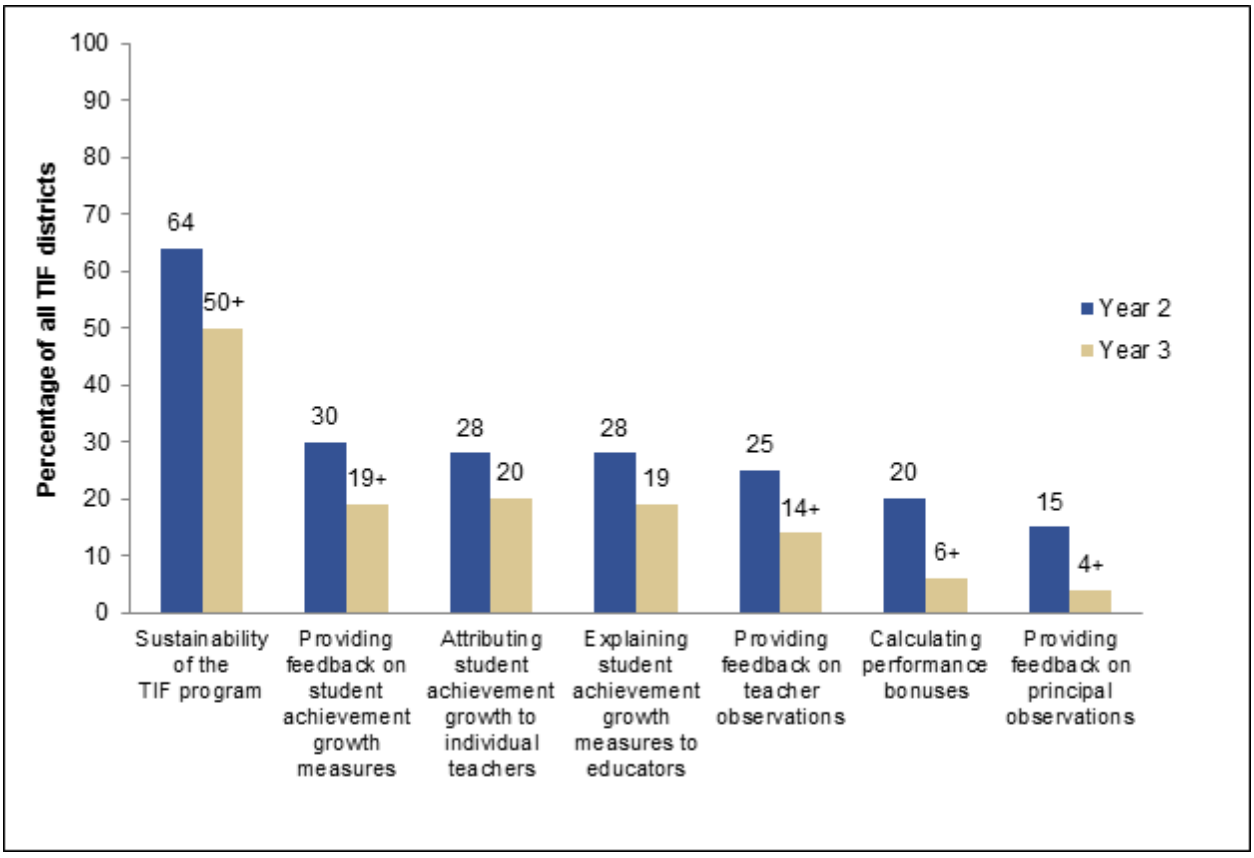
[24] Surveys of district administrators did not ask about professional development for principals.

[25] 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. Since the 2013 district survey was the first survey that included questions about challenges districts faced, that survey asked generally if districts found these issues challenging to implement. The 2014 survey asked districts to report if they had found these activities challenging to implement during the 2013–2014 school year.

than one-fifth of TIF districts in Year 3 (Appendix C, Table C.3). For example, about 20 percent of the districts reported that explaining student achievement growth to teachers or attributing student achievement growth to individual teachers was a major challenge (Figure III.2). Fewer than 15 percent of districts reported that calculating bonuses or providing feedback based on observations was a major challenge.

Districts were less likely to report major challenges in program implementation in Year 3 than in Year 2. By Year 3, districts had evaluated educators, calculated bonuses, and provided feedback to educators on their performance multiple times, which could have reduced the challenging nature of implementing the program. Compared to Year 2, fewer districts reported major challenges in Year 3 (Figure III.2). For example, significantly fewer districts reported major challenges with providing feedback on student achievement growth measures (19 versus 30 percent), teacher observations (14 versus 25 percent) or principal observations (4 versus 15 percent), and calculating performance bonuses (6 versus 20 percent). This was true for the range of potential challenges that we asked about (Appendix C, Table C.3), and in no case did significantly more districts report an item to be a major challenge in Year 3 than in Year 2.

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



Source: District survey (2013 and 2014).

Notes: Between 134 and 139 TIF districts responded to these survey questions in both Years 2 and 3. Further details about survey results, including results for activities that districts reported as a "minor challenge" or "not a challenge," can be found in Appendix C, Table C.3.

Figure reads: In Year 2, 64 percent of all TIF districts reported that sustainability of their TIF program was a major challenge. In Year 3, 50 percent reported that sustainability of their TIF program was a major challenge.

+Difference between Year 2 and Year 3 is statistically significant at the 0.05 level, two-tailed test.

We also asked districts if they felt that sustaining their program would be challenging. Half of districts reported that sustainability of the TIF program was a major challenge (Figure III.2). Although concerns about sustainability stand out among the potential challenges, fewer districts in Year 3 than in Year 2 (50 versus 64 percent) reported sustainability to be a major challenge. It is unclear why, as grantees were closer to the end of their TIF grant, fewer districts reported challenges with sustaining their TIF program. On the one hand, this finding might reflect that districts had begun to secure funding for their program after the grant ends. On the other hand, it could indicate that districts did not intend to continue their TIF program. To address this question, the fourth and final report will examine districts' plans for sustaining the TIF program after the grant period ends.

Summary

As a comprehensive program for reforming educator compensation and improving educator effectiveness, TIF programs were designed to have multiple, interrelated components. Our analysis of implementation in all 144 TIF districts sought to determine whether they could put into place such a comprehensive system, and whether they faced particular challenges doing so.

Similar to Year 2, in Year 3 the 2010 TIF districts implemented most required components of a comprehensive performance-based compensation system—without major, widespread challenges. Nevertheless, many districts still 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. Near the end of the third year of implementation, fewer districts reported major challenges to implementing their TIF program than previously reported. This particular finding suggests that it may take multiple years before some districts can implement a comprehensive compensation reform program without experiencing major challenges.

IV. TIF IMPLEMENTATION IN EVALUATION DISTRICTS

In this chapter, we describe the implementation of TIF by the evaluation districts—those that were awarded 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 the implementation of TIF for pay-for-performance to be able to improve educator effectiveness and student achievement. The components of the program needed to provide incentives and supports for educators to improve their effectiveness, information about those components needed to be communicated to educators, and educators needed 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 whether pay-for-performance bonuses were differentiated, substantial, and challenging to earn. 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 the implementation of the TIF grant in evaluation districts is 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 three years of TIF implementation during the period covered by this report. We refer to each year of implementation—2011–2012, 2012–2013, and 2013–2014—as Years 1, 2, and 3, respectively.[26] 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.

[26] 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. Cohort 2 districts completed two years of implementation, 2012–2013 and 2013–2014, referred to as Years 1 and 2 for this cohort. In Appendix D, we present key implementation findings from Years 1 and 2 for Cohorts 1 and 2 combined—that is, findings from 2011–2012 and 2012–2013 for Cohort 1 and 2012–2013 and 2013–2014 for Cohort 2.

Key Findings on TIF Implementation in Evaluation Districts

- **Most evaluation districts reported implementing all required components for teachers, and all districts reported meeting at least three of the four required components. The only component not consistently implemented continued to be professional development.**
- **All evaluation districts reported using school achievement growth to evaluate teachers, and some also chose to evaluate teachers based on the achievement growth of the students they teach.**
- **Most teachers received similar performance ratings and bonus amounts in Year 3 as they did in Year 2, with many teachers receiving higher ratings on classroom observations than on student achievement growth.**
- **The highest-performing teachers earned a pay-for-performance bonus about four times the average bonus. Yet, most teachers received a bonus, which, on average, was smaller than suggested by the TIF grant guidance.**
- **Teachers' understanding of performance measures continued to improve between the second and third year of implementation, but their understanding of their eligibility for bonuses did not.**
- **Many teachers in schools that offered pay-for-performance bonuses still 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 it 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 that all the individual components be implemented 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, and all districts reported meeting at least three of the four required components. The only component not consistently implemented continued to be professional development. In Year 3, 60 percent of evaluation districts implemented all four required components for teachers. All evaluation districts reported using a measure of effectiveness that included students' 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 of 10 evaluation districts reported providing the required professional development (similar to all 2010 TIF districts). The percentage of districts meeting each requirement was similar in Years 1, 2, and 3.

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)
Teachers			
Requirement 1: Measures of educator effectiveness[a]	100	100	100
Requirement 2: Pay-for-performance bonus	100	100	100
Requirement 3: Additional pay opportunities[b]	100	100	100
Requirement 4: Professional development	70	70	60
Implemented requirements 1, 2, and 3	100	100	100
Implemented all requirements	70	70	60
Principals			
Requirement 1: Measures of educator effectiveness[a]	70	100	100
Requirement 2: Pay-for-performance bonus	100	100	100
Requirement 3: Additional pay opportunities[b]	100	100	100
Implemented requirements 1, 2, and, 3[c]	70	100	100
Number of Districts	**10**	**10**	**10**

Source: District survey (2012, 2013, and 2014) and district interviews (2012, 2013, and 2014).

[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.

All evaluation districts also reported meeting three of the four required components for principals. All districts reported evaluating principals using student achievement growth and at least two observations by trained observers and offering pay-for-performance bonuses to principals (Table IV.1). Districts could meet the third requirement—additional pay opportunities—by providing opportunities to either teachers or principals; as discussed above, all districts fulfilled this requirement. We were unable to assess whether districts implemented the fourth required component for principals—professional development—because we did not have such data for principals. The percentage of districts meeting each requirement was identical in Years 2 and 3.

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 had discretion in choosing 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.[27]

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 some also chose to evaluate teachers based on classroom achievement growth. To evaluate teachers in Year 3, all evaluation districts reported using school achievement growth, 70 percent reported using classroom achievement growth, and 40 percent reported using achievement growth of student subgroups (Table IV.2).

Table IV.2. Measures of Student Achievement and Observations of Practices Used to Evaluate Teachers and Principals, as Reported by Evaluation Districts, Year 3 (Percentages)

Performance Measure	Teachers	Principals
Student Achievement		
Student achievement level	30	30
Student achievement growth	100	100
School achievement growth	100	100
Achievement growth of student subgroups[a]	40	60
Classroom achievement growth	70	NA
Observation Measure		
Conducting at least two observations by trained observer	100	100
Number of Districts	**10**	**10**

Source: District survey, 2014.

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.

NA is not applicable.

To evaluate principals, all evaluation districts used school achievement growth, and 60 percent used achievement growth of student subgroups (Table IV.2).

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 growth measures provided by the state and four districts used models developed by private vendors.

[27] 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 may have implemented these measures regardless of receiving a TIF grant.

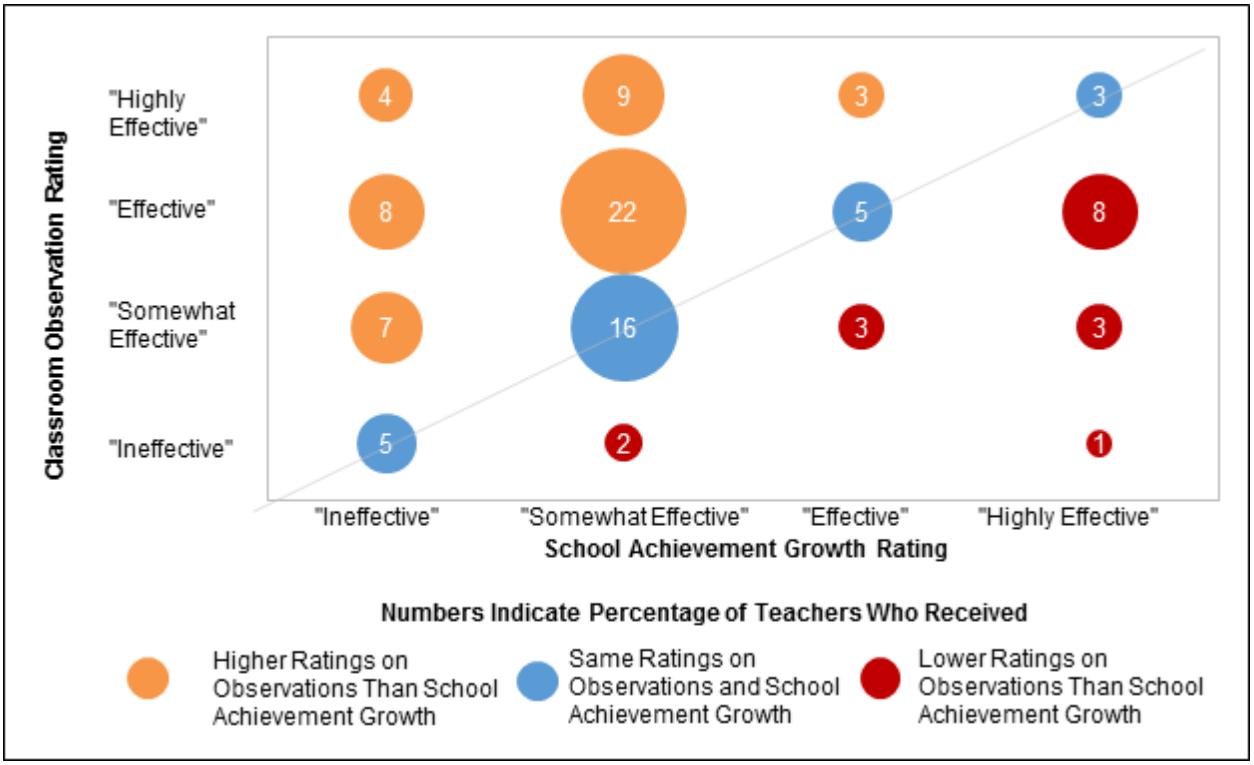
Districts also had discretion in meeting the requirement to conduct observations of classroom or school practices. For example, districts could decide which rubrics they wanted to use to observe teachers and principals, the number of observations in a year (as long as there were at least two), and which staff 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. On average in Year 3, evaluation districts reported conducting four classroom observations per year, each about 40 minutes long (Appendix D, Table D.1). Most often, evaluation districts reported that Year 3 classroom observations were conducted by the principal or other administrators at the teacher's school (78 percent), although about half of the districts (44 percent) also reported that teacher leaders or peer observers conducted classroom observations.

Teachers may 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 may 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 may be designed to evaluate different aspects of performance, so they do not necessarily need 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. The blue circles (those on the diagonal) show the percentages of teachers who received similar ratings on the two measures. For example, 16 percent of teachers received a rating of "somewhat effective" on both classroom observations and school achievement growth. The gold circles (those above the diagonal) represent teachers who received a higher rating on classroom observations than on school achievement growth. For example, nine 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. In Year 3, 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 teachers—slightly more than half (53 percent)—received a higher rating on classroom observations than on school achievement growth (represented by the gold 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 1–4 rating scale—might be expected since these measures could be measuring different aspects of teacher effectiveness. But a difference of two rating levels could be sending a mixed message to teachers about their effectiveness. One-fifth (21 percent) of teachers received a classroom observation rating that was at least two levels above their school achievement growth rating, whereas only 4 percent received a school achievement growth rating at least two levels above their observation rating. These patterns were even more pronounced among principals. More than two thirds of principals (69 percent) received a higher rating based on observations of their practices than on school achievement growth, and 44 percent received observation ratings that were at least two levels higher than their school achievement growth ratings (Appendix D, Table D.2).

Figure IV.1. Comparison of Teachers' Ratings on Classroom Observations and School Achievement Growth in Year 3 (Percentages)



Source: Educator administrative data (N = 3,642 teachers).

Notes: Categories are study-constructed labels to represent quarters of a 1-to-4 rating scale. "Ineffective" = bottom quarter (1 to 1.75); "Somewhat Effective" = second quarter (1.75 to 2.5); "Effective" = third quarter (2.5 to 3.25); "Highly Effective" = top quarter (3.25 to 4). The figure is based on teachers with ratings on both classroom observations and school achievement growth in Year 3.

Figure reads: In Year 3, 4 percent of teachers received a classroom observation rating of "highly effective" and a school achievement growth rating of "ineffective".

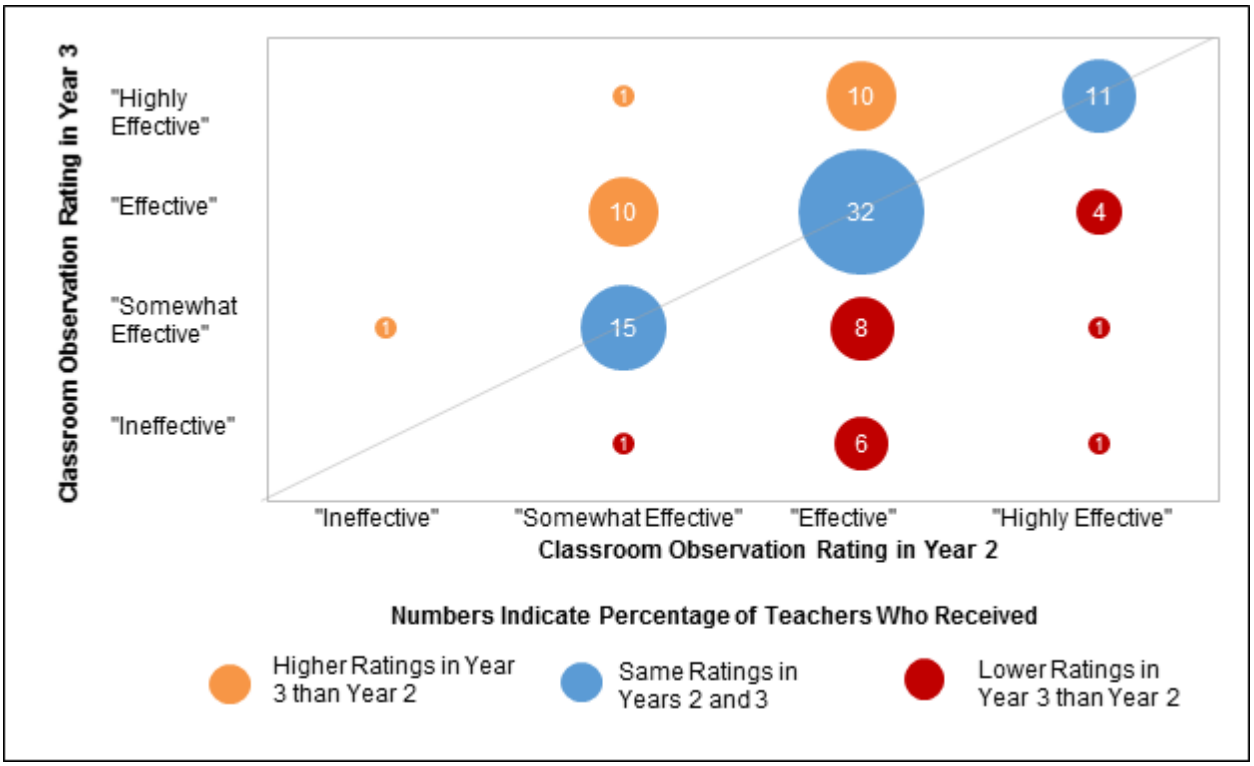
Many teachers received higher ratings on classroom observations than on classroom achievement growth. Although it might be expected that school achievement growth ratings would differ from individual observation ratings since school achievement growth is based on the collective work of school staff, we found similar patterns among teachers who were evaluated on classroom achievement growth.[28] For example, 28 percent of these teachers received similar observation and classroom achievement growth ratings, and 50 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.

Figure IV.2 illustrates the percentages of teachers who received each possible combination of ratings based on classroom observations for Years 2 and 3. Similar to Figure IV.1, the blue circles

[28] Within the seven districts that used classroom achievement growth in Year 3, 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).

(those on the diagonal) show the percentages of teachers who received similar classroom observation ratings in Years 2 and 3. For example, 15 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 3 than Year 2. For example, eight percent of teachers received a rating of "somewhat effective" in Year 3 and a rating of "effective" in Year 2.

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



Source: Educator administrative data (N = 2,575).

Notes: Categories are study-constructed labels to represent quarters of a 1-to-4 rating scale. "Ineffective" = bottom quarter (1 to 1.75); "Somewhat Effective" = second quarter (1.75 to 2.5); "Effective" = third quarter (2.5 to 3.25); "Highly Effective" = top quarter (3.25 to 4). The figure is based on teachers with classroom observations ratings in both Years 2 and 3.

Figure reads: One percent of teachers received a classroom observation rating of "ineffective" in Year 2 and "somewhat effective" in Year 3.

On each performance measure, most teachers received similar ratings in Year 3 as they did in Year 2. More than half of teachers received similar ratings, based on a 1-to-4 rating scale, in Years 2 and 3. Specifically, 58 percent of teachers received a similar rating based on classroom observations, 56 percent received a similar rating based on student achievement growth in their schools, and 55 percent received a similar rating based on classroom achievement growth (Figure IV.2 and Appendix Tables D.4 and D.5). Of those teachers who received different ratings for the same measure in Years 2 and 3, they 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-fourth to one-third of teachers earned a rating in Year

3 that was one level higher or lower than their rating in Year 2, whereas fewer than 20 percent of teachers earned ratings that differed by two or more levels from one year to the next.[29]

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, teachers' responses may depend on whether they had to meet a minimum classroom observation rating to receive a bonus based on student achievement growth or they could receive separate bonuses for each performance measure. Their responses 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—in particular, how closely the bonuses aligned with the guidance provided in the TIF grant. The evaluation design was based on random assignment of the pay-for-performance bonus component of the TIF program to some schools (the treatment schools) and not others (control schools).

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 need to perform well on only one measure to earn a bonus. Educators might even choose to focus improving their performance only on the measure (or measures) that they believed they could change most easily.[30]

All evaluation districts met the TIF grant requirement to offer teachers pay-for-performance bonuses, and all chose to offer separate bonuses for different performance measures. In Year 3, 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 40 percent provided bonuses for achievement growth

[29] Findings were similar when comparing the ratings that teachers received in Years 1 and 2. For example, nearly or more than half of teachers received similar ratings in Year 2 as they did in Year 1 on classroom observations (66 percent), school achievement growth (56 percent), and classroom achievement growth (48 percent).

[30] The 2012 TIF competition required grantees to assign educators one overall evaluation rating that combines 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.)

of student subgroups.[31] Most districts set an absolute maximum bonus that could be earned for each measure, but in some districts, the maximum bonus that could be earned depended on the number of bonus recipients (Table IV.3).[32]

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

	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 in tested grades and subjects could receive a bonus for classroom achievement growth			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			X	X	X	X		X	X	X
District changed some aspect of its program between the 2012–2013 and 2013–2014 school years			X						X	X

Source: District interviews (2012, 2013, and 2014); 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.

As discussed in Chapter I, although districts had discretion to specify the structure of performance bonuses, the TIF grant notice provided guidance to these districts by giving examples of bonuses that were *substantial* (with an average bonus worth 5 percent of the average educator salary), *differentiated* (with at least some educators receiving a payout worth three times the average bonus), and *challenging* to earn (with only those performing significantly better than the average receiving bonuses). This guidance was intended to encourage districts to structure bonuses in a way that would motivate teachers to improve their effectiveness. For example, teachers may pay little attention to a bonus program that only offers small bonuses. Even if bonuses were generally large, teachers would have

[31] In contrast, most (67 percent) of the Cohort 2 districts used a single, combined performance rating to determine bonuses.

[32] Appendix D, Tables D.6 and D.7 provide summary and detailed information, respectively, on teacher pay-for-performance programs for Cohorts 1 and 2.

little monetary incentive to improve if nearly everyone got a bonus or if higher and lower performers received similar bonuses.

At least half of the districts awarded the highest-performing teachers a pay-for-performance bonus at least three times the average bonus. Fifty percent of evaluation districts met the guidance for awarding differentiated performance bonuses for teachers in Year 3, and 60 to 70 percent met this guidance in the previous years (Table IV.4).[33] On average across evaluation districts, the maximum bonus ($7,743 in Year 3) was about four times the average bonus ($1,851 in Year 3) in treatment schools (Figure IV.3).[34]

Most teachers received a bonus, which, on average, was smaller than suggested by the TIF grant guidance. One-fifth or fewer of the districts met the guidance for awarding bonuses that were challenging to earn (Table IV.4). Across districts, on average, more than 70 percent of treatment teachers received a bonus, and the distribution of performance bonuses remained relatively stable across years (Figure IV.4).[35] In each year, 20 percent of evaluation districts met the guidance for awarding substantial bonuses for teachers (Table IV.4). Across evaluation districts, the average bonus for treatment teachers was about $1,850, or about 4 percent of the average teacher salary (Figure IV.3).[36]

Table IV.4. Evaluation Districts Meeting TIF Grant Goals for Pay-for-Performance Bonuses for Teachers (Percentages)

TIF Grant Goal	Year 1	Year 2	Year 3
Substantial: Average bonus was at least 5 percent of average salary	20	20	20
Differentiated: Highest bonus was at least three times the average bonus	70	60	50
Challenging: Fewer than 50 percent of teachers received a pay-for-performance bonus	20	20	10
Number of Districts	**10**	**10**	**10**

Source: Educator administrative data.

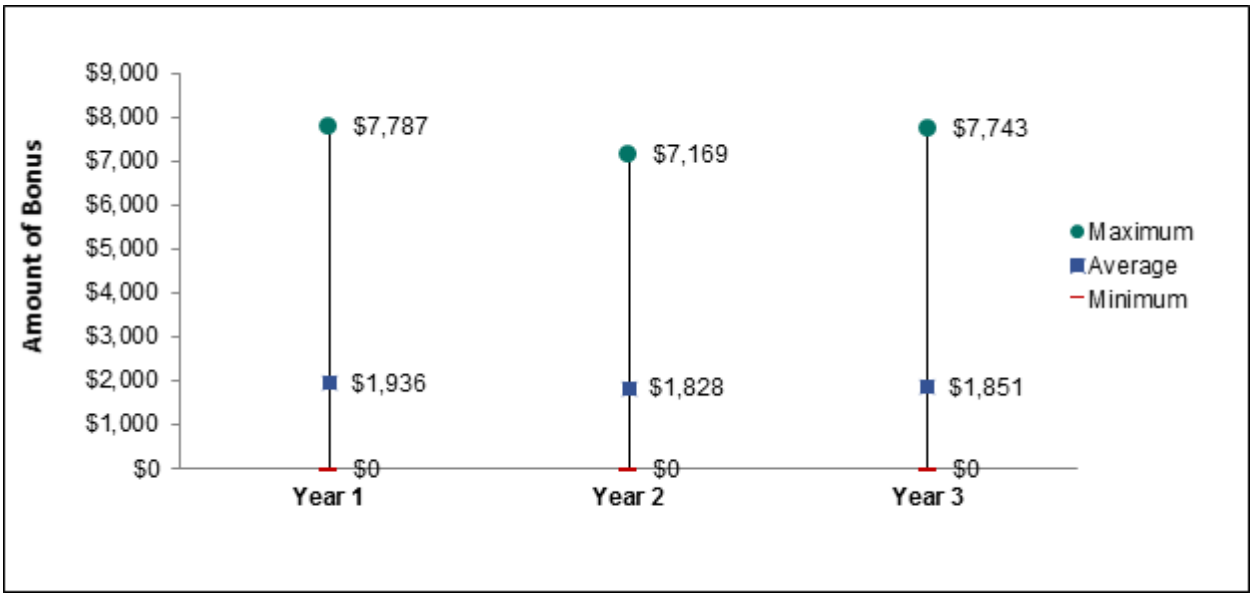
[33] In Years 1 and 2, when findings were based on both Cohorts 1 and 2, the percentages of districts meeting the guidance for awarding bonuses that were substantial, differentiated, or challenging to earn were similar to the percentages of Cohort 1 districts meeting the guidance (Appendix D, Table D.8).

[34] When Year 2 findings were based on both Cohorts 1 and 2, the average ($1,837) and maximum ($6,846) performance bonus amounts were similar to the average and maximum bonus amounts for Cohort 1 only (Appendix D, Figure D.1).

[35] Appendix D, Figures D.3 and D.4 show the percentage of teachers who earned a bonus, by district. In Year 3, all but one Cohort 1 district awarded performance bonuses to 60 percent or more of its treatment teachers (Appendix D, Figure D.3). In Year 2, most Cohort 1 and 2 districts (10 out of 13) awarded performance bonuses to at least 50 percent of their treatment teachers (Appendix D, Figure D.4).

[36] We calculated whether bonuses were substantial using the average teacher salary that districts specified during interviews. The average salary across the 10 evaluation districts in Year 3 was about $49,000 for teachers and $90,000 for principals.

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



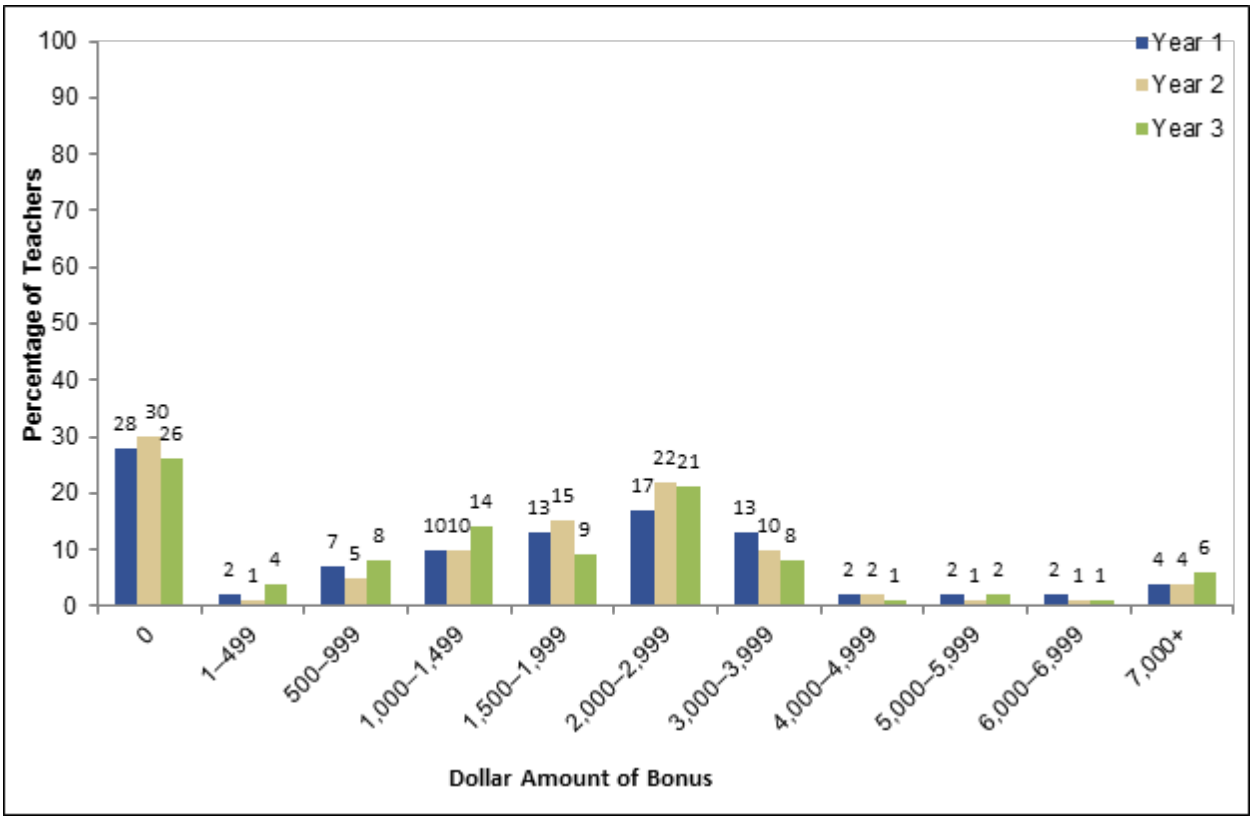
Source: Educator administrative data (N = 2,183 in Year 1, N = 2,193 in Year 2, and N = 2,260 in Year 3).

Note: 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 $1,936, and the maximum pay-for-performance bonus was $7,787.

Districts may have opted to award most teachers a performance bonus—perhaps to gain teachers' support for the program—yet only award large bonuses to a small number of teachers. Across districts, about half of teachers (48 percent) in Year 3 received a performance bonus of at least $1,500, which is about three times the automatic 1 percent bonus that control teachers received (Figure IV.4). However, only 7 percent received a performance bonus of at least $5,000, or approximately 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



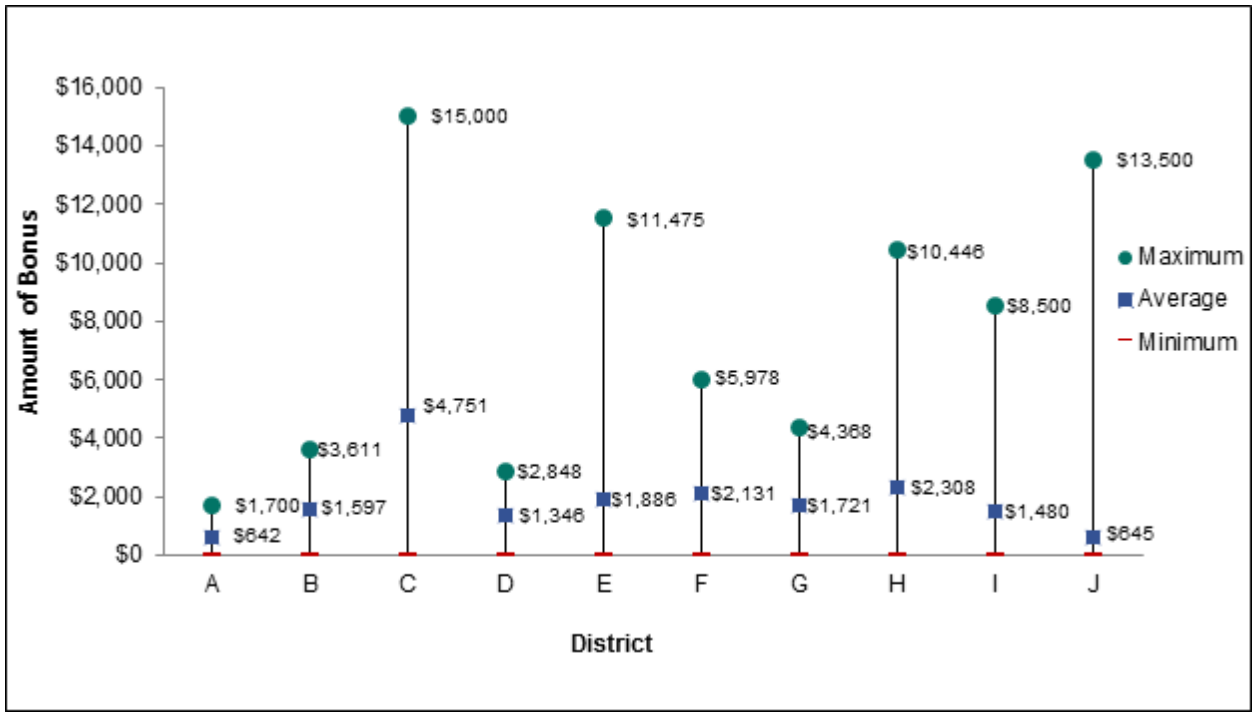
Source: Educator administrative data (N = 2,189 teachers in Year 1, N = 2,191 teachers in Year 2, and N = 2,260 teachers in Year 3).

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 between $1 and $499.

Maximum performance bonus amounts for teachers varied substantially across districts. The average of the 10 districts' maximum performance bonus amounts (Figure IV.3) masks considerable differences across districts in the maximum bonus that teachers earned. In Year 3, maximum performance bonus amounts were at least $10,000 in four districts, between $4,000 and $8,500 in three districts, and less than $3,700 in three districts (Figure IV.5).[37] This variation suggests that setting the range of performance bonuses was an important dimension on which the evaluation districts could exercise discretion in designing their TIF program, and this led to substantially different maximum bonus amounts.

[37] Maximum performance bonus amounts varied to a similar extent across all Cohort 1 and 2 districts in Year 2. In particular, the maximum bonus amounts for the three districts in Cohort 2 ranged from $5,300 to $6,000 (Appendix D, Figure D.5). 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 Tables D.6 and D.7.

Figure IV.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).

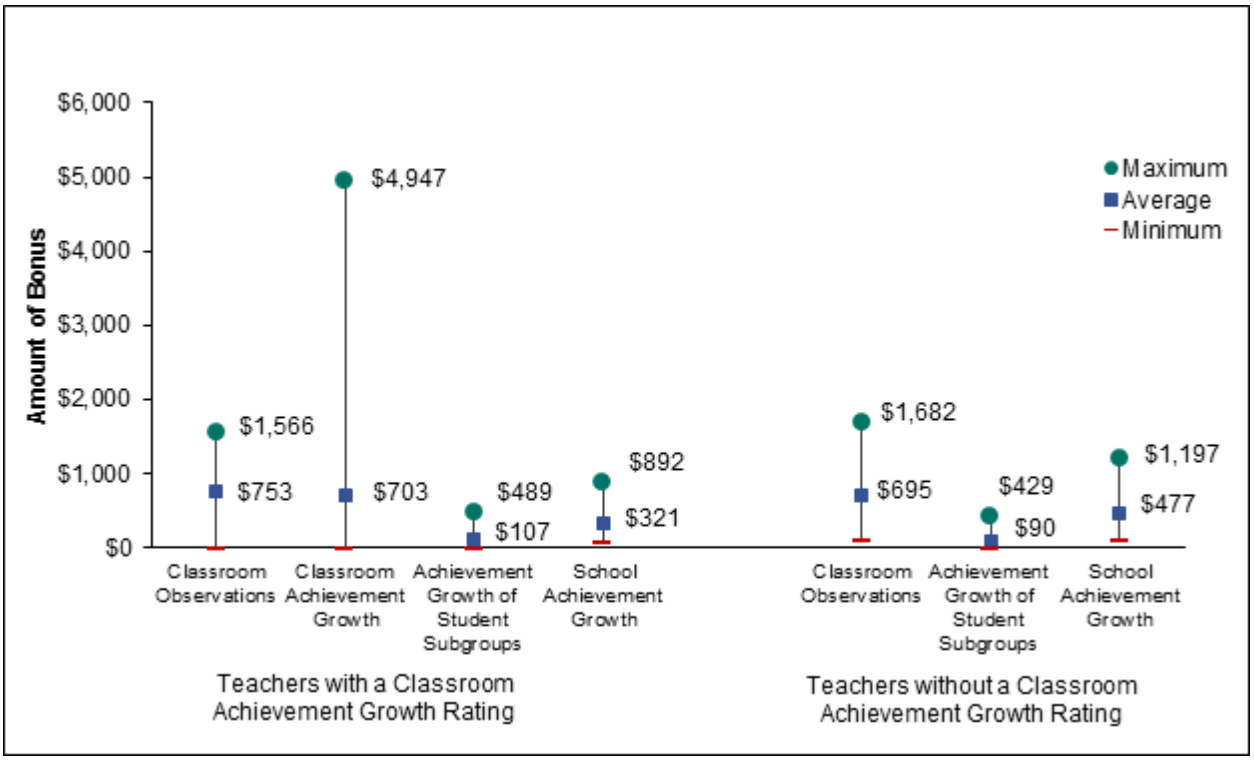
Figure reads: For District A in Year 3, the minimum pay-for-performance bonus was $0, the average pay-for-performance bonus was $642, and the maximum pay-for-performance bonus was $1,700.

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 needed 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 needed 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. Within the seven districts that used classroom achievement growth, 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 $5,000 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, teachers received a total bonus of $1,884 ($753 for classroom observations, $703 for classroom achievement growth, $321 for school achievement growth, and $107 for achievement growth of student subgroups). Therefore, these teachers on average earned almost 3.5 times as large a bonus based on their individual performance ($1,456) than based on a group's performance ($428). They also earned, on average, more of their bonus based on student achievement growth than on classroom observations ($1,131 versus $753).

Figure IV.6. Minimum, Average, and Maximum Performance Bonus for Each Performance Measure, in Districts Using Classroom Achievement Growth Measures, Year 3



Source: Educator administrative data (N = 944 teachers with a classroom achievement growth rating, and N = 415 teachers without a classroom achievement growth rating).

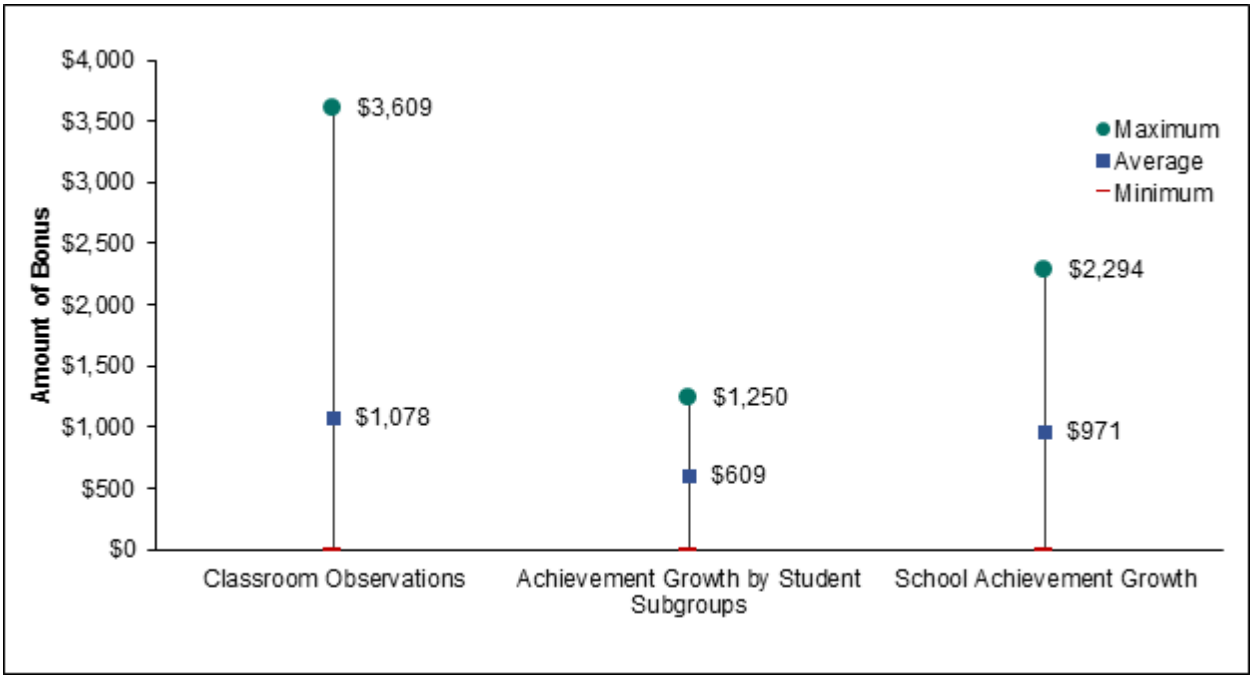
Notes: Seven districts used classroom achievement growth measures in Year 3. Figure is based on teachers in those districts who received classroom observation ratings.

Figure reads: On average across districts that used classroom achievement growth measures in Year 3, among teachers with a classroom achievement growth rating, the minimum bonus for classroom observation ratings was $0, the average was $753, and the maximum was $1,566.

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 ($1,262) that was about two-thirds the size of the total bonus earned by teachers who were assessed on classroom achievement growth ($1,884; Figure IV.6). On average, these teachers earned more for classroom observations ($695) than for the group performance measures based on student achievement growth ($567).

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 slightly more than $2,600 across the three main performance measures—classroom observations, school achievement growth, and the achievement growth of student subgroups (Figure IV.7). Almost 60 percent ($1,580) of their overall bonus came from group performance measures based on student achievement growth ($971 for school achievement growth and $609 for achievement growth of student subgroups).

Figure IV.7. Minimum, Average, and Maximum Performance Bonus for Each Performance Measure, in Districts Not Using Classroom Achievement Growth Measures, Year 3



Source: Educator administrative data (N = 460).

Notes: Three districts did not use classroom achievement growth measures in Year 3. 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 3, the minimum bonus for classroom observation ratings was $0, the average was $1,078, and the maximum was $3,609.

To explore whether the bonuses that teachers earned were providing them with consistent messages about their performance over time, we examined the total bonus amounts that teachers received in Years 2 and 3.

Most teachers received similar performance bonus amounts in Years 2 and 3. Fifty-seven percent of teachers received a bonus amount in the same range in Years 2 and 3. In Years 2 and 3, 18 percent of teachers did not earn a bonus, 11 percent earned a bonus ranging of $1,500 or less, 18 percent earned a bonus ranging from $1,501 to $3,000, and 10 percent earned a bonus above $3,000 (Appendix Table D.9). For each range of bonus amounts, teachers were also most likely to remain in

the same range. For example, most of the teachers who earned more than a $3,000 bonus in Year 2 continued to do so in Year 3.[38]

Most evaluation districts met the TIF guidance for awarding substantial bonuses for principals, but few awarded bonuses that were differentiated and challenging to earn. 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. Sixty percent of evaluation districts met the guidance for awarding substantial bonuses for principals in Year 3, an increase from the 30 percent that met this guidance in Years 1 and 2 (Table IV.5). Across districts, the average bonus for treatment principals in Year 3 ($4,039) was slightly less than 5 percent of the average principal salary (Figure IV.8). However, as in Years 1 and 2, only 10 percent of districts met the guidance for differentiated bonuses in Year 3 (Table IV.5). On average across districts, the maximum bonus ($7,307 in Year 3) was less than twice the average bonus (Figure IV.8).[39] Likewise, in each year, no more than 20 percent of the districts met the guidance for awarding bonuses for principals that were challenging to earn, and at least three-fourths of treatment principals received a bonus (Table IV.5 and Appendix D, Figure D.8).

Table IV.5. Evaluation Districts Meeting TIF Grant Goals for Pay-for-Performance Bonuses for Principals, (Percentages)

TIF Grant Goal	Year 1	Year 2	Year 3
Substantial: Average bonus was at least 5 percent of average salary	30	30	60
Differentiated: Highest bonus was at least three times the average bonus	10	10	10
Challenging: Fewer than 50 percent of teachers received a pay-for-performance bonus	20	20	10
Number of Districts	**10**	**10**	**10**

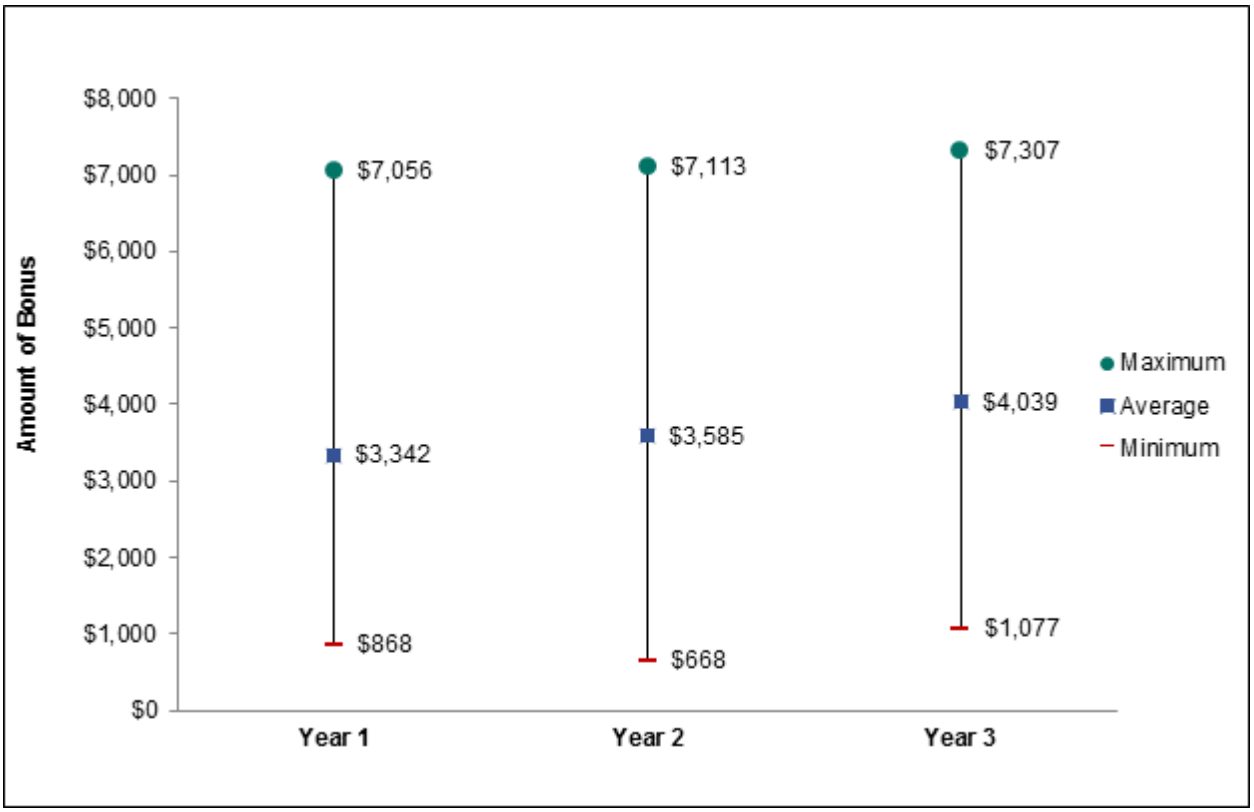
Source: Educator administrative data.

As intended by the study design, the automatic 1 percent bonus provided to teachers and principals in control schools was small and 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 $433 in Year 3, and the maximum automatic bonus was only slightly higher ($672 in Year 3; Appendix D, Figure D.9). For principals in control schools, the average automatic bonus was $764 in Year 3, with a maximum automatic bonus of $957. Both teachers and principals in control schools received automatic bonuses that were, on average, approximately 20 percent of the average amount of the performance bonuses that their counterparts in treatment schools received.

[38] Likewise, most teachers (53 percent) received a similar bonus amount in Year 2 as they did in Year 1.

[39] When Year 2 findings were based on both Cohorts 1 and 2, the average ($3,444) and maximum ($6,442) performance bonus amounts were slightly lower than the corresponding amounts for Cohort 1 only (Appendix D, Figure D.6).

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



Source: Educator administrative data (N = 65 principals in Year 1, N = 68 principals in Year 2, and N = 65 principals in Year 3).

Note: 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.7).

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 $3,342 and the maximum pay-for-performance bonus was $7,056.

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 or mentor/lead teacher opportunities. All districts reported offering additional pay for teachers to take on extra roles and responsibilities, and three reported offering similar opportunities to principals. Districts most commonly reported offering teachers additional pay for the roles of master and mentor teachers—70 percent of evaluation districts reported offering these roles (Table IV.6). During interviews, officials from districts that

offered these roles reported that the number of master teacher positions available within districts ranged from 6 to 61 (depending on the number of study schools) and that the number of mentor teacher positions at each school ranged from 1 to 6. Districts noted that master teachers might lead professional development sessions and mentor teachers might provide day-to-day coaching or modeling of lessons.

Table IV.6. Additional Pay Opportunities, as Reported by Evaluation Districts, Year 3

	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	70	$2,857
Master or lead teacher	70	$7,714
Department chair or head	10	—
Lead curriculum specialist	22	—
Serving on a schoolwide committee or task force	0	—
Leadership team member	10	—
Additional Factors		
Teaching in a hard-to-staff school or high-need subject area	30	$5,333
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	33	$5,500
Number of Districts—Range[a]	**9–10**	**3–7**

Source: District survey, 2014.

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.

We compared the amount of money teachers could earn for these additional pay opportunities to the amount they could earn for pay-for-performance bonuses. According to the theory of change (Chapter I), pay-for-performance is expected to 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 in order to be qualified for these additional pay opportunities. If so, these additional pay opportunities could have diminished the potential impacts of pay-for-performance.

Although teachers could potentially earn as much money from taking on additional responsibilities as from pay-for-performance bonuses, they actually earned less, on average, from these additional pay opportunities. In Year 3, the reported maximum additional pay of $7,714 for serving as a master or lead teacher (among evaluation districts offering this type of pay) was about the same amount as the maximum pay-for-performance bonus of $7,743 (among all evaluation districts; Table IV.6). However, the average actual pay for additional roles and responsibilities in Year 3 was $498, less than 30 percent of the average performance bonus for teachers of $1,851 (Appendix D, Table D.10). This is because only a small fraction of teachers (17 percent) received additional pay for extra

work.[40] Additional pay opportunities may also 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.

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.[41] To describe this required component, we used data from the district survey and interviews with district administrators.

Most districts offered teachers advice on how to improve their observation rating, but few offered professional development to improve teachers' ratings based on student achievement growth. During interviews, eight districts reported offering teachers professional development that was specific to the teacher's classroom observation rating, but only three districts reported providing professional development specific to the teacher's rating on the student achievement growth measure(s) (Appendix D, Table D.12). Almost all evaluation districts (90 percent) offered professional development to help teachers understand the performance measures used for their TIF program (Table IV.7).

Table IV.7. Professional Development Activities for Teachers Planned Under TIF, as Reported by Evaluation Districts, Year 3 (Percentages)

	Evaluation Districts
Focus of Professional Development	
Understanding performance measures of TIF program	90
Feedback based on TIF performance ratings	70
Number of Districts	**10**

Source: District survey, 2014.

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 was communicated to educators. We focus on two types of information that districts needed to communicate in Year 3: general information about the program and specific information to individual teachers about the performance ratings and bonuses they earned in Year 2. Data for this section come from the district survey and interviews with district administrators.

District or grantee staff, rather than school staff, typically communicated general information about TIF programs to teachers. Deciding who communicates about TIF involves a trade-off. Communication by district or grantee staff might help ensure uniformity and accuracy of information, but communication by school staff (for example, asking principals to explain the program

[40] Average amounts of additional pay for roles and responsibilities did not differ between teachers in treatment and control schools (see Appendix D, Table D.11).

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

to their teachers) uses staff who might have closer relationships with the teachers.[42] In Year 3, 6 of 10 districts reported that communication about TIF came from district or grantee staff (Table IV.8).

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 chose to communicate information through a variety of approaches. On average, districts used about three approaches to communicate aspects of their TIF program to educators in Year 3 (Table IV.8). Most districts (90 percent) reported using written materials, group presentations, and the district website to explain how teachers would be evaluated. 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 presentations (90 percent), the district website (80 percent), and written materials (70 percent) to inform educators of the potential amount they could earn in performance bonuses.

Table IV.8. Districts' Communication Activities in Year 3 (Percentages Unless Otherwise Noted)

	Evaluation Districts
Responsible for Majority of Communication about TIF	
District or grantee official	60
School-level staff (such as a principal or lead teacher)	30
TIF coach	10
Average Number of Communication Methods Used by Districts About How Teachers Would be Evaluated	2.9
Average Number of Communication Methods Used by Districts About Potential Amounts of TIF Bonuses	2.6
Communication Methods on How Teachers Would be Evaluated	
Written materials (including letters, email, brochures, program manuals, newsletters)	90
Group presentation	90
Individual meetings	10
District website	90
Communication Methods on Potential Amounts of TIF Bonuses	
Written materials (including letters, email, brochures, program manuals, newsletters)	70
Group presentation	90
Individual meetings	10
District website	80
Used Survey or Focus Group to Assess if Teachers Understood their Eligibility for a Bonus	60
Number of Districts	**10**

Source: District interviews, 2014.

Most districts assessed teachers' understanding of their eligibility to earn a performance bonus. 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

[42] 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.

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. Sixty percent of districts reported using a survey or focus group to check if teachers understood their eligibility for a bonus (Table IV.8).[43]

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 and what they earned in prior years, teachers also may 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 what they received in the past. Nevertheless, only 3 of the 10 districts reported informing educators about the percentage of teachers who received a bonus in Year 2, and only one district reported providing information about the maximum or average bonuses awarded in Year 2 (Table IV.9).

Table IV.9. Information Districts Provided to Teachers About Actual Pay-for-Performance Bonuses from Year 2 (Percentages)

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

Source: District interviews, 2014.

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

NA is not applicable.

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 is also likely to 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 rating may be more aware of their measured effectiveness and how to change their practices to improve their ratings. Letters or e-mails 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

[43] Although this feedback might lead to improved communication and better educator understanding, teachers in districts that used a survey or focus group to check if teachers understood their eligibility did not have a better understanding than teachers in other districts (Appendix D, Table D.18).

the information. In-person communication also allows teachers the opportunity to discuss with their principal (or supervisor) the basis for the rating and how to improve on it. Ninety percent of the districts reported that they used in-person meetings to inform teachers about both their observation rating and their student achievement growth rating from Year 2 (Appendix D, Table D.13). More than half of the districts also reported using an online system to communicate these ratings.

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 may 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 awards; 60 percent reported informing nonrecipients that they did not earn a bonus (Table IV.9). In practice, because most teachers received a bonus, relatively few treatment teachers (about 10 percent) were not 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 may 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. Most of the districts (80 percent) reported informing bonus recipients of their individual performance bonus by sending a letter. Thirty percent reported holding individual meetings with teachers to discuss the bonus amount they received (Appendix D, Table D.14).

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). There were differences among evaluation districts in the timing of notifying teachers of their bonuses from Year 2 (2012–2013) and paying out those bonuses. Of the nine districts that paid out bonuses within 12 months after the end of 2012–2013, only three reported notifying and paying any teachers before the start of the 2013–2014 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 six districts reported notifying and paying teachers between October 2013 and January 2014.

Teacher and Principal Perspectives Regarding TIF Implementation

Teachers' and principals' understanding of the TIF program is important because it reflects how well the program's incentives were communicated and in turn can determine how the program may influence educators' behaviors and ultimately student achievement (as described by the theory of change discussed in Chapter I). Moreover, educators' reports about program features can identify ways

in which their understanding of 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 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 perspectives 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 Year 2 and Year 3 surveys after bonuses had been awarded for one and two 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 need to understand how they are being evaluated, as a first step toward figuring out how to improve their performance.

Teachers' understanding of performance measures was fairly high and continued to improve. More than 75 percent of teachers in the third year reported being evaluated on student achievement growth, and over 85 percent reported being evaluated on at least two classroom observations (Table IV.10). Furthermore, a higher percentage of teachers in Year 3 reported being evaluated on student achievement growth (84 percent of treatment teachers and 78 percent of control teachers) than in Year 2 (78 percent of treatment teachers and 72 percent of control teachers; Table IV.10). Teachers' improved awareness about performance measures in Year 3 continued a trend that had begun earlier. For example, a higher percentage of teachers in Year 2 reported being evaluated on at least two classroom observations (87 percent of treatment teachers and 83 percent of control teachers) than in Year 1 (74 percent of treatment teachers and 76 percent of control teachers).

Educators in treatment schools continued 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, in Years 2 and 3, treatment teachers were 6 percentage points more likely to report being evaluated on student achievement growth than control teachers (for example, 84 percent of treatment teachers and 78 percent of control teachers in Year 3; Table IV.10). Treatment principals also were more likely than control principals (99 versus 81 percent in Year 3) to report being evaluated on student achievement growth (Table IV.11). This suggests that the offer of pay-for-performance led educators in treatment schools to be more aware of how they were evaluated.

Table IV.10. Teachers' Reports of the Measures Used to Evaluate Teachers (Percentages)

	Year 1		Year 2		Year 3	
	Treatment	Control	Treatment	Control	Treatment	Control
Student Achievement Measures						
Student achievement level	56	61	69 +	67	74	71
Student achievement growth	71	70	78*+	72	84*+	78+
School achievement growth	62	63	73*+	68	79*+	74
Achievement growth of student subgroups[a]	55	56	66*+	60	67	68+
Classroom achievement growth	60	62	57	58	65 +	64
Classroom Observation Measure						
At least two classroom observations by trained observers	74	76	87 +	83+	91	88
Number of Teachers—Range[b]	**382–384**	**393–398**	**432–437**	**432–434**	**413–419**	**427–431**

Source: Teacher survey (2012, 2013, and 2014).

[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.11. Principals' Reports of the Measures Used to Evaluate Principals (Percentages)

	Year 1		Year 2		Year 3	
	Treatment	Control	Treatment	Control	Treatment	Control
Student Achievement Measure						
Student achievement level	89	93	85*	69+	75	75
Student achievement growth	88	92	91*	67+	99*	81
School achievement growth	89	90	90*	65+	97*	81
Achievement growth of student subgroups[a]	83	90	83*	64+	79	69
Observation Measure						
At least two observations by trained observer	—	—	44	59	62+	51
Number of Principals—Range[b]	**59–63**	**58–60**	**63–64**	**57–58**	**58–59**	**57–59**

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

[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.

— is not available.

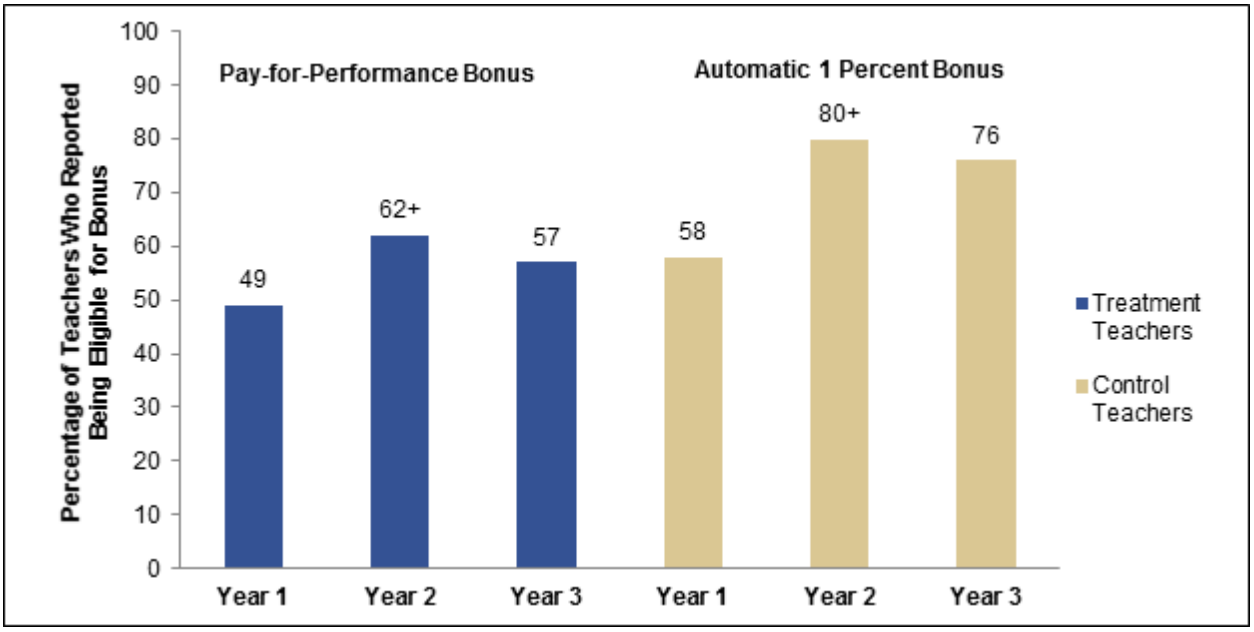
*Difference between treatment and control group is statistically significant at the .05 level, two-tailed test.
+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

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 need to 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.

More educators in treatment schools understood they were eligible for a performance bonus in Year 2 than Year 1, but there was no further improvement in Year 3. For example, a higher percentage of teachers in treatment schools correctly reported their bonus eligibility between Year 1 (49 percent) and Year 2 (62 percent; Figure IV.9). However, the percentage of teachers in treatment schools that reported being eligible for a performance bonus in Year 3 (57 percent) was similar to the percentage in Year 2 (62 percent; Figure IV.9). Among principals in treatment schools, a lower percentage reported being eligible for a performance bonus in Year 3 (78 percent) compared to Year 2 (90 percent; Figure IV.10).[44]

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



Source: Teacher survey (2012, 2013, and 2014).

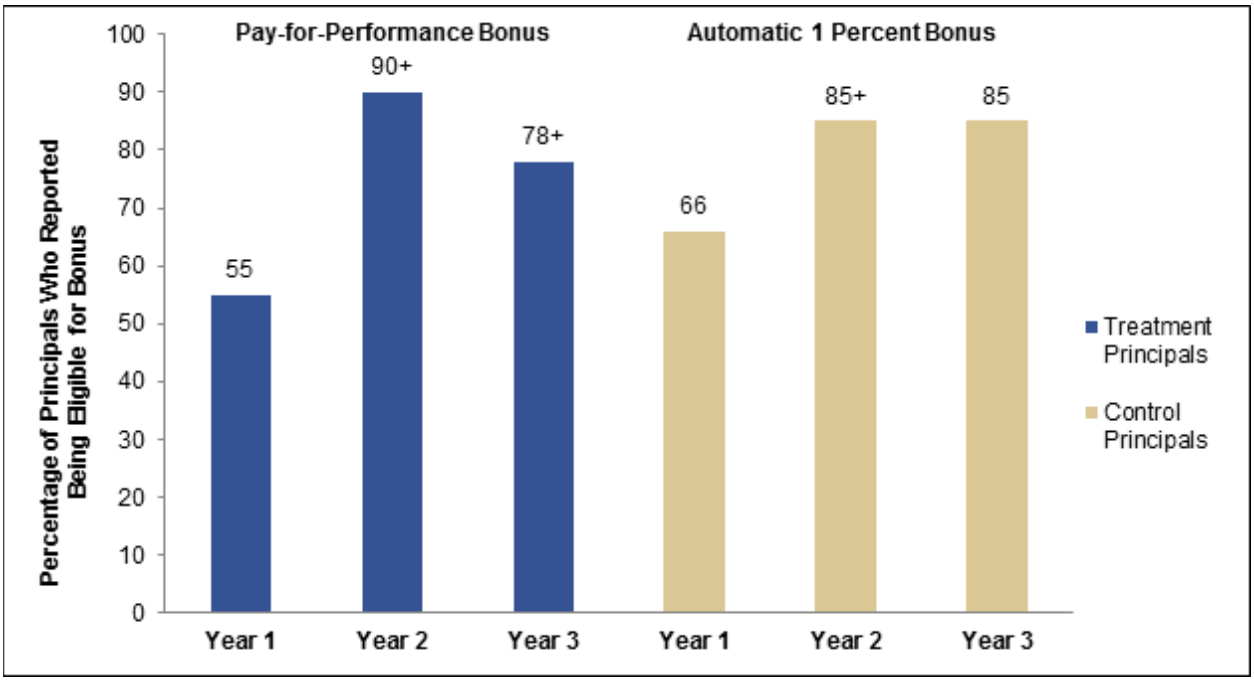
Notes: A total of 377 treatment teachers in Year 1, 444 in Year 2, and 424 in Year 3 responded to the question about eligibility for a pay-for-performance bonus. A total of 381 control teachers in Year 1, 445 in Year 2, and 448 in Year 3 responded to the question about eligibility for an automatic 1 percent bonus.

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.

[44] When we restricted the sample to principals who responded to the survey in both years, the results were similar. Therefore, the drop in the percentage of treatment principals reporting that they were eligible for pay-for-performance bonuses was not due to a change in which principals responded to the survey.

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



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

Notes: A total of 64 treatment principals in Year 1, 63 in Year 2, and 58 in Year 3 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, and 61 in Year 3 responded to the question about eligibility for an automatic 1 percent bonus.

Figure reads: In Year 1, 55 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.

Many teachers and some principals in treatment schools still did not understand they were eligible to earn a performance bonus. By the third year of TIF implementation, about 40 percent of treatment teachers were still unaware that they could potentially earn a performance bonus (57 percent of treatment teachers reported being eligible for a bonus in Year 3, implying that 43 percent of treatment teachers did not report being eligible for one; Figure IV.9.) Although understanding of eligibility was better among principals than teachers, about 20 percent of principals in Year 3 still did not know they were eligible to earn a bonus based on their performance (78 percent reported being eligible and 22 percent reported not being eligible; Figure IV.10).[45,46]

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

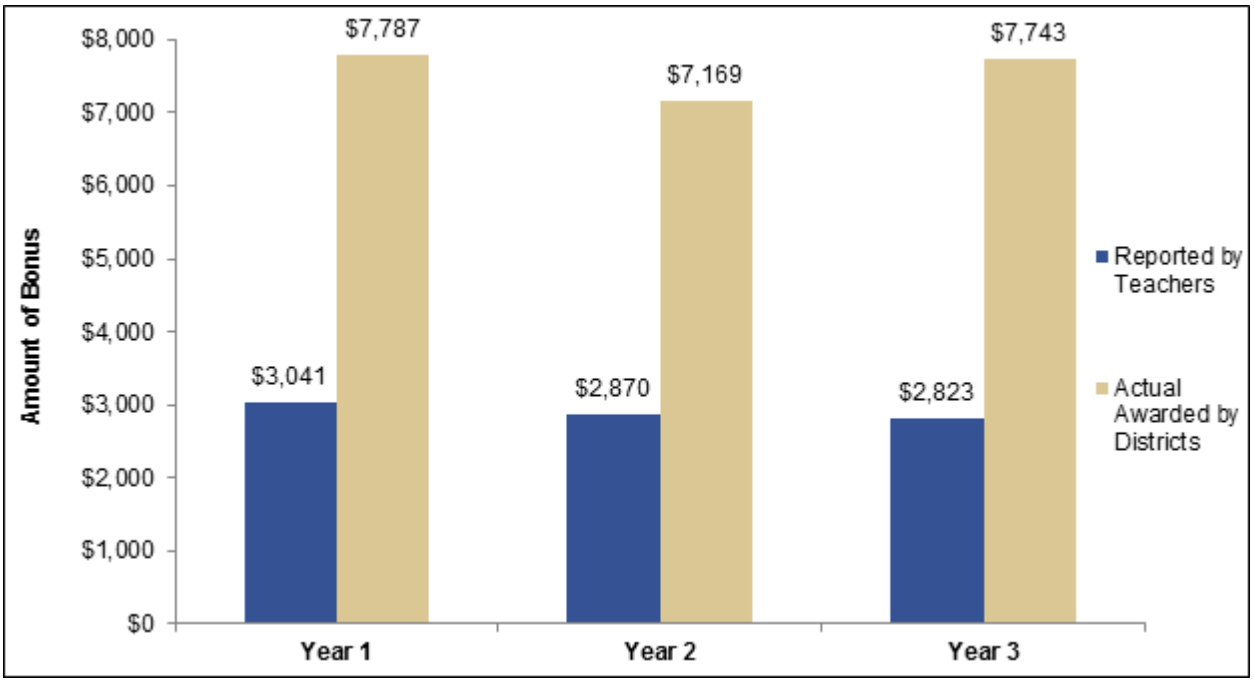
For performance bonuses to provide an incentive for teachers to change their behaviors, teachers not only need to understand they are eligible for a bonus but also must believe the potential amount of the bonus is enough to change their teaching practices or effort. Figure IV.11 shows, on average, the maximum performance bonus that teachers believed was available and the actual maximum

[45] When analyses for Years 1 and 2 were based on Cohorts 1 and 2, similar but somewhat smaller percentages of teachers and principals reported being eligible for the correct type of bonus (Appendix D, Figures D.10 and D.11).

[46] Some educators thought they were eligible for the wrong bonus. For example, in Year 3, 18 percent of control teachers and 13 percent of control principals thought they were eligible for pay-for-performance bonuses (Appendix D, Table D.15).

performance bonus that districts awarded to teachers. Teachers' expectations in Year 1 would have been primarily shaped by how well districts communicated the design of the pay-for-performance component to their teachers. In Years 2 and 3, however, teachers' expectations could also have been influenced by the actual bonuses awarded after Years 1 and 2.

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



Source: Teacher survey (2012, 2013, and 2014) 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.12 shows that our results are similar if districts are weighted by the number of schools when calculating districts' payouts.

A total of 196 treatment teachers in tested grades and subjects responded to this survey question in Year 1, a total of 218 in Year 2, and a total of 217 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 27 additional responses for treatment teachers in Year 1, 14 additional responses in Year 2, and 15 additional responses in Year 3. See Appendix B for additional discussion on the imputation methods. Appendix D, Table D.16 shows that our results are similar if we do not impute the missing bonus amounts.

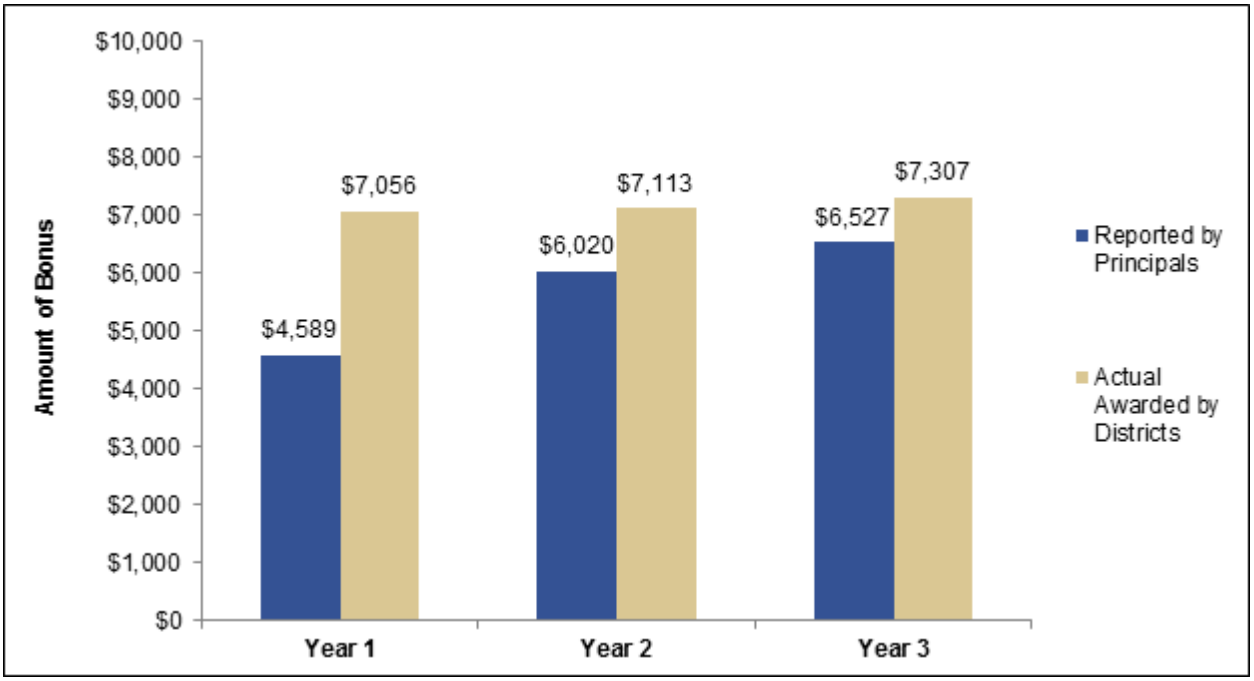
Figure reads: In Year 1 on average, the actual maximum pay-for-performance bonus that evaluation districts awarded to teachers was $7,787, and the maximum pay-for-performance bonus teachers reported they could earn was $3,041.

Teachers continued to underestimate 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 two-fifths the size of the actual maximum bonus districts awarded (Figure IV.11). For example, in Year 3, teachers in treatment schools, on average, reported that the maximum pay-for-performance bonus that teachers in their teaching position could receive was $2,823, whereas the actual maximum bonus awarded by districts was $7,743. The maximum bonuses reported by teachers include a maximum bonus of $0 for teachers who did not believe they were eligible for a performance bonus. Therefore, the maximum bonus reported by teachers, on average, may 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 that, on average, the maximum performance bonus they could receive was about $3,600 in all three years (not shown).

Principals also continued to underestimate the potential amount of performance bonuses they could receive, but their expectations were better aligned with actual bonus payouts than were teachers' expectations. In Year 3, principals in treatment schools, on average, reported that the maximum pay-for-performance bonus they could receive was $6,527, whereas the actual maximum bonus districts awarded to principals was $7,307 (Figure IV.12). Principals who correctly reported their eligibility for a performance bonus believed the maximum bonus they could receive was about $7,300 (not shown), nearly identical to the actual maximum awarded.[47]

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



Source: Principal survey (2012, 2013, and 2014) and educator administrative data.

Note: 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 ($8,344 in Year 1, $8,369 in Year 2, and $8,489 in Year 3), implying a somewhat wider gap between principals' reported maximum bonus amounts and the actual amounts (Appendix D, Figure D.13).

A total of 56 treatment principals responded to this survey question in Year 1, a total of 61 in Year 2, and a total of 58 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 8 additional responses for treatment principals in Year 1, 2 additional responses in Year 2, and 0 additional responses in Year 3. See Appendix B for additional discussion on the imputation methods. Appendix D, Table D.16 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 $4,589.

[47] Findings for teachers and principals were similar when analyses for Years 1 and 2 were based on Cohorts 1 and 2 (Appendix D, Figures D.14 and D.15).

Many treatment teachers were not aware they received a performance bonus. About 45 percent of the teachers who received a bonus based on their Year 2 performance did not report receiving one (Table IV.12). Almost all teachers (94 percent) who did not receive a bonus correctly reported not getting a performance bonus.

Table IV.12. Actual and Reported Receipt of Pay-for-Performance Bonus from Year 2 for Teachers in Treatment Schools in Year 3

	Actual Bonus Receipt in Year 2	
Belief About Bonus Receipt in Year 2	Percentage of Teachers Who Received a Bonus	Percentage of Teachers Who Did Not Receive a Bonus
Reported Receiving a Bonus	55	6
Reported Not Receiving a Bonus	45	94
Percentage of Treatment Teachers in Year 3	**100**	**100**

Source: Teacher survey (2014) and educator administrative data.

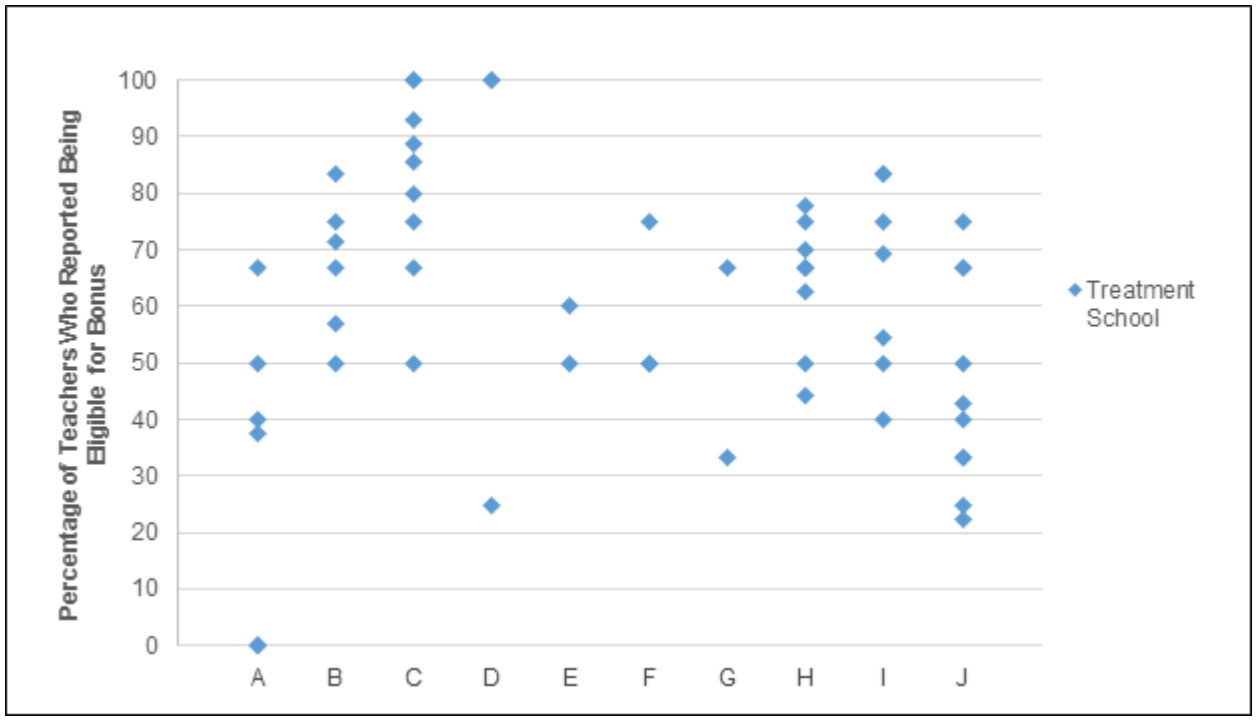
Notes: A total of 420 treatment teachers responded to the survey question about whether they received a bonus based on their performance last year. Of those, 229 teachers received an actual Year 2 bonus and 191 did not.

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 whether bonuses were included 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 may be explained by differences in teachers' characteristics, such as whether the teacher had ever received a bonus or whether the teacher attended TIF-related professional development sessions.

Most of the differences in teachers' understanding occurred among teachers in the same school. Figure IV.13 displays the variation in treatment teachers' understanding of eligibility for pay-for-performance bonuses for each evaluation district. 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. A diamond at the top of the figure (100 percent) indicates that all the teachers in that school correctly reported being eligible for a pay-for-performance bonus. As the figure shows, teacher understanding varied within districts and within schools. In fact, 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 more than 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.17). Similarly, most of the variation (more than 70 percent) in teachers' understanding of the maximum bonus they could earn occurred among teachers in the same school (Appendix D, Table D.17).

Figure IV.13. Treatment Teachers' Reported Pay-for-Performance Bonus Eligibility by School and by District, Year 3 (Percentages)



Source: Teacher survey, 2014 (N = 412 teachers in 58 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 3, of the five schools in District A that offered teachers pay-for-performance bonuses, no teachers in one school reported being eligible for a bonus, 38 percent of teachers in one school reported being eligible for a bonus, 40 percent of teachers in one school reported being eligible for a bonus, 50 percent of teachers in one school reported being eligible for a bonus, and 67 percent of teachers in one school reported being eligible for a bonus.

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. Since the maximum bonus amount awarded varied by district (ranging from $1,700 to $15,000, as shown in 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) assessed teachers' understanding using focus groups or surveys, (3) expected at least 75 percent of teachers to attend TIF-required professional development, (4) used classroom achievement growth to determine performance bonuses, (5) awarded an average pay-for-performance bonus that was at least 5 percent of the average teacher salary, (6) paid pay-for-performance bonuses through a separate check (rather than teachers' regular paycheck), and (7) told all treatment teachers the total bonus amount that they earned in Year 2 (including $0 for those who did not receive one). 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 2, (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 characteristics 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 key district implementation characteristics, with two exceptions (Appendix D, Tables D.18 and D.19). Teachers in districts that offered a high average bonus had a better understanding of their eligibility for performance bonuses than teachers in districts that offered a low average bonus (72 versus 50 percent; Table IV.13). Unexpectedly, teachers' understanding of their bonus eligibility and of the maximum bonus was worse in districts that informed all treatment teachers (including nonrecipients) of their bonus amounts than in districts that did not.

Teachers who received or reported receiving a bonus based on the prior year's performance had a better understanding of their eligibility for a performance bonus and of the maximum bonus they could earn. Because many teachers were unaware that they received a performance bonus, we examined how teachers' understanding in Year 3 varied with both their actual and reported receipt of a bonus based on Year 2 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 (Table IV.13). Teachers' belief of bonus receipt was, however, more strongly associated with understanding than actual bonus receipt. For example, 90 percent of the teachers who believed they received a bonus last year correctly reported being eligible for a performance bonus, compared to two-thirds of the teachers who actually received a bonus.

Teachers who were TIF mentors had a better understanding of their eligibility for performance bonus. Treatment teachers who reported being mentors as part of their TIF program were more likely to report they were eligible for a pay-for-performance bonus than those who were not TIF mentors (71 versus 56 percent; Table IV.13). TIF mentors also reported potential bonus amounts that were closer to the actual awarded compared to teachers who were not TIF mentors, but the difference was not statistically significant. Interestingly, having a mentor or being a mentor, but not as part of the TIF program, was not associated with understanding of their eligibility or of the maximum possible bonus (Appendix D, Table D.19).

None of the other characteristics we examined could account for the variation in teachers' understanding (Appendix D, Tables D.18 and D.19).

Educators' Understanding of and Experiences with Other Required 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 may also shed light on 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 required components identically in treatment and control schools.

Table IV.13. Treatment Teachers' Reported Eligibility for Pay-for-Performance Bonuses and Reported Maximum Bonuses in Year 3, by Selected District and Teacher Characteristics

	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)	**57**	**30**	**424**
Subgroup Analyses By District Characteristics			
District's Average Pay-for-Performance Bonus from Prior Year			
(1) High—at least 5 percent of average salary	72	30	**145**
(2) Low—less than 5 percent of average salary	50	29	**279**
Difference, (1) – (2)	22*	1	
District Communication of Prior Year Actual Bonuses			
(1) Told all treatment teachers the total bonus amount that they earned (including $0 for nonrecipients)	54	22	**250**
(2) Did not tell all treatment teachers the total bonus amount that they earned	73	41	**174**
Difference, (1) – (2)	-19*	-19*	
Subgroup Analyses By Teacher Characteristics			
Report About Receiving a Pay-for-Performance Bonus Based on Prior Year's Performance			
(1) Reported receiving a pay-for-performance bonus	91	48	**140**
(2) Reported not receiving a pay-for-performance bonus	43	23	**283**
Difference, (1) – (2)	48*	25*	
Actual Receipt of a Pay-for-Performance Bonus Based on Prior Year's Performance			
(1) Received a pay-for-performance bonus	66	37	**229**
(2) Did not receive a pay-for-performance bonus	46	22	**192**
Difference, (1) – (2)	20*	14*	
Mentoring Role			
(1) Teacher mentored other teachers as part of TIF	71	42	**56**
(2) Teacher did not mentor other teachers as part of TIF	56	28	**367**
Difference, (1) – (2)	15*	14	

Source: Teacher and district surveys (2014), district interviews (2014), and educator administrative data.

Notes: For teachers who reported being eligible for a bonus but left the amount missing, bonus amounts were imputed through multiple imputation methods. See Appendix B for additional discussion on the imputation methods. In the row for "All Teachers," results differ from those presented in Figure IV.11 because Figure IV.12 pertains only to teachers in tested grades and subjects.

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

Most teachers were aware of the opportunity to take on additional roles and responsibilities. In Year 3, more than 80 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 between Years 1 and 2 for both treatment and control teachers, there was no improvement between Years 2 and 3. 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). As a result, teachers' awareness of additional pay opportunities in Year 3 was higher in treatment schools than in control schools (Table IV.14).

Table IV.14. Eligibility for Additional Pay Opportunities, as Reported by Teachers and Principals (Percentages)

	Year 1		Year 2		Year 3	
	Treatment	Control	Treatment	Control	Treatment	Control
Teachers						
Teachers Could Receive Additional Pay for Taking on Extra Roles or Responsibilities	57	56	89 +	88+	88*	82+
Roles or Responsibilities						
Mentor teacher	44	40	72 +	74+	60*+	55+
Master or lead teacher	40	39	54 +	57+	61*+	54
Department chair or head	18	20	22*	29+	24	28
Lead curriculum specialist	26	25	35 +	38+	35	34
Schoolwide committee or task force member	11	11	18 +	21+	18	22
Leadership team member	35	29	23 +	27	19	18+
Additional Factors						
Teach in a hard-to-staff or high-need school	25	23	30 +	31+	30	29
Attend professional development activities or enroll in graduate level courses	30	28	25	24	30	27
Number of Teachers—Range[a]	**246–385**	**234–393**	**436–440**	**438–444**	**402–421**	**425–447**
Principals						
Principals Could Receive Additional Pay for Taking on Extra Roles or Responsibilities	—	14	20 +	16	24	21
Number of Principals	**64**	**63**	**64**	**61**	**59**	**61**

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

Note: The finding for treatment principals in Year 1 was suppressed due to the small number of principals in the category.

[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.

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 Year 3 (Table IV.14). Although only one of the districts reported offering extra pay for principals to accept additional responsibilities in Year 3, 20 percent of the districts reported offering principals extra pay for working in a hard-to-staff school, attending professional development, or enrolling in graduate courses (not shown). Some principals may have interpreted their eligibility for earning extra pay for these other factors as extra pay for additional roles or responsibilities.

More than half of teachers reported they received the professional development required under the TIF grant but indicated they received only a few hours of it. In Year 3, approximately two-thirds of teachers reported that they received or expected to receive professional development focused on understanding performance measures used in TIF, and about 58 percent reported receiving or expecting to receive feedback based on their performance ratings (Appendix D, Table D.20). Of those who expected to receive any professional development on these two topics, the expected amount of time on each topic was three hours (Appendix D, Table D.21).

Summary

According to the theory of change presented in Chapter I, some key steps needed to occur in the implementation of TIF for pay-for-performance to be able 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 the implementation of the TIF grant in evaluation districts is useful context for interpreting findings presented later in this report on the program's impacts on educator and student outcomes.

The findings from this chapter indicate that evaluation districts' third year of implementation of TIF was very similar to their second year, with several possible factors continuing to dampen the potential for pay-for-performance to improve educator effectiveness. For example, many teachers in treatment schools continued to believe they were ineligible for a performance bonus or underestimated how much they could earn from these bonuses. Principals' understanding of their eligibility worsened. As in previous years, 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, the actual structure of the bonuses may not have provided educators with an incentive to change their behavior.

If educators were motivated to change their practices, they still may have found it difficult to determine how to do so. Although most districts provided professional development to help teachers improve their classroom observation rating, few provided professional development to help teachers improve their ratings based on student achievement growth. Furthermore, the amount of professional development provided may have been inadequate to help teachers improve their practices. Among teachers who expected to receive professional development to understand how they were evaluated or how to improve their performance ratings, the expected amount of time on each topic was about three 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 may 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.

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 after two and three years of TIF implementation. Educators in treatment schools were eligible for pay-for-performance bonuses, and 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 schools and control schools can be attributed to the impacts of pay-for-performance.[48]

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

- **Most teachers and principals reported being satisfied with their professional opportunities, factors associated with how they were evaluated, and their school environment.**
- **In contrast to prior years, teachers in treatment schools in Year 3 were at least as satisfied as teachers in control schools with their professional opportunities, how they were evaluated, and their school environment.**
- **Principals in treatment schools were more satisfied than were principals in control schools with their opportunities to earn extra pay, and their job satisfaction improved on several dimensions compared to the prior year.**
- **Most teachers and principals had positive attitudes toward their TIF program, and by Year 3, teachers in treatment schools felt at least as positively toward TIF as teachers in control schools did.**

The chapter is based on 10 evaluation districts that completed three years of TIF implementation during the period covered by this report. Because the impacts of pay-for-performance on educators' attitudes and behaviors were generally similar between the first and second years of

[48] 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 if all educators correctly understood their eligibility or 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 individual educators' attitudes and behaviors or having enabled schools to attract or retain more educators with particular attitudes and behaviors.

implementation (Chiang et al. 2015), we only present findings for Year 2 (2012–2013) and Year 3 (2013–2014).[49] Although attitudes and behaviors in Year 3 are a key focus of this chapter, we also examined how these outcomes evolved between Years 2 and 3. Surveys in Years 2 and 3 were administered after educators had received one and two years of bonuses based on their prior year of performance. Therefore, these data provide an opportunity to examine whether educators' initial impressions of performance-based compensation changed after two years of bonuses were awarded and educators gained more experience with the program components.

Impact of Pay-for-Performance on Educators' Attitudes

In this section, we present the impacts of pay-for-performance on educators' satisfaction with and attitudes toward their jobs and the TIF program.

Satisfaction with Job and Factors Associated with Evaluation System

Most teachers and principals in treatment and control schools were satisfied with their professional opportunities, how they were evaluated, and their school environment. In Year 3, about 80 percent of teachers reported being somewhat or very satisfied with their opportunities to enhance their skills, the feedback on their performance, the quality of interaction with colleagues, and colleagues' efforts (Table V.1). In addition, about three-quarters reported being satisfied with their job overall. Teachers reported being least satisfied with opportunities to earn extra pay (61 percent of treatment teachers and 50 percent of control teachers) and school morale (62 percent of treatment teachers and 53 percent of control teachers). In Year 3, the percentage of principals satisfied with aspects of their professional opportunities, evaluation system, and school environment ranged from 54 to 96 percent (Table V.2).

In contrast to prior years, teachers in treatment schools in Year 3 were at least as satisfied as teachers in control schools with their professional opportunities, how they were evaluated, and their school environment. 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 recognition of their accomplishments and factors associated with how they were evaluated (Table V.1). Treatment teachers in Year 2 only reported being more satisfied than control teachers with their opportunity to earn extra pay. But in Year 3, treatment teachers reported being more satisfied than control teachers with school morale (62 versus 53 percent), the quality of their interaction with colleagues (83 versus 79 percent), and their opportunities to earn extra pay (61 versus 50 percent; Table V.1). They also responded similarly to control teachers on the other satisfaction questions.

[49] 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. By the time of this report, Cohort 2 districts had completed only two years of implementation. In Appendix E, Tables E.1 through E.4, we present impacts on educators' satisfaction and attitudes from Year 2 for Cohorts 1 and 2 together—that is, findings from 2012–2013 for Cohort 1 and from 2013–2014 for Cohort 2.

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

	Year 2			Year 3		
Satisfaction Dimension	Treatment	Control	Impact	Treatment	Control	Impact
Opportunities for Pay and Development						
Opportunities for professional advancement	72	74	-3	74	74	0
Opportunities to enhance skills	80	81	-1	78	81	-3
Opportunities to earn extra pay	62	54	9*	61	50	11*
Factors Associated with Evaluation System						
Use of student achievement scores to assess performance	60	69	-9*	66	66	1
Feedback on my performance	75	80	-5*	79	81	-2
School Environment						
Recognition of accomplishments	60	66	-6*	67	62	5
Quality of interaction with colleagues	82	82	0	83	79	4*
Colleagues' efforts	84	83	0	84	83	0
School morale	58	59	-1	62	53	9*
Job Satisfaction						
Overall job satisfaction	73	74	-1	76	75	0
Number of Teachers—Range[a]	**444–448**	**446–449**		**426–430**	**455–459**	

Source: Teacher survey (2013 and 2014).

Notes: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding. None of the differences between Years 2 and 3 within treatment status 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.

*Impact 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 by 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 stem from several hypotheses. Teachers in tested grades and subjects could feel more pressure from the TIF program than do teachers in nontested grades because they could be 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 feel they have 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 may lead to these teachers being less supportive of pay-for-performance or their TIF program. Separating teachers by their level of experience is of interest because teachers who had been teaching longer under a different evaluation and compensation system could have been less receptive to the new system.

The results of the subgroup analyses should be interpreted carefully. The impact estimate within each subgroup, which is based purely on the study's experimental 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, they can inform best practices for designing or targeting future pay-for-performance programs.

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

	Year 2			Year 3		
Satisfaction Dimension	Treatment	Control	Impact	Treatment	Control	Impact
Opportunities for Pay and Development						
Opportunities for professional advancement	86	89	-3	93	90	3
Opportunities to enhance skills	87	85	2	96	89	7
Opportunities to earn extra pay	63	64	-1	84+	54	29*
Factors Associated with Evaluation System						
Use of observations to assess skills	61	85	-24*	96+	85	11
Use of student achievement scores to assess performance	66	82	-16*	87	81	6
Feedback on my performance	67	80	-13	83	82	1
School Environment						
Recognition of accomplishments	64	75	-12	81+	69	11
Quality of interaction with colleagues	86	90	-4	95	89	7
Colleagues' efforts	90	85	5	94	94+	1
School morale	75	82	-7	90+	84	7
Number of Principals—Range[a]	**63–64**	**60–61**		**58–59**	**61–62**	

Source: Principal survey (2013 and 2014).

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.

In contrast to Year 2, veteran teachers in Year 3 did not respond consistently less favorably toward pay-for-performance than less experienced teachers. In Year 2, we found that veteran teachers—those with more than 15 years of experience—tended to respond least favorably to pay-for-performance on factors associated with their evaluation and school environment (Chiang et al. 2015). However, in Year 3, there is no consistent pattern of veteran teachers responding more or less favorably to pay-for-performance compared to less experienced teachers. For example, the impact on teachers' satisfaction with their opportunities to earn extra pay tended to be the least positive among veteran teachers. However, the impact of pay-for-performance on teachers' satisfaction with school morale was more positive for veteran teachers than for teachers with fewer than five years of experience. Likewise, for the dimensions on which pay-for-performance changed teachers' satisfaction, the impacts were similar for teachers in tested and nontested grades and subjects (Appendix E, Table E.5).

Principals in treatment schools were more satisfied than principals in control schools with their opportunities to earn extra pay, and their satisfaction improved on several dimensions compared to the prior year. Treatment principals were more likely than control principals to report being satisfied on all aspects of their professional opportunities, evaluation system, and school environment, but the differences were generally not statistically significant (Table V.2). The one significant difference was that a higher percentage of principals in treatment schools than in control schools were satisfied with their opportunities to earn extra pay (84 versus 54 percent). This overall pattern was a reversal from Year 2, when treatment principals had tended to be less satisfied than control principals. The difference between Years 2 and 3 was mainly driven by an increase in satisfaction among treatment principals. For example, more treatment principals in Year 3 compared to Year 2 reported being satisfied with their opportunities to earn extra pay (84 versus 63 percent), the use of observations to assess their skills (96 versus 61 percent), recognition of their accomplishments (81 versus 64 percent), and school morale (90 versus 75 percent).

Educators' Attitudes Toward TIF

Most teachers were glad to be participating in TIF, and by Year 3, teachers in treatment schools felt at least as positively toward TIF as teachers in control schools. In Year 3, as in prior years, most teachers were glad to be participating in TIF. Approximately two-thirds of teachers in Year 3 were glad they were participating in TIF, and nearly 60 percent felt TIF was fair (Table V.3). However, in contrast to Year 2, treatment teachers in Year 3 no longer felt less favorably than control teachers about the effect of TIF on teacher collaboration, their freedom to teach the way they like, and the use of student test scores to measure student learning. And for the first time, treatment teachers were more likely than control teachers to report that their job satisfaction increased due to the TIF program (39 versus 33 percent in Year 3). However, pay-for-performance continued to cause a higher percentage of treatment teachers than control teachers to feel increased pressure to perform (by 14 and 10 percentage points in Years 2 and 3, respectively).

Unlike in Year 2, veteran teachers in Year 3 had similar or more favorable attitudes toward TIF than less experienced teachers. As with satisfaction, we examined the impacts of pay-for-performance on teachers' attitudes toward TIF separately within subgroups defined by teaching assignment and level of experience. In Year 2, we found that pay-for-performance had a stronger, less favorable impact on veteran teachers—those with more than 15 years of experience (Chiang et al. 2015). However, in Year 3, the impact of pay-for-performance was generally similar among veteran and less experienced teachers. On two dimensions, veteran teachers even responded more favorably than less experienced teachers. The impact of pay-for-performance on teachers reporting that TIF increased their job satisfaction was most positive for veteran teachers, and the impact on teachers reporting they felt increased pressure to perform because of TIF was smallest for veteran teachers. 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 (Appendix E, Table E.7).

We found no clear evidence that attitudes toward TIF differed between principals in treatment and control schools. 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. Treatment and control principals reported similar attitudes toward their nearly all aspects of their TIF program (Table V.4). The one exception was that treatment principals in Year 3 were more likely than control principals to agree that they played an important role in implementing the TIF program at their school.

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

	Year 2			Year 3		
Statement	Treatment	Control	Impact	Treatment	Control	Impact
Teachers who do the same job should receive the same pay	61	66	-4	68	66	2
Standardized student test scores in my district measure what students have learned	34	41	-7*	30	34	-5
My principal is a good judge of teacher talent	74	74	0	79	77	2
I am glad that I am participating in the TIF program	66	71	-5	69	68	1
My job satisfaction has increased due to the TIF program	38	38	0	39	33	6*
I feel increased pressure to perform due to the TIF program	65	51	14*	63	53	10*
I have less freedom to teach the way I would like to teach due to the TIF program	40	30	10*	37	35	2
The TIF program has harmed the collaborative nature of teaching	29	21	8*	26	25	1
The TIF program has caused teachers to work more effectively	50	56	-6	56	51	5
The TIF program is fair	54	59	-5	57	60	-3
The process used to determine how bonuses are determined was adequately explained to me	66	62	4	70	63	7*
Number of Teachers—Range[a]	**397–440**	**383–442**		**386–425**	**383–447**	

Source: Teacher survey (2013 and 2014).

Notes: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding. None of the differences between Years 2 and 3 within treatment status 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.

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

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

	Year 2			Year 3		
Statement	Treatment	Control	Impact	Treatment	Control	Impact
The TIF program has been clearly communicated to me	93	97	-4	97	90	7
This school has less chance of earning a bonus because of the characteristics of our student population	38	24	14	29	31	-2
The evaluation system omits important aspects of school administration that should be considered	54	48	6	41	47	-6
The TIF program contributes to greater collegiality and professionalism among the staff at this school	56	68	-12	64	57	6
Teachers at this school are more comfortable with frequent formal observations of their teaching because of the TIF program	58	68	-10	72	67	5
Parents and the school community believe the TIF program is important	50	43	7	43	53	-11
The TIF program is likely to continue for the foreseeable future	71	73	-2	58+	51+	7
I played an important role in implementing the TIF program at my school	86	84	2	91	77	14*
Number of Principals—Range[a]	**59–63**	**58–60**		**58–59**	**58–61**	

Source: Principal survey (2013 and 2014).

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.

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, by the end of the third year of TIF implementation, teachers in treatment schools reported attitudes toward their job and their TIF program that were at least as favorable as those reported by teachers in control schools. On some dimensions, treatment teachers reported being even more satisfied than control teachers. One possible explanation for this shift in attitudes may be that treatment teachers' attitudes improved because they experienced multiple years of performance bonuses. As noted in Chapter IV, more than 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 bonus receipt might be just as important, if not more important, in predicting their satisfaction as actual bonus receipt. 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 believe they received) bonuses may 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, we found no significant difference in any measure of teacher satisfaction in Year 3 between teachers who had been awarded a bonus in Year 2 and those who had not (Appendix E, Table E.6). Similarly, we found no evidence that those who received a Year 2 performance bonus had more favorable attitudes toward TIF than those who did not (Table V.5). 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 most differences were not statistically significant (Table V.5 and Appendix E, Table E.6). Given the large number of relationships examined, even the findings in which there was a significant difference between the responses of teachers who reported receiving a bonus and those who did not could have occurred by chance.

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

In this section, we present the impacts of pay-for-performance on teacher recruitment. As shown in the theory of change in Chapter I, principals can influence the effectiveness of the teacher workforce through their recruitment efforts.[50]

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, so pay-for-performance had the potential to influence the principals' approaches to teacher recruitment. 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, being able to offer larger bonuses might help principals recruit more higher-performing teachers.

Principals in treatment schools were more likely to use components of TIF to recruit teachers than were principals in control schools. When recruiting teachers in Year 3, principals in treatment schools were more likely than principals in control schools to report emphasizing opportunities for earning performance-based pay (43 versus 10 percent), career advancement (37 versus 17 percent), and professional development (77 versus 56 percent; Table V.6). A higher

[50] 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. We found little evidence that principals made decisions on teacher assignments or nonmonetary benefits differently in response to pay-for-performance (Appendix E, Tables E.9 and E.10). We also examined whether pay-for-performance affected how teachers reported spending their time. The teacher survey asked teachers to estimate the hours they spent on school-related activities during the most recent full week of school. Unlike other measures of teachers' time use (for example, a daily time log), this measure only may have been capable of detecting large changes in how teachers used their time. Based on teachers' responses, we found no evidence that performance bonuses impacted teachers' time on school-related activities (Appendix E, Table E.11).

percentage of treatment than control principals also reported emphasizing the TIF program in particular as a recruitment incentive, although the difference was not statistically significant.

Table V.5. Treatment Teachers' Attitudes Toward TIF Program by Bonus Receipt and Report of Bonus Receipt, Year 3 (Percentages Who "Agree" or "Strongly Agree")

	Actual Year 2 Bonus Receipt			Report of Year 2 Bonus Receipt		
Statement	Received a Bonus	Did Not Receive a Bonus	Difference	Reported Receiving Bonus	Reported Not Receiving a Bonus	Difference
Teachers who do the same job should receive the same pay	70	64	7	71	65	6
Standardized student test scores in my district measure what students have learned	37	32	5	40	30	10
My principal is a good judge of teacher talent	75	76	-2	83	75	8
I am glad that I am participating in the TIF program	67	64	2	77	66	11
My job satisfaction has increased due to the TIF program	37	37	-1	48	36	12*
I feel increased pressure to perform due to the TIF program	70	61	9	66	60	7
I have less freedom to teach the way I would like to teach due to the TIF program	32	41	-10	29	40	-11
The TIF program has harmed the collaborative nature of teaching	31	26	5	22	27	-5
The TIF program has caused teachers to work more effectively	57	55	3	68	54	14
The TIF program is fair	59	56	3	61	56	5
The process used to determine how bonuses are determined was adequately explained to me	75	60	15	78	65	14*
Number of Teachers—Range[a]	**217–228**	**166–194**		**131–137**	**251–283**	

Source: Teacher survey (2014) and educator administrative data.

Notes: Pay-for-performance bonus receipt information comes from Year 2 educator administrative data. The difference between those that received (or reported receiving) a bonus and those that 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. Incentives Used to Recruit Teachers (Percentages Who Reported They Were "Always" or "Often" Used)

	Year 2			Year 3		
Incentives	Treatment	Control	Impact	Treatment	Control	Impact
TIF-Related Incentives						
Opportunities to earn performance-based pay	33	17	16	43	10	33*
Opportunities for career advancement	27	28	-1	37	17	20*
Opportunities for professional development	66	57	9	77	56	21*
The TIF program	45	40	5	56	39	17
Other Job-Related Incentives						
Salary	21	22	0	29	18	11
The level of teacher involvement in school decision making	53	52	0	56	56	0
Collegiality of teaching staff	79	88	-9	85	74+	11
The school culture and/or educational philosophy	81	92	-11	87	79+	8
The school's reputation	64	77	-12	73	66	7
The school's location or neighborhood	29	28	1	35	46	-11
The level of student achievement at the school	45	44	1	50	37	12
Number of Principals—Range[a]	**61–64**	**60–61**		**56–59**	**59–61**	

Source: Principal survey (2013 and 2014).

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.

Pay-for-performance had no impact on principals' success in filling teacher vacancies. Similar to prior years, principals of treatment and control schools reported having similar recruitment experiences in terms of interviews per vacancy and acceptances per offer made. Based on the principals' reports, there were no statistically significant differences between treatment and control schools in the number of candidates interviewed per vacancy or the number of acceptances per job offer made (Table V.7). Although 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 6.

Table V.7. Teaching Vacancies and Hiring Experiences (Averages Unless Otherwise Noted)

	Year 2			Year 3		
	Treatment	Control	Impact	Treatment	Control	Impact
Classroom with teacher vacancies	4	5	-1	5	5	0
Applications school reviewed for positions	33	33	-1	49+	40+	9
Applicants school interviewed	11	17	-6*	17+	17	0
Offers school made	4	5	-1	6	5	0
Offers that were accepted	4	4	-1	5	5	0
Interview ratio (number of applicants interviewed per classroom vacancy)	3	4	-1	4	4	-1
Acceptance rate (percentage of offers accepted out of offers made)	81	84	-3	85	82	3
Number of Principals—Range[a]	**61–64**	**58–61**		**56–59**	**55–59**	

Source: Principal survey (2013 and 2014).

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.

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 by the third 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, and sometimes more satisfied, with aspects of their job and their TIF program as those in control schools. 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 after three years of TIF implementation.

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

- **Pay-for-performance had a positive impact on teachers' and principals' performance ratings based on student achievement growth in the first year of implementation, but by the third year educators in treatment and control schools received similar ratings.**
- **Pay-for-performance led to slightly, but not statistically significantly, higher classroom observation ratings for teachers in each year.**
- **After three years, pay-for-performance had small, positive impacts on students' math and reading achievement that were equivalent to about four weeks of additional learning.**
- **The impacts of pay-for-performance on student achievement differed across districts, but differences in impacts were not related to differences in key program characteristics measured by this study.**

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. Pay-for performance could lead to greater educator effectiveness by enabling schools to attract and retain more effective educators or motivating educators to improve their effectiveness. 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.

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 and 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.[51] Data for this chapter come from districts' administrative records on educators and students.

[51] 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

The chapter is based on 10 evaluation districts that completed three years of TIF implementation during the period covered by this report. We refer to the first, second, and third years of implementation—2011–2012, 2012–2013, and 2013–2014—as Years 1, 2, and 3.[52] Examining impacts over three years provided an opportunity to see whether impacts evolved over time. For example, impacts could have been larger in Year 3 than earlier years for several reasons. Educators' understanding of their evaluation measures increased over time (see Chapter IV), and educators could also have been more motivated to improve in Year 3 after seeing the first and second rounds of performance bonuses that were awarded after Years 1 and 2 were completed. In addition, in contrast to earlier years, by Year 3 teachers in treatment schools were no longer less satisfied than teachers in control schools (see 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 those measures 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.

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

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 if all educators correctly understood their eligibility or the amount they could earn in a bonus.

[52] 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. Cohort 2 districts completed only two years of implementation, 2012–2013 and 2013–2014, referred to as Years 1 and 2 for this cohort. In Appendix F, we present two years of impacts on educator effectiveness and student achievement for Cohorts 1 and 2 together—that is, Year 1 findings from 2011–2012 for Cohort 1 and 2012–2013 for Cohort 2 and Year 2 findings from 2012–2013 for Cohort 1 and 2013–2014 for Cohort 2.

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 findings below capture 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 Classrooms and Schools

The two most common student achievement growth measures that districts used to evaluate educators were those that measured achievement growth of all students in a school and in teachers' specific classrooms (see Chapter IV). School achievement growth 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 7 of the 10 districts in Year 3, some teachers were also evaluated on student achievement growth in their own classrooms. In those districts, teachers who received classroom achievement growth ratings were typically those who taught grades and subjects that were tested using annual state assessments.

Pay-for-performance had a positive impact on teachers' and principals' performance ratings based on student achievement growth in Year 1, but by Year 3 educators in treatment and control schools received similar ratings. In Year 1, educators in treatment schools had school achievement growth ratings that were 0.34 points higher on a 1-to-4 rating scale than those of educators in control schools (Table VI.1).[53,54] Likewise, 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 Year 1.

[53] 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 3.

[54] The impacts of pay-for-performance on educators' performance ratings based on student achievement growth were somewhat sensitive to the inclusion of Cohort 2. In Year 1, when Cohorts 1 and 2 were included in the analyses, the impacts of pay-for-performance on school and classroom achievement growth ratings were no longer statistically significant. In Year 2, the impacts were not significant with or without including Cohort 2 (Appendix F, Table F.5).

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						
Ratings in Year 1	2.60	2.25	0.34*	0.04	**NA**	**124**[a]
Ratings in Year 2	2.55	2.27	0.27	0.07	**NA**	**131**
Ratings in Year 3	2.41	2.37	0.04	0.74	**NA**	**132**
Classroom Achievement Growth[b]						
Ratings in Year 1	2.26	2.08	0.18*	0.03	**1,092**	**73**
Ratings in Year 2	2.22	2.17	0.05	0.38	**1,339**	**73**
Ratings in Year 3	2.54	2.53	0.01	0.81	**2,049**	**91**

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 they did not place educators into performance categories or onto a numeric scale.

[b]Classroom achievement growth ratings were available only for districts that evaluated teachers based on classroom achievement growth. In Year 1 and Year 2, six districts evaluated teachers based on classroom achievement growth. In Year 3, seven districts evaluated teachers based on classroom achievement growth.

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

NA is not applicable.

However, the impacts of pay-for-performance on both school and classroom achievement growth ratings diminished over the three years of TIF implementation. In Year 2, educators in treatment schools continued to earn higher school achievement growth ratings than those in control schools, but the difference was no longer statistically significant (*p*-value = 0.07). By Year 3, educators in treatment and control schools earned similar school achievement growth ratings. Impacts on classroom achievement growth ratings dissipated even earlier; by Year 2, teachers in treatment and control schools earned similar classroom achievement growth ratings.[55]

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 to slightly, but not statistically significantly, higher classroom observation ratings for teachers in each year. Although differences between the classroom observation ratings of teachers in treatment schools and those in control schools were not statistically significant, they were positive and similar in all three years and almost significant by Year 3 (*p*-value =

[55] In Year 3, seven districts evaluated teachers based on classroom achievement growth, compared to six districts in Years 1 and 2. Findings are similar in Year 3 when only the six districts that evaluated teachers based on classroom achievement growth in Years 1 and 2 were included in the analysis (Appendix F, Table F.7).

0.24 in Year 1, *p*-value = 0.09 in Year 2, *p*-value = 0.07 in Year 3; Table VI.2).[56] One possible explanation for these differences is that teachers could have slightly changed their practices in response to the opportunity to earn performance bonuses. Another possible explanation is that the observers in treatment schools could have been more lenient in their ratings than those in control schools; as discussed in Chapter IV, teachers were typically observed by principals at their schools.

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	0.24	**3,622**	**132**
Ratings in Year 2	2.98	2.93	0.04	0.09	**3,612**	**132**
Ratings in Year 3	2.96	2.91	0.04	0.07	**3,642**	**132**
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	0.49	**121**	**119**

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 differences were statistically significant at the .05 level, two-tailed test.

[a]Analyses of observation ratings for principals included fewer than 132 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.

Pay-for-performance had no impact on the observation ratings principals earned. In all three years, there were no statistically significant differences between observation ratings for principals in treatment and control schools (Table VI.2).

Educator Performance Ratings for Returning and Newly Hired Teachers

Although we found no clear evidence that pay-for-performance, on average, raised teachers' performance ratings in Years 2 and 3, those average findings could have masked larger impacts on specific groups of teachers. For example, pay-for-performance led to only a slight—and not statistically significant—increase in classroom observation ratings on average, but this slight average increase could have represented a mix of stronger impacts on some teachers and no impacts on others. Here, we focus on two specific groups of teachers—returning and newly hired teachers—whose effectiveness could have responded differently to pay-for-performance, and we assess whether pay-for-performance had stronger impacts on either of these groups.[57]

Examining impacts on returning and newly hired teachers can suggest the ways in which pay-for-performance leads to greater teacher effectiveness. According to the theory of change from Chapter

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

[57] Findings for returning and newly hired principals were imprecise because of small numbers of principals but can be found in Appendix F (Tables F.12 and F.13). We found no significant impacts of pay-for-performance on performance ratings for these subgroups of principals.

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 allowed schools to keep more effective teachers, then returning teachers ought to have higher performance ratings in treatment schools than control schools. In fact, between Years 1 and 4, a slightly higher percentage of teachers stayed in treatment schools than control schools (51 percent 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.[58]

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.[59] As discussed in Chapter V, throughout the first three years, principals at treatment schools were more likely to report using pay-for-performance as a recruitment tool for hiring teachers than control principals. Therefore, 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, in Chapter V, we found that returning teachers in treatment schools were 18 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

[58] As explained in Chapter II, the study design required that half of the participating schools within a district would implement pay-for-performance bonuses and the other half would not. This design could have led to larger mobility impacts than if pay-for-performance had been implemented districtwide. 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 (Appendix F, Table F.10).

[59] Newly hired teachers at treatment schools may be more effective than new hires at control schools if treatment schools more successfully recruited teachers who had previously worked at treatment schools and these teachers, already subject to performance bonuses, were more effective compared to other newly hired teachers. Using data collected for an earlier report in this evaluation (Chiang et al. 2015), we found no evidence for this possibility. In both Years 2 and 3, only about 5 percent of newly hired teachers in the study schools had worked in a treatment school in the previous year, and this percentage was similar for new hires in treatment and control schools.

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 3, returning teachers were those who had stayed in their school between Years 2 and 3, and newly hired teachers were those who were new to their school in Year 3. Findings that defined returning teachers as those who stayed in their school since Year 1 were similar (Appendix F, Table F.11). Since returning and new teachers at the same school earned the same school achievement growth ratings, these analyses focused on classroom observation ratings and classroom achievement growth that captured individual teachers' performance.

In Year 3, pay-for-performance had a small, positive impact on the classroom observation ratings of returning teachers but no impact on their classroom achievement growth ratings. In Year 3, classroom observation ratings were higher among returning teachers in treatment schools than in control schools by 0.06 points—a difference about 6 percent of the way between two performance levels on the four-level rating scale (Table VI.3). In Year 2, the impact on observation ratings was similar in magnitude (0.05 points), but not statistically significant (*p*-value = 0.11). Among returning teachers who were evaluated on classroom achievement growth, those in treatment and control schools earned similar classroom achievement growth ratings in both Years 2 and 3.

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	0.11	-0.03	0.52	**2,893**	**719**	**132**
Classroom achievement growth ratings	0.06	0.36	-0.04	0.73	**1,021**	**318**	**73**
Year 3							
Classroom observation ratings	0.06*	0.05	-0.01	0.82	**2,863**	**779**	**132**
Classroom achievement growth ratings	0.02	0.74	-0.01	0.92	**1,597**	**452**	**91**

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 3, returning teachers were those who had stayed in their school between Years 2 and 3, and newly hired teachers were those who were new to their school in Year 3.

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

Pay-for-performance had no impacts on the performance ratings of newly hired teachers. In Years 2 and 3, newly hired teachers in treatment and control schools had similar performance ratings based on classroom observations and classroom achievement growth (Table VI.3).

Impact of Pay-For-Performance on Student Achievement

Improving student achievement is the ultimate objective of the TIF grants. Although the grants were designed to accomplish this objective by enhancing educator effectiveness, the analyses in the previous section suggest that, by Year 3, pay-for-performance had few impacts on educators' performance ratings. 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 factored 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 made use of 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.

In each year, this analysis examined the cumulative impact of pay-for-performance on schools' average student achievement since the beginning of the study. After Year 3, average student achievement could be cumulatively higher in treatment schools than control schools if pay-for-performance raised educator effectiveness—and therefore students' growth—in any of the three years. For example, if students in treatment schools experienced higher growth in Year 1 due to more effective teaching, but teaching quality (and student growth) in subsequent years was similar in treatment and control schools, students in the treatment schools would, cumulatively, still have higher achievement.[60]

As discussed in Chapter II, we standardized test scores from different states and grades into z-scores, which reflected how well each student scored when compared with the average student in his or her state and grade. Below we report the impact of pay-for-performance on the average student achievement of schools, which we characterize simply as the impact on student achievement.

After three years, pay-for-performance led to slightly higher student achievement in both math and reading. At the end of Year 3, students in treatment schools scored, on average, 0.05 standard deviations higher on math assessments than did students in control schools (Table VI.4). Evidence for a positive impact in math was stronger at the end of Year 3 than at the end of the previous two years, when the impacts (0.02 standard deviations in Year 1 and 0.04 standard deviations in Year 2) were not statistically significant (p-value = 0.36 in Year 1 and p-value = 0.08 in Year 2). In reading, student achievement at the end of Year 3 was higher by 0.04 standard deviations in treatment

[60] In supplemental analyses, we calculated our own measure of annual school achievement growth using the administrative data on students' reading and math scores. The impacts on our measure of annual school achievement growth were similar to impacts on districts' measures of school achievement growth (see Appendix B for technical details and Appendix F for results).

schools than control schools. This impact was similar in size to impacts on reading achievement at the end of the previous two years (0.03 standard deviations in both Years 1 and 2). As the negative *z*-scores indicate, the average achievement of students in both treatment and control schools was below the statewide mean, reflecting the fact that study schools were low-performing schools.[61]

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	0.36	**40,847**	**132**
Reading	-0.37	-0.40	0.03*	0.05	**40,571**	**132**
Year 2						
Math	-0.39	-0.43	0.04	0.08	**40,708**	**132**
Reading	-0.36	-0.39	0.03*	0.04	**40,390**	**132**
Year 3						
Math	-0.37	-0.42	0.05*	0.02	**40,037**	**132**
Reading	-0.33	-0.37	0.04*	0.02	**39,807**	**132**

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.

There are several ways to interpret the magnitudes of the impacts on student achievement. First, the impacts can be expressed as a difference in percentiles of achievement. In Year 3, the average student in a control school earned a math score at approximately the 34th percentile of student achievement statewide (Figure VI.1).[62] The average student in a treatment school earned a math score at approximately the 36th percentile—a gain of 2 percentile points. Similarly, the impact on reading achievement after Year 3 lifted the average student in these schools from the 36th to the 37th percentile.

The impacts can also be compared with the average one-year gain in achievement for students nationwide. For example, 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

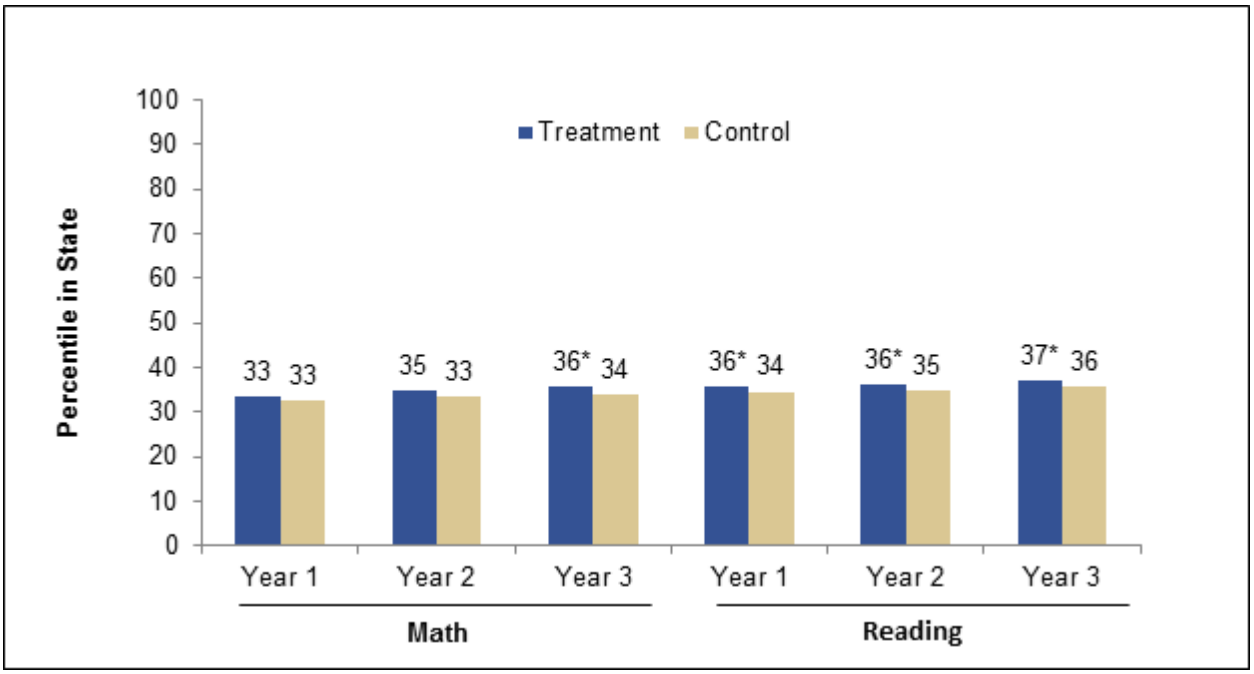
[61] The estimated impacts of pay-for-performance on student achievement were consistent across a variety of alternative analytic models (see Appendix F, Tables F.14 and F.15). The only exception is that a model that did not account for preexisting differences between treatment and control schools produced different findings. As discussed in Chapter II and Appendix B, our main analysis adjusted the impact findings to account for the fact that treatment schools had 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. As expected, when we did not account for these preexisting differences, the estimated impacts of pay-for-performance on math and reading achievement were smaller and not statistically significant. In addition, when both Cohorts 1 and 2 were included in the analysis, the estimated impacts in Years 1 and 2 were generally similar to the estimated impacts based on Cohort 1 only (Appendix F, Table F.16). The only notable difference was that the estimated impact on reading achievement in Year 1 was statistically significant based on Cohort 1 only, but not significant based on both cohorts.

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

achievement. Therefore, after Year 3, the increase of 0.05 standard deviations in math achievement resulting from pay-for-performance was equivalent to an additional 10 percent of a year of learning, or 4 weeks of additional 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, after Year 3, the increase of 0.04 standard deviations in reading achievement resulting from pay-for-performance was equivalent to an additional 11 percent of a year of learning, amounting again to about 4 weeks of additional learning.

Pay-for-performance could affect elementary and middle school grades differently, so we examined impacts separately by grade span. We found no statistically significant differences in impacts on elementary and middle school grades (Appendix F, Table F.17).

Figure VI.1. Average Student Achievement in Treatment and Control Schools After Years 1, 2, and 3 (Percentiles)



Source: Student administrative data (N = 40,847 students for Year 1 math; N = 40,708 students for Year 2 math; N = 40,037 for Year 3 math; N = 40,571 students for Year 1 reading; N = 40,390 students for Year 2 reading; N = 39,807 for Year 3 reading).

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

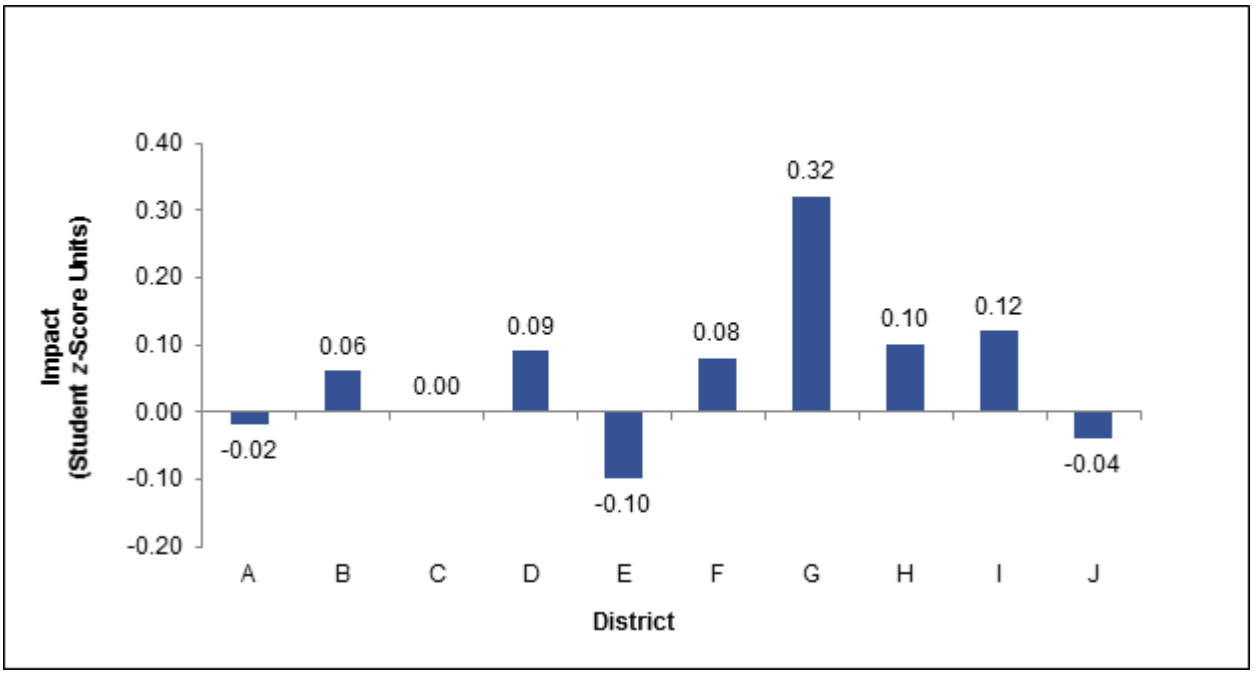
*Difference between treatment and control schools is statistically significant at the .05 level, two-tailed test.

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. 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. District-specific impacts on math achievement after Year 3 ranged from -0.10 to 0.32 standard deviations and, without considering their statistical significance, impacts were positive in 6 of the 10 districts, negative in 2 districts, and about zero (within 0.02 standard deviations) in the other 2 (Figure VI.2). Impacts on reading achievement after Year 3 also varied across districts, ranging from -0.06 to 0.18 standard deviations (Figure VI.3). Without considering their statistical significance, impacts in reading were positive in six districts, negative in one district, and about zero in the remaining three districts.[63]

Figure VI.2. Impact of Pay-for-Performance on Student Achievement in Math After Year 3, by District (Student z-Score Units)



Source: Student administrative data (N = 40,037).

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 0.02 student *z*-score units after Year 3.

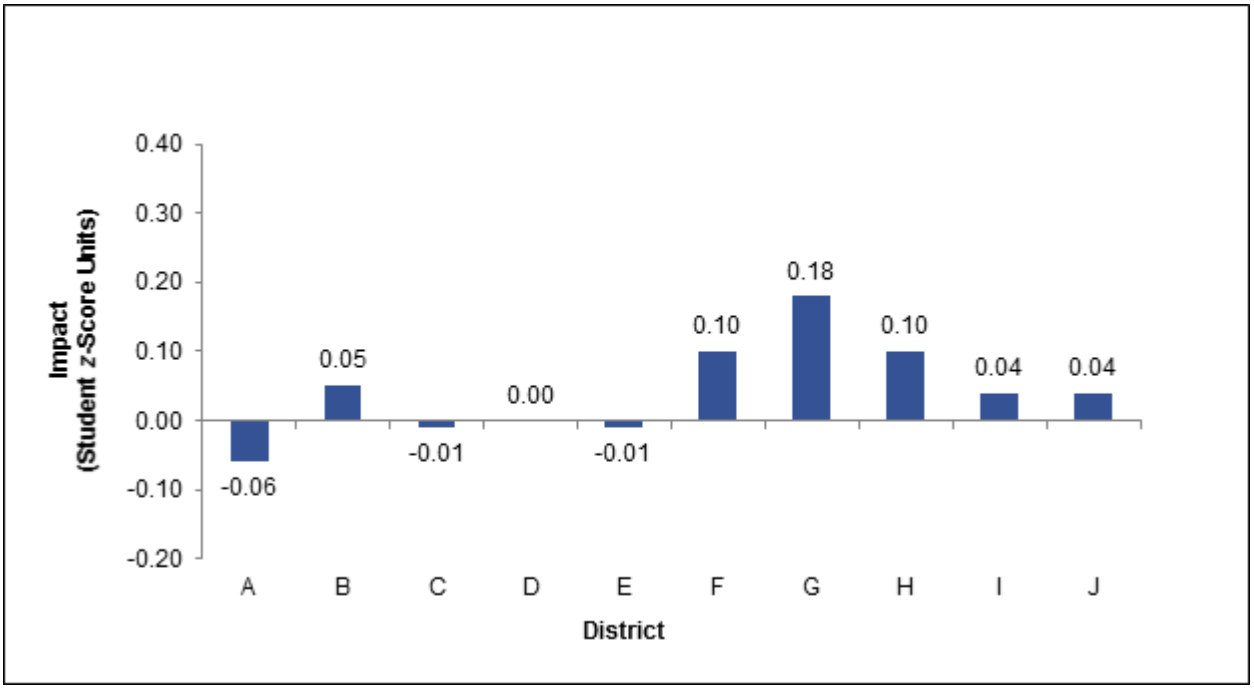
We sought to identify explanations for why impacts differed across districts. In particular, as discussed in Chapter IV, both the design and implementation of TIF programs also differed across districts. Therefore, we examined whether impacts were systematically larger or smaller in districts that designed or implemented their programs in particular ways.

TIF program and implementation characteristics measured by this study did not explain differences across districts in the impacts of pay-for-performance on student achievement.

[63] 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 in Years 1 and 2 also varied across all 13 districts in Cohorts 1 and 2.

The impacts of pay-for-performance on reading and math achievement were not related to a variety of program and implementation characteristics, 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 level of differentiation of bonuses, (4) the degree to which earning a bonus was challenging, (5) the timing of awarding bonuses based on the prior year, and (6) teachers' understanding of their pay-for-performance eligibility (see Appendix G for details).[64]

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



Source: Student administrative data (N = 39,807).

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 0.06 student *z*-score units after Year 3.

Differences in Student Achievement Impacts Across Schools

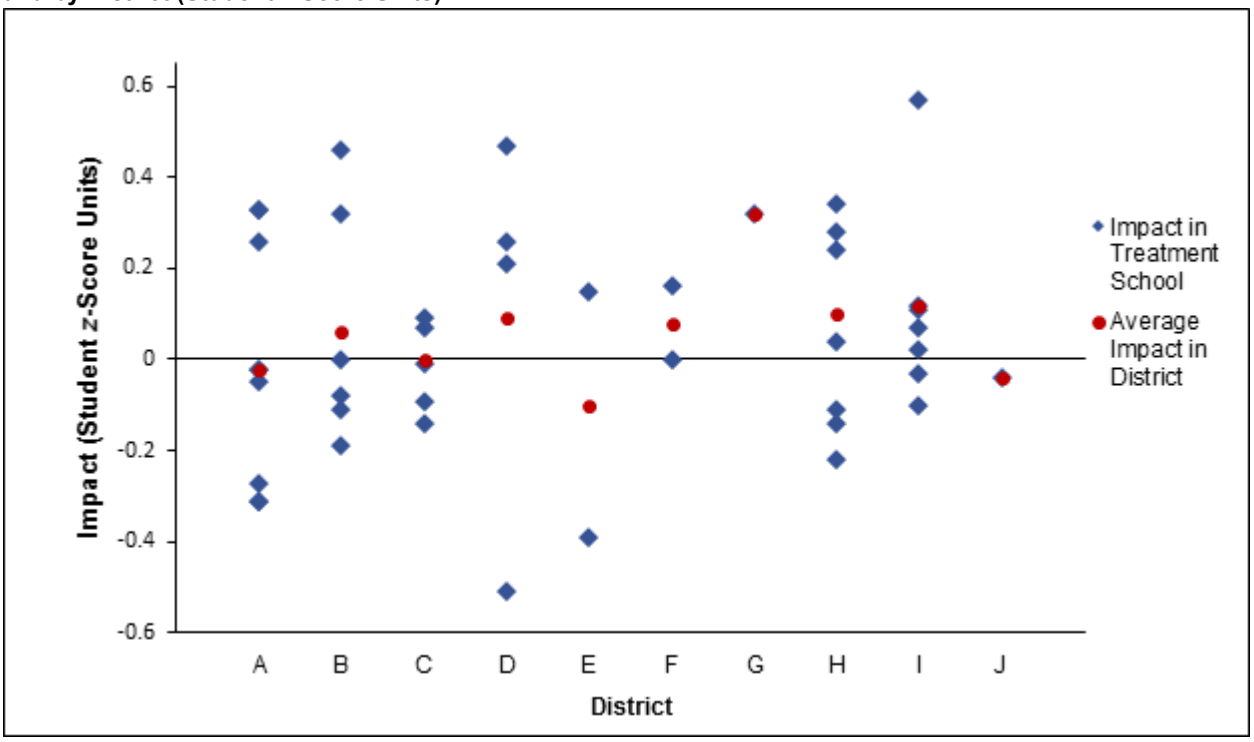
Within each of the 10 districts in the study, treatment schools may have also differed in the degree to which pay-for-performance affected student achievement. Although treatment schools within a district participated in programs with the same design and possibly the same implementation, pay-for-performance may have affected teacher and principal behaviors differently across schools, leading to differences in impacts on student achievement. For example, in schools with teachers who were more motivated by earning a bonus, pay-for-performance may have had stronger impacts on teaching

[64] We also examined the impacts of pay-for-performance on math and reading achievement when excluding District G. Although District G was similar to other districts in terms of their program characteristics and implementation experiences, we explored if the relatively large impacts in District G could be driving the average impacts reported in Table VI.4. Math and reading impacts in Year 3 were similar when we excluded this district. Pay-for-performance had a significant impact of 0.03 standard deviations in reading, and an impact of 0.04 standard deviations in math (*p*-value=0.07).

practices and, therefore, student achievement. 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. In both math and reading, 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 (see Figures VI.4 and VI.5, in which each blue diamond represents the achievement impact in an individual treatment school, and each red circle represents the average achievement impact in a district).[65] Statistically, most of the variation in impacts across schools occurred within the same district (85 percent in math and 93 percent in reading) rather than across districts.

Figure VI.4. Impact of Pay-for-Performance on Student Math Achievement After Year 3, by Treatment School and by District (Student *z*-Score Units)



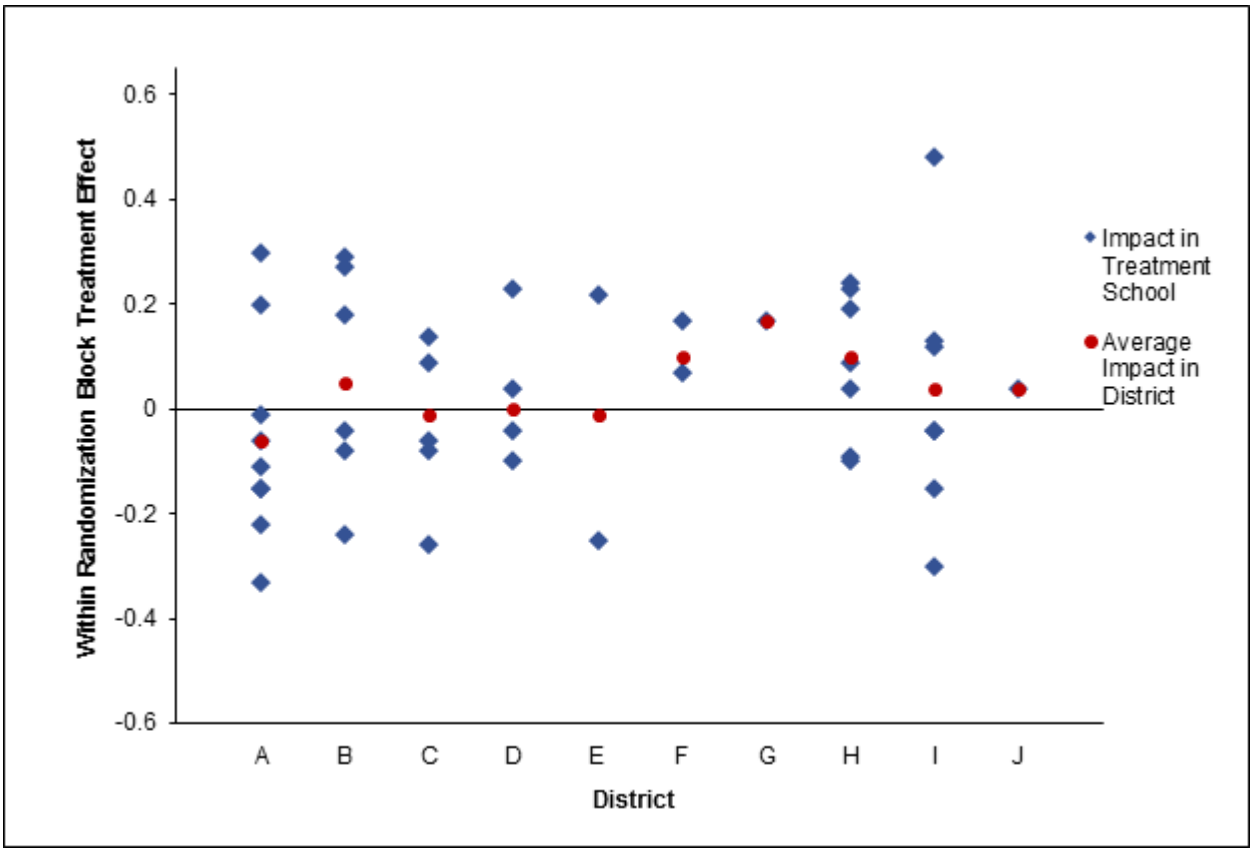
Source: Student administrative data (N = 40,037).

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 Year 3, within the six treatment schools in District A, pay-for-performance raised student math achievement by 0.33 standard deviations in one school and 0.26 standard deviations in one school. It lowered student math achievement by 0.02 standard deviations, 0.05 standard deviations, 0.27 standard deviations, and 0.31 standard deviations in the other schools. On average, pay-for-performance lowered student math achievement by 0.02 standard deviations in District A.

[65] 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 (see Chapter II).

Figure VI.5. Impact of Pay-for-Performance on Student Reading Achievement After Year 3, by Treatment School and by District (Student z-Score Units)



Source: Student administrative data (N = 39,807).

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 Year 3, within the eight treatment schools in District A, pay-for-performance raised student reading achievement by 0.33 standard deviations in one school and 0.22 standard deviations in one school. It lowered student reading achievement by 0.01 standard deviations, 0.06 standard deviations, 0.15 standard deviations, 0.22 standard deviations, and 0.33 standard deviations in the other schools. On average, pay-for-performance lowered student reading achievement by 0.06 standard deviations in District A.

Given that impacts on student achievement differed across schools, we sought to determine whether those differences were related to differences in impacts on teacher and principal behaviors. If schools with larger impacts of pay-for-performance on certain behaviors also had larger impacts on student achievement, this would provide suggestive (correlational) evidence that pay-for-performance might affect student achievement by way of influencing those behaviors.

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 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 a direct 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. Of the eighteen relationships we examined between impacts on educator behaviors and impacts on student achievement (nine measures of behaviors and two subjects), only one was statistically significant. Given the large number of relationships examined, the single significant finding could have occurred just by chance (see Appendix G for details).

Summary

A primary objective of TIF grants is to raise student achievement in high-need schools. The evidence in this chapter indicates that the pay-for-performance component of TIF made a small contribution toward achieving this objective. After three years of TIF implementation, pay-for-performance slightly improved student achievement in both math and reading. In each subject, the cumulative impact of pay-for-performance was equivalent to about four additional weeks of learning. Most of the difference in achievement between treatment and control schools emerged in the first two years, and this difference was sustained—but did not significantly grow—in the third year.

As depicted in the theory of change (Chapter I), the ability of pay-for-performance to improve student achievement depends on several factors. First, educators must understand their eligibility for a performance bonus. In Year 3, many educators continued to misreport their eligibility, and their understanding was no better than it was in the previous year (Chapter IV). This (mis)understanding may help explain why student achievement impacts did not grow between the second and third years.

Second, pay-for-performance needs to provide educators with the motivation to improve and enable schools to be an appealing place to work for effective educators. In contrast to previous years, by Year 3 teachers who were eligible for pay-for-performance were at least as satisfied with their jobs as those who were not eligible (Chapter V). However, this improvement in satisfaction was not accompanied by a larger impact on student achievement in Year 3. The improvement in satisfaction may not have been large enough to trigger changes in educator effectiveness, or it may take time for more favorable attitudes to translate into better classroom and school practices. Moreover, as discussed in Chapter IV, the bonuses continued to be small on average and generally not challenging to earn, which may have dampened the motivation for teachers to improve. Even more importantly, teachers still underestimated how much they could earn from the bonuses, so they may not have perceived a compelling monetary incentive to become a high performer.

Third, educators need to know how to change their practices in ways that improve student achievement. We found that pay-for-performance did have small (although insignificant), positive impacts on teachers' classroom observation ratings each year and had a small, positive impact on the classroom observation ratings of returning teachers. This suggests that teachers may have changed their practices slightly in response to pay-for-performance. However, the changes in practices captured by the observation measures were not related to higher student achievement. That is, schools with

larger impacts of pay-for-performance on observation ratings did not have larger impacts on student achievement. From this evidence, it is unclear whether teachers could really identify the changes to their practices that would most effectively improve their performance and raise student achievement.

Although the overall impact of pay-for-performance on student achievement was small, impacts were larger in some districts than in others. This raises the question of whether particular ways of designing or implementing their TIF programs could lead to larger impacts. For example, we examined whether differences in districts' average or maximum bonuses, or the timing of awarding those bonuses, were related to differences in student achievement impacts across districts. None of the characteristics we examined could help explain observed differences in student achievement impacts across districts.

In fact, impacts differed even more across treatment schools in the same district than across districts. This suggests that the size of the impacts may be shaped less by district-level program characteristics than by the ways in which the teachers in individual schools choose to change their behaviors. However, changes in teachers' self-reported behaviors—especially those that could reflect strategic actions to raise their ratings, larger effort on the job, or changes in practices—were unrelated to the student achievement impacts. Therefore, although pay-for-performance led to small improvements in student achievement, it continues to be unclear what factors caused this improvement.

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.

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.

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.

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.

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.

Liang, Kung-Yee, and Scott L. Zeger. "Longitudinal Data Analysis Using Generalized Linear Models." Biometrika, vol. 73, no. 1, 1986, pp. 13–22.

Marsh, J.A., M.G. Springer, D.F. McCaffrey, K. Yuan, S. Epstein, J. Koppich, and A. 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.

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.

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.

THIS PAGE IS INTENTIONALLY BLANK

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. Cohort 1 completed three years of implementation during the period covered by this report, 2011–2012, 2012–2013, and 2013–2014, referred to as Years 1, 2, and 3. Cohort 2 districts completed two years of implementation, 2012–2013 and 2013–2014, referred to as Years 1 and 2 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 three full years (Year 1 is 2011–2012, Year 2 is 2012–2013, and Year 3 is 2013–2014). Of the 138 Cohort 1 schools that were randomly assigned, 6 schools were dropped from all analyses to keep a constant analysis sample of 132 schools each year. After the first year of TIF implementation, four schools either closed, chose to drop out of the study, or were consolidated. These schools, along with their matched pair, are excluded from the main analysis. The results based on Cohorts 1 and 2 (shown in later appendices) include schools that have implemented TIF for at least two years. These supplemental analyses of Years 1 and 2 are based on a constant analysis sample of 173 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 5.1 percent in analyses for Cohort 1 and 5.5 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	132	66	66	NA
Number of schools in year 2 analyses	132	66	66	NA
Number of schools in year 3 analyses	132	66	66	NA
Attrition rate year 1	4.3	4.3	4.3	0
Attrition rate year 2	4.3	4.3	4.3	0
Attrition rate year 3	4.3	4.3	4.3	0
Analyses of Principal Survey Data				
Number of schools in year 1 analyses	131	66	65	NA
Number of schools in year 2 analyses	132	66	66	NA
Number of schools in year 3 analyses	132	66	66	NA
Attrition rate year 1	5.1	4.3	5.8	-1.5
Attrition rate year 2	4.3	4.3	4.3	0
Attrition rate year 3	4.3	4.3	4.3	0
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	173	87	86	NA
Number of schools in year 2 analyses	173	87	86	NA
Attrition rate year 1	5.5	5.4	5.5	-0.1
Attrition rate year 2	5.5	5.4	5.5	-0.1
Analyses of Principal Survey Data				
Number of schools in year 1 analyses	172	87	85	NA
Number of schools in year 2 analyses	173	87	86	NA
Attrition rate year 1	5.5	5.4	6.6	-1.2
Attrition rate year 2	5.5	5.4	5.5	-0.1

Notes: The primary analyses in the main body of the report are based on schools that implemented the program for three years (Cohort 1). Supplemental analyses are based study schools that implemented the program for at least two 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. The samples sizes in these tables are smaller than the full sample sizes because of missing data. For example, districts did not provide data on educator or student characteristics for some schools in our study, so school sample sizes in these tables are smaller than the full sample of Cohort 1 and 2 schools (183 schools).

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 132 Cohort 1 schools in the final analysis sample, 20 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	-0.55	-0.51	-0.04*
Reading	-0.49	-0.47	-0.02
Race/Ethnicity			
White, non-Hispanic	25	27	-3*
African American, non-Hispanic	47	46	1
Hispanic	22	20	2*
Other	6	7	0
Other Characteristics			
Female	48	49	-1
Eligible for free/reduced-price lunch	80	79	2
Disabled or has an Individualized Education Program	14	14	0
Overage for grade	13	13	0
English language learner	8	8	0
Grade Span			
Grades 3–5	62	62	0
Grades 6–8	38	38	0
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value			0.08
Number of Students—Range[a]	**19,220-30,023**	**18,725-29,153**	
Number of Schools—Range[a]	**60-87**	**59-86**	

Source: Student administrative data.

Notes: The table is based on the 173 Cohort 1 and Cohort 2 study schools. The pre-implementation year is 2010–2011 for Cohort 1 and 2011–2012 for Cohort 2. One school did not provide data for the pre-implementation year, so we excluded this school and its matched school from this analysis.

[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	84	83	1	63	59	4
Race/ethnicity						
White, non-Hispanic	76	75	1	60	56	4
Black, non-Hispanic	17	19	-1	32	37	-5
Hispanic	4	3	1	4	2	1
Other	3	3	0	4	5	-1
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
Less than 5 years	22	23	-1	20	14	6
5-15 years	45	46	-1	31	42	-10
More than 15 years	33	31	2	48	44	4
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value			0.25			0.44
Number of Educators—Range[a]	**2,222-2,956**	**2,180-2,851**		**49-85**	**55-88**	
Number of Schools—Range[a]	**69-87**	**68-86**		**47-83**	**53-84**	

Source: Educator administrative data.

Notes: Year 1 is 2011–2012 for Cohort 1 and 2011–2012 for Cohort 2. 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, and 2014 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/ethnicity						
White, non-Hispanic	76	73	3*	70	61	9
Black, non-Hispanic	17	20	-3*	25	32	-8
Hispanic	3	3	0	2	2	0
Other	4	4	0	4	5	-1
Age (average years)	43	43	0	48	48	0
Education						
Master's degree or higher	58	59	-2	99	90	8
Experience in K–12 Education						
Total experience (average years)	13	13	0	16	15	1
Less 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[a]	**729-1,812**	**770-1,790**		**25-54**	**28-56**	
Number of Schools—Range[a]	**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 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.

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, and 2014 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 and 2014 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. 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,and 113 in 2013, and 120 in 2014 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 and 2014 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 Year 2 and Year 3 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 88 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 2014 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, 2013–2014 School Year, Cohorts 1 and 2

	Overall	Non-Evaluation Districts	Evaluation Districts
All Districts			
Number of districts	158	145	13
Number of respondents	144	131	13
Response rate (respondents over total)	91	90	100

Source: District survey, 2014.

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, 2013–2014 School Year, Cohorts 1 and 2 (Percentages Unless Otherwise Noted)

	Respondents	Nonrespondents
Student Racial/Ethnic Distribution		
White, non-Hispanic	49	42
Black, non-Hispanic	27	26
Hispanic	19	19
Student Socioeconomic Status		
Eligible for free/reduced-price lunch	64	70
Title 1 eligible schools (schoolwide)	78	80
Enrollment (averages)		
Total enrollment	21,558	18,722
District Location[a]		
Urban	35	25
Suburban	19	33
Town	22	8
Rural	25	33
District Census Bureau Region		
Northeast	9	7
Midwest	28	29
South	45	57
West	18	7
Number of Districts	**138-144**	**12-14**

Source: District survey (2014) and Common Core of Data for 2012–2013 school year.

Notes: Seven TIF non-evaluation districts are not included in the 2012–2013 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.

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)		
	Total	Treatment	Control	Total	Treatment	Control	Total	Treatment	Control
Teachers									
Number of Sampled Teachers[a]	961	478	483	950	471	479	1,016	506	510
Number of respondents	880	433	447	872	433	439	917	441	476
Response rate (percentage)	92	91	93	92	92	92	90	87	93*
Number of Teachers in the Final Analysis Sample[b]	**795**	**393**	**402**	**904**	**451**	**453**	**892**	**431**	**461**
Principals									
Number of Sampled Principals	131	66	65	132	66	66	132	66	66
Number of respondents	129	65	64	126	64	62	122	59	63
Response rate (percentage)	98	98	98	95	97	94	92	89	95
Number of Principals in the Final Analysis Sample[c]	**129**	**65**	**64**	**125**	**64**	**61**	**121**	**59**	**62**

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

[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)		
	Total	Treatment	Control	Total	Treatment	Control
Teachers						
Number of Sampled Teachers[a]	254	135	119	305	154	151
Number of respondents	232	120	112	256	136	120
Response rate (percentage)	91	89	94	84	88	79
Number of Teachers in the Final Analysis Sample[b]	**231**	**126**	**105**	**263**	**140**	**123**
Principals						
Number of Sampled Principals	41	21	20	41	21	20
Number of respondents	35	18	17	37	20	17
Response rate (percentage)	85	86	85	90	95	85
Number of Principals in the Final Analysis Sample[c]	**35**	**18**	**17**	**36**	**19**	**17**

Source: Teacher and principal surveys (2013 and 2014).

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		
Grade Taught	Total	Treatment	Control	Total	Treatment	Control	Total	Treatment	Control
1st Grade Only	226	109	117	302	157	145	314	146	168
4th Grade Only	222	111	111	220	105	115	215	104	111
7th-Grade English/Language Arts and/or Math Only	203	100	103	199	98	101	172	83	89
7th-Grade Science Only	66	37	29	60	34	26	54	25	29
More than One Targeted Grade or Subject	78	36	42	123	57	66	137	73	64
Total	**795**	**393**	**402**	**904**	**451**	**453**	**892**	**431**	**461**

Source: Teacher survey (2012, 2013 and 2014).

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	
	Respondents	Nonrespondents	Respondents	Nonrespondents	Respondents	Nonrespondents
Demographic Characteristics						
Female	88	85	88	84	89	84
Race/Ethnicity						
White, non-Hispanic	73	67	68	68	67	69
Black, non-Hispanic	22	27	27	24	26	23
Hispanic	2	1	3	2	3	2
Other	3	6	3	6	4	6
Age (average years)	40	41	40	41	41	41
Education						
Master's degree or higher	44	32	47	47	45	44
Experience in K–12 Education						
Total experience (average years)	11	12	10	9	12	11
Less than 5 years	25	26	33	34	25	37*
5-15 years	45	38	43	49	41	34
More than 15 years	30	36	23	17*	34	29
Number of Teachers—Range[a]	**566-802**	**72-106**	**787-1,058**	**90-136**	**839-1,060**	**89-139**

Source: Teacher survey (2012, 2013, and 2014) 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	
	Treatment	Control	Treatment	Control	Treatment	Control
Demographic Characteristics						
Female	89	85*	90	86*	90	88
Race/Ethnicity						
White, non-Hispanic	75	68*	73	69	73	76
Black, non-Hispanic	18	23	20	25*	22	17
Hispanic	4	4	4	3	2	3
Other	3	5*	3	3	3	4
Age (average years)	40	39	41	41	40	41
Education						
Master's degree or higher	42	53*	48	55	49	48
Experience in K–12 Education						
Total Experience (average years)	11	10	11	10	11	10
Less than 5 years	27	27	26	30	29	30
5-15 years	45	50	47	45	44	47
More than 15 years	28	23	27	25	27	23
Number of Teachers—Range[a]	**240-355**	**277-372**	**318-431**	**319-426**	**334-415**	**344-428**

Source: Teacher survey (2012, 2013, and 2014) 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	
	Treatment	Control	Treatment	Control	Treatment	Control
Demographic Characteristics						
Female	58	67	61	67	66	62
Race/Ethnicity						
White, non-Hispanic	66	60	61	53	61	55
Black, non-Hispanic	27	33	34	36	34	33
Hispanic	2	0	2	4	0	5
Other	5	7	4	7	6	7
Age (average years)	49	48	48	49	47	49
Education						
Master's degree or higher	95	92	100	95	95	95
Experience in K–12 Education						
Total experience (average years)	16	15	17	14	18	16
Less than 5 years	19	13	17	17	6	14
5-15 years	30	38	26	43*	39	41
More than 15 years	50	48	57	40*	55	45
Number of Principals—Range[a]	**37-60**	**39-60**	**47-62**	**37-55**	**42-58**	**41-58**

Source: Principal survey (2012, 2013, and 2014) 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 132 Cohort 1 schools included 4,333 full-time teachers in Year 1, 4,433 full-time teachers in Year 2, and 4,545 full-time teachers in Year 3. 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	67	70
Full-time principals at the beginning of the year (*eligible to be included in analysis*)	65	69
Year 2 (2012–2013)		
All principals at the beginning of the year	69	71
Full-time principals at the beginning of the year (*eligible to be included in analysis*)	68	70
Year 3 (2013–2014)		
All principals at the beginning of the year	68	70
Full-time principals at the beginning of the year (*eligible to be included in analysis*)	65	69
Schools Included in the Analyses of Principal Outcomes		
Year 1 (2011–2012)		
All Cohort 1 schools	66	66
Schools with principals at the beginning of the year	65	66
Schools with full-time principals at the beginning of the year	63	65
Year 2 (2012–2013)		
All Cohort 1 schools	66	66
Schools with principals at the beginning of the year	66	65
Schools with full-time principals at the beginning of the year	65	64
Year 3 (2013–2014)		
All Cohort 1 schools	66	66
Schools with principals at the beginning of the year	65	66
Schools with full-time principals at the beginning of the year	63	65

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, and 3, all 132 Cohort 1 schools provided classroom observations ratings for at least some teachers. One district (with 20 schools) did not provide principal observation ratings for Year 1; all 10 Cohort 1 districts provided principal observation ratings for Years 2 and 3. 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	0.71	**4,333**	**132**
Had Classroom Achievement Growth Rating[a]	38	39	-1	0.38	**2,884**	**73**
Year 2						
Had Classroom Observation Rating	84	83	1	0.57	**4,433**	**132**
Had Classroom Achievement Growth Rating[a]	44	43	1	0.74	**2,954**	**73**
Year 3						
Had Classroom Observation Rating	84	83	1	0.51	**4,545**	**132**
Had Classroom Achievement Growth Rating[a]	59	58	1	0.58	**3,600**	**91**

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 Year 3, 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	94	85	9*	0.04	**138**	**129**
Year 3						
Had Observation Rating	97	87	9	0.06	**134**	**128**

Source: Educator administrative data.

Note: The number of principals exceeds the number of schools in the analysis sample because a few schools had more than one principal.

[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 and 3 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 Year 3.

Table A.16. Characteristics of Teachers with and Without Classroom Observation Ratings, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 1		Year 2		Year 3	
	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	84	86	85
Race/ethnicity						
White, non-Hispanic	66	65	66	66	64	65
Black, non-Hispanic	29	29	29	30	31	29
Hispanic	2	4	3	2	3	2
Other	2	2	2	2	2	3
Age (average years)	40	41	41	42	40	41
Education						
Master's degree or higher	41	43	42	44	40	44*
Total Experience in K–12 Education (average years)	10	11	10	11	10	10
Less than 5 years	30	32	32	32	34	36
5-15 years	47	41*	44	39*	44	42
More than 15 years	24	27	24	29*	23	22
Number of Teachers—Range[a]	**2,585-3,586**	**370-686**	**2,755-3,597**	**371-781**	**2,892-3,625**	**553-811**
Number of Schools—Range[a]	**98-132**	**65-99**	**100-132**	**73-106**	**106-132**	**84-103**

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.17. Characteristics of Teachers with and Without Classroom Achievement Growth Ratings, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 1		Year 2		Year 3	
	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	86	84	86	85	87	83*
Race/ethnicity						
White, non-Hispanic	63	66	65	66	65	66
Black, non-Hispanic	32	28*	28	28	29	27
Hispanic	3	4	5	4	5	5
Other	2	2	2	2	2	2
Age (average years)	39	40	38	40*	39	41*
Education						
Master's degree or higher	36	39	38	40	36	46*
Total Experience in K–12 Education (average years)	9	10	8	10*	8	10*
Less than 5 years	34	33	39	35	39	35*
5-15 years	47	42*	47	43	45	40*
More than 15 years	19	25	14	22*	16	25*
Number of Teachers—Range[a]	**631-1,073**	**1,337-1,751**	**934-1,324**	**1,210-1,576**	**1,487-2,046**	**1,150-1,458**
Number of Schools—Range[a]	**56-73**	**56-73**	**59-73**	**59-73**	**73-91**	**72-88**

Source: Educator administrative data.

Note: 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 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.18. Characteristics of Principals with and Without Observation Ratings, Cohort 1 (Percentages Unless Otherwise Noted)

	Year 2		Year 3	
	Principals with Observation Ratings	Principals Without Observation Ratings	Principals with Observation Ratings	Principals Without Observation Ratings
Demographic Characteristics				
Female	75	74	73	81
Race/ethnicity				
White, non-Hispanic	54	60	54	66
Black, non-Hispanic	42	40	41	34
Hispanic	0	0	0	0
Other	4	0	5	0
Age (average years)	47	52	45	55*
Education				
Master's degree or higher	93	100	95	89
Total Experience in K–12 Education (average years)	13	14	12	13
Less than 5 years	23	35	16	13
5-15 years	48	13*	60	68
More than 15 years	30	52	24	19
Number of Principals—Range[a]	**83-117**	**12-19**	**86-121**	**5-12**
Number of Schools—Range[a]	**82-116**	**12-16**	**85-119**	**5-11**

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		
	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Demographic Characteristics									
Female	87	85	2*	86	85	2*	86	85	1
Race/ethnicity									
White, non-Hispanic	74	73	1	74	72	2	73	72	1
Black, non-Hispanic	20	21	-1	19	22	-2	21	20	0
Hispanic	3	2	0	3	3	0	3	4	-1
Other	4	4	0	4	3	1	3	4	0
Age (average years)	42	41	0	42	41	0	42	42	0
Education									
Master's degree or higher	51	49	1	49	51	-2	47	48	-1
Total Experience in K–12 Education (average years)	12	11	0	11	11	0	11	11	0
Less than 5 years	23	25	-2	27	29	-2	27	30	-3
5-15 years	47	47	0	46	45	2	45	44	1
More than 15 years	30	28	2	27	27	1	28	26	2
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value			0.09			0.01			0.36
Number of Teachers—Range[a]	**1,268-1,799**	**1,317-1,787**		**1,334-1,786**	**1,421-1,811**		**1,445-1,810**	**1,447-1,815**	
Number of Schools—Range[a]	**49-66**	**49-66**		**50-66**	**50-66**		**53-66**	**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		
	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Demographic Characteristics									
Female	89	86	2	88	87	2	88	87	1
Race/ethnicity									
White, non-Hispanic	63	64	-1	64	62	2	65	64	1
Black, non-Hispanic	32	30	1	29	32	-3	30	30	1
Hispanic	4	4	0	5	4	1	3	4	-1
Other	1	2	-1	1	1	0	1	2	-1
Age (average years)	40	38	2*	40	38	1*	41	40	1
Education									
Master's degree or higher	36	38	-2	37	37	0	34	35	-1
Total Experience in K–12 Education (average years)	10	8	2*	9	8	2*	10	10	1
Less than 5 years	31	36	-5*	35	42	-6*	33	34	-1
5-15 years	45	48	-3	44	47	-3	41	44	-2
More than 15 years	24	16	8*	20	11	9*	26	22	4*
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value			0.00			0.00			0.03
Number of Teachers—Range[a]	**299-537**	**332-536**		**440-651**	**494-676**		**753-1,027**	**734-1,019**	
Number of Schools—Range[a]	**28-37**	**28-36**		**30-37**	**29-36**		**37-46**	**36-45**	

Source: Educator administrative data.

Note: 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			Year 2			Year 3		
	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Demographic Characteristics									
Female	59	62	-3	59	65	-5	60	64	-4
Race/ethnicity									
White, non-Hispanic	65	62	4	59	54	6	64	49	15
Black, non-Hispanic	24	29	-5	33	34	-1	30	39	-9
Hispanic or Other	10	10	1	8	13	-4	6	12	-6
Age (average years)	49	47	1	48	48	0	49	48	0
Education									
Master's degree or higher	93	91	1	96	95	1	96	95	1
Total Experience in K–12 Education (average years)	16	15	1	15	14	1	18	15	3
Less than 5 years	15	11	4	18	15	3	8	16	-8
5-15 years	38	41	-3	37	47	-10	38	41	-3
More than 15 years	47	48	-1	45	38	7	54	43	11
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value			0.72			0.42			0.03
Number of Principals—Range[a]	**35-59**	**34-52**		**44-61**	**39-56**		**46-62**	**40-59**	
Number of Schools—Range[a]	**35-59**	**34-52**		**44-61**	**38-55**		**45-61**	**40-58**	

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. 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.

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	**44,791**	**132**
Reading	92	92	0	**44,791**	**132**
Year 2					
Math	92	92	0	**44,906**	**132**
Reading	91	92	-1	**44,906**	**132**
Year 3					
Math	93	93	0	**43,342**	**132**
Reading	93	93	0	**43,342**	**132**

Source: Student administrative data.

Note: Differences are not statistically significant at the 0.05 level, two-tailed test.

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	
	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	-0.44	-0.79*	-0.44	-0.80*	-0.66	-0.96*
Reading	-0.39	-0.71*	-0.39	-0.69*	-0.59	-1.01*
Race/Ethnicity						
White, non-Hispanic	28	29	28	29	27	29
African American, non-Hispanic	42	45*	42	45*	42	42
Hispanic	24	19*	24	20*	25	22*
Other	6	7	6	7	6	7
Other Characteristics						
Female	50	43*	49	45*	50	44*
Eligible for free/reduced-price lunch	77	79	77	78	81	80
Disabled or has an Individualized Education Program	12	30*	13	34*	11	32*
Overage for grade	12	24*	12	22*	11	24*
English language learner	8	8	7	7	12	11
Grade Span						
Grades 3–5	64	66	64	65	67	66
Grades 6–8	36	34	36	35	33	34
Number of Students—Range[b]	**23,835-40,877**	**1,511-3,914**	**20,962-40,719**	**1,153-4,187**	**13,711-40,038**	**568-3,304**
Number of Schools—Range[b]	**84-132**	**80-129**	**84-132**	**72-123**	**106-132**	**60-124**

Source: Student administrative data.

[a]These averages are only calculated for students who were tested in the pre-implementation year, so they exclude 3rd graders in Year 1; 3rd and 4th graders in Year 2; and 3rd, 4th, and 5th graders in Year 3.

[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 0.05 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	
	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	-0.83*	-0.44	-0.82*	-0.64	-0.97*
Reading	-0.39	-0.72*	-0.39	-0.73*	-0.59	-0.95*
Race/Ethnicity						
White, non-Hispanic	28	28	28	28	27	28
African American, non-Hispanic	43	43	42	44*	42	41
Hispanic	24	21*	24	22	25	24
Other	6	8	6	7	6	7
Other Characteristics						
Female	50	43*	50	44*	50	43*
Eligible for free/reduced-price lunch	77	79	77	80	81	81
Disabled or has an Individualized Education Program	12	31*	13	33*	11	31*
Overage for grade	12	23*	12	22*	11	24*
English language learner	8	9	7	7	12	13
Grade Span						
Grades 3–5	64	67	64	65	66	67
Grades 6–8	36	33	36	35	34	33
Number of Students—Range[b]	**23,674-40,584**	**1,558-4,207**	**20,925-40,400**	**1,190-4,506**	**13,716-39,810**	**565-3,532**
Number of Schools—Range[b]	**84-132**	**81-130**	**84-132**	**80-131**	**105-132**	**68-127**

Source: Student administrative data.

[a]These averages are only calculated for students who were tested in the pre-implementation year, so they exclude 3rd graders in Year 1; 3rd and 4th graders in Year 2; and 3rd, 4th, and 5th graders in Year 3.

[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 0.05 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		
Characteristic	Treatment	Control	Treatment-Control	Treatment	Control	Treatment-Control	Treatment	Control	Treatment-Control
Achievement in the Pre-Implementation Year (average *z*-score)[a]									
Math	-0.46	-0.41	-0.05*	-0.46	-0.44	-0.03	-0.65	-0.72	0.07
Reading	-0.40	-0.38	-0.02	-0.40	-0.40	0.00	-0.58	-0.67	0.09
Race/Ethnicity									
White, non-Hispanic	27	29	-2*	27	29	-2	26	29	-2
African American, non-Hispanic	42	41	1	42	41	1	42	40	1
Hispanic	24	23	2	25	23	2	26	24	2
Other	6	6	0	6	7	0	6	7	-1
Other Characteristics									
Female	49	50	-1	49	49	0	49	50	-1
Eligible for free/reduced-price lunch	77	78	-1	77	76	1	81	82	0
Disabled or has an Individualized Education Program	12	12	1	13	13	0	11	11	0
Overage for grade	12	11	0	12	11	0	11	10	0
English language learner	8	9	0	7	8	0	12	12	0
Grade Span									
Grades 3–5	64	64	0	64	64	-1	67	66	0
Grades 6–8	36	36	0	36	36	1	33	34	0
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value			0.01*			0.12			0.10
Number of Students—Range[b]	**11,904-20,525**	**11,848-20,322**		**10,263-20,251**	**10,693-20,457**		**6,710-20,026**	**7,015-20,011**	
Number of Schools—Range[b]	**42-66**	**42-66**		**42-66**	**42-66**		**53-66**	**53-66**	

Source: Student administrative data.

[a]These averages are only calculated for students who were tested in the pre-implementation year, so they exclude 3rd graders in Year 1; 3rd and 4th graders in Year 2; and 3rd, 4th, and 5th graders in Year 3.

[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 0.05 level, two-tailed test.

Table A.26. Characteristics of Students in the Reading Analysis Sample, Cohort 1 (Percentages Unless Otherwise Noted)

Characteristic	Year 1			Year 2			Year 3		
	Treatment	Control	Treatment-Control	Treatment	Control	Treatment-Control	Treatment	Control	Treatment-Control
Achievement in the Pre-Implementation Year (average *z*-score)[a]									
Math	-0.46	-0.41	-0.05*	-0.46	-0.43	-0.03	-0.64	-0.71	0.07
Reading	-0.39	-0.38	-0.02	-0.40	-0.39	0.00	-0.57	-0.67	0.09
Race/Ethnicity									
White, non-Hispanic	27	29	-2*	27	29	-2	26	29	-2
African American, non-Hispanic	43	42	1	42	41	1	42	41	1
Hispanic	24	23	2	25	23	2	26	24	2
Other	6	6	0	6	7	-1	6	7	-1
Other Characteristics									
Female	50	50	0	49	50	0	49	50	-1
Eligible for free/reduced-price lunch	77	78	-1	77	76	1	81	82	0
Disabled or has an Individualized Education Program	12	12	0	13	13	0	11	11	0
Overage for grade	12	11	0	12	11	0	11	10	0
English language learner	8	9	0	7	8	0	12	12	0
Grade Span									
Grades 3–5	64	64	0	64	64	0	67	66	0
Grades 6–8	36	36	0	36	36	0	33	34	0
Test of Whether Characteristics Jointly Predict Treatment Status: *p*-value			0.02*			0.12			0.15
Number of Students—Range[b]	**11,803-20,343**	**11,803-20,228**		**10,223-20,031**	**10,696-20,359**		**6,710-19,880**	**7,021-19,927**	
Number of Schools—Range[b]	**42-66**	**42-66**		**42-66**	**42-66**		**52-66**	**53-66**	

Source: Student administrative data.

[a]These averages are only calculated for students who were tested in the pre-implementation year, so they exclude 3rd graders in Year 1; 3rd and 4th graders in Year 2; and 3rd, 4th, and 5th graders in Year 3.

[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 0.05 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 of the analytic methods used to estimate impacts of pay-for-performance on educator and student outcomes. Fourth, 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. Fifth, 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 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. At the time of this report, Cohort 1 had completed three years of implementation—2011–2012, 2012–2013, and 2013–2014—referred to as Years 1, 2 and 3. Cohort 2 districts had completed only two years of implementation, 2012–2013 and 2013–2014, referred to as Years 1 and 2 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, 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.

One district in Cohort 2 rated educators on a continuous scale for each performance measure and assigned a total score (also on a continuous scale) equal to the sum of the scores from each performance measure. These districts divided the range of the total performance scale (0 to 100) into four intervals, each corresponding to a different performance category. For analysis purposes, we translated educators' scores on each performance measure into the same four categories by dividing the continuous scale of each measure into four intervals, using the same proportional division as the district used for the total scale. We then standardized the categorical ratings by using the approach described earlier.

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 all scores into *z*-scores by subtracting the statewide grade-specific mean and dividing by the statewide grade-specific standard deviation.

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 one-half of 1 percent of all scores.

Table B.1. Test Scores That Were Dropped or Recoded, Cohort 1 (Percentages)

	Year 1			Year 2			Year 3		
Type of Exclusion or Recoding	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
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	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	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	0.1	0.1	0.0	0.2	0.2	0.0	0.1	0.2	-0.1*
Number of Students with Test Scores	**20,529**	**20,323**		**20,252**	**20,458**		**20,026**	**20,011**	
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	0.0
Dropped because score was more than 5 standard deviations above or below the statewide mean	0.1	0.1	0.0	0.1	0.1	0.0	0.1	0.0	0.1*
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	0.0	0.2	0.2	0.0	0.2	0.2	0.0
Number of Students with Test Scores	**20,354**	**20,238**		**20,045**	**20,374**		**19,894**	**19,932**	

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.

*Difference is statistically significant at the 0.05 level, two-tailed test.

Describing the Average Distribution of Performance Ratings and Payouts

In Chapter IV, we described the distribution—averaged across the 10 Cohort 1 districts—of performance ratings and payouts (including performance bonuses, automatic 1 percent bonuses, and additional pay) that educators received from their TIF programs. 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 of the 10 districts. 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 (see Appendix D, Figures D.2, D.7, D.12, and D.13). 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.

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. We then discuss how we estimated impacts within subgroups defined by educator or student characteristics (presented in Chapters V and VI) or districts' program characteristics (presented in Appendix G) and assessed the differences in impacts between subgroups. Finally, we discuss how we estimated the association between impacts on educator behaviors and impacts on student outcomes (presented in Appendix G). For simplicity, we refer primarily to impacts on educator and student outcomes, but we used the same analytic methods to estimates 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) $Y_{js} = \beta T_s + X'_{js}\delta + Z'_s\gamma + B'_s\pi + \varepsilon_{js}$,

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.

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 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 differences in the percentages of teachers or principals in treatment and control schools with missing covariate values, with one exception: 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		
Missing Data on:	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Achievement in the Pre-Implementation Year[a]									
Math	34	34	1	57	56	1	78	77	2*
Reading	35	34	1	58	57	1	79	77	2*
Race/Ethnicity									
Missing race characteristics	0	0	0	0	0	0*	0	0	0
Other Characteristics									
Female	0	0	0	0	0	0	0	0	0
Eligible for free/reduced-price lunch	37	37	0*	37	37	0	20	20	0
Disabled or has an Individualized Education Program	0	0	0*	16	16	0	16	16	0
Overage for grade	1	1	0	1	1	0	5	6	0
English language learner	0	0	0*	16	16	0	18	17	1
Number of Students	**20,525**	**20,322**		**20,251**	**20,457**		**20,026**	**20,011**	
Number of Schools	**66**	**66**		**66**	**66**		**66**	**66**	

Source: Student administrative data.

Notes: The difference between the treatment and control estimates may not equal the difference shown in the table because of rounding. Some differences less than 0.5 were statistically significant, and are reported as 0*.

[a]This characteristic is only defined for students who were tested in the pre-implementation year, so it is missing for 3rd graders in Year 1; 3rd and 4th graders in Year 2; and 3rd, 4th, and 5th graders in Year 3.

*Difference is statistically significant at the 0.05 level, two-tailed test.

Table B.3. Students in the Reading Analysis Sample with Missing Covariate Data (Percentages)

	Year 1			Year 2			Year 3		
Missing Data on:	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Achievement in the Pre-Implementation Year[a]									
Math	34	34	0	57	56	1	78	77	2*
Reading	34	34	1	58	56	1	79	77	2*
Race/Ethnicity									
Missing race characteristics	0	0	0*	0	0	0*	0	0	0
Other Characteristics									
Female	0	0	0	0	0	0	0	0	0
Eligible for free/reduced-price lunch	37	37	0*	37	37	0	20	20	0
Disabled or has an Individualized Education Program	0	0	0*	16	16	0	16	16	0
Overage for grade	1	1	0	1	1	0	5	6	0
English language learner	0	0	0*	16	16	0	18	17	1
Number of Students	**20,343**	**20,228**		**20,031**	**20,359**		**19,880**	**19,927**	
Number of Schools	**66**	**66**		**66**	**66**		**66**	**66**	

Source: Student administrative data.

Notes: The difference between the treatment and control estimates may not equal the difference shown in the table because of rounding. Some differences less than 0.5 were statistically significant, and are reported as 0*.

[a]This characteristic is only defined for students who were tested in the pre-implementation year, so it is missing for 3rd graders in Year 1; 3rd and 4th graders in Year 2; and 3rd, 4th, and 5th graders in Year 3.

*Difference is statistically significant at the 0.05 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	1
Race/Ethnicity	4	3	0
Age	4	3	1*
Education	33	32	1
Experience in K–12 Education	14	12	2
Number of Teachers	**2,181**	**2,152**	
Number of Schools	**66**	**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	25	18	7*
Number of Principals	**65**	**69**	
Number of Schools	**63**	**65**	

Source: Educator administrative data.

Notes: The difference between the treatment and control estimates may not equal the difference shown in the table because of rounding. We also examined the percentages of principals with missing data by sex, race/ethnicity and age. Missing data for these categories was rare, and there was no significant difference between treatment and control principals.

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

Estimation Model for Subgroup Analyses

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 Appendix G, we assessed how impacts on student achievement differed by districts' program characteristics.

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:

$$(3)\quad \begin{aligned} Y_{js} &= \beta_1 T_s + \gamma_2 Group2_j + \gamma_3 Group3_j + \beta_2 (T_s \times Group2_j) + \beta_3 (T_s \times Group3_j) \\ &+ X'_{js}\delta + Z'_s\gamma + B'_s\pi + \varepsilon_{js}. \end{aligned}$$

In equation (3), 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 (3) 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 differed across subgroups. For scenarios in which individuals were partitioned into two (rather than three) subgroups, equation (3) was identical except that it did not include indicators and interaction terms involving *Group3*$_j$.

When examining how impacts varied with districts' program characteristics, our main approach divided districts into two subgroups that differed on that characteristic, allowing us to follow the basic subgroup model shown in equation (3). However, some program characteristics could be expressed as a continuous variable (such as the average size or amount of differentiation in teachers' pay-for-performance bonuses). For those characteristics, we also estimated a variant of equation (3) that used this continuous measure of the program characteristic. In that model, we did not include subgroup indicators, and we replaced the two interaction terms with an interaction between the treatment indicator and the continuous measure of the program characteristic.

Assessing Variation in Impacts Across Districts

To assess whether impacts varied across districts (Chapter VI), we estimated a modified version of equation (2) for student achievement outcomes as follows:

$$(4)\quad Y_{js} = \sum_{d=1}^{10} \beta_d (T_s \times I_s^{(d)}) + X'_{js}\delta + Z'_s\gamma + B'_s\pi + \varepsilon_{js},$$

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 (4) 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.

Assessing Variation in 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 this, 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:

$$(5)\ Y_{js} = \sum_{b=1}^{44} \beta_{bd}(T_s \times B_s^{(bd)}) + X'_{js}\delta + Z'_s\gamma + B'_s\pi + \varepsilon_{js}.$$

In equation (5), $B_s^{(bd)}$ 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:

$$(6)\ \hat{\beta}_{bd} = u_d + \omega_{bd},$$

where $\hat{\beta}_{bd}$ was the estimated block-specific impact of pay-for-performance on student achievement from equation (5), u_d was 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, $Var(\hat{\beta}_{bd})$.

Estimating Associations Between Impacts on Educator Behaviors and Impacts on Student Outcomes

In Appendix G, we estimated associations between impacts on educator behaviors and 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 [5]). 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 this regression, 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 educator behaviors and 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 (85 percent in math and 93 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.[67]

Comparing Impacts on Districts' Ratings of School Achievement Growth to Impacts on Student Math and Reading Achievement

In Chapter VI, we reported the impacts of pay-for-performance on districts' ratings of school achievement growth and on student test scores in math and reading. Because districts' ratings of school achievement growth were supposed to be based on the same student test scores that we collected for the study, we sought to compare the impacts of pay-for-performance on these two types of outcomes. However, impacts on these two types of outcomes were not initially comparable for two reasons. First, impacts on school achievement growth ratings were expressed in points on a 1-to-4 rating scale, whereas impacts on student test scores were expressed in student z-score units. Second, school achievement growth ratings captured annual student growth, so impacts on those ratings in a particular year reflected whether growth in that year was higher in treatment schools than control schools. In contrast, impacts on student test scores reflected the cumulative impacts of exposure to pay-for-performance for up to three years. Pay-for-performance could have a positive cumulative

[67] 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.

impact after three years if it increased schools' annual student achievement growth in the first year, even if it led to no additional growth over the next two years.

We used the following method to compare impacts on districts' ratings of school achievement growth with impacts on student test scores. First, we used the student test scores to estimate impacts of pay-for-performance on our own measure of annual school achievement growth. Like impacts on student test scores, impacts on this measure were expressed in student z-score units. Second, we converted impacts on school achievement growth ratings from a 1-to-4 point rating scale into student z-score units. This allowed us to compare impacts on districts' ratings of school achievement growth, now expressed in student z-score units, to impacts on our own measure of annual school achievement growth.

We estimated impacts on our own measure of annual school achievement growth using a regression model that differed from our main model for student achievement (equation [2]) in two ways. First, it controlled for prior-year test scores instead of pre-intervention test scores, so the outcome was effectively growth (or value added) from one year to the next. Second, it pooled math and reading test scores together to estimate one average school growth measure. Accordingly, the model included a binary indicator for math (rather than reading) outcomes and interactions between the math indicator and all student-level and school-level covariates. Since the model required prior-year test score data, we restricted the sample to students in grades 4 through 8 for whom prior-year scores were available.

We used a two-step method to convert impacts on districts' ratings of school achievement growth from points on a 1-to-4 rating scale into student z-score units. In the first step, we converted impacts on school achievement growth ratings into school-level standard deviation units, by dividing the impacts by the pooled within-district standard deviation of school achievement growth ratings. In the second step, we multiplied these impacts by the pooled within-district standard deviation of our own measure of annual school achievement growth, expressed in student z-score units. For example, in Appendix F, we found that pay-for-performance raised districts' ratings of school achievement growth in Year 2 by 0.28 points on the 1-to-4 rating scale (Appendix F, Table F.18). Because the standard deviation of school achievement growth ratings was 0.99 points, this impact could be expressed as $0.28/0.99 = 0.28$ standard deviations of school achievement growth. Because one standard deviation of our measure of school achievement growth was 0.09 student z-score units, this impact could then be expressed as $0.28*0.09 = 0.03$ student z-score units.

To use a consistent sample of schools for these conversions, we restricted the sample of schools in these analyses to those with nonmissing values of school achievement growth ratings and our measure of annual school achievement growth (that is, schools with students in grades 4 through 8).

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 and from Years 2 to 3) 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:

$$(7)\quad \overline{Y}_{st} = \frac{1}{N_{st}} \sum_{j=1}^{N_{st}} Y_{jst}$$

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 $\overline{Y}_{st}$ in the two years being compared (Years 1 versus 2 or Years 2 versus 3). 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 or Year 3 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:

$$(8)\quad \overline{Y}_{st} = \delta Later_t + \sum_{h=1}^{N} \varphi_h I_s^{(h)} + \omega_{st}$$

where $Later_t$ was an indicator for the later year (Year 2 when comparing Years 1 and 2, and Year 3 when comparing Years 2 and 3) 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 answered the eligibility question, 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.6 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.

Table B.6. Realized Values of Minimum Detectable Impacts

Outcome	Units	Minimum Detectable Impact
School Achievement Growth Ratings, Year 1	Points on 1-to-4 scale	0.47
School Achievement Growth Ratings, Year 2	Points on 1-to-4 scale	0.41
School Achievement Growth Ratings, Year 3	Points on 1-to-4 scale	0.37
Teachers' Classroom Observation Ratings, Year 1	Points on 1-to-4 scale	0.07
Teachers' Classroom Observation Ratings, Year 2	Points on 1-to-4 scale	0.07
Teachers' Classroom Observation Ratings, Year 3	Points on 1-to-4 scale	0.07
Teachers' Classroom Achievement Growth Ratings, Year 1	Points on 1-to-4 scale	0.22
Teachers' Classroom Achievement Growth Ratings, Year 2	Points on 1-to-4 scale	0.15
Teachers' Classroom Achievement Growth Ratings, Year 3	Points on 1-to-4 scale	0.13
Observation Ratings for Principals, Year 1	Points on 1-to-4 scale	0.22
Observation Ratings for Principals, Year 2	Points on 1-to-4 scale	0.27
Observation Ratings for Principals, Year 3	Points on 1-to-4 scale	0.20
Student Math Achievement, Year 1	Student *z*-score units	0.05
Student Math Achievement, Year 2	Student *z*-score units	0.06
Student Math Achievement, Year 3	Student *z*-score units	0.07
Student Reading Achievement, Year 1	Student *z*-score units	0.04
Student Reading Achievement, Year 2	Student *z*-score units	0.04
Student Reading Achievement, Year 3	Student *z*-score units	0.04

Source: Educator and student administrative data.

APPENDIX C

SUPPLEMENTAL FINDINGS ON PROGRAMS AND EXPERIENCES OF ALL 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 final sample for these analyses consisted of 144 TIF districts—13 evaluation and 131 non-evaluation districts—that participated in TIF in 2013–2014 and responded to the 2014 district survey. The 2013–2014 school year, which we refer to as Year 3, was the third year of implementation for nearly all those districts. We refer to the 2011–2012 and 2012–2013 school years as Year 1 and Year 2, respectively.

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, Year 3 (Percentages Unless Otherwise Noted)

	All TIF Districts
Teachers	
Average Number of Classroom Observations per School Year	3
Average Length of Classroom Observations (in minutes)	45
Conducting Observations by a Trained Observer	96
Classroom Observations are Conducted by:	
Principal or other administrators at the teacher's school	93
Teacher leaders or peer observers[a]	52
District administrative staff	49
Externally hired observers (Non-district employees)	7
Number of Districts—Range[b]	**135–138**
Principals	
Average Number of Observations per School Year	3
Average Length of Observations (in minutes)	47
Observations are conducted by:	
Superintendent	49
Other central office administrator from the same district	51
Administrator from another district	4
Number of Districts—Range[b]	**138–141**

Source: District survey, 2014.

[a]Department heads, coaches, other senior teachers (at or outside school).

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

Table C.2. Additional Pay Opportunities for Teachers and Principals for Additional Factors, Year 3

	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	33	$3,129
Attending professional development activities or enrolling in graduate-level courses	30	$960
Number of Districts—Range[a]	**140-141**	**31-45**
Principals		
Additional Factors		
Working in a hard-to-staff school	12	$6,750
Attending professional development activities or enrolling in graduate-level courses	23	$958
Number of Districts—Range[a]	**140–143**	**14–45**

Source: District survey, 2014.

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 and 3. The sample was restricted to the 140 districts that responded to both the 2013 and 2014 district survey.

Table C.3. Challenges Implementing TIF in Year 2 and Year 3 (Percentages)

	In Year 2, Percentage of All TIF Districts Reporting Activity Was:			In Year 3, Percentage of All TIF Districts Reporting Activity Was:		
Activity	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	23	50	20	32	48
Attributing student achievement growth to individual teachers	28	28	44	20	35	45
Explaining student achievement measures to educators	28	45	28	19	57	24
Providing useful and timely feedback on student achievement measures to educators	30	41	29	19+	45	36
Collecting and storing data linking teachers to student achievement data	22	38	40	18	37	45
Teacher Classroom Observations						
Choosing a classroom observation tool	7	20	72	1+	7+	91+
Finding a tool that is ready for implementation	8	17	75	1+	7+	92+
Hiring observers	2	20	78	3	10+	87
Training observers to use the tool	11	47	42	4+	38	58+
Scheduling and/or conducting observations	24	51	24	16	54	30
Providing useful and/or timely feedback from observations	25	47	28	14+	49	37
Collecting and storing observation data	14	34	52	4	35	62
Principal Observations						
Choosing a principal observation tool	14	34	52	3+	18+	79+
Finding a tool that is ready for implementation	16	29	56	3+	17+	80+
Hiring observers	2	14	84	3	9	88
Training observers to use the tool	5	39	56	4	30	66
Scheduling and/or conducting observations	14	48	38	12	39	50
Providing useful and/or timely feedback from observations	15	41	44	4+	46	50
Pay-for-Performance Bonuses						
Defining the criteria for earning a pay-for-performance bonus or the amount of the bonus	22	42	35	10+	29+	60+
Calculating pay-for-performance bonuses	20	29	51	6+	31	63
Distributing pay-for-performance bonuses	9	35	56	4	28	68

Activity	In Year 2, Percentage of All TIF Districts Reporting Activity Was:			In Year 3, Percentage of All TIF Districts Reporting Activity Was:		
	Major Challenge	Minor Challenge	Not a Challenge	Major Challenge	Minor Challenge	Not a Challenge
Communicating the TIF Program to Educators or Other Stakeholders						
Communicating the TIF program to educators	14	48	39	6+	45	49
Communicating bonus payouts to educators	14	42	44	5+	41	54
Communicating with other stakeholders	13	52	35	9	49	43
Obtaining or Maintaining Support for the TIF Program						
Teachers or teachers' union or association	12	31	57	5+	30	65
Principals or principals' union or association	1	18	81	2	16	82
Superintendent	2	14	83	1	14	86
School board	1	30	69	4	20	76
Parents or broader community	2	26	72	3	25	72
Other TIF Issues						
Choosing educators for additional roles and responsibilities	7	46	47	4	34	61+
Sustainability of the TIF program	64	29	7	50+	32	18+
Number of Districts—Range[a]	**132–140**	**132–140**	**132–140**	**132–140**	**132–140**	**132–140**

Source: District survey (2013 and 2014).

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

+Difference between Year 2 and Year 3 is significant at the .05 level, two-tailed test.

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 three years of implementation during the period covered by this report. Year 1 represents the first year of implementation (2011–2012), Year 2 the second (2012–2013), and Year 3 the third year of implementation (2013–2014). Cohort 2 districts completed only two years of implementation, 2012–2013 and 2013–2014, referred to as Years 1 and 2 for this cohort.

The analyses in Chapter IV were based on Cohort 1 only and, in general, focused on findings for Year 3. 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 for Year 2 based on 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 in Year 3. Table D.1 shows additional details on teacher classroom observations as reported by the Cohort 1 districts.

Figure IV.1 in Chapter IV illustrates that school achievement growth and classroom observations sometimes identified the same teachers as high-performing in Year 3, 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 3. More than two-thirds of the 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 3, about one-quarter (28 percent) of teachers received similar ratings based on classroom observations and classroom achievement growth, and 50 percent received a higher classroom observation rating than classroom achievement growth rating.

Table D.1. Classroom Observations to Evaluate Teachers in Year 3, Cohort 1 (Percentages Unless Otherwise Noted)

	Evaluation Districts
Average Number of Observations per School Year	4
Average Length of Observations (in minutes)	42
Observations are Conducted by a Trained Observer	100
Observations are Conducted by	
Principal or other administrators at the teacher's school	78
Teacher leaders or peer observers[a]	44
District administrative staff	33
Externally hired observers (nondistrict employees)	11
Number of Districts	**9-10**

Source: District survey, 2014.

[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 3, Cohort 1 (Percentages)

	Principal Observation Rating			
School Achievement Growth Rating	"Ineffective" or "Somewhat Effective"	"Effective"	"Highly Effective"	**Number of Principals**
"Ineffective"	4	11	5	**29**
"Somewhat Effective"	4	18	28	**52**
"Effective"	1	8	3	**16**
"Highly Effective"	1	8	8	**24**
Number of Principals	**14**	**50**	**57**	**121**

Source: Educator administrative data.

Notes: Categories are study-constructed labels to represent quarters of a 1-to-4 rating scale. "Ineffective" = bottom quarter (1 to 1.75); "Somewhat Effective" = second quarter (1.75 to 2.5); "Effective" = third quarter (2.5 to 3.25); "Highly Effective" = top quarter (3.25 to 4). The table is based on principals with ratings on both observations and school achievement growth in Year 3. "Ineffective" and "Somewhat Effective" categories were combined due to the small number of principals who received ratings in these categories.

Table D.3. Comparison of Teachers' Ratings on Classroom Observations and Classroom Achievement Growth in Year 3, Cohort 1 (Percentages)

	Classroom Observation Rating				
Classroom Achievement Growth Rating	"Ineffective"	"Somewhat Effective"	"Effective"	"Highly Effective"	**Number of Teachers**
"Ineffective"	1	14	15	2	**574**
"Somewhat Effective"	0	13	12	5	**531**
"Effective"	0	5	8	2	**224**
"Highly Effective"	0	5	12	6	**560**
Number of Teachers	**23**	**488**	**1,014**	**364**	**1,889**

Source: Educator administrative data.

Notes: Categories are study-constructed labels to represent quarters of a 1-to-4 rating scale. "Ineffective" = bottom quarter (1 to 1.75); "Somewhat Effective" = second quarter (1.75 to 2.5); "Effective" = third quarter (2.5 to 3.25); "Highly Effective" = top quarter (3.25 to 4). The table is based on teachers with ratings on both classroom observations and classroom achievement growth in Year 3.

Figure IV.2 in Chapter IV compares teachers' classroom observation ratings in Years 2 and 3. Tables D.4 and D.5 compare teachers' school achievement ratings and classroom achievement growth ratings in Years 2 and 3, respectively. More than half of teachers received similar ratings on these measures in both years (56 percent on school achievement growth and 55 percent for classroom achievement growth).

Table D.4. Comparison of Teachers' School Achievement Growth Ratings in Years 2 and 3, Cohort 1 (Percentages)

	Teacher's School Achievement Growth Rating in Year 3				
Teacher's School Achievement Growth Rating in Year 2	"Ineffective"	"Somewhat Effective"	"Effective"	"Highly Effective"	**Number of Teachers**
"Ineffective"	13	7	2	1	**906**
"Somewhat Effective"	8	34	3	6	**1,556**
"Effective"	0	5	3	0	**308**
"Highly Effective"	3	5	3	6	**582**
Number of Teachers	**810**	**1,388**	**496**	**658**	**3,352**

Source: Educator administrative data.

Notes: Categories are study-constructed labels to represent quarters of a 1-to-4 rating scale. "Ineffective" = bottom quarter (1 to 1.75); "Somewhat Effective" = second quarter (1.75 to 2.5); "Effective" = third quarter (2.5 to 3.25); "Highly Effective" = top quarter (3.25 to 4). The table is based on teachers with school achievement growth ratings in both Years 2 and 3.

Table D.5. Comparison of Teachers' Classroom Achievement Growth Ratings in Years 2 and 3, Cohort 1 (Percentages)

	Classroom Achievement Growth Rating in Year 3				
Classroom Achievement Growth Rating in Year 2	"Ineffective"	"Somewhat Effective"	"Effective"	"Highly Effective"	**Number of Teachers**
"Ineffective"	21	10	2	4	**393**
"Somewhat Effective"	7	17	7	2	**228**
"Effective"	3	3	6	4	**150**
"Highly Effective"	0	1	3	11	**164**
Number of Teachers	**346**	**208**	**152**	**229**	**935**

Source: Educator administrative data.

Notes: Categories are study-constructed labels to represent quarters of a 1-to-4 rating scale. "Ineffective" = bottom quarter (1 to 1.75); "Somewhat Effective" = second quarter (1.75 to 2.5); "Effective" = third quarter (2.5 to 3.25); "Highly Effective" = top quarter (3.25 to 4). The table is based on teachers with classroom achievement growth ratings in both Years 2 and 3.

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 for Year 2 on 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 in Year 3 (or Year 2) when the findings for that year 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 3. 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 Year 3, Cohorts 1 and 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 in tested grades and subjects could receive a bonus for their students' achievement growth			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			X	X	X	X		X	X	X	X		
District changed some aspect of its program between the 2012–2013 and 2013–2014 school years			X						X	X		X	X

Source: District interviews (2012, 2013, and 2014); 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.

Table D.7. Detailed Information on Measures and Criteria Used for Evaluation Districts' Teacher Pay-for-Performance Bonus Programs in Year 3, Cohorts 1 and 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 (SLTs)
    - 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

Source: District interviews (2012, 2013, and 2014), 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

Table IV.4 in Chapter IV shows the percentage of the Cohort 1 districts in Years 1, 2, and 3 that met the TIF grant goals for substantial, differentiated, and challenging to earn bonuses. Table D.8 compares the percentage of Cohort 1 districts that met these criteria to the percentage of both cohorts (Cohorts 1 and 2) that met these criteria in Years 1 and 2.

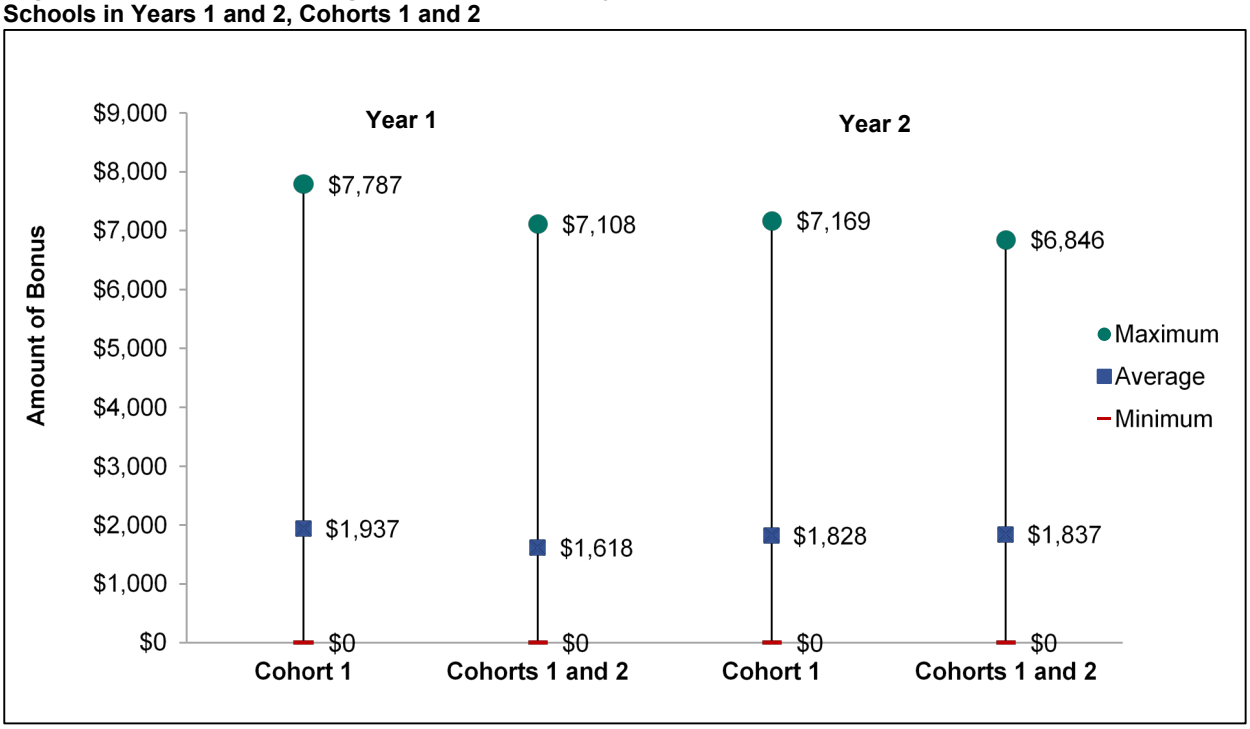
Table D.8. Evaluation Districts Meeting TIF Grant Goals for Pay-for-Performance Bonuses for Teachers in Years 1 and 2, Cohorts 1 and 2 (Percentages)

TIF Grant Goal	Cohort 1	Cohorts 1 and 2
Year 1		
Substantial: Average Bonus Was at Least 5 Percent of Average Salary	20	15
Differentiated: Highest Bonus Was at Least Three Times the Average Bonus	70	77
Challenging: Less Than 50 Percent Of Teachers Received a Pay-for-Performance Bonus	20	31
Year 2		
Substantial: Average Bonus Was at Least 5 Percent of Average Salary	20	23
Differentiated: Highest Bonus Was at Least Three Times the Average Bonus	60	62
Challenging: Less Than 50 Percent of Teachers Received a Pay-for-Performance Bonus	20	23
Number of Districts	**10**	**13**

Source: Educator administrative data.

Figure IV.3 shows the minimum, average, and maximum pay-for-performance bonuses in Years 1, 2, and 3 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 and 2 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, 2, and 3 for teachers in Cohort 1, with districts weighted by the number of schools.

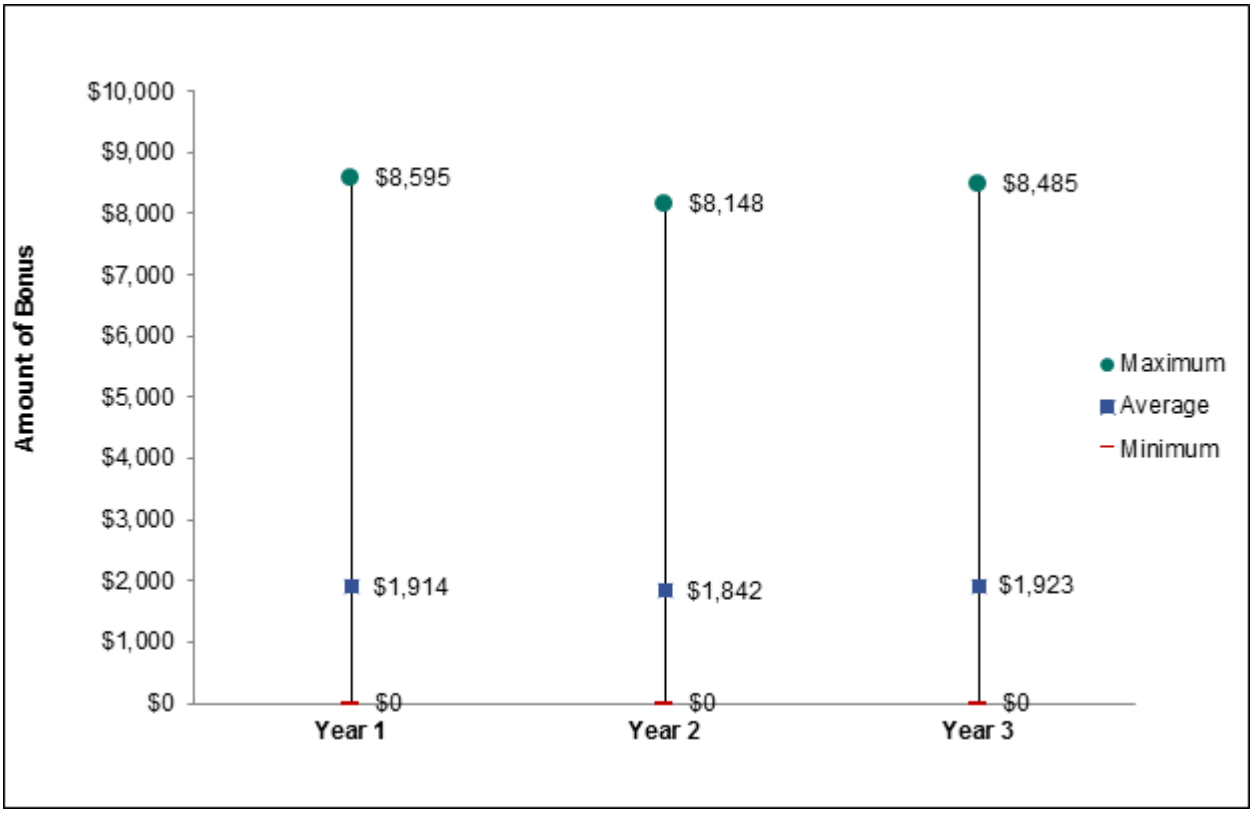
Figure D.1. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools in Years 1 and 2, Cohorts 1 and 2



Source: Educator administrative data (N = 2,181 teachers for Year 1 Cohort 1, N = 3,001 teachers for Year 1 Cohorts 1 and 2, N = 2,191 teachers for Year 2 Cohort 1, and N = 3,097 teachers for Year 2 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.

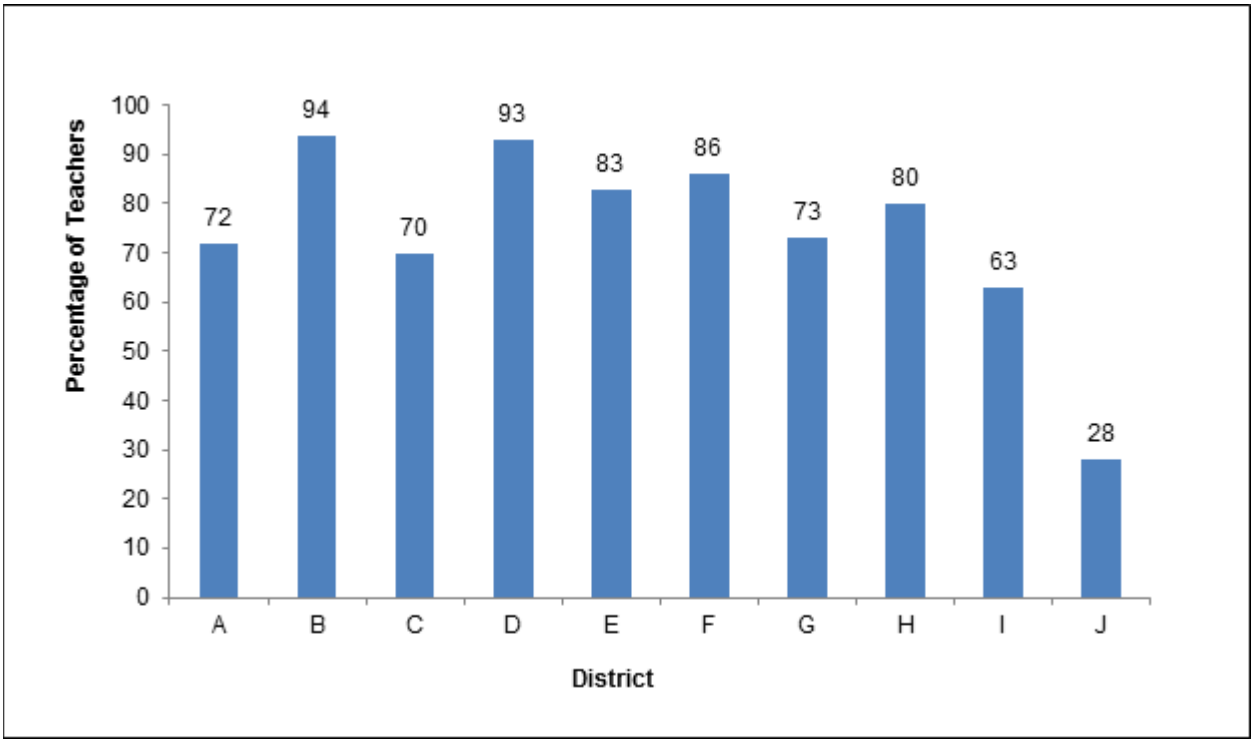
Figure D.2. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools, with Districts Weighted by the Number of Schools, Cohort 1



Source: Educator administrative data (N = 2,181 teachers in Year 1, N = 2,191 teachers in Year 2, and N = 2,260 teachers in Year 3).

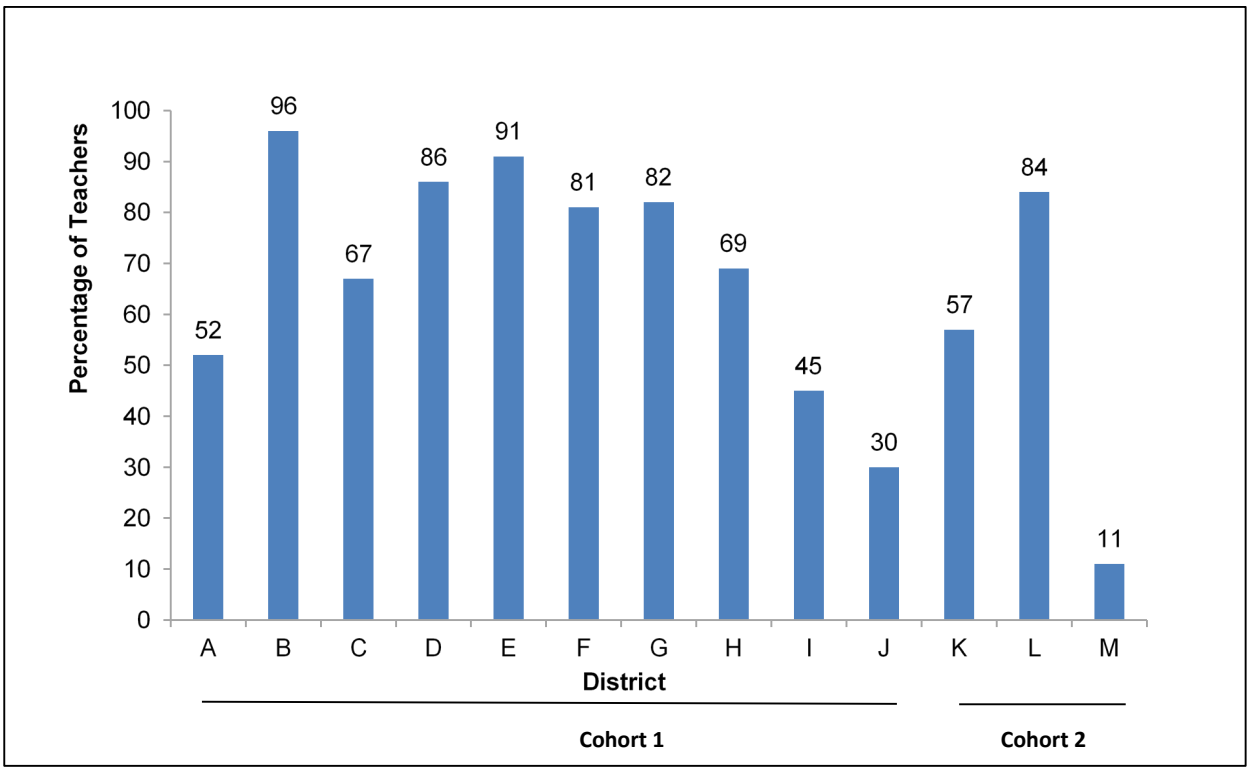
Applicants for the evaluation grants received guidance on the structure of their pay-for-performance bonus, including the example of *challenging* to earn bonuses, in which only those performing significantly better than the average (therefore, fewer than 50 percent) would receive a bonus. Figure IV.4 shows that across districts, on average, more than 70 percent of treatment teachers received a bonus each year. Figure D.3 shows the percentage of teachers earning pay-for-performance bonuses in Year 3, by district, for Cohort 1. Figure D.4 shows the percentage of teachers earning pay-for-performance bonuses in Year 2, by district, for Cohorts 1 and 2.

Figure D.3. Percentage of Treatment Teachers Earning a Pay-for-Performance Bonus in Year 3, By District, Cohort 1



Source: Educator administrative data (N ranges from 68 teachers in District D to 460 in District J).

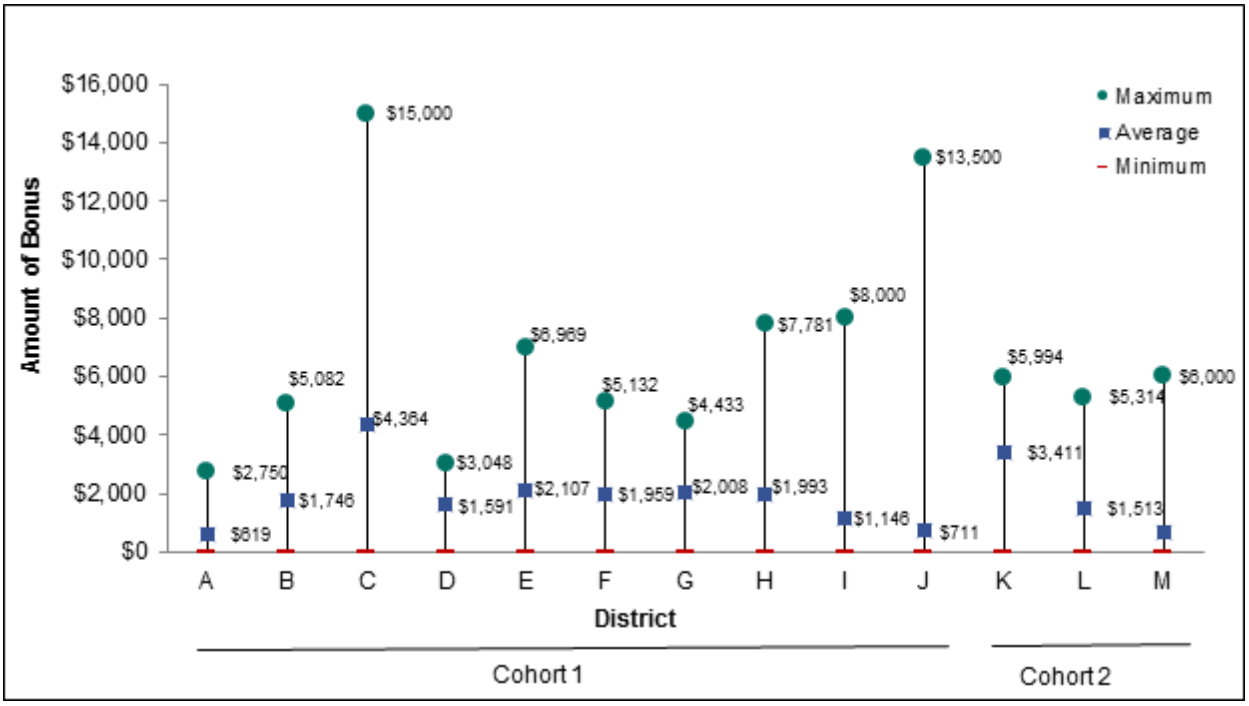
Figure D.4. Percentage of Treatment Teachers Earning a Pay-for-Performance Bonus in Year 2, by District, Cohorts 1 and 2



Source: Educator administrative data (N ranges from 46 teachers in District L to 467 in District M).

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 3. For comparison, we show the distribution of teachers' Year 2 pay-for-performance bonuses by district for Cohorts 1 and 2 (Figure D.5).

Figure D.5. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Teachers in Treatment Schools in Year 2, by District, Cohorts 1 and 2



Source: Educator administrative data (N ranges from 49 teachers in District L to 432 in District J).

Table D.9 compares the amount teachers in treatment schools received in performance bonuses in Years 2 and 3 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 (57 percent) received similar bonus award amounts in Years 2 and 3.

Table D.9. Comparison of Teachers' Performance Bonus Amounts in Years 2 and 3, Cohort 1

	Performance Bonus, Year 3				
Performance Bonus, Year 2	$0	$1-1,500	$1,501-3,000	Above $3,000	**Number of Teachers**
$0	18	5	1	2	**637**
$1-1,500	3	11	4	1	**362**
$1,501-3,000	5	10	18	4	**430**
Above $3,000	3	1	4	10	**380**
Number of Teachers	**648**	**457**	**342**	**362**	**1,809**

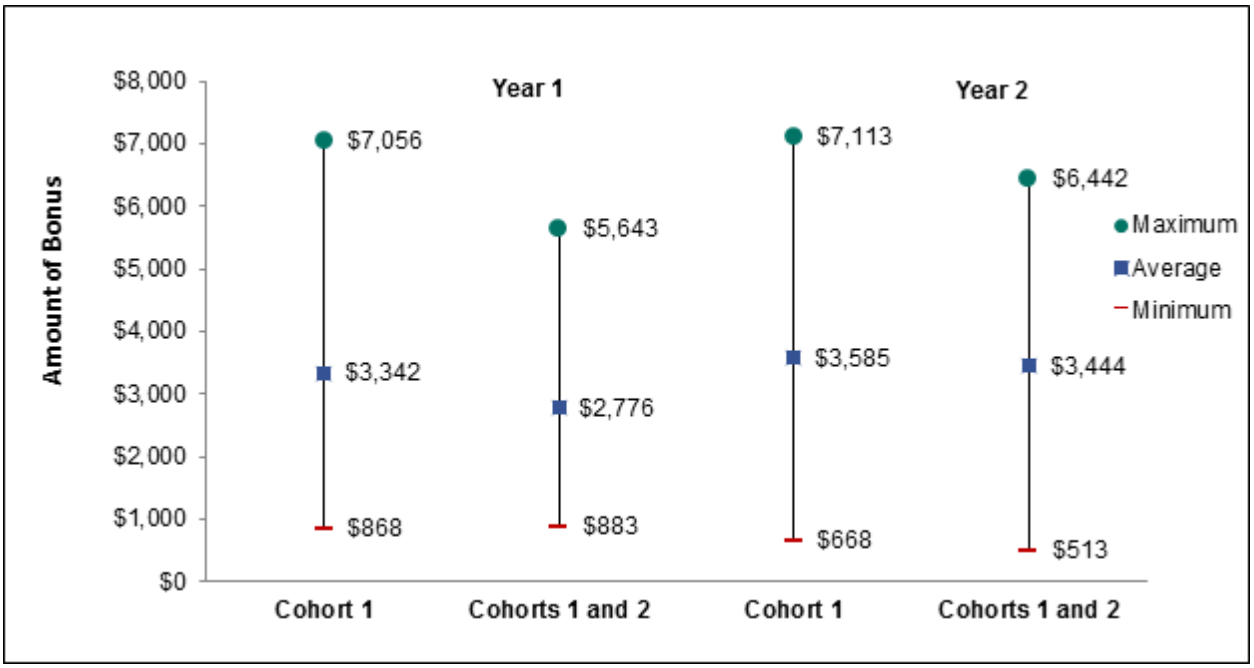
Source: Educator administrative data.

Note: Table is based on teachers who worked in treatment schools in both Years 2 and 3.

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, 2, and 3 for principals in Cohort 1, with each district equally weighted. Figure D.6 presents the minimum, average, and maximum pay-for-performance bonuses in Years 1 and 2 for principals in Cohorts 1 and 2, also with districts weighted equally. Figure D.7 shows the minimum, average, and maximum pay-for-performance bonuses for principals in Cohort 1 with districts weighted by the number of schools.

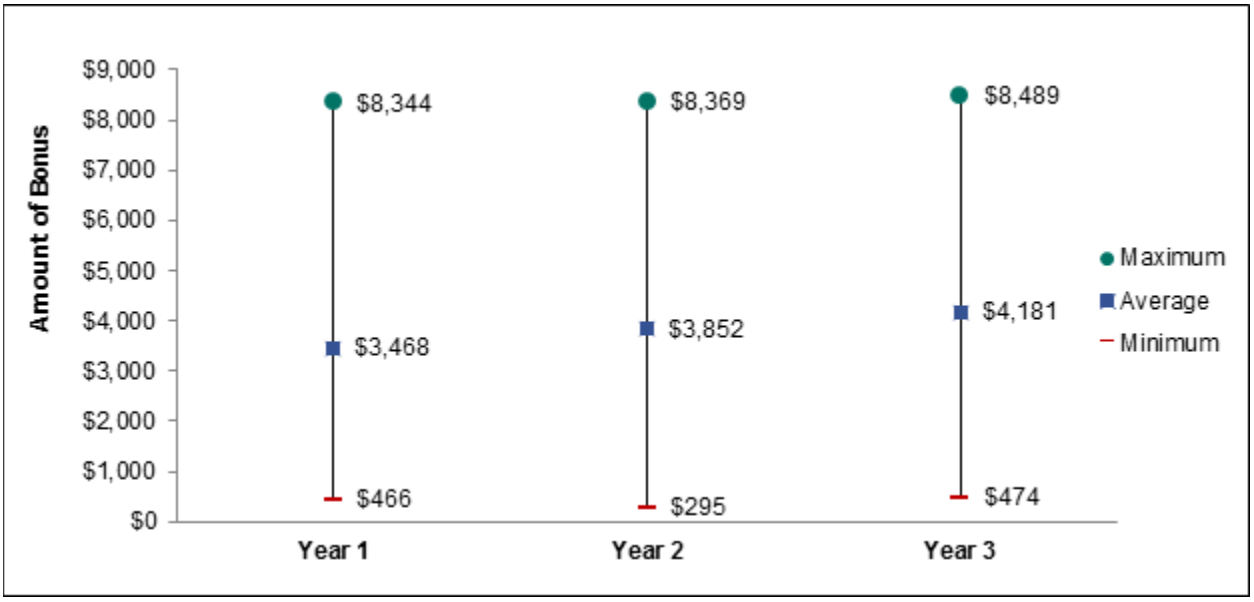
Figure D.6. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Principals for Years 1 and 2, Cohorts 1 and 2



Source: Educator administrative data (N = 65 principals in Year 1, Cohort 1 and N = 91 principals in Year 1, Cohorts 1 and 2; N = 68 principals in Year 2, Cohort 1 and N = 88 principals in Year 2, 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.

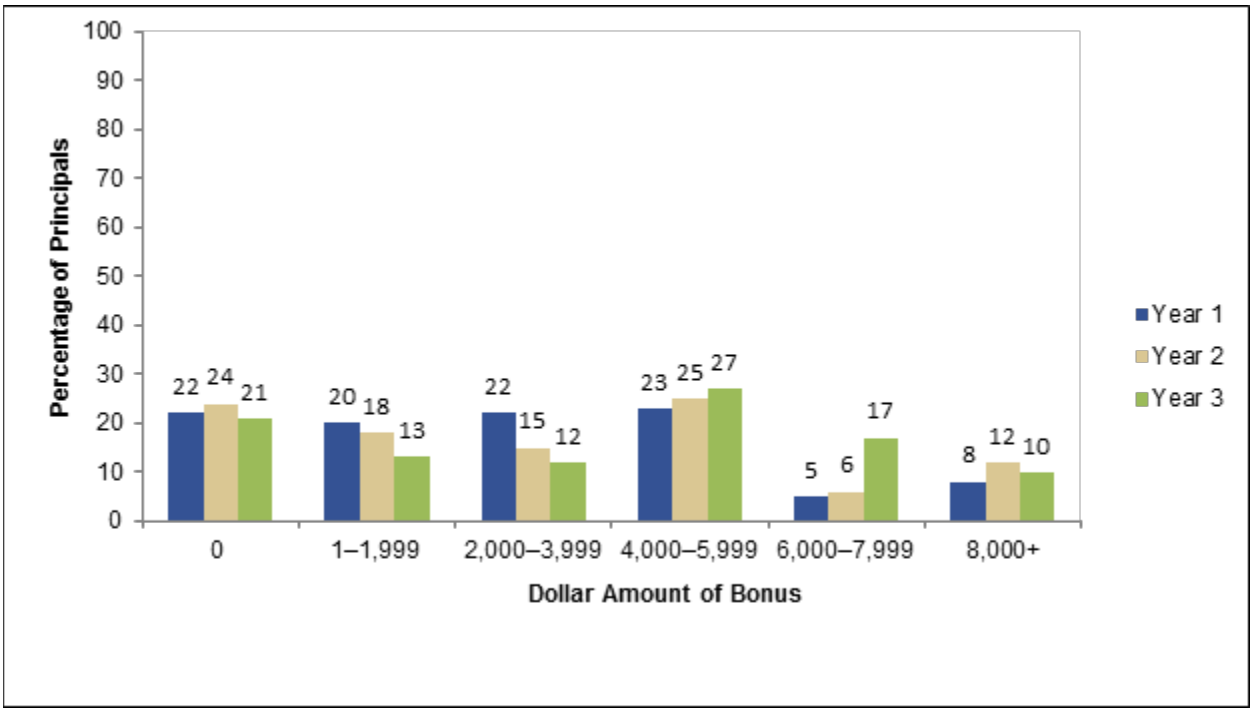
Figure D.7. Minimum, Average, and Maximum Pay-for-Performance Bonuses for Principals in Treatment Schools, with Districts Weighted by the Number of Schools, Cohort 1



Source: Educator administrative data (N = 65 principals in Year 1, N = 68 principals in Year 2, and N = 65 principals in Year 3).

In Chapter IV, we noted that 20 percent of the districts met the guidance for challenging bonuses in Years 1 and 2 (Table IV.5). Figure D.8 illustrates the distribution of principals' pay-for-performance bonuses in Years 1, 2, and 3 for Cohort 1. At least 75 percent of principals in each year received a bonus.

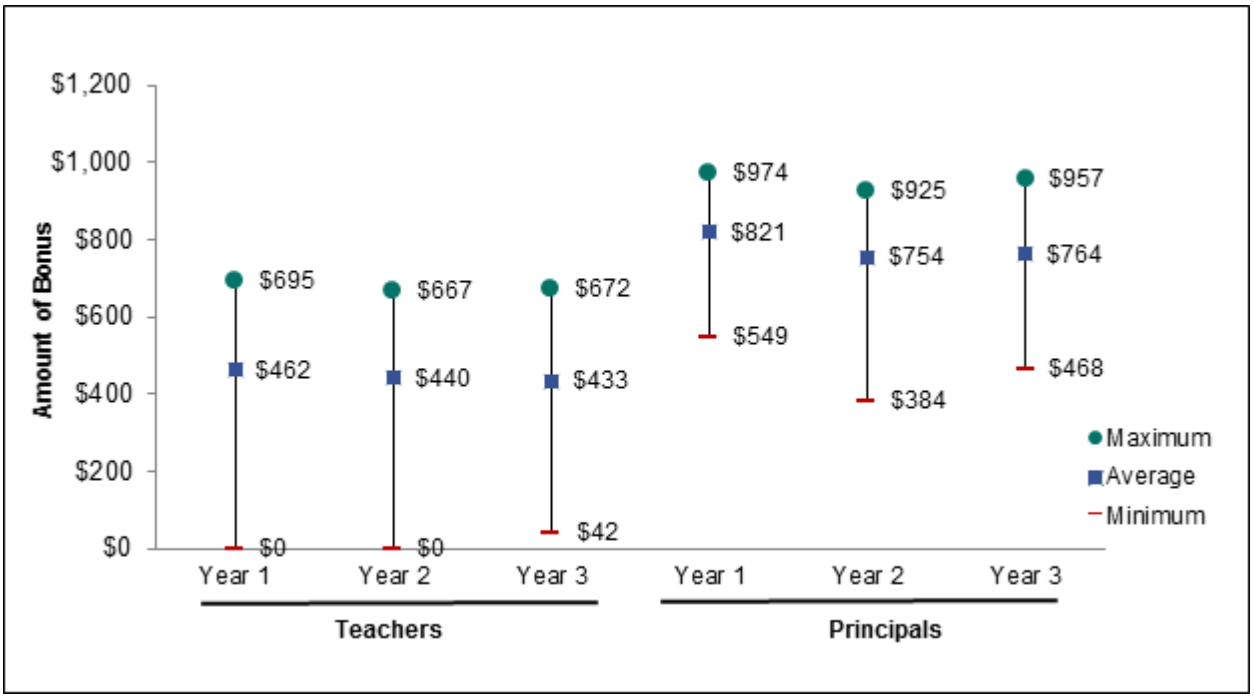
Figure D.8. Distribution of Pay-for-Performance Bonuses for Principals in Treatment Schools, Cohort 1



Source: Educator administrative data (N = 65 principals in Year 1, N = 68 principals in Year 2, and N = 65 principals in Year 3).

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.9 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 did not vary substantially.

Figure D.9. Minimum, Average, and Maximum Automatic 1 Percent Bonuses for Teachers and Principals in Control Schools, Cohort 1



Source: Educator administrative data (Year 1: N = 2,152 teachers and N = 69 principals; Year 2: N = 2,242 teachers and N = 70 principals; Year 3: N = 2,285 teachers and N = 69 principals).

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.10 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, 2, and 3. 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 ($1,851 in Year 3; Figure IV.3)

Table D.11 compares the amount of additional pay received by Cohort 1 teachers in treatment and control schools in Years 1, 2, and 3. 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.10. Average and Maximum Amounts of Additional Pay Opportunities for Teachers, Cohort 1

Additional Pay Opportunities	Year 1	Year 2	Year 3
Average Amount for Additional Pay Opportunities (dollars)	449	492	498
Maximum Amount for Additional Pay Opportunities (dollars)	4,766	5,594	5,275
Percentage of Districts Offering Additional Pay Opportunities	100	100	100
Percentage of Teachers Receiving Additional Pay Opportunities	12	17	17
Number of Teachers	**4,333**	**4,433**	**4,545**

Source: Educator administrative data.

Table D.11. Actual Amounts of Teachers' Additional Pay, Cohort 1

	Year 1		Year 2		Year 3	
	Treatment	Control	Treatment	Control	Treatment	Control
Roles and Responsibilities						
Average additional pay (dollars)	489	504	510	532	514	500
Received pay (percentage)	12*	14	15*	18	17	17
Other Additional Pay[a]						
Average additional pay (dollars)	344	329	351	388	360	392
Received pay (percentage)	22	22	13*	18	14	19
Number of Teachers	**2,181**	**2,152**	**2,191**	**2,242**	**2,260**	**2,285**

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 0.05 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.12 shows that all or almost all districts reported providing professional development on how to improve their performance on classroom observations and achievement growth. 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.12. Professional Development Based on Observation and Student Achievement Growth Ratings, As Reported by Evaluation Districts in Year 3, Cohort 1 (Percentages Unless Otherwise Noted)

	Professional Development Based on Classroom Observations	Professional Development Based on Student Achievement Growth
Teachers Received Professional Development on How to Improve their Performance on the Measure	100	90
Teachers Received Professional Development Based on their Individual Performance on the Measure	80	30
Average Hours of Professional Development	28	38
Frequency		
Throughout year	60	80
1–4 times per year	30	20
Varies	10	0
Teachers Were Required to Participate	80	60
Number of Districts	**10**	**10**

Source: District interviews, 2014.

Communication of TIF Program

We asked district administrators more detailed information on their communication activities during the district interviews. Table D.13 shows districts' activities to communicate information about ratings based on classroom observations and student achievement growth to teachers. Table D.14 provides information on what districts told Cohort 1 teachers about their individual bonus for Year 2 and what districts informed teachers about Year 2 bonuses more generally.

Table D.13. Districts' Activities to Communicate to Teachers Their Classroom Observation and Student Achievement Growth Ratings for Year 2, Cohort 1 (Percentages)

	Evaluation Districts
Activities to Communicate Classroom Observation Rating	
In-person meeting	90
Online system	80
Letter to individual	40
E-mail to individual	20
Activities to Communicate Student Achievement Growth Rating	
In-person meeting	90
Online system	60
Letter to individual	50
E-mail to individual	20
Number of Districts	**10**

Source: District interviews, 2014.

Table D.14. Communication Methods Used to Inform Teachers About Individual Pay-for-Performance Bonuses Based on the Second Year of TIF Implementation (Percentages)

	Evaluation Districts
Letter or Email to Each Teacher with Individual Bonus Amount	80
Individual Meeting with Each Teacher to Discuss Bonus Amount	30
Number of Districts	**10**

Source: District survey, 2014.

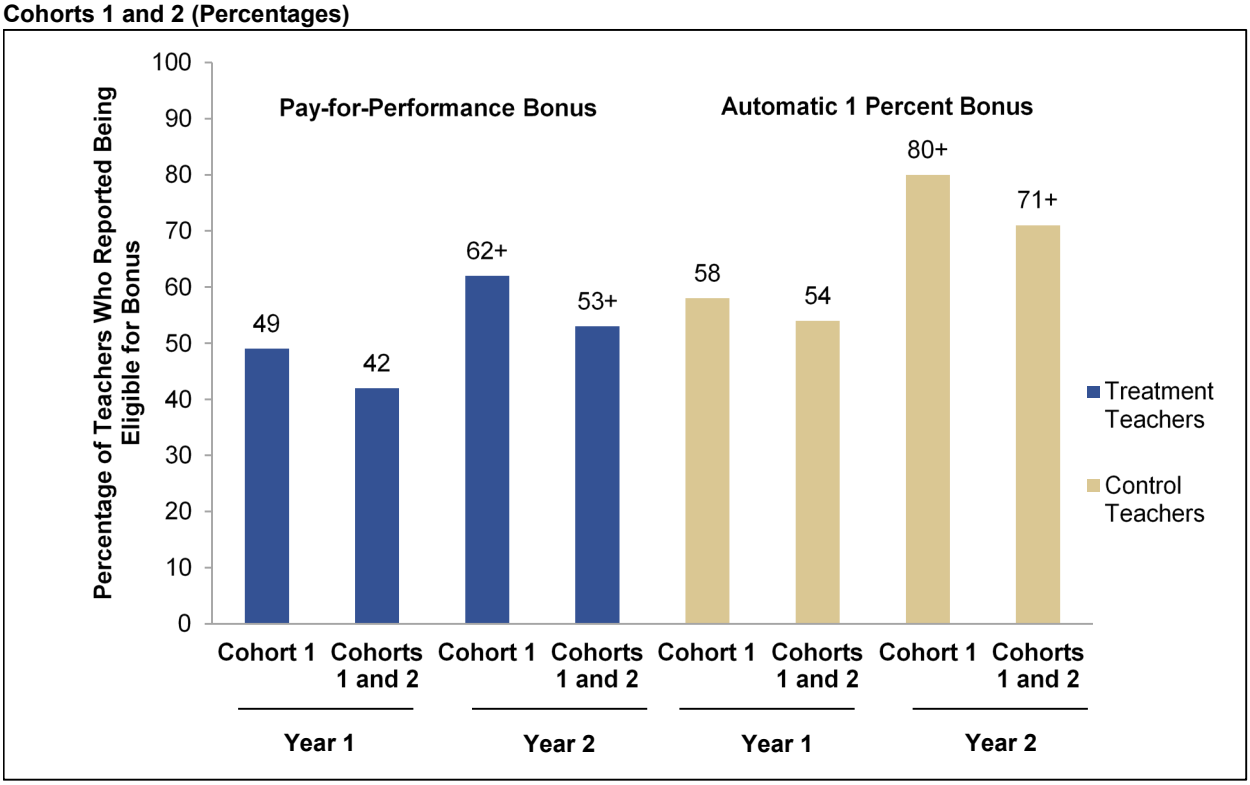
Teacher and Principal Perspectives Regarding TIF Implementation

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

Figures IV.9 and IV.10 show the percentages of Cohort 1 educators in treatment and control schools who reported they were eligible for either bonus in Years 1 through 3. Figure D.10 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 and 2 for Cohort 1 compared to Cohorts 1 and 2 combined. Figure D.11 shows the same information for principals. When Years 1 and 2 analyses were based on Cohorts 1 and 2, similar but somewhat 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.10. Teachers' Pay-for-Performance Bonus Eligibility in Years 1 and 2, as Reported by Teachers, Cohorts 1 and 2 (Percentages)

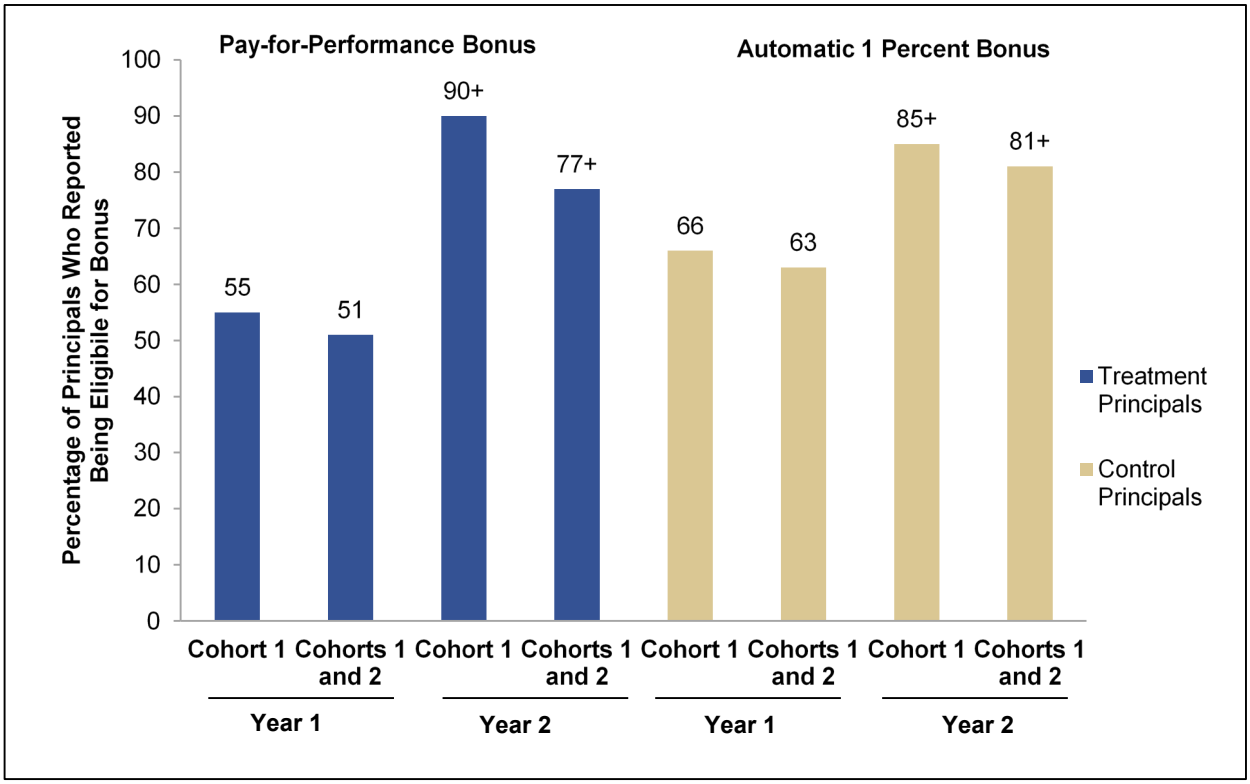


Source: Teacher survey (2012, 2013, and 2014).

Notes: A total of 377 treatment teachers in Cohort 1 and 495 in Cohorts 1 and 2 responded to the question about eligibility for a pay-for-performance bonus in Year 1. A total of 444 treatment teachers in Cohort 1 and 582 in Cohorts 1 and 2 responded to the question about eligibility for a pay-for-performance bonus in Year 2. A total of 381 control teachers in Cohort 1 and 474 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 565 in Cohorts 1 and 2 responded to the question about eligibility for an automatic 1 percent bonus in Year 2.

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

Figure D.11. Principals' Pay-for-Performance Bonus Eligibility in Years 1 and 2, as Reported by Principals, Cohorts 1 and 2 (Percentages)



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

Notes: A total of 64 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 1. A total of 63 treatment principals in Cohort 1 and 82 in Cohorts 1 and 2 responded to the question about eligibility for a pay-for-performance bonus in Year 2. A total of 64 control principals in Cohort 1 and 81 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 78 in Cohorts 1 and 2 responded to the question about eligibility for an automatic 1 percent bonus in Year 2.

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

Table D.15 shows the percentage of Cohort 1 educators who correctly reported their bonus eligibility as intended by the study design (also shown in Figures D.10 and D.11), 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 between Years 2 and 3 (Figures IV.9 and IV.10). Furthermore, by the third year of TIF implementation, many educators continued to misreport their eligibility. For example, in Year 3, 43 percent of treatment teachers did not report being eligible for a pay-for-performance bonus, 40 percent of treatment teachers believed they were eligible for an automatic 1 percent bonus, and 18 percent of control teachers believed they were eligible for a pay-for-performance bonus.

Table D.15. Bonus Eligibility as Reported by Teachers and Principals, Cohort 1 (Percentages)

	Year 1		Year 2		Year 3	
	Treatment	Control	Treatment	Control	Treatment	Control
Teachers						
Pay-for-Performance Bonus	49*	17	62*+	17	57*	18
Automatic 1 Percent Bonus	39*	58	40*	80+	40*	76
Number of Teachers—Range[a]	**377–378**	**379–381**	**434–444**	**445–449**	**412–424**	**448–455**
Principals						
Pay-for-Performance Bonus	55*	13	90*+	15	78*+	13
Automatic 1 Percent Bonus	27*	66	32*	85+	21*	85
Number of Principals—Range[a]	**63–64**	**63–64**	**63–64**	**61**	**58–59**	**61**

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

[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.

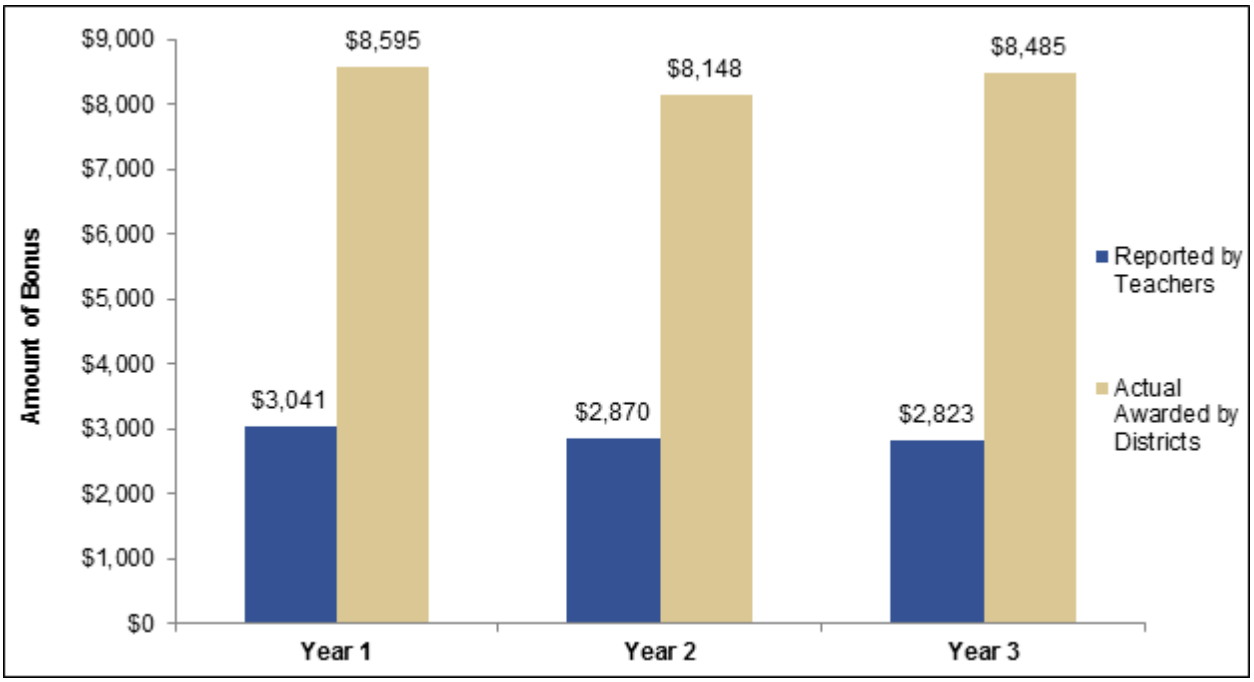
+Difference with prior year within treatment status is statistically significant at the .05 level, two-tailed test.

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

Figures IV.11 and IV.12 show the reported and actual and maximum pay-for-performance bonuses for teachers and for principals, respectively, for Cohort 1 in Years 1, 2, and 3. 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 and 2 for Cohorts 1 and 2.

Figures D.12 and D.13 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.12 and D.13 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.12. Reported and Actual Maximum Pay-for-Performance Bonus for Teachers in Treatment Schools, with Districts Weighted by the Number of Schools, Cohort 1

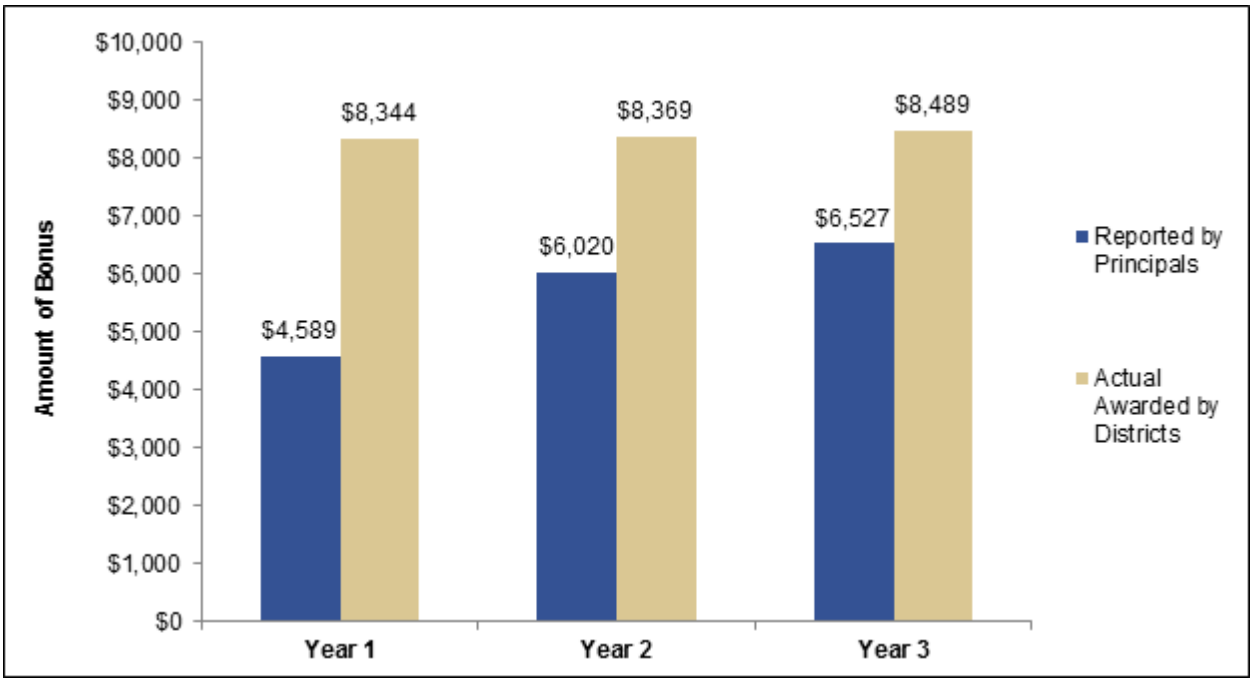


Source: Teacher survey (2012, 2013, and 2014); district interviews; and educator administrative data.

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 196 treatment teachers in tested grades and subjects responded to this survey question in Year 1, a total of 218 in Year 2, and a total of 217 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 27 additional responses for treatment teachers in Year 1, 14 additional responses in Year 2, and 15 additional responses in Year 3. See Appendix B for additional discussion on the imputation methods. Appendix D, Table D.16 shows that our results are similar if we do not impute the missing bonus amounts.

Figure D.13. Reported and Actual Maximum Pay-for-Performance Bonus for Principals in Treatment Schools, with Districts Weighted by the Number of Schools, Cohort 1



Source: Principal survey (2012, 2013, and 2014); 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 56 treatment principals responded to this survey question in Year 1, a total of 61 in Year 2, and a total of 58 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 8 additional responses for treatment principals in Year 1, 2 in Year 2, and 0 in Year 3. See Appendix B for additional discussion on the imputation methods. Appendix D, Table D.16 shows that our results are similar if we do not impute the missing bonus amounts.

Table D.16 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.16 shows that our results are similar if we do not impute the missing bonus amounts.

Table D.16. Educators' Reports of the Maximum Possible Bonus Amount: Imputed and Non-Imputed Bonus Amounts, Cohort 1

	Year 1		Year 2		Year 3	
	Treatment	Control	Treatment	Control	Treatment	Control
Teachers						
Pay-for-Performance						
With imputed amounts	$3,041*	$395	$2,870*	$506	$2,823*	$134
Only non-Imputed amounts	$2,802*	$293	$2,823*	$460	$2,776*	$122
Automatic 1 Percent Bonus						
With imputed amounts	$831	$1,086	$994	$872	$1,151	$1,928
Only non-imputed amounts	$578	$970	$749	$764	$1,554	$2,602
Number of Teachers—Range[a]	**196–224**	**190–222**	**194–232**	**185–252**	**177–217**	**186–234**
Principals RT 27						
Pay-For-Performance						
With imputed amounts	$4,589*	$652	$6,020*	$321	NA	NA
Only non-Imputed amounts	$4,316*	$207	$5,960*	$321	$6,527*	$374
Automatic 1 Percent Bonus						
With imputed amounts	$1,859	$1,060	$1,107	$1,214	$788	$1,338
Only non-Imputed amounts	$1,751	$979	$851	$992	$837	$1,286
Number of Principals—Range[a]	**56–64**	**58–64**	**60–64**	**46–61**	**58–59**	**53–61**

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

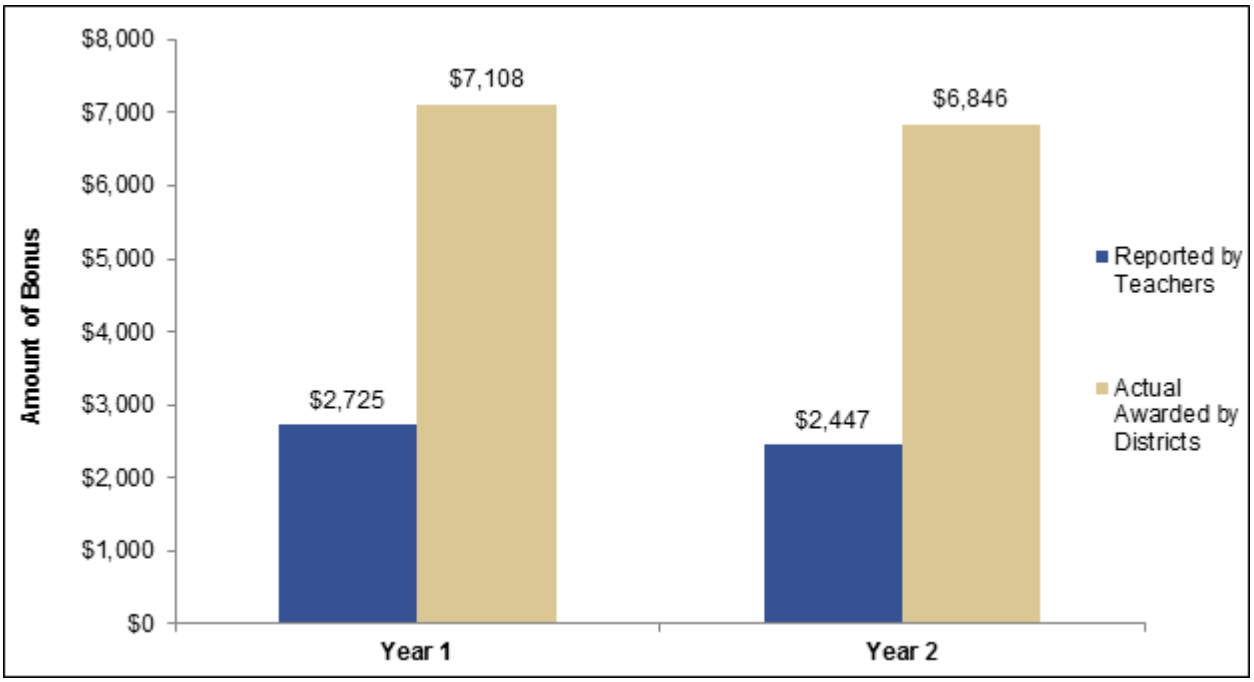
Notes: All principals who reported being eligible for pay-for-performance bonuses in Year 3 responded to the survey question about maximum possible bonus amount. Therefore, no multiple imputation was needed for principals' maximum possible pay-for-performance bonus amount in Year 3.

[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.14 and D.15 show the actual and reported maximum pay-for-performance bonuses for teachers and for principals for Years 1 and 2 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.14. Reported and Actual Maximum Pay-for-Performance Bonus for Teachers in Treatment Schools, Cohorts 1 and 2

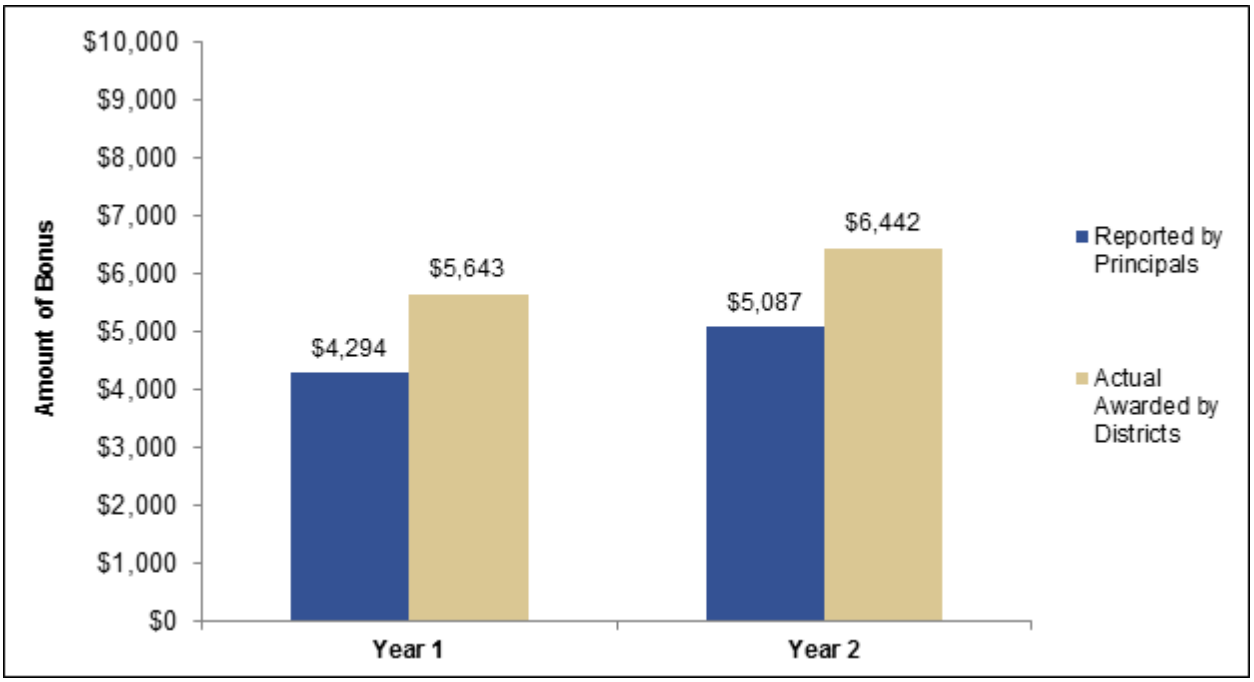


Source: Teacher survey (2012, 2013, and 2014) and educator administrative data.

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 264 treatment teachers in tested grades and subjects in Cohort 1 and 2 schools responded to this survey question in Year 1 and a total of 294 in Year 2. 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 29 additional responses for treatment teachers in Year 1 and 22 additional responses in Year 2.

Figure D.15. Reported and Actual Maximum Pay-for-Performance Bonus for Principals in Treatment Schools, Cohorts 1 and 2



Source: Principal survey (2012, 2013 and 2014) 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 72 treatment principals in Cohorts 1 and 2 responded to this survey question in Year 1 and a total of 77 in Year 2. 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 and 5 additional responses in Year 2.

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.17 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.[68] 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.

[68] 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.17. 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 1	Year 2	Year 3
Variation Across Districts	12	5	11	13	15	12
Variation Across Schools Within Districts	3	3	1	5	13	11
Variation Across Teachers Within Schools	86	91	88	82	72	78
Number of Teachers	**377**	**444**	**424**	**377**	**444**	**424**
Number of Schools	**66–77**	**66–444**	**66–424**	**66–377**	**66–444**	**66–424**

Source: Teacher survey (2012, 2013, and 2014).

Note: Percentages may not add up to 100 because of rounding.

Tables D.18 and D.19 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.

Table D.18. Treatment Teachers' Reported Eligibility for Pay-for-Performance Bonuses and Reported Maximum Bonuses in Year 3, 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)	**57**	**30**	**424**
District Communication Approach			
(1) Centralized—relied primarily on district staff	44	30	**240**
(2) Decentralized—relied primarily on school staff	59	22	**184**
Difference, (1) – (2)	-14	8	
District Assessment of Teachers' Understanding of TIF			
(1) Assessed understanding	52	18	**227**
(2) Did not assess understanding	54	30	**197**
Difference, (1) – (2)	-2	-12	
District Expectations of Teachers' Participation in Professional Development for Current Year			
(1) At least 75 percent of teachers will participate	52	22	**125**
(2) Fewer than 75 percent of teachers will participate	61	32	**299**
Difference, (1) – (2)	-9	-11	
Districts' Use of Classroom Achievement Growth to Determine Pay-for-Performance Bonuses			
(1) Used classroom achievement growth	49	35	**305**
(2) Did not use classroom achievement growth	76	41	**119**
Difference between (1) – (2)	-27	-6	
Districts' Average Prior Year Pay-for-Performance Bonuses			
(1) High—at least 4.5 percent of average salary	72	30	**145**
(2) Low—less than 4.5 percent of average salary	50	29	**279**
Difference, (1) – (2)	22*	1	
Districts' Prior Year Pay-for-Performance Bonus Distribution Method			
(1) Pay-for-performance bonus paid in separate check	64	34	**189**
(2) Pay-for-performance bonus paid in regular paycheck	59	31	**235**
Difference, (1) – (2)	5	4	
District Communication of Prior Year Actual Bonuses			
(1) Told all treatment teachers the total bonus amount that they earned (including $0 for nonrecipients)	54	22	**250**
(2) Did not tell all treatment teachers the total bonus amount that they earned	73	41	**174**
Difference, (1) – (2)	-19*	-19*	

Source: Teacher and district surveys (2014) and district interviews (2014).

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.19. Treatment Teachers' Reported Eligibility for Pay-for-Performance Bonuses and Reported Maximum Bonuses in Year 3, 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)	**57**	**30**	**424**
Subgroup Analysis by Principal Understanding			
Principal Understanding of Teachers' Eligibility			
(1) Principal correctly reported teachers' eligibility	55	37	**151**
(2) Principal incorrectly reported teachers' eligibility	53	28	**222**
Difference, (1) – (2)	2	10	
Subgroup Analyses by Teacher Characteristics			
Teacher Experience in Current School			
(1) More than one year in school	62	29	**344**
(2) First year in school	43	21	**78**
Difference, (1) – (2)	18	8	
Teaching Assignment			
(1) Tested grade and subject	64	32	**217**
(2) Nontested grade and subject	50	27	**207**
Difference, (1) – (2)	13	5	
Report About Receiving a Pay-for-Performance Bonus Based on Prior Year's Performance			
(1) Reported receiving a pay-for-performance bonus	91	48	**140**
(2) Reported not receiving a pay-for-performance bonus	43	23	**283**
Difference, (1) – (2)	48*	25*	
Actual Receipt of a Pay-for-Performance Bonus Based on Prior Year's Performance			
(1) Received a pay-for-performance bonus	66	37	**229**
(2) Did not receive a pay-for-performance bonus	46	22	**192**
Difference, (1) – (2)	20*	14*	
Participation in Professional Development About TIF Performance Measures			
(1) Teacher participated in professional development	52	31	**252**
(2) Teacher did not participate in professional development	50	27	**150**
Difference, (1) – (2)	2	5	
Mentoring Role			
(1) Teacher had a mentor teacher	52	27	**206**
(2) Teacher did not have a mentor teacher	55	30	**217**
Difference, (1) – (2)	-3	-3	
(1) Teacher mentored other teachers	62	36	**114**
(2) Teacher did not mentor other teachers	57	28	**309**
Difference, (1) – (2)	5	8	
(1) Teacher mentored other teachers as part of TIF	71	42	**56**
(2) Teacher did not mentor other teachers as part of TIF	56	28	**367**
Difference, (1) – (2)	15*	14	

Source: Teacher and principal surveys (2014) 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.20 and D.21 show that teachers in treatment and control schools reported similar professional development experiences. These tables also support the finding discussed in Chapter IV that more than half of teachers reported they received the professional development required under the TIF grant but indicated they received only a few hours.

Table D.20. Professional Development Teachers Reported Receiving or Expecting to Receive During Year 3, Cohort 1 (Percentages)

Professional Development Topics	Treatment	Control	Difference
Understanding Components of TIF	61	57	3
Understanding Performance Measures of TIF	67	62	5
Feedback Based on TIF Performance Ratings	59	57	3
Differentiated Instructional Strategies Based on Student Assessments	79	78	0
Instructional Techniques and Strategies	89	92	-4*
Aligning Curricula to State or District Standards	80	79	1
Number of Teachers—Range[a]	**408–412**	**442–448**	

Source: Teacher survey, 2014.

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 D.21. Hours of Expected Professional Development in Year 3, as Reported by Teachers, Cohort 1 (Averages)

	Expected Hours Among Teachers Who Expected to Receive Any Professional Development in the Specified Topic		
Professional Development Topics	Treatment	Control	Difference
Understanding Components of TIF	4	4	0
Understanding Performance Measures of TIF	3	3	0
Feedback Based on TIF Performance Ratings	3	3	0
Differentiated Instructional Strategies Based on Student Assessments	9	8	0
Instructional Techniques and Strategies	12	13	-1
Aligning Curricula to State or District Standards	9	9	0
Number of Teachers—Range[a]	**229–362**	**237–391**	

Source: Teacher survey, 2014.

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

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. At the time of this report, Cohort 1 had completed three years of implementation, 2011–2012, 2012–2013, and 2013-2014, referred to as Years 1, 2, and 3. Cohort 2 districts had completed only two years of implementation, 2012–2013 and 2013-2014, referred to as Years 1 and 2 for this cohort.

Tables E.1 through E.7 present impact estimates for the second year 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.

Year 2 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 three full years. Here, we show estimates for the second year of implementation (Year 2) for all study schools that have implemented the program, combining Cohorts 1 and 2. These tables also include the Year 2 estimates for Cohort 1 only for easy comparison with the Year 2 estimates based on both cohorts.

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 2 (Cohort 1)			Year 2 (Cohorts 1 and 2)		
Satisfaction Dimension	Treatment	Control	Impact	Treatment	Control	Impact
Opportunities for Pay and Development						
Opportunities for professional advancement	72	74	-3	72	73	-2
Opportunities to enhance skills	80	81	-1	80	80	0
Opportunities to earn extra pay	62	54	9*	65	59	6*
Evaluation System						
Use of student achievement scores to assess performance	60	69	-9*	55	65	-10*
School Environment						
Recognition of accomplishments	60	66	-6*	61	63	-2
Quality of interaction with colleagues	82	82	0	80	81	-1
Colleagues' efforts	84	83	0	82	83	-1
School morale	58	59	-1	54	56	-3
Job Satisfaction						
Overall job satisfaction	73	74	-1	69	71	-1
Number of Teachers—Range[a]	**444–447**	**446–449**		**581–585**	**567–571**	

Source: Teacher survey (2013 and 2014).

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.2. Principals' Satisfaction with Professional Opportunities, Evaluation System, and School Environment, Cohorts 1 and 2 (Percentages Who Are "Somewhat" or "Very" Satisfied)

	Year 2 (Cohort 1)			Year 2 (Cohorts 1 and 2)		
Satisfaction Dimension	Treatment	Control	Impact	Treatment	Control	Impact
Opportunities for Pay and Development						
Opportunities to enhance skills	87	85	2	88	83	4
Opportunities to earn extra pay	63	64	-1	57	59	-2
Evaluation System						
Feedback on my performance	67	80	-13	70	74	-5
School Environment						
Recognition of accomplishments	64	75	-12	60	69	-9
Quality of interaction with colleagues	86	90	-4	86	88	-2
Colleagues' efforts	90	85	5	88	87	1
School morale	75	82	-7	75	79	-4
Number of Principals—Range[a]	**63–64**	**61**		**82–83**	**77–78**	

Source: Principal survey (2013 and 2014).

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

Table E.3. Teachers' Attitudes Toward TIF Program, Cohorts 1 and 2 (Percentages Who "Agree" or "Strongly Agree")

	Year 2 (Cohort 1)			Year 2 (Cohorts 1 and 2)		
Statement	Treatment	Control	Impact	Treatment	Control	Impact
Teachers who do the same job should receive the same pay	61	66	-4	62	67	-4*
Standardized student test scores in my district measure what students have learned	34	41	-7*	30	35	-5
My principal is a good judge of teacher talent	74	74	0	72	73	-2
I am glad that I am participating in the TIF program	66	71	-5	65	71	-6
My job satisfaction has increased due to the TIF program	38	38	0	36	37	-1
I feel increased pressure to perform due to the TIF program	65	51	14*	61	47	14*
I have less freedom to teach the way I would like to teach due to the TIF program	40	30	10*	38	30	8*
The TIF program has harmed the collaborative nature of teaching	29	21	8*	32	22	10*
The TIF program has caused teachers to work more effectively	50	56	-6	48	52	-4
The TIF program is fair	54	59	-5	53	57	-4
The process used to determine how bonuses are determined was adequately explained to me	66	62	4	59	54	5
Number of Teachers—Range[a]	**397–440**	**383–442**		**484–573**	**472–560**	

Source: Teacher survey (2013 and 2014).

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 2 (Cohort 1)			Year 2 (Cohorts 1 and 2)		
Statement	Treatment	Control	Impact	Treatment	Control	Impact
The TIF program has been clearly communicated to me	93	97	-4	86	91	-5
This school has less chance of earning a bonus because of the characteristics of our student population	38	24	14	38	25	13
The evaluation system omits important aspects of school administration that should be considered	54	48	6	58	50	8
The TIF program contributes to greater collegiality and professionalism among the staff at this school	56	68	-12	54	64	-10
Teachers at this school are more comfortable with frequent formal observations of their teaching because of the TIF program	58	68	-10	57	62	-5
Parents and the school community believe the TIF program is important	50	43	7	45	35	10
The TIF program is likely to continue for the foreseeable future	71	73	-2	70	68	1
I played an important role in implementing the TIF program at my school	86	84	2	79	79	1
Number of Principals—Range[a]	**59–63**	**58–60**		**77–81**	**74–77**	

Source: Principal survey (2013 and 2014).

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

Additional Findings on Teachers' Attitudes by Subgroup

Tables E.5 through E.7 show supplementary analyses of teachers' satisfaction and attitudes in the third 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 2 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 3, Cohort 1 (Percentage Points)

	Impacts on Whether Teachers Were "Somewhat" or "Very" Satisfied with…										
Subgroup	Feedback on My Performance	Use of Student Achievement Scores to Assess Teacher Effectiveness	Opportunities for Professional Advancement	Opportunities to Enhance My Skills	Opportunities to Earn Extra Pay	Recognition of Accomplishments	Quality of Interactions with Colleagues	Colleagues' Efforts	School Morale	Overall Job Satisfaction	**Number of Teachers**
All Teachers (primary analysis)	-2	1	0	-3	11*	5	4*	0	9*	0	**881–888**
Teaching Assignment											
(1) Tested grades and subjects	-1	3	4	-1	11	8	5	1	10	5	**454–457**
(2) Nontested grades and subjects	-4	-2	-5	-5	11*	3	3	-1	8	-4	**427–432**
Difference between (1) - (2)	**3**	**6**	**8**	**5**	**0**	**4**	**2**	**2**	**2**	**9**	
Teacher Experience											
(1) Less than 5 years	-10	5	4	-6	13	4	8	1	3	-1	**222–223**
(2) 5 to 15 years	2	1	-1	-4	13*	5	6	3	12*	2	**436–440**
(3) Greater than 15 years	-3	-4	-4	4	4	8	-3	-5	11	0	**222–226**
Difference between (1) - (2)	**-12**	**4**	**5**	**-2**	**-1**	**-2**	**3**	**-2**	**-9**	**-3**	
Difference between (3) - (2)	**-4**	**-5**	**-4**	**8**	**-9**	**3**	**-8**	**-8**	**-1**	**-2**	

Source: Teacher survey, 2014.

[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 3, Cohort 1 (Percentages who "Agree" or "Strongly Agree")

	Actual Year 2 Bonus Receipt			Report of Year 2 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	78	72	6	75	73	2
Opportunities to enhance skills	78	75	3	80	74	5
Opportunities to earn extra pay	63	59	5	65	58	7
Evaluation System						
Use of student achievement scores to assess teacher effectiveness	73	65	7	79	64	14*
Feedback on teacher performance	82	76	6	81	74	7
School Environment						
Recognition of accomplishments	66	64	2	73	61	12
Quality of interaction with colleagues	85	84	0	86	83	4
Colleagues' efforts	85	82	3	85	82	3
School morale	55	60	-5	57	57	0
Job Satisfaction						
Overall job satisfaction	74	74	-1	82	71	10
Number of Teachers—Range[a]	**228–230**	**195–197**		**138–140**	**282–286**	

Source: Teacher survey (2014) and educator administrative data.

Notes: Pay-for-performance bonus receipt information comes from Year 2 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 3, Cohort 1 (Percentage Points)

	Impacts on Whether Teachers Responded They "Agreed" or "Strongly Agreed" with…											
Subgroup	Teachers Who Do the Same Job Should Receive the Same Pay	Standardized Student Test Scores in My District Measure What Students Have Learned	My Principal Is a Good Judge of Teacher Talent	I Am Glad I Am Participating in the TIF Program	My Job Satisfaction Has Increased due to the TIF Program	I Feel Increased Pressure to Perform due to the TIF Program	I Have Less Freedom to Teach the Way I Would Like to Teach 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	The TIF Program Is Fair	The Process Used to Determine How Bonuses Are Determined Was Adequately Explained to Me	**Number of Teachers**
All Teachers (primary analysis)	2	-5	2	1	6*	10*	2	1	5	-3	7*	**769–872**
Teaching Assignment												
(1) Tested grades and subjects	2	-1	3	1	7	9*	5	8	7	-7	6	**402–449**
(2) Nontested grades and subjects	1	-8	0	1	6	10*	-2	-5	2	1	8	**367–423**
Difference between (1) - (2)	**1**	**7**	**2**	**0**	**1**	**-1**	**8**	**13**	**4**	**-8**	**-3**	
Teacher Experience												
(1) Less than 5 years	6	-1	7	2	8	14*	10	10	8	-4	10	**168–213**
(2) 5 to 15 years	-3	-8	-3	3	2	11*	2	-2	6	1	10*	**394–435**
(3) Greater than 15 years	6	-2	5	-6	13*	5	-7	-2	0	-9	-1	**203–224**
Difference between (1) - (2)	**8**	**6**	**10**	**0**	**6**	**3**	**8**	**12**	**2**	**-5**	**1**	
Difference between (3) - (2)	**8**	**6**	**8**	**-8**	**11**	**-6**	**-10**	**0**	**-6**	**-10**	**-11**	

Source: Teacher survey, 2014.

[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 (over 95 percent) in both treatment and control schools reported having input in hiring decisions, 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.

Table E.8. Principals' Autonomy in Hiring Teachers, Cohort 1 (Percentages)

	Year 2			Year 3		
	Treatment	Control	Impact	Treatment	Control	Impact
Principal has complete autonomy over teacher hiring	22	15	7	20	13	7
Principal is part of a school-level team responsible for teacher hiring	47	57	-11	55	60	-5
Principal receives a set of prescreened candidates from the district office as the pool from which he or she can interview and hire	27	25	3	23	24	-2
Number of Principals	**64**	**61**		**59**	**62**	

Source: Principal survey (2013 and 2014).

Notes: None of the impacts are statistically significant at the .05 level, two-tailed test. The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding. None of the differences between Years 2 and 3 within treatment status 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.

We found no evidence that principals determine teacher assignments or compensations differently in response to pay-for-performance. In Year 3, pay-for-performance had no significant impact on any measure of principals' staffing decisions (Table E.9). Similar to prior years, treatment and control principals were equally likely to report that they use teacher's ability to produce high test scores when making decisions, suggesting that pay-for-performance is not inducing principals to make strategic assignments of teachers.

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 (46 percent of treatment principals and 35 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). However, control principals in Year 3 were more likely than treatment principals were to use one particular nonmonetary benefit. Control principals were more likely than treatment principals to give teachers priority in teaching assignments (18 versus 11 percent).

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 2			Year 3		
	Treatment	Control	Impact	Treatment	Control	Impact
The teacher's experience in a grade level or subject area	89	90	-1	88	82	6
The teacher's seniority	13	3	9*	15	13+	1
The teacher's content knowledge	92	93	-1	96	93	3
The teacher's ability to produce high test scores in grades/classes in which state or federal assessments are administered	64	66	-2	62	64	-2
The teacher's ability to work with certain student populations	84	81	3	92	85	6
To balance teacher experience and expertise in a grade level or subject	69	73	-4	72	75	-2
Number of Principals—Range[a]	**62–63**	**58–59**		**58–59**	**59–61**	

Source: Principal survey (2013 and 2014).

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 E.10. Nonmonetary Benefits Used to Recognize Teachers' Performance or Responsibilities, Cohort 1 (Percentages)

	Year 2			Year 3		
	Treatment	Control	Impact	Treatment	Control	Impact
Use of nonmonetary benefits	40	37	3	46	35	11
Type of nonmonetary benefits:						
Release from classroom teaching for mentoring or other leadership activities	28	32	-3	33	32	2
Decision-making authority on issues such as hiring staff or adopting curriculum	32	30	2	36	30	6
Priority in teaching assignments	9	18	-9	11	18	-8*
Priority in student assignments	3	7	-4	8	5	3
Number of Principals—Range[a]	**63–64**	**60**		**59**	**60**	

Source: Principal survey (2013 and 2014).

Notes: The difference between the treatment and control estimates may not equal the impact shown in the table because of rounding. None of the differences between Years 2 and 3 within treatment status 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.

*Impact 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, teachers in Year 3 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.

Table E.11. Teachers' Time Spent on School-Related Activities in the Most Recent Full Week (Average Hours)

	Year 2			Year 3		
	Treatment	Control	Impact	Treatment	Control	Impact
Time Spent During School Hours on						
Teaching students in the classroom, small groups, or individually	28	28	0	28	29	0
Supervising students in other activities	4	4	0	3	3	0
Preparation on your own (e.g., lessons, grading, assignment)	8	7	1	8	8	0
Preparation and professional development with colleagues (e.g., common lesson planning, workshops, staff meetings, mentoring)	4	4	0	4	4	-1
Other activities	2	2	0	2	2	0
Total hours during school hours (calculated)	45	44	0	45	46	0
Time Spent During Nonschool Hours on						
Academic-related activities with students	3	4	-1*	2	3+	0
Other activities with students	1	1	0	1	1	0
Preparation on your own	10	8	2*	9+	9	0
Preparation and professional development with colleagues	3	3	0	3	3	0
Other school-related activities	1	1	0	1	1	0
Total hours during nonschool hours (calculated)	18	17	1	16	16	-1
Number of Teachers—Range[a]	**434–447**	**443–448**		**374–430**	**398–460**	

Source: Teacher survey (2013 and 2014).

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.

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 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 three years of implementation during the period covered by this report. Year 1 represents the first year of implementation (2011–2012), Year 2 the second (2012–2013), and Year 3 the third year of implementation (2013–2014). Cohort 2 districts completed only two years of implementation, 2012–2013 and 2013–2014, referred to as Years 1 and 2 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 quality 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
Outcomes Examined in Table VI.1								
School Achievement Growth Ratings								
Year 1	44[a]	44[a]	41	41	62	62	62	62
Year 2	48	48	45	44	66	65	66	65
Year 3	48	48	45	45	66	66	66	66
Classroom Achievement Growth Ratings								
Year 1	23[b]	23[b]	21	21	37	36	37	36
Year 2	23[b]	23[b]	21	21	37	36	37	36
Year 3	33[c]	33[c]	30	30	46	45	46	45
Outcomes Examined in Table VI.2								
Teachers' Classroom Observation Ratings								
Year 1	48	48	45	45	66	66	66	66
Year 2	48	48	45	45	66	66	66	66
Year 3	48	48	45	45	66	66	66	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	61	58
Outcomes Examined in Table VI.4								
Student Math Achievement								
Year 1	48	48	45	45	66	66	66	66
Year 2	48	48	45	45	66	66	66	66
Year 3	48	48	45	45	66	66	66	66
Student Reading Achievement								
Year 1	48	48	45	45	66	66	66	66
Year 2	48	48	45	45	66	66	66	66
Year 3	48	48	45	45	66	66	66	66

Source: 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 Year 3. 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 quality 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 findings on student achievement captured the cumulative impacts of pay-for-performance on schools' average student achievement after three years. In Chapter II, 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.

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, for the teacher observation ratings, teachers in large schools received less weight than those in small schools. We explored two 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 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.

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
Outcomes Examined in Table VI.1						
School Achievement Growth Ratings						
Year 1	2.60	1.00	2.25	0.99	0.34*	0.04
Year 2	2.55	1.05	2.27	0.95	0.27	0.07
Year 3	2.41	1.03	2.37	0.93	0.04	0.74
Classroom Achievement Growth Ratings						
Year 1	2.26	0.96	2.08	0.95	0.18*	0.03
Year 2	2.22	1.01	2.17	1.04	0.05	0.38
Year 3	2.54	1.11	2.53	1.13	0.01	0.81
Outcomes Examined in Table VI.2						
Teachers' Classroom Observation Ratings						
Year 1	2.94	0.51	2.91	0.55	0.03	0.24
Year 2	2.98	0.48	2.93	0.53	0.04	0.09
Year 3	2.96	0.71	2.91	0.69	0.04	0.07
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	0.49
Outcomes Examined in Table VI.4						
Student Math Achievement (student *z*-score units)						
Year 1	-0.43	0.93	-0.45	0.93	0.02	0.36
Year 2	-0.39	0.92	-0.43	0.92	0.04	0.08
Year 3	-0.37	0.94	-0.42	0.93	0.05*	0.02
Student Reading Achievement (student *z*-score units)						
Year 1	-0.37	0.95	-0.40	0.96	0.03*	0.05
Year 2	-0.36	0.95	-0.39	0.95	0.03*	0.04
Year 3	-0.33	0.95	-0.37	0.95	0.04*	0.02

Source: Educator and student administrative data.

Note: 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.

For school achievement growth ratings, neither the main model nor the model that gave districts equal weight found significant impacts of pay-for-performance (Table F.3, model 1). For teacher observation ratings, estimates from the models using alternative weighting approaches were similar to those from the main model (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. As expected, excluding covariates reduced the precision of the estimates, resulting in *p*-values slightly greater than the main model. For both school achievement growth and teacher observation ratings, neither the main model nor this specification found significant impacts of pay-for-performance in Year 3 (Tables F.3 and F.4).

Table F.3. Impacts of Pay-for-Performance on School Achievement Growth Ratings in Year 3 Using Alternative Specifications, Cohort 1 (Points on 1-to-4 Scale)

Model	Treatment Schools	Control Schools	Impact	*p*-value	**Number of Schools**
Main Model	2.41	2.37	0.04	0.74	**132**
Alternative Specifications					
Weights					
(1) Districts are weighted equally	2.35	2.39	-0.04	0.76	**132**
Covariates					
(2) No covariates except randomization block indicators	2.41	2.37	0.04	0.77	**132**

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.4. Impacts of Pay-for-Performance on Teachers' Classroom Observation Ratings in Year 3 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	2.96	2.91	0.04	0.07	**3,642**	**132**
Alternative Specifications						
Weights						
(1) Teachers are weighted equally	2.95	2.91	0.05	0.07	**3,642**	**132**
(2) Districts are weighted equally	2.82	2.78	0.04	0.15	**3,642**	**132**
Covariates						
(3) No covariates except randomization block indicators	2.94	2.91	0.03	0.25	**3,642**	**132**

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.

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 two 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 only Cohort 1, the estimated impacts of pay-for-performance on school achievement growth ratings and classroom achievement growth ratings in Year 1 were no longer found to be significant (*p*-values = 0.13 and 0.08) when both cohorts were included in the analysis (Table F.5). In Year 2, impacts on school achievement growth ratings and classroom achievement growth ratings were not significant when based on Cohort 1 only or when based on both cohorts. Likewise, estimated impacts on observation ratings in Cohorts 1 and 2 were similar to those in Cohort 1 only and not significant (Table F.6).

Table F.5. Student Achievement Growth Ratings in Years 1 and 2, 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	2.60	2.25	0.34*	0.04	NA	**124**
Classroom Achievement Growth Ratings	2.26	2.08	0.18*	0.03	**1,092**	**73**
Year 1, Cohorts 1 and 2						
School Achievement Growth Ratings	2.44	2.24	0.20	0.13	NA	**165**
Classroom Achievement Growth Ratings	2.37	2.27	0.10	0.08	**2,270**	**110**
Year 2, Cohort 1						
School Achievement Growth Ratings	2.55	2.27	0.27	0.07	NA	**131**
Classroom Achievement Growth Ratings	2.22	2.17	0.05	0.38	**1,339**	**73**
Year 2, Cohorts 1 and 2						
School Achievement Growth Ratings	2.75	2.60	0.15	0.22	NA	**172**
Classroom Achievement Growth Ratings	2.24	2.30	-0.06	0.23	**2,626**	**112**

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 only available for the six districts in Cohort 1 and three districts in Cohort 2 that evaluated teachers based on classroom achievement growth in Years 1 and 2. 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.

NA is not applicable.

Table F.6. Observation Ratings for Teachers and Principals in Years 1 and 2, 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	0.24	**3,622**	**132**
Observation Ratings for Principals	3.08	3.18	-0.10	0.20	**105**	**105**
Year 1, Cohorts 1 and 2						
Teachers' Classroom Observation Ratings	2.99	2.97	0.02	0.43	**4,960**	**173**
Observation Ratings for Principals	3.05	3.13	-0.09	0.23	**150**	**150**
Year 2, Cohort 1						
Teachers' Classroom Observation Ratings	2.98	2.93	0.04	0.09	**3,612**	**132**
Observation Ratings for Principals	3.14	3.01	0.13	0.19	**118**	**117**
Year 2, Cohorts 1 and 2						
Teachers' Classroom Observation Ratings	3.08	3.08	0.00	0.93	**4,990**	**173**
Observation Ratings for Principals	3.29	3.19	0.11	0.17	**156**	**155**

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

Findings for a Consistent Sample of Schools

The number of districts and schools included in the main impact estimates varies 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 that did not place school achievement growth into performance categories or onto a numeric scale in that year, but they included this district in the other years. 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 that evaluated teachers on this measure in Year 3. Thus, differences in the main impact estimates across years could reflect both changes in the impacts over time and changes in the composition of districts and schools included in the analyses.

In Table F.7, we present the impact of pay-for-performance on educator performance ratings based on a consistent sample of districts and schools across years. In this table, differences in the impact estimates across years only reflect changes in impacts over time. Findings based on a consistent sample of schools were similar to the main impact findings, suggesting that the decline in these impacts over time are not driven by changes in the samples of districts and schools included in the analyses.

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[a]						
Ratings in Year 1	2.60	2.25	0.34*	0.04	**NA**	**124**
Ratings in Year 2	2.51	2.26	0.24	0.08	**NA**	**124**
Ratings in Year 3	2.42	2.39	0.04	0.78	**NA**	**124**
Classroom Achievement Growth[b]						
Ratings in Year 1	2.26	2.08	0.18*	0.03	**1,092**	**73**
Ratings in Year 2	2.22	2.17	0.05	0.38	**1,339**	**73**
Ratings in Year 3	2.38	2.36	0.02	0.72	**1,601**	**73**

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 three years, these analyses exclude the district in which school achievement growth ratings did not place educators into performance categories or onto a numeric scale in Year 1.

[b]For all three years, these analyses include only the six districts that evaluated teachers based on classroom achievement growth in Years 1 and 2.

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

NA is not applicable.

Overall Educator Retention Rates

Overall retention rates—that is, 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 differences in impacts across these two groups to shape the overall effectiveness of the school's staff. For example, if a large school had only one teacher depart each year, then the effectiveness of the departing teacher's replacement (even if vastly different from the effectiveness of the returning teachers) would have little influence on overall effectiveness at the school.

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 25 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, leaving the potential for differences in impacts across returning and newly hired educators an important way of shaping overall educator effectiveness.

We 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 Year 1, 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. 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	0.13	**4,333**	**132**
Between Years 2 and 3	77	77	0	0.74	**4,433**	**132**
Between Years 3 and 4	73	70	3	0.07	**4,545**	**132**
Two-Year Period						
Between Years 1 and 3	66	64	2	0.16	**4,333**	**132**
Between Years 2 and 4	60	57	3	0.06	**4,433**	**132**
Three-Year Period						
Between Years 1 and 4	51	49	3*	0.04	**4,333**	**132**

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.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	**134**	**128**
Between Years 2 and 3	78	79	-2	0.84	**138**	**129**
Between Years 3 and 4	71	75	-4	0.49	**134**	**128**
Two-Year Period						
Between Years 1 and 3	64	59	5	0.63	**134**	**128**
Between Years 2 and 4	57	58	-1	0.92	**138**	**129**
Three-Year Period						
Between Years 1 and 4	41	44	-3	0.80	**134**	**128**

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.

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 3, educators working in treatment and control schools had similar demographic characteristics and professional background, with one exception: principals in treatment schools were more likely to be white than those in control schools (Table F.10).

Table F.10. Characteristics of Teachers and Principals in Year 3, Cohort 1 (Percentages Unless Otherwise Noted)

	Teachers			Principals		
	Treatment	Control	Difference	Treatment	Control	Difference
Demographic Characteristics						
Female	86	85	1	63	66	-3
Race/ethnicity						
White, non-Hispanic	74	72	2	65	50	15*
Black, non-Hispanic	20	21	-1	29	39	-10
Hispanic or Other	7	7	-1	6	11	-5
Age (average years)	42	42	0	48	49	-1
Education						
Master's degree or higher	47	49	-2	95	94	1
Experience in K–12 Education						
Total experience (average years)	11	11	0	18	16	2
Less than 5 years	27	30	-3	9	17	-8
5–15 years	45	44	1	39	41	-2
More than 15 years	28	26	1	53	42	10
Number of Educators—Range[a]	**1,704–2,200**	**1,741–2,236**		**48–65**	**43–68**	
Number of Schools—Range[a]	**53–66**	**53–66**		**47–63**	**42–64**	

Source: Educator administrative data.

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

*Difference between treatment and control educators 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 differences in performance ratings between teachers who stayed in treatment schools and those who stayed in control schools and between newly hired teachers in those two groups of schools (see Table VI.3). In those main analyses, 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. Findings were similar in analyses that classified returning teachers as those who had stayed in their school since Year 1 and newly hired teachers as those who were new to their school since Year 1 (Table F.11).

Table F.11. Performance Ratings in Year 3 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
Classroom Observation Rating	0.07*	0.03	-0.01	0.71	2,307	1,335	132
Classroom Achievement Growth Rating	0.05	0.41	-0.05	0.51	1,272	777	91

Source: Educator administrative data.

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

We also examined differences in impacts of pay-for-performance on the performance ratings for returning and newly hired principals. We found no impacts of pay-for-performance on observation ratings or school achievement growth ratings of returning principals, when they were defined as either those who stayed in their school since the previous year (Table F.12) or since Year 1 (Table F.13). Similarly, we found no impacts of pay-for-performance on the performance ratings of newly hired principals, when they were defined as those who were new to their school since the previous year (Table F.12) or since 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	0.44	0.45	110	27	128
Year 3							
Observation Ratings	0.06	0.49	-0.05	0.81	96	25	119
School Achievement Growth Ratings	0.15	0.35	-0.50	0.21	108	26	128

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

Table F.13. Performance Ratings in Year 3 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
Classroom Observation Rating	0.05	0.66	0.04	0.75	76	45	119
Classroom Achievement Growth Rating	0.17	0.39	-0.20	0.39	83	51	128

Source: Educator administrative data.

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

Student Achievement

This section presents three 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, and (3) subgroup analyses that assess impacts within elementary and middle grades separately.

Sensitivity Analyses

We explored the sensitivity of the main impact estimates to several changes to 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.

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. 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 to the main impact estimates (Tables F.14 and F.15, models 2 and 3), with one exception. The positive impact of pay-for-performance on math

achievement in Year 3 was not significant when we gave each district the same weight (Table F.14, model 3).

Table F.14. Impacts of Pay-for-Performance on Student Achievement in Math Using Alternative Specifications in Year 3, Cohort 1 (Student *z*-Score Units)

	Impact	*p*-value	Number of Students	Number of Schools
Main Model	0.05*	0.02	40,037	132
Alternative Specifications				
Standardizing Test Scores				
(1) Compute *z*-scores using sample means/standard deviations	0.06*	0.02	40,037	132
Weights				
(2) Students weighted equally	0.05*	0.02	40,037	132
(3) Districts weighted equally	0.05	0.08	40,037	132
Covariates				
(4) No covariates except randomization block indicators	0.02	0.44	40,037	132
(5) Only covariates are school-level pre-implementation means of student achievement and student race/ethnicity and randomization block indicators	0.06*	0.01	40,037	132
(6) All covariates interacted with state indicators	0.06*	0.04	40,037	132
(7) Include student pretests interacted with grade indicators	0.05*	0.02	40,037	132
(8) Include student pretests, squared and cubed	0.06*	0.02	40,037	132
(9) Include baseline teacher characteristics	0.05*	0.02	40,037	132

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. For math, the main model found statistically significant impacts in Year 3, but the impacts from the alternative model were smaller (0.02 versus 0.05) and insignificant. For reading, the main model also found statistically significant impacts in Year 3, whereas estimates from the alternative model were smaller (0.01 versus 0.04) and insignificant.

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. This model produced impact estimates that were similar in magnitude and precision to those produced by the main 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).

Controlling for baseline teacher characteristics. As discussed in Chapter II, there were some differences in teacher characteristics between treatment and control schools, though none of these differences were statistically significant. Our main analyses do not control for baseline teacher characteristics, so we explore the sensitivity of our results to their inclusion. In both math and reading, estimates are unaffected by the inclusion of baseline teacher characteristics (Table F.14 and F.15, model 9).

Table F.15. Impacts of Pay-for-Performance on Student Achievement in Reading Using Alternative Specifications in Year 3, Cohort 1 (Student *z*-Score Units)

	Impact	*p*-value	Number of Students	Number of Schools
Main Model	0.04*	0.02	39,807	132
Alternative Specifications				
Standardizing Test Scores				
(1) Compute *z*-scores using sample means/standard deviations	0.04*	0.03	39,807	132
Weights				
(2) Students weighted equally	0.04*	0.01	39,807	132
(3) Districts weighted equally	0.04*	0.05	39,807	132
Covariates				
(4) No covariates except randomization block indicators	0.01	0.63	39,807	132
(5) Only covariates are school-level pre-implementation means of student achievement and student race/ethnicity and randomization block indicators	0.04*	0.01	39,807	132
(6) All covariates interacted with state indicators	0.04*	0.01	39,807	132
(7) Include student pretests interacted with grade indicators	0.04*	0.02	39,807	132
(8) Include student pretests, squared and cubed	0.04*	0.01	39,807	132
(9) Include baseline teacher characteristics	0.04*	0.02	39,807	132

Source: Student administrative data.

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

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 and 2 for Cohorts 1 and 2, as well as the main impact estimates from Chapter VI, which only included Cohort 1 schools. When Cohort 2 schools were included in the analysis, pay-for-performance no longer had a significant impact on achievement in reading in Year 1. Impacts for math were not significant in Year 1 in either the sample that only included Cohort 1 or the sample that included both cohorts. Results in Year 2 that include Cohort 1 and 2 schools are similar to those that only include Cohort 1 schools. In both analyses, the impact on reading is positive and statistically significant, though the effect is smaller with Cohort 1 and 2 schools (0.02 versus 0.03). As in Year 1, impacts for math in Year 2 were not significant with either the Cohort 1 sample or the Cohort 1 and 2 sample.

Table F.16. Student Achievement in Math and Reading, Cohorts 1 and 2 (Student *z*-Score Units)

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	0.36	**40,847**	**132**
Reading	-0.37	-0.40	0.03*	0.05	**40,571**	**132**
Year 1, Cohorts 1 and 2						
Math	-0.54	-0.55	0.01	0.46	**54,027**	**173**
Reading	-0.48	-0.50	0.02	0.20	**53,547**	**173**
Year 2, Cohort 1						
Math	-0.39	-0.43	0.04	0.08	**40,708**	**132**
Reading	-0.36	-0.39	0.03*	0.04	**40,390**	**132**
Year 2, Cohorts 1 and 2						
Math	-0.50	-0.53	0.03	0.06	**52,880**	**173**
Reading	-0.46	-0.49	0.02*	0.05	**52,679**	**173**

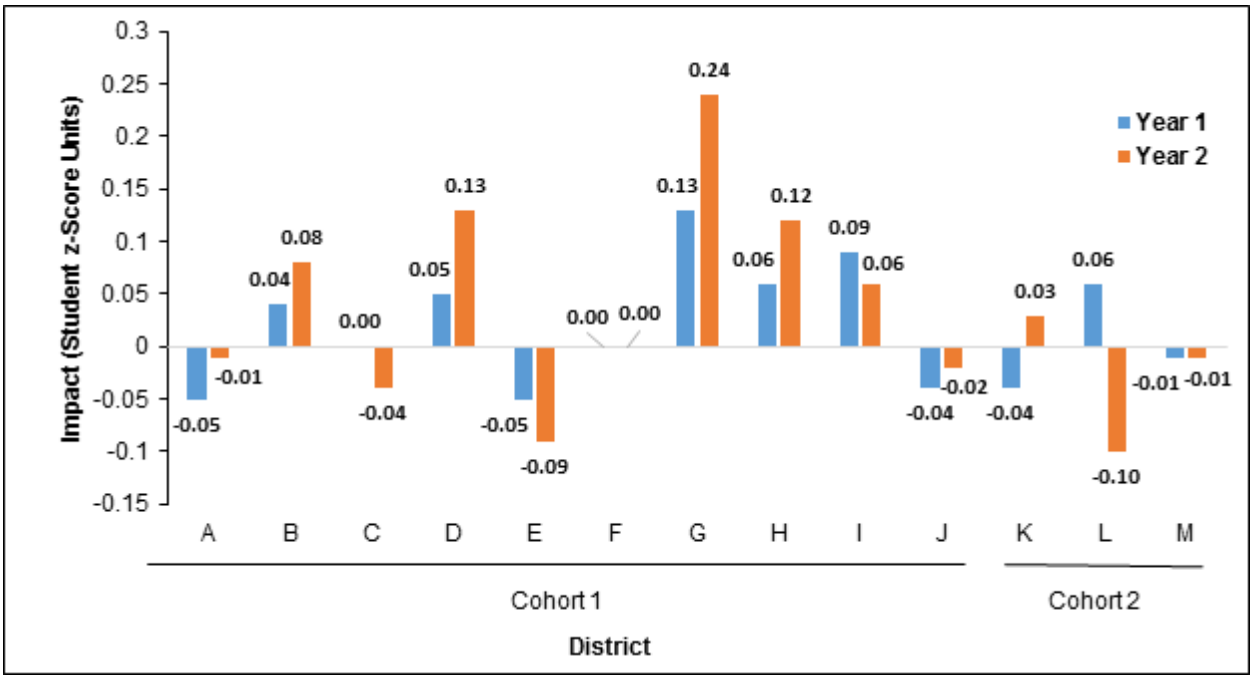
Source: Student administrative data.

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 district-level math and reading impacts in Years 1 and 2 across all 13 districts in Cohorts 1 and 2. Similar to the Year 3 findings for Cohort 1 (Figures VI.2 and VI.3), these figures illustrate that impacts in Years 1 and 2 also varied across all 13 districts.

Figure F.1. Impact of Pay-for-Performance on Student Math Achievement After Years 1 and 2, by District, Cohorts 1 and 2 (Student *z*-Score Units)

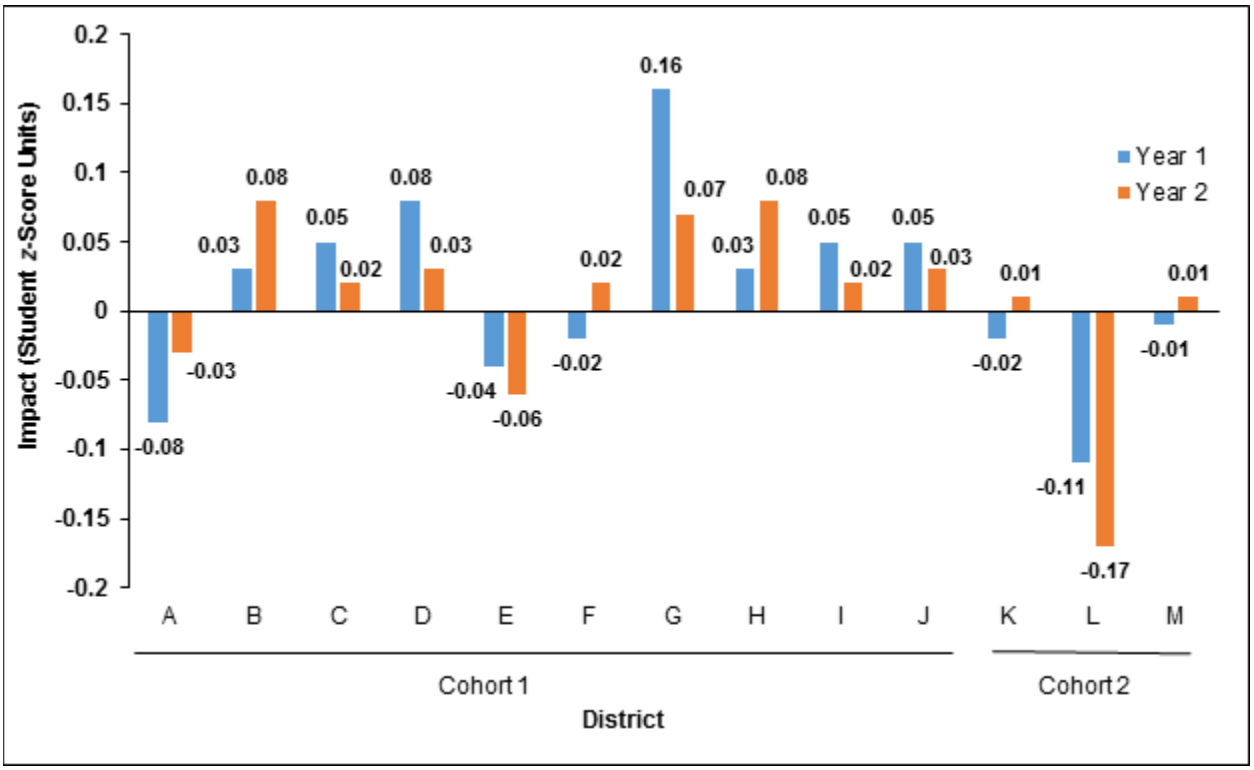


Source: Student administrative data (N = 54,027 in Year 1 and 52,880 in Year 2).

Note: An F-test of the null hypothesis that impacts are equal across districts has a *p*-value of less than 0.01 in Year 1 and Year 2.

Figure reads: In District A, pay-for-performance lowered math achievement by 0.05 standard deviations after Year 1 and by 0.01 standard deviations after Year 2.

Figure F.2. Impact of Pay-for-Performance on Student Reading Achievement After Years 1 and 2, by District, Cohorts 1 and 2 (Student *z*-Score Units)



Source: Student administrative data (N = 53,547 in Year 1 and 52,679 in Year 2).

Note: An F-test of the null hypothesis that impacts are equal across districts has a *p*-value of less than 0.01 in Year 1 and Year 2.

Figure reads: In District A, pay-for-performance lowered reading achievement by 0.08 standard deviations after Year 1 and by 0.03 standard deviations after Year 2.

Subgroup Findings

In Table F.17, we present the impacts of pay-for-performance on student achievement separately within elementary grades (grades 3 through 5) and middle grades (grades 6 through 8). For math achievement, impacts in middle and elementary schools in Years 1 and 2 are statistically insignificant, as in the main specification that pools these groups. In Year 3 the impact on math achievement is larger in middle school grades than in elementary grades (0.08 versus 0.04), but the difference between them is not statistically significant. For reading achievement, effects are larger in middle school grades than in elementary school grades in Year 1 (0.06 versus 0.02), similar in Year 2 (0.03 and 0.02, both insignificant), and larger in middle school grades than elementary grades in Year 3 (0.06 versus 0.03). However, in each of Years 1 through 3 the difference between reading impacts in middle school grades and elementary school grades is statistically insignificant.

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	0.91	-0.40	-0.42	0.02	0.30
(2) Grades 6–8	-0.40	-0.45	0.05	0.10	-0.32	-0.37	0.06*	0.02
Difference between (1) and (2)			-0.04	0.12			-0.04	0.17
Number of Students	**20,525**	**20,322**			**20,343**	**20,228**		
Number of Schools	**66**	**66**			**66**	**66**		
Year 2								
Grades 3–5	-0.41	-0.45	0.04	0.16	-0.38	-0.42	0.03	0.06
Grades 6–8	-0.34	-0.39	0.05	0.11	-0.32	-0.34	0.02	0.33
Difference between (1) and (2)			-0.01	0.73			0.01	0.76
Number of Students	**20,251**	**20,457**			**20,031**	**20,359**		
Number of Schools	**66**	**66**			**66**	**66**		
Year 3								
Grades 3–5	-0.39	-0.43	0.04	0.13	-0.37	-0.39	0.03	0.23
Grades 6–8	-0.32	-0.40	0.08*	0.03	-0.26	-0.32	0.06*	0.01
Difference between (1) and (2)			-0.03	0.40			-0.04	0.27
Number of Students	**20,026**	**20,011**			**19,880**	**19,927**		
Number of Schools	**66**	**66**			**66**	**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.

Reconciling Impacts on Growth Ratings and Student Achievement

Impacts of pay-for-performance on student achievement are linked to impacts on the rate at which student achievement grows. For example, if pay-for-performance causes a widening gap in achievement between the same group of treatment and control students over time, pay-for-performance bonuses must at some point have caused an increase in student growth. The measures of student growth used to evaluate teachers are, however, not directly comparable to impacts on student achievement for several reasons. First, grade spans included may differ between the two because typically school- or classroom-growth measures are only estimable for grades that have a pretest. Third grade classrooms thus tend to be excluded from growth measures unless the school uses a second grade exam. Second, not all schools use a school- or classroom-growth measure, so the samples used to estimate impacts on growth ratings are by necessity smaller than those used to

estimate student achievement impacts. Third, our analyses of impacts on growth ratings employ a four-point scale to describe growth, which is a different scale from the student *z*-score units employed in our analysis of student achievement impacts. Fourth, 9 of the 10 Cohort 1 districts combined math and reading achievement into one school achievement growth rating, whereas our analyses examined the impact of pay-for-performance separately on math and reading achievement.

To allow for a direct comparison of impacts on growth and impacts on student achievement, we developed our own measure of growth, which (1) is restricted to grades 4 through 8, (2) is restricted to districts using school-growth as a measure of teacher performance, and (3) uses student *z*-scores. Our growth model uses test scores as the outcome, where math and reading are pooled together, regressed on an indicator for treatment status. We also include the primary covariates used to estimate student achievement impacts with two changes: prior year test scores replace pre-implementation test scores, and these covariates are interacted with a subject indicator. When district measures of school achievement growth are converted to student *z*-score units and schools with all grades below fourth are excluded, the results are similar to the impacts on our constructed growth measure (Table F.18).

Table F.18. Impacts of Pay-for-Performance on District-Constructed and Study-Constructed Measures of School Achievement Growth, Cohort 1

	Impacts of Pay-for-Performance on		
Year	School Achievement Growth Ratings from Districts' Measures (Points on 1-to-4 Scale)	School Achievement Growth Ratings from Districts' Measures (Student *z*-Score Units)	School Achievement Growth from Study-Constructed Measure (Student *z*-Score Units)
Year 1	0.34*	0.04*	0.03
Year 2	0.28	0.03	0.03*
Year 3	0.05	0.00	0.01
Number of Schools			
Year 1[a]	122	122	122
Year 2	129	129	129
Year 3	130	130	130

Source: Educator and student administrative data.

Note: Analyses are based on schools that contained any of the grades from 4 to 8—the grades in which the study could measure students' achievement growth from the previous year.

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

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

THIS PAGE IS INTENTIONALLY BLANK

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 in Year 3. We examine whether the characteristics of districts' TIF programs and their implementation were associated with impacts on student achievement. We also examine whether schools that had greater impacts of pay-for-performance on measures of educators' strategic behavior, effort, and practices also had greater impacts on student achievement. Such a relationship would suggest that pay-for-performance improved student achievement by way of affecting those educator behaviors.

In this appendix, we provide (1) the rationale for choosing the district characteristics we examined and information on how we characterized districts into subgroups based on their characteristics, (2) findings on the association between district characteristics and impacts on student achievement, (3) information on measures of educators' strategic behavior, effort, and practices, and (4) findings on the relationship between impacts on educator behaviors and impacts 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 three years of implementation during the period covered by this report—2011–2012, 2012–2013, and 2013-2014—referred to as Years 1, 2, and 3, respectively.

Explaining Differences in Impacts Across Districts

This section discusses the relationships between districts' program and implementation characteristics and impacts on student achievement. First, we provide details on each characteristic and the way in which we categorized districts into two subgroups that differed on the characteristic. Second, we compare the impacts of pay-for-performance on student achievement between these subgroups. Third, for characteristics that could be measured on a continuous scale, we report findings from a sensitivity analysis that examined the relationships between the continuous measures of those characteristics and impacts on student achievement.

Program Characteristics Examined

We examined six district-level program and implementation characteristics (Table G.1). Four of these characteristics—the use of classroom achievement growth to measure teacher effectiveness and award bonuses, the size of the average bonus, the amount of differentiation in bonuses, and the degree to which earning a bonus was challenging—pertain to how the programs were designed. Two of these 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.

For each characteristic, we identified a subgroup of districts that had higher levels of the characteristic (or, in the case of classroom achievement growth 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 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.

We used data from the teacher survey, district interview, and administrative data to categorize districts into subgroups. Since teachers in Year 3 (2013-2014) would be responding to information about actual bonus awards from the prior year bonus payouts, we used bonus data from Year 2 (2012-2013) to place districts into subgroups for characteristics related to bonuses. For the remaining characteristics, we used information from the 2014 district interview and the 2014 teacher survey.

Use of classroom achievement growth. Districts' use of classroom achievement growth to measure teacher effectiveness and award bonuses shaped the degree to which teachers' bonuses were determined by their own performance or that of a larger group of teachers. In the seven districts that used these measures in Year 3, teachers who were evaluated on classroom achievement growth earned, on average, most of their total bonus (77 percent) based on measures of individual performance—specifically, classroom observations and classroom achievement growth (Chapter IV, Figure IV.6). In the remaining three districts, teachers earned, on average, most of their total bonus (59 percent) based on group performance measures, including school achievement growth and the achievement growth of student subgroups defined by grades or subject areas (Chapter IV, Figure IV.7).

Table G.1. Program and Implementation Characteristics Used for Subgroup Analyses

Characteristic	Reason for Examining This Characteristic	Subgroup Definition	Number of Districts in Subgroup
Use of Classroom Achievement Growth to Measure Teacher Effectiveness and Award Bonuses[a]	Measure increases emphasis on individual over group performance, which may enhance teachers' control over their own ratings but discourage collaboration.	Districts that 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 large average bonus if the average bonus in Year 2 was at least 5 percent of average teacher salary.	3
Amount of Differentiation in Bonuses[b]	More differentiation implies a larger monetary gain from performing well on the performance ratings.	Districts had a large amount of differentiation if the standard deviation of bonuses in Year 2 was at least 4 percent of average teacher salary.	4
Degree to Which Earning a Bonus was Challenging[b]	If nearly everyone receives a bonus, teachers may perceive less monetary incentive to improve.	Districts awarded bonuses that were challenging to earn if fewer than 50 percent of teachers in Year 2 received a bonus.	3
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 early awarding of bonuses from Year 2 if they awarded at least one component of the bonuses no later than the August after Year 2.	3
Teachers' Understanding of Their Pay-for-Performance Eligibility[c]	Understanding of eligibility is necessary for bonuses to affect behavior.	Districts had high 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 3.	4

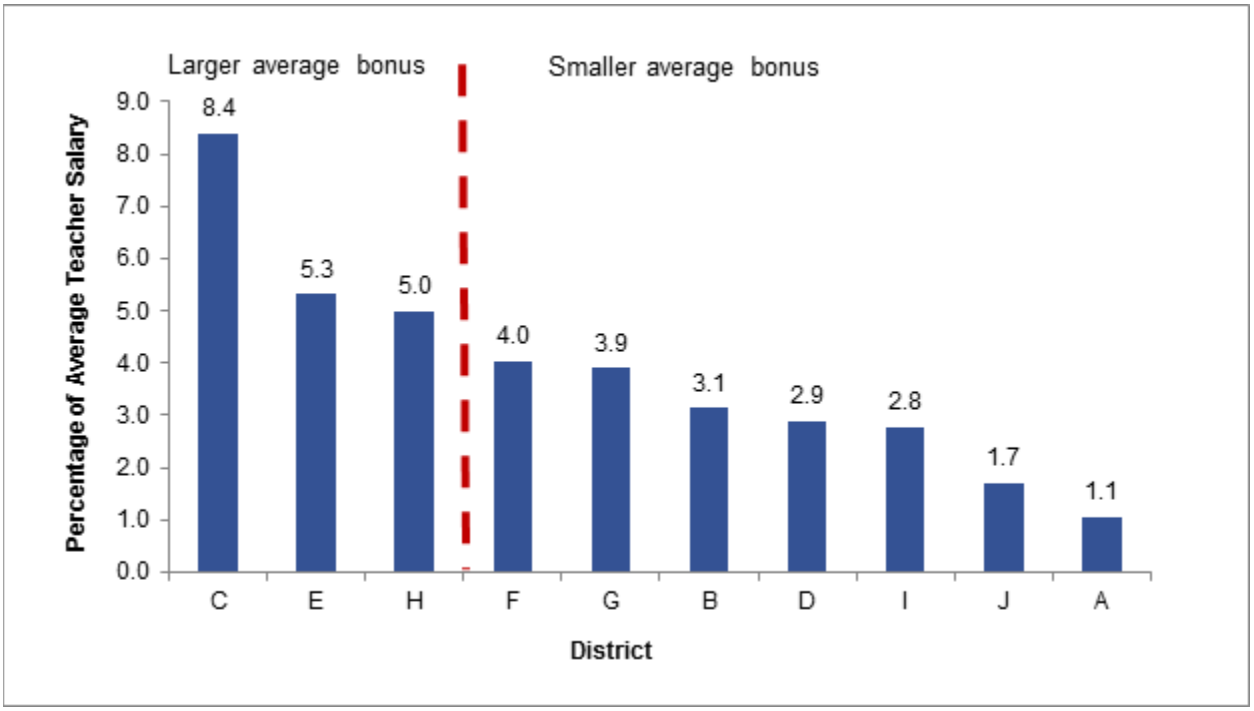
[a]Based on district interviews, 2014.
[b]Based on educator administrative data from Year 2.
[c]Based on teacher survey, 2014.

An emphasis on individual, rather than group, incentives could have a positive, negative, or no association with impacts on student achievement. On the one hand, teachers might be more motivated to respond to individual incentives because they have more control over their own performance than that of their colleagues. On the other hand, individual incentives may involve comparing teachers with each other and could harm teacher collaboration.

Size of average bonus. Teachers may pay more attention to bonuses that are larger on average. For example, bonuses may be more likely to be a topic of discussion among teachers if they constitute a larger portion of the teachers' compensation. Consistent with this possibility, teachers' awareness of their eligibility for pay-for-performance in Year 3 was higher in districts with a larger average bonus in Year 2 (Chapter IV). To the extent that larger bonuses are more salient to teachers, they may lead to larger impacts of pay-for-performance.

Because the analysis was aimed at explaining differences in impacts in Year 3, we measured the size of the average bonus in the prior year (Year 2), the most recent year of actual bonuses that teachers experienced. Average performance bonuses for treatment teachers ranged from one to eight percent of average salary (Figure G.1). The three districts that awarded an average bonus at least 5 percent of average teacher salary were classified as having a larger average bonus.

Figure G.1. Average Performance Bonus Earned by Teachers in Treatment Schools in Year 2 as a Percentage of Average Teacher Salary

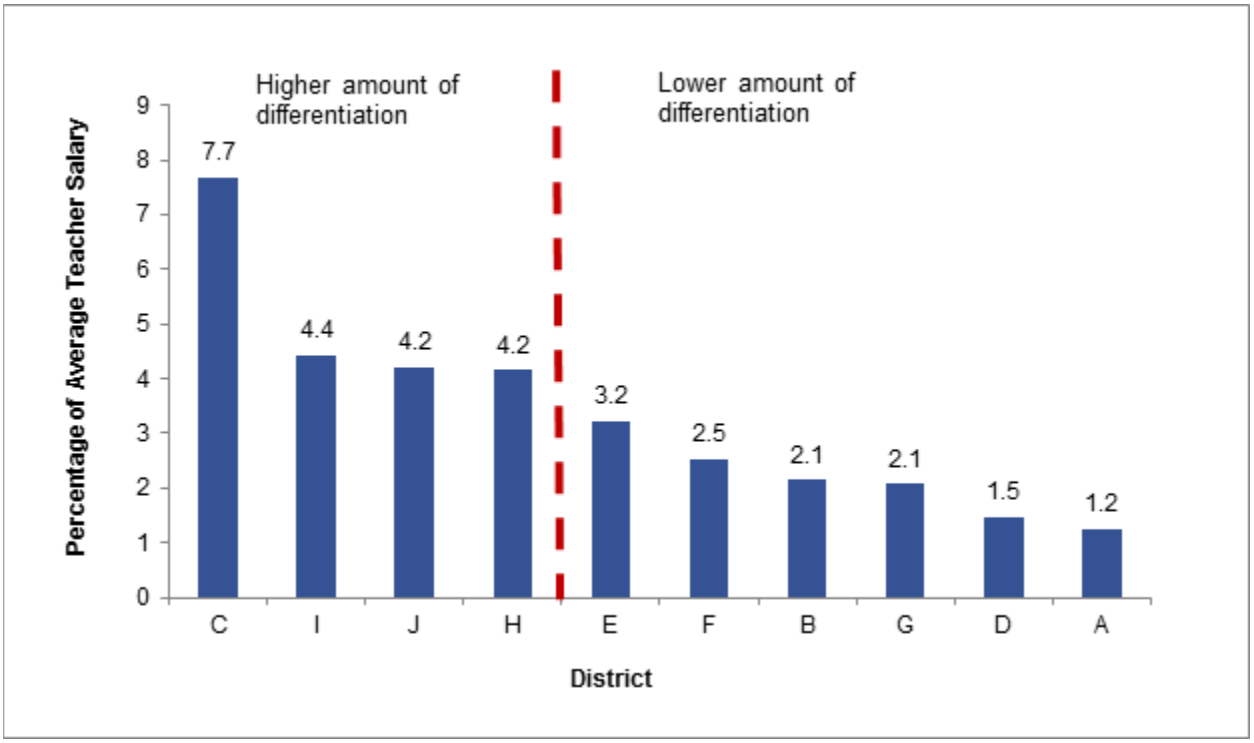


Source: Educator administrative data, Year 2 (N = 2,225 teachers).

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.

We measured the amount of differentiation in each district's bonuses by calculating the standard deviation of the bonuses in Year 2, which captured how extensively below- and above-average bonuses differed in dollar value from the average bonus. In four of the ten districts, the standard deviation of treatment teachers' bonuses exceeded 4 percent of average teacher salary, and we classified these districts as having a higher amount of differentiation (Figure G.2). 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.2. Standard Deviation of Pay-for-Performance Bonuses Earned by Teachers in Treatment Schools in Year 2 as a Percentage of Average Teacher Salary

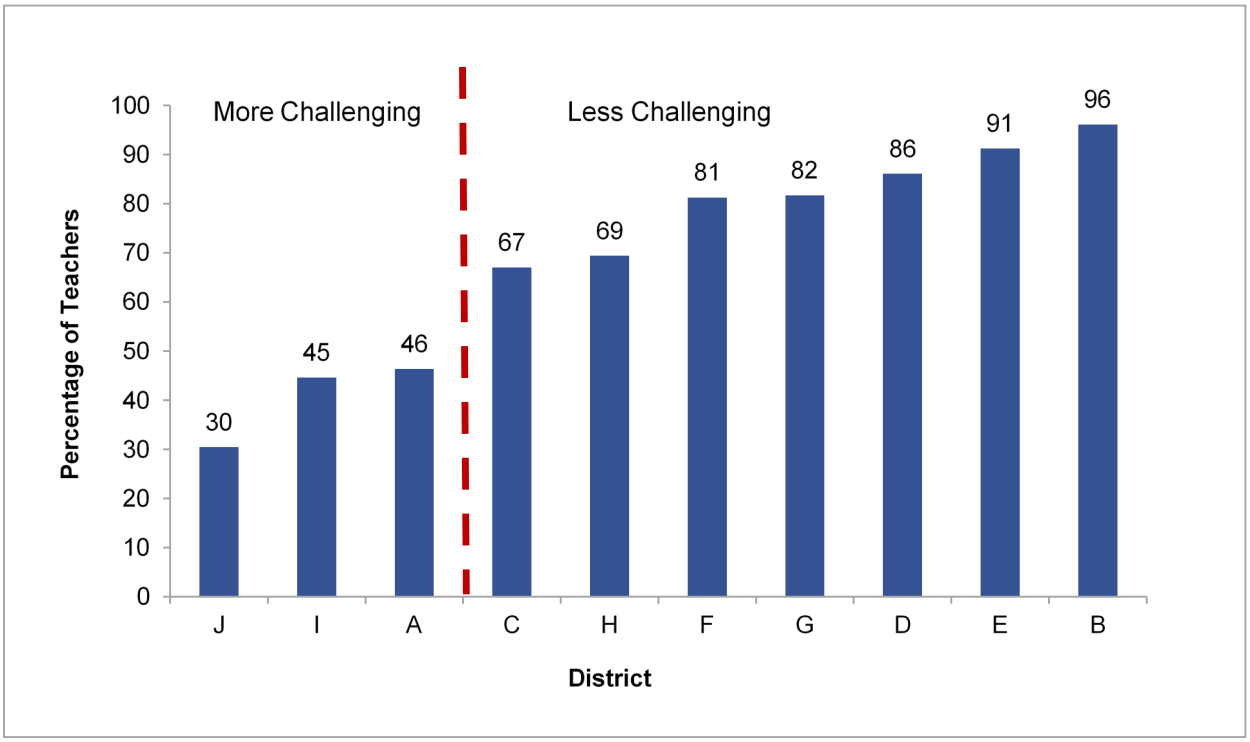


Source: Educator administrative data, Year 2 (N = 2,225 teachers).

Degree to which earning a bonus was challenging. If earning a bonus is not challenging—with nearly everyone receiving a bonus—teachers may perceive little monetary incentive to improve since they might expect to earn a bonus without changing their practices. However, awarding bonuses to a large percentage of teachers could increase teachers' acceptance of the program and their job satisfaction. This in turn, may increase teachers' effort and increase student achievement.

We measured the degree to which earning a bonus was challenging by calculating the percentage of treatment teachers that earned a bonus in Year 2. This percentage ranged from 30 to 96 percent across districts (Figure G.3). Three districts awarded bonuses to fewer than 50 percent of its teachers. We classified these districts as having bonuses that were more challenging to earn.

Figure G.3. Percentage of Teachers in Treatment Schools Earning a Pay-for-Performance Bonus in Year 2



Source: Educator administrative data, Year 2 (N = 2,225 teachers).

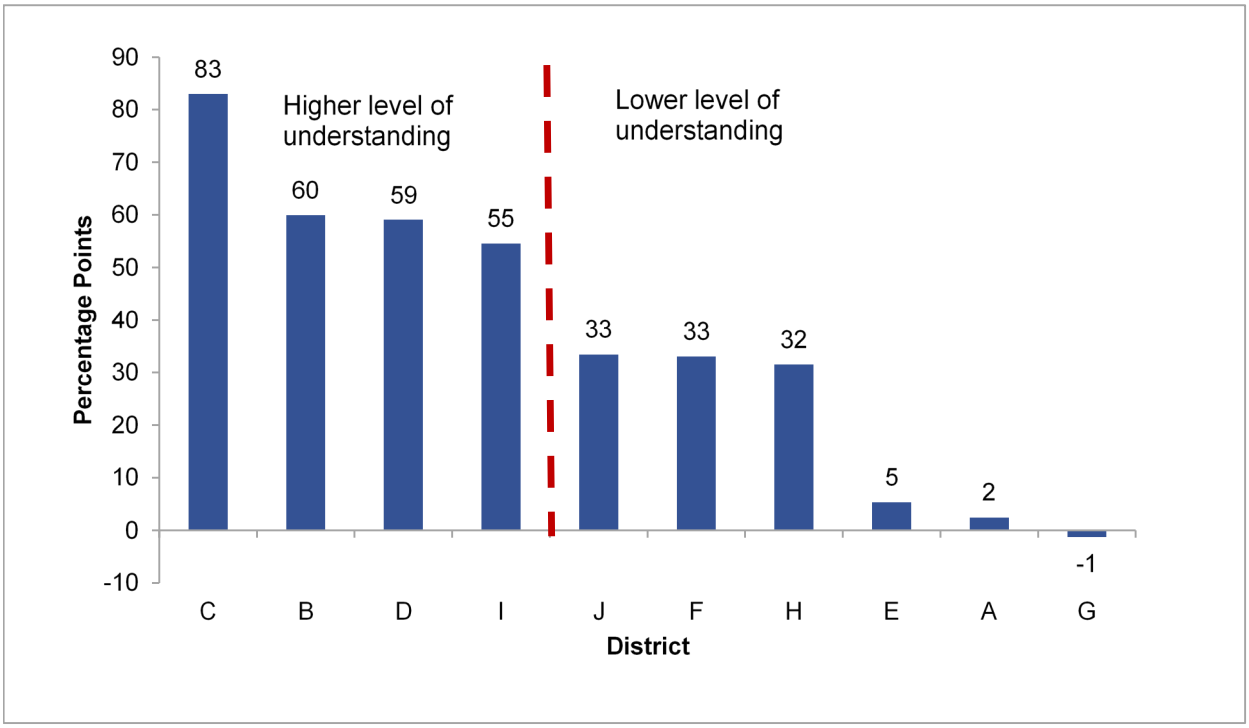
Timing of awarding bonuses. Bonuses may affect teacher productivity by encouraging teachers to change their practices or to change schools. Districts that distribute awards earlier allow their teachers more time to respond in these ways.

As discussed in Chapter IV, there were differences among evaluation districts in the timing of awarding bonuses from Year 2 (2012–2013). Three districts paid out at least some components of the bonuses before the start of the 2013–2014 school year, and we classified those districts as having awarded bonuses earlier. Among the remaining districts, six reported paying teachers between October 2013 and January 2014, and one paid teachers after the end of the 2013–2014 school year. 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, we measured teachers' understanding of their pay-for-performance eligibility by calculating 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 83 percentage points across districts in Year 3 (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 3



Source: Educator survey data, Year 3 (N = 892 teachers).

Findings on Differences in Student Achievement Impacts Between District Subgroups

For each pair of subgroups that differed on a particular characteristic, we estimated the impacts of pay-for-performance on student achievement in Year 3 within the two subgroups and examined whether the impacts differed between the subgroups. A statistically significant difference in impacts between the two subgroups would represent an association between the characteristic and impacts. 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.

TIF program and implementation characteristics measured by this study did not explain differences across districts in the impacts of pay-for-performance on student achievement (Table G.2). None of the six characteristics that we examined had a statistically significant relationship with impacts on student achievement in math or reading in Year 3.

Table G.2. Differences in the Impacts of Pay-for-Performance on Student Achievement in Year 3 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	0.02	0.66	0.03	0.38
Large Average Bonus	-0.02	0.62	0.01	0.87
High Amount of Differentiation in Bonuses	-0.04	0.48	0.01	0.73
Earning a Bonus was Challenging	-0.07	0.08	-0.06	0.07
Early Awarding of Bonuses	0.01	0.85	0.00	0.88
High Level of Teacher Understanding of Pay-for-Performance Eligibility	0.01	0.84	-0.03	0.39
Number of Students	**40,037**		**39,807**	
Number of Schools	**132**		**132**	

Source: Student administrative data.

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

Sensitivity Analysis

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 in Year 3. Consistent with the subgroup findings reported earlier, the program characteristics were generally not related to student achievement impacts (Table G.3). The only exception was that the percentage of teachers who received a bonus in Year 2 was positively associated with impacts on math achievement in Year 3, though the relationship was very small in magnitude).

Table G.3. Association Between Continuous Measures of Program Characteristics and the Impacts of Pay-for-Performance on Student Achievement in Year 3

	Math		Reading	
Program Characteristic	Association	*p*-value	Association	*p*-value
Average Performance Bonus in Year 2 as a Percentage of Average Teacher Salary	0.003	0.66	0.003	0.62
Standard Deviation of Performance Bonuses in Year 2 as a Percentage of Average Teacher Salary	-0.010	0.28	-0.001	0.89
Percentage of Teachers That Received a Performance Bonus in Year 2	0.002*	0.05	0.001	0.13
Number of Months Since May 2013 When District Paid Out Year 2 Performance Bonuses[a]	0.006	0.61	0.006	0.37
Difference Between the Percentages of Teachers in Treatment and Control Schools Who Believed They Were Eligible for Performance Bonuses in Year 3	-0.001	0.57	0.000	0.68
Number of Students—Range[b]	**36,538–40,037**		**36,358–39,807**	
Number of Schools—Range[b]	**114–132**		**114–132**	

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.

[a]Estimates exclude one district that awarded bonuses from Year 2 after the end of Year 3.

[b]Sample sizes are 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.

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. To identify potential explanations for these differences in impacts, 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.

Educator Behaviors Examined

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 2014 survey questions on topics that could reflect strategic behavior, effort, and teaching practices. In addition, we used observation ratings (from 2014 administrative data) as a direct measure of teaching practices. Table G.4 details the specific items used.

Table G.4. 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	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	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.

Findings on Associations Between Impacts on Behaviors and Impacts on Achievement.

Changes in educators' reported behaviors and observation ratings due to pay-for-performance did not explain differences across schools in impacts on student achievement**.** Of the eighteen relationships we examined between impacts on educator behaviors and impacts on student achievement (nine measures of behaviors and two subjects), only one was statistically significant. Given the large number of relationships examined, the single significant finding could have occurred just by chance (Table G.5).

Table G.5. Association Between Impacts on Educator Behaviors and Impacts on Student Achievement in Year 3

	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)	-0.02	0.70	0.00	0.98
Hours Spent by Teachers During School Day on Instructional Activities	0.00	0.42	0.00	0.46
Effort				
Hours Spent by Teachers Outside the School Day on Instructional Activities	0.00	0.58	0.00	0.76
Teachers Believe that TIF Caused Them to Work More Effectively (percentage points)	-0.02	0.88	-0.10	0.30
Teachers Feel Increased Pressure to Perform due to TIF (percentage points)	-0.05	0.74	0.09	0.38
Teaching Practices				
Teachers Believe that TIF Harmed the Collaborative Nature of Teaching (percentage points)	0.06	0.69	-0.04	0.66
Teachers Feel They Have Less Freedom in Teaching due to TIF (percentage points)	-0.26*	0.04	-0.03	0.77
Teachers Believe that Students Will Benefit from the Feedback Received from Classroom Observations (percentage points)	0.12	0.49	-0.09	0.46
Observation Ratings (points on 1-to-4 scale)	0.09	0.61	0.00	0.98
Number of Random Assignment Blocks	**44**		**44**	

Source: Educator and student administrative data (Year 3) and principal and teacher surveys (2014).

Note: 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.

THIS PAGE IS INTENTIONALLY BLANK


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA