Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment

November 2013

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

U.S. Department of Education

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment

November 2013

Steven Glazerman
Ali Protik
Bing-ru Teh
Julie Bruch
Jeffrey Max
Mathematica Policy Research

Elizabeth Warner
Project Officer
Institute of Education Sciences

NCEE 2014-4003
U.S. DEPARTMENT OF EDUCATION

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

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Evaluation and Regional Assistance
Ruth Curran Neild
Commissioner

November 2013

The report was prepared for the Institute of Education Sciences under Contract No. ED-04-CO-0112/007. The project officer is Elizabeth Warner in the National Center for Education Evaluation and Regional Assistance.

Glazerman, S., A. Protik, B. Teh, J. Bruch, J. Max. (2013). *Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Experiment* (NCEE 2014-4003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

ACKNOWLEDGMENTS

This study is the product of many people's efforts. We are deeply grateful to the many teachers, principals, district leaders, and central office staff whose hard work and patience made both the intervention and the research possible, although we unfortunately cannot acknowledge them by name. Spearheading the implementation of the intervention were staff from The New Teacher Project (TNTP), including Coral Jenrette, Mónica Vásquez, Emma Cartwright, Latricia Barksdale, and Kristen Rasmussen. At Mathematica, Tim Silva played an important role in overseeing implementation and working closely with TNTP and the districts. Monica Leal Priddy at Optimal Solutions Group led a team, including Kimberly Hahn, Carolina Herrera, Grace Hong, and Mark Partridge, that collected extensive school records data and played an important role in gathering data needed for program implementation.

This report also relies heavily on teacher and principal surveys. At Mathematica, Nancy Carey and Kristina Rall led the survey research effort with invaluable assistance from Theresa Boujada and her team at Mathematica's Survey Operations Center.

The evaluation team at Mathematica benefited from expert programming and research assistance from Alena Davidoff-Gore, Maureen Higgins, and Christopher Jones. John Deke, Duncan Chaplin, and Neil Seftor read and provided helpful comments on earlier versions of the report. A technical working group (TWG) provided useful input on program design and the research. TWG members included Dale Ballou, Lisa Barrow, Jason Kamras, Robert Meyer, Anthony Milanowski, Jeffrey Smith, and Jacob Vigdor. Sharon Peters edited the report, and Jackie McGee prepared it for publication.

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST[1]

The research team for this evaluation consists of a prime contractor, Mathematica Policy Research of Princeton, New Jersey, and a subcontractor, Optimal Solutions Group of College Park, Maryland. Neither of these organizations or their key staff members have financial interests that could be affected by findings from the evaluation. No one on the technical working group, convened by the research team to provide advice and guidance, has financial interests that could be affected by findings from the evaluation.

[1] Contractors carrying out research and evaluation projects for IES frequently need to obtain expert advice and technical assistance from individuals and entities whose other professional work may not be entirely independent of or separable from the tasks they are carrying out for the IES contractor. Contractors endeavor not to put such individuals or entities in positions in which they could bias the analysis and reporting of results, and their potential conflicts of interest are disclosed.

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

CONTENTS

EXECUTIVE SUMMARY xxv

I INTRODUCTION 1

A. Policy Problem: Unequal Access to High-Performing Teachers 1

B. One Policy Response: Selective Transfer Incentives 2

1. Overview of the TTI 3
2. Logic Model: How Can Teacher-Transfer Incentives Affect Student Achievement? 5

C. Studying Teacher-Transfer Incentives 9

1. Research Questions and Study Design 9
2. Survey Data 12
3. Administrative Data 13

II THE STUDY SAMPLE 15

A. School Districts 15

1. How Districts Were Selected 15
2. Description of the Districts and the Study Context 16

B. Schools 20

1. Identifying Potential Sending and Receiving Schools 22
2. Defining “Participating” Sending and Receiving Schools 24

C. Students 25

III IMPLEMENTATION PROCESS AND PLACEMENT RESULTS 31

A. How Were the Highest-Performing Teachers Identified and Recruited? 31

1. Value-Added Analysis to Identify the Highest-Performing Teachers 31
2. Identifying and Filling Teaching Vacancies 32

B. How Did Teachers React to the Transfer Incentive? 34

1. Take-Up Rates 34
2. Retention of Transfer and Retention-Stipend Teachers Over Two Years 37

III *(continued)*

C. Where Did TTI Transfer Teachers Come From? 37

D. Who Filled the Study Vacancies? 41

1. Control-Group Vacancies 42
2. Treatment-Group Vacancies 44

IV INTERMEDIATE IMPACTS 47

A. Assignment of Teachers to Students and Grades 47

1. Assignment of Students to Teachers 47
2. Assignment of Teachers to Grades 50

B. Teachers' Mentoring and Leadership Roles 51

1. Mentoring Received 51
2. Mentoring Provided and Other Leadership Roles 52

C. Teacher Attitudes 53

D. Principal Reports on School Climate and Teacher Contributions 55

E. Summary of Intermediate Impact Findings 56

V IMPACTS ON STUDENT ACHIEVEMENT 57

A. Data and Methods 57

B. Impacts in Elementary and Middle Schools 58

C. Impacts by District 61

D. Combined Elementary and Middle School Impacts 63

E. Interpreting the Impact Estimates 64

1. TTI Impacts Versus Effectiveness of Transfer Teachers 64
2. Resource-Allocation Effects 65

VI IMPACTS ON TEACHER RETENTION 67

A. Data and Methods 67

B. Retention Impacts 68

VII COST-EFFECTIVENESS 73

A. Cost-Effectiveness Methods 73

B. Costs of TTI 75

C. Cost Comparison 77

D. Unmeasured Effects of TTI 79

REFERENCES 81

APPENDIX A SUPPLEMENTAL MATERIALS FOR CHAPTERS I AND II A.1

APPENDIX B VALUE-ADDED ANALYSIS TO IDENTIFY HIGHEST-PERFORMING TEACHERS B.1

APPENDIX C SUPPLEMENTAL MATERIALS FOR CHAPTER III C.1

APPENDIX D IDENTIFICATION OF FOCAL TEACHERS D.1

APPENDIX E SUPPLEMENTAL MATERIALS FOR CHAPTER IV E.1

APPENDIX F SUPPLEMENTAL MATERIALS FOR CHAPTER V F.1

APPENDIX G SUPPLEMENTAL MATERIALS FOR CHAPTER VI G.1

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

TABLES

II.1 Demographic Characteristics of Potential Sending and Receiving Schools (percentages)....23

II.2 Student Characteristics for Elementary School Math Analysis Sample, by Treatment Status....27

II.3 Student Characteristics for Elementary School Reading Analysis Sample, by Treatment Status....28

II.4 Student Characteristics for Middle School Math Analysis Sample, by Treatment Status....28

II.5 Student Characteristics for Middle School Reading Analysis Sample, by Treatment Status....29

III.1 Percentage of Teachers Receiving Study Payments After the First Program Year....37

III.2 Characteristics of TTI Transfer Teachers' Students Before and After Transferring....41

III.3 How Study Schools Filled Their Vacancies in the Absence of a Transfer Program (Focal Teachers, Control Group Only)....43

III.4 Characteristics of Control Focal Teachers (percentages)....44

III.5 How Study Schools Filled Their Vacancies Using Transfer Program (Focal Teachers, Treatment Group Only)....45

III.6 Characteristics of Treatment and Control Focal Teachers (percentages)....45

IV.1 How Students Were Assigned to Classrooms, Principal Report (percentages)....48

IV.2 Characteristics of Nonfocal Stayers and Movers....50

IV.3 Mentoring Received by Focal and Nonfocal Teachers (percentages)....51

IV.4 Mentoring and Other Leadership Provided by Focal and Nonfocal Teachers (percentages unless indicated otherwise)....53

IV.5 Teacher-Reported Challenges Associated with First Year at the School, by Focal and Nonfocal Teachers....54

IV.6 Teacher Reports on Supportive Environment, by Focal and Nonfocal Teachers....55

IV.7 Summary of Findings on Intermediate Impacts....56

V.1 Test-Score Impacts in Elementary Schools....59

V.2 Test-Score Impacts in Middle Schools 60

V.3 Combined Test-Score Impacts, Benchmark and Vacancy Weighted 64

VI.1 One-Year Impacts on Retention in School 70

VI.2 One-Year Impacts on Retention in School for Focal Teachers, by Grade Span 71

VII.1 Cost of TTI Relative to Estimated CSR Alternative 78

A.1 Maximum Number of Schools per Batch with Vacancies in the Same Grade and Subject A.3

A.2 Number of Teacher Teams per Block A.3

A.3 Summary of Response Rates, by Instrument and Treatment Status A.6

A.4 Respondents Versus Full Sample of Respondents and Nonrespondents (percentages) A.7

A.5 Survey Response Rates by Subgroup, Teacher Survey (percentages) A.8

A.6 Survey Completion Rates by Subgroup, Principal Survey (percentages) A.9

A.7 Respondents Versus Full Sample of Respondents and Nonrespondents, Teacher Survey (percentages) A.10

A.8 Respondents Versus Full Sample of Respondents and Nonrespondents, Principal Survey (percentages) A.11

B.1 Value-Added Scores: Highest-Performing Versus Other Eligible Teachers B.8

B.2 Student Characteristics: Highest-Performing Versus Other Eligible Teachers (percentages) B.9

C.1 Hiring Rates in the Treatment and Control Teacher Teams C.3

C.2 Candidate Interview Process and Perceptions, by Transfer Status (percentages) C.3

C.3 Structure of Candidate Interview, by Transfer Status (percentages) C.4

C.4 Top Self-Reported Reasons for Not Applying to TTI (percentages) C.4

C.5 Characteristics of Candidates, by Application Status (percentages unless otherwise noted) C.5

C.6 Factors Related to the Probability of Applying C.7

C.7 Factors Related to the Probability of Transferring C.8

C.8 Value-Added Scores and Student Characteristics of Candidates, by Application Status (percentages except for value-added scores) C.9

D.1 Example of Focal Teacher Identification D.6

D.2 Characteristics of Focal Teachers Under Alternative Definitions D.7

E.1 Teacher Satisfaction with School, by Focal and Nonfocal Teachers E.10

E.2 Principal Reports on Team Climate E.11

E.3 Principal Ratings of Teacher Contributions E.12

F.1 Test-Score Impacts in Cohort 1 Districts Only, Program Year 1 ..... F.4

F.2 Team-Level Test-Score Comparisons, by Grade F.17

F.3 Focal-Teacher Test-Score Comparisons, by Grade F.17

F.4 Nonfocal-Teacher-Level Test-Score Comparisons, by Grade F.18

F.5 Percentage of 8th-Grade Students on Study Teams Taking General Math and Algebra I Post-Tests, by Treatment Status, Program Year 1 ..... F.19

F.6 Test-Score Impacts in Middle Schools, Omitting 8th-Grade Math Students from One District, Program Year 1 ..... F.20

F.7 Test-Score Impacts in Elementary Schools, Adjustment for Taking Test in Spanish in One District F.22

F.8 Test-Score Impacts in Middle Schools, Adjustment for Taking Test in Spanish in One District F.23

F.9 Test-Score Impacts in Elementary Schools, Benchmark Model, Blocks Containing Schools with Single-Treatment Status Only F.24

F.10 Test-Score Impacts in Middle Schools, Benchmark Model, Blocks Containing Schools with Single-Treatment Status Only F.25

F.11 Test-Score Impacts in Elementary Schools, Benchmark Model, Blocks Containing Schools with Both Treatment and Control Teams F.26

F.12 Test-Score Impacts in Middle Schools, Benchmark Model, Blocks Containing Schools with Both Treatment and Control Teams F.27

F.13 Complete List of Sensitivity Analyses F.28

F.14 Test-Score Impacts in Elementary Schools, Benchmark Model, Selective Method for Identifying Focal Teachers F.30

F.15 Test-Score Impacts in Middle Schools, Benchmark Model, Selective Method for Identifying Focal Teachers F.31

F.16 Test-Score Impacts in Elementary Schools, Benchmark Model Excluding Students with Imputed Pre-Test F.32

F.17 Test-Score Impacts in Middle Schools, Benchmark Model Excluding Students with Imputed Pre-Test F.33

F.18 Test-Score Impacts in Elementary Schools, Benchmark Model Plus Opposite-Subject Pre-Test F.34

F.19 Test-Score Impacts in Middle Schools, Benchmark Model Plus Opposite-Subject Pre-Test F.35

F.20 Test-Score Impacts in Elementary Schools, Benchmark Model with Interactions of Student-Background Covariates and District Dummies F.36

F.21 Test-Score Impacts in Middle Schools, Benchmark Model with Interactions of Student-Background Covariates and District Dummies F.37

F.22 Test-Score Impacts in Elementary Schools, Benchmark Model Without Pre-Test Variable and Student-Background Covariates F.38

F.23 Test-Score Impacts in Middle Schools, Benchmark Model Without Pre-Test Variable and Student-Background Covariates F.39

F.24 Test-Score Impacts in Elementary Schools, Benchmark Model Replacing Treatment Variable with Treatment Multiplied by Percent of Year Enrolled in School F.40

F.25 Test-Score Impacts in Middle Schools, Benchmark Model Replacing Treatment Variable with Treatment Multiplied by Percent of Year Enrolled in School F.41

F.26 Test-Score Impacts in Elementary Schools, Benchmark Model Adding Pilot Teams to the Analysis Sample F.42

F.27 Test-Score Impacts in Middle Schools, Benchmark Model Adding Pilot Teams to the Analysis Sample F.43

F.28 Sample-Size Information for Table V.1. Test-Score Impacts in Elementary Schools F.44

F.29 Sample-Size Information for Table V.2. Test-Score Impacts in Middle Schools F.45

F.30 Sample-Size Information for Table V.3. Test-Score Impacts in Elementary and Middle Schools Combined F.46

G.1 Logit-Model Retention Results Versus Benchmark-Model Retention Results G.4

G.2 Student-Characteristic-Model Retention Results Versus Benchmark-Model Retention Results G.5

G.3 Impacts on Retention in School, Comparison of Alternative Samples G.6

G.4 Impacts on Retention in School, Inclusive and Selective Focal Teacher Samples G.7

G.5 Two-Year Impacts on Retention in School, Cohort 1 Districts Only G.8

G.6 One-Year Impacts on Retention in School, by Grade Span G.11

G.7 Two-Year Impacts on Retention in School, by Grade Span G.11

G.8 One-Year Retention Rates of Transfer and Retention-Stipend Teachers G.14

G.9 Two-Year Retention Rates of Transfer and Retention-Stipend Teachers G.14

G.10 Sample Sizes for Table VI.1, Table G.1, Table G.2, and Table G.5 ..... G.15

G.11 Sample Sizes for Table VI.2 ..... G.15

G.12 Sample Sizes for Table G.3 ..... G.15

G.13 Sample Sizes for Table G.4 ..... G.16

G.14 Sample Sizes for Table G.6 ..... G.16

G.15 Sample Sizes for Table G.7 ..... G.17

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

FIGURES

ES.1 Random Assignment Study Design xxix

ES.2 Percentage of TTI Vacancies Assigned and Filled, by Month xxxiii

ES.3 Test-Score Impacts in Elementary Schools xxxv

ES.4 Test-Score Impacts in Middle Schools xxxvi

ES.5 Impacts on Retention in School, Cohort 1 Districts Only xxxvii

I.1 Logic Model: How Transfer Incentives Affect Teachers and Students in Receiving Schools 6

I.2 Random Assignment Study Design 10

II.1 Percentage of Students in Lowest- and Highest-Poverty Elementary Schools Who Are Low Income (FRL), by District 17

II.2 Percentage of Students in Lowest- and Highest-Poverty Middle Schools Who Are Low-Income (FRL), by District 17

II.3 Percentage of Teachers Offered Bonuses and Stipends 20

II.4 Bonuses and Stipend Amounts Offered, Average Across All Teachers 20

II.5 Selection of Schools and Teams into the Study Sample 21

II.6 Sending and Receiving Schools 25

III.1 Percentage of TTI Vacancies Assigned and Filled, by Month 33

III.2 Take-Up Rates Among TTI Transfer Candidates, by Grade Span and Subject 34

III.3 Contrast in TTI Transfer Teachers' Sending- and Receiving-School Achievement Rank 39

III.4 Contrast in TTI Transfer Teachers' Sending- and Receiving-School Poverty Rank 40

IV.1 Classroom Assignment of Students Who Are More or Less Academically Challenging, Teachers' Perceptions, by Their Treatment and Focal Status 49

V.1 Year 1 Impacts on Math Scores, Elementary Focal Teachers, by District (cohorts 1 and 2) 62

V.2 Year 2 Impacts on Math Scores, Elementary Focal Teachers, by District (cohort 1) 62

VI.1 Impacts on Retention in School, Cohort 1 Districts Only 70

VII.1 Costs of TTI per Team 75

A.1 Random Assignment Study Design A.5

C.1 Percentage of Elementary-Level TTI Vacancies Assigned and Filled, by Month C.10

C.2 Percentage of Middle School-Level TTI Vacancies Assigned and Filled, by Month C.11

C.3 Types of Elementary-Level Transfers, by Achievement Rank C.12

C.4 Types of Middle School-Level Transfers, by Achievement Rank C.12

C.5 Types of Elementary-Level Transfers, by Poverty Ranks C.13

E.1 Students' Prior Math Scores in Focal Teachers' Classrooms E.4

E.2 Students Prior Reading Scores in Focal Teachers' Classrooms E.4

E.3 Low-Income Students in Focal Teachers' Classrooms E.5

E.4 ELLs, Relative Percentage in Focal Teachers' Classrooms E.6

E.5 SPED Students, Relative Percentage in Focal Teachers' Classrooms E.6

E.6 White Students, Relative Percentage in Focal Teachers' Classrooms E.7

E.7 Black Students, Relative Percentage in Focal Teachers' Classrooms E.7

E.8 Hispanic Students, Relative Percentage in Focal Teachers' Classrooms E.8

E.9 Classroom Assignment of Students Who Are More or Less Behaviorally Challenging, Teachers' Perceptions, by Their Treatment and Focal Status E.9

F.1 Year 1 Impacts on Math Scores, Elementary Teams, by District (cohorts 1 and 2) F.5

F.2 Year 2 Impacts on Math Scores, Elementary Teams, by District (cohort 1) F.6

F.3 Year 1 Impacts on Math Scores, Elementary Nonfocal Teachers, by District (cohorts 1 and 2) F.6

F.4 Year 2 Impacts on Math Scores, Elementary Nonfocal Teachers, by District (cohort 1) F.7

F.5 Year 1 Impacts on Reading Scores, Elementary Teams, by District (cohorts 1 and 2) F.7

F.6 Year 2 Impacts on Reading Scores, Elementary Teams, by District (cohort 1) F.8

F.7 Year 1 Impacts on Reading Scores, Elementary Focal Teachers, by District (cohorts 1 and 2) F.8

F.8 Year 2 Impacts on Reading Scores, Elementary Focal Teachers, by District (cohort 1) F.9

F.9 Year 1 Impacts on Reading Scores, Elementary Nonfocal Teachers, by District (cohorts 1 and 2) F.9

F.10 Year 2 Impacts on Reading Scores, Elementary Nonfocal Teachers, by District (cohort 1) F.10

F.11 Year 1 Impacts on Math Scores, Middle Teams, by District (cohorts 1 and 2) F.10

F.12 Year 2 Impacts on Math Scores, Middle Teams, by District (cohort 1) F.11

F.13 Year 1 Impacts on Math Scores, Middle Focal Teachers, by District (cohorts 1 and 2) F.11

F.14 Year 2 Impacts on Math Scores, Middle Focal Teachers, by District (cohort 1) F.12

F.15 Year 1 Impacts on Math Scores, Middle Nonfocal Teachers, by District (cohorts 1 and 2) F.12

F.16 Year 2 Impacts on Math Scores, Middle Nonfocal Teachers, by District (cohort 1) F.13

F.17 Year 1 Impacts on Reading Scores, Middle Teams, by District (cohorts 1 and 2) F.13

F.18 Year 2 Impacts on Reading Scores, Middle Teams, by District (cohort 1) F.14

F.19 Year 1 Impacts on Reading Scores, Middle Focal Teachers, by District (cohorts 1 and 2) F.14

F.20 Year 2 Impacts on Reading Scores, Middle Focal Teachers, by District (cohort 1) F.15

F.21 Year 1 Impacts on Reading Scores, Middle Nonfocal Teachers, by District (cohorts 1 and 2) F.15

F.22 Year 2 Impacts on Reading Scores, Middle Nonfocal Teachers, by District (cohort 1) F.16

G.1 One-Year Team-Level Retention Impacts, by District G.9

G.2 Two-Year Team-Level Retention Impacts, by District G.10

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

EXECUTIVE SUMMARY

One way to improve struggling schools' access to effective teachers is to use selective transfer incentives. Such incentives offer bonuses for the highest-performing teachers to move into schools serving the most disadvantaged students. In this report, we provide evidence from a randomized experiment that tested whether such a policy intervention can improve student test scores and other outcomes in low-achieving schools.

The intervention, known to participants as the Talent Transfer Initiative (TTI), was implemented in 10 school districts in seven states. The highest-performing teachers in each district—those who ranked in roughly the top 20 percent within their subject and grade span in terms of raising student achievement year after year (an approach known as value added)—were identified. These teachers were offered $20,000, paid in installments over a two-year period, if they transferred into and remained in designated schools that had low average test scores. The main findings from the study follow.

Main Findings

- **The transfer incentive successfully attracted high value-added teachers to fill targeted vacancies**. Almost 9 out of 10 targeted vacancies (88 percent) were filled by the high-performing teachers who had been identified as candidates eligible for the transfer intervention. To achieve those results, a large pool of high-performing teachers was identified (1,514) relative to the number of vacancies filled (81). The majority of candidates did not attend an information session (68 percent) or complete an online application to participate in the transfer intervention (78 percent).
- **The transfer incentive had a positive impact on test scores (math and reading) in targeted elementary classrooms**. These impacts were positive in each of the two years after transfer, between 0.10 and 0.25 standard deviations relative to each student's state norms. This is equivalent to moving up each student by 4 to 10 percentile points relative to all students in their state. In middle schools, we did not find evidence of impacts on student achievement. When we combined the elementary and middle school data, the overall impacts were positive and statistically significant for math in year 1 and year 2, and for reading only in year 2. Our calculations suggest that this transfer incentive intervention in elementary schools would save approximately $13,000 per grade per school compared with the cost of class-size reduction aimed at generating the same size impacts. However, overall cost-effectiveness can vary, depending on a number of factors, such as what happens after the last installments of the incentive are paid out after the second year. We also found there was significant variation in impacts across districts.
- **The transfer incentive had a positive impact on teacher-retention rates during the payout period; retention of the high-performing teachers who transferred was similar to their counterparts in the fall immediately after the last payout**. We followed teachers during both the period when they were receiving bonus payments and afterward. Retention rates were significantly higher during the payout period—93 versus 70 percent. After the payments stopped, the difference between cumulative retention of the high-performing teachers who transferred and their counterparts (60 versus 51 percent) was not statistically significant.

Background

There is growing concern that the nation's most effective teachers are not working in the schools with the most disadvantaged students (Goldhaber 2008; Peske and Haycock 2006; Tennessee Department of Education 2007; Sass et al. 2012; Glazerman and Max 2011). One strategy to remedy this situation is the use of monetary incentives to recruit teachers who have demonstrated success in raising student test scores ("value added") to teach in low-achieving schools. This strategy has been tried in some fashion in several places, such as Mobile, Alabama; Chattannooga, Tennessee; Palm Beach, Florida; and the states of California and Virginia (Max et al. 2007). It has the potential to redistribute some of the highest-performing teachers in a district from higher-achieving schools to lower-achieving schools. However, there is a need for research that addresses several questions related to such a policy, including whether teachers would be willing to transfer if offered incentives, and, if they do transfer, how their presence will change the dynamics in their new schools, how long they would stay in those schools, and whether they would improve student achievement in those schools.

The U.S. Department of Education's Institute of Education Sciences (IES) contracted with Mathematica Policy Research to study the effectiveness of an intervention that is based on this strategy. The intervention, known to participating districts as the TTI (and described in Box ES.1), offers $20,000 to the highest-performing teachers in tested grades and subjects within each district who agree to transfer into one of the lowest-achieving schools in their district and stay for at least two years. Highest-performing teachers were identified based on their value-added scores[2] because some of the study districts were already using value added as one of the measures of teacher performance and because pay-for-performance policies like TTI are likely to use value-added scores as performance measures. We used whatever value-added measure the district was using because that is what would have been used in the absence of the study. In cases where such a measure was not in use, we calculated it ourselves. Teachers were eligible to transfer through TTI if, based on value-added measures, they were among the top 20 percent of teachers in the district and were not currently teaching in the lowest-achieving schools. Teachers who were in the top 20 percent but were already teaching in the lowest-achieving schools were offered $10,000 in retention stipends to continue teaching at those schools for two years. In this final report, we cover implementation and impacts for 10 districts that agreed to participate in the study. Seven began implementation in 2009 (cohort 1); an additional three began implementation in 2010 (cohort 2). In an earlier report (Glazerman et al. 2012), we presented early implementation and intermediate impacts for the first 7 districts.

Research Questions and Study Design

In this study, we address a set of specific research questions related to this transfer incentive policy, including both implementation and impact questions:

- What can we learn from the **implementation** of TTI? Specifically, what can we learn about the timing and scale of implementation, who transfers, and from where they transfer?

[2] Value-added measures seek to describe the contribution that teachers make (the value that they add) to student achievement growth, holding constant factors outside the teacher's control, such as student background and prior learning (McCaffrey et al. 2004; Lipscomb et al. 2010).

- What were the **intermediate impacts** in schools receiving the transfer teachers (referred to as receiving schools)? Specifically, how did TTI affect the dynamics within those schools, such as the allocation of resources, staffing patterns, assignment of students to teachers and courses, and school climate?
- What was TTI's **impact on student test scores** in receiving schools?
- What was TTI's **impact on teacher retention** in receiving schools?

The impact questions relate to the effect of the transfer incentive policy relative to the absence of such a policy. In other words, we sought to measure effects relative to the outcomes that would have been realized had the school not had the opportunity to use the $20,000 incentive to fill its vacancy with a teacher designated as highest performing.

Box ES.1. How the Talent Transfer Initiative Works

The intervention is designed to proceed within each district according to the following steps. The first step is to conduct a value-added analysis of student test scores to identify the highest-performing teachers, defined as the top 20 percent based on a value-added measure of teachers in tested grades and subjects in each district. The second step is to classify schools as "potential receiving" or "potential sending" schools. Potential receiving schools are those with the lowest achievement in the district, based on school-average test scores in the most recent year, and, in some cases, rankings on school accountability. The rare exceptions that are already participating in a comparable intervention are exempted. The rest of the schools in the district are potential sending schools.

The third step is recruitment of (1) eligible high-performing teachers in sending schools, whom we refer to as "transfer candidates," and, simultaneously, (2) principals of receiving schools. The highest-performing teachers (identified in the first step) in potential sending schools are offered a series of transfer incentive payments, totaling $20,000 over two years, to transfer into and remain in one of the receiving schools in their district. The offer is made to these teachers, known as "transfer candidates," in the spring, at which point they are invited to apply to the program.

At the same time, principals of potential receiving schools are invited to an information session and asked to identify likely teaching vacancies in targeted grades and subjects. To be considered for inclusion in TTI, principals must volunteer a vacancy. Eligibility is based on grade level and subject of the vacancy. A site manager in each district helps principals fill the targeted vacancies by providing information about transfer candidates and arranging and encouraging interviews. This extra hiring support is in addition to the TTI transfer incentive.

Next, applicants must interview with and be offered a position by the receiving-school principal and then voluntarily transfer to qualify for the transfer incentive. To improve the probability of matching high-performing teachers with low-achieving schools, the implementation team works with each district to finalize offers and acceptances by early summer.

Finally, the transfer teachers participate in a half-day orientation just before the start of the school year. Because they are selected on the basis of their performance in the classroom, it is assumed that they do not require additional formal support beyond what teachers normally receive. To facilitate the transition, however, the site manager provides informal support and answers any questions throughout the two school years of the intervention period. TTI teachers who fill study-assigned vacancies receive their first incentive payment after the orientation, and those who remain during the intervention period in the positions into which they transferred receive incentive payments in December and June, for a total of $20,000.

Teachers who are identified as highest-performing but who are already teaching in low-achieving (potential receiving) schools are not eligible to transfer, but they are offered a retention stipend of $10,000 for staying at their schools over the same two-year period as transfer teachers.

The methods for answering these questions included descriptive tabulations (for implementation questions) and causal analysis (for impact questions). The causal analysis relied on an experimental design in which we used random assignment to form equivalent groups of classrooms with and without the intervention to compare outcomes after one and two years. The units we assigned were teacher teams, which we defined as all teachers in a specific grade and subject. In elementary schools, the teacher team consisted of all classroom teachers in the grade. In middle schools, the teacher team consisted of all the math or English/language arts (ELA) teachers who taught at least one class in the grade level of interest. For example, all teachers responsible for teaching 7th-grade math in the same school made up one team. All teachers in the school who were responsible for 8th-grade ELA were considered another team. Thus, teacher teams were based on the grade span (elementary and middle school) as well as the subject (math or reading) the teachers were teaching. Because our unit of assignment was teacher teams, we present impact estimates by grade span and subject.

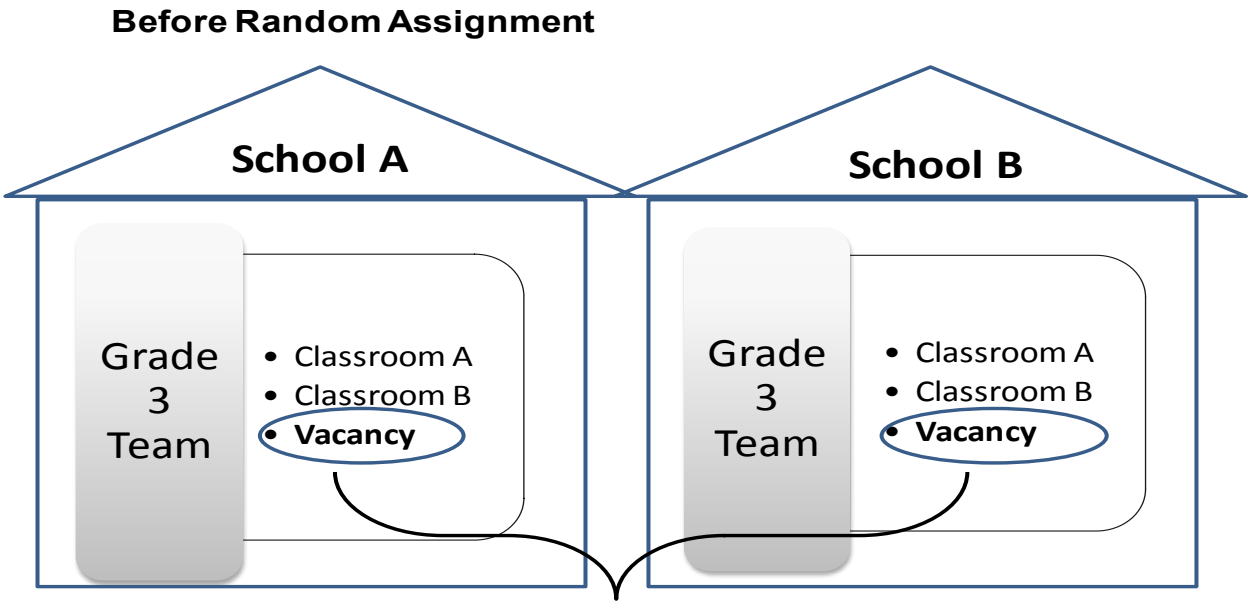
Random Assignment

The research team randomly assigned study subjects to a treatment or control group in the following way. First, we identified low-achieving schools that had a vacancy within a teaching team. If we learned of multiple eligible teacher teams in the same district at approximately the same time, we matched schools with vacancies in the same grade (and subject, in the case of middle school teams) within the same district. When possible, we also matched schools with vacancies based on their student achievement ranking and the percentage of students eligible for free or reduced-priced lunch (FRL). These matched schools formed blocks, and we then randomly assigned teacher teams within each block to either treatment status (with the opportunity to fill the team's vacancy with a TTI teacher) or control status (in which vacancies were filled through whatever process the school would normally use). For example, consider two schools, A and B, each of which had one vacancy in the grade 3 teacher team (see Figure ES.1). The grade 3 teacher team in school A was randomly assigned to treatment and is eligible to fill its vacancy through TTI; the grade 3 teacher team in school B was consequently assigned to control status, so normal hiring practices were to be followed. Teams that could not be assigned in pairs were assigned in blocks containing an odd number of teams. This situation arose either because pairs of vacancies were not available at about the same time or because a close match—in terms of student achievement ranking and/or percentage of students eligible for FRL—could not be found among schools with vacancies.

Repeated for many blocks, this process created two groups of teacher teams that were, on average, similar in terms of student characteristics and school context. The only systematic difference between the two groups was whether, in hiring for the vacancy, there was the opportunity to use the TTI policy with the associated $20,000 transfer incentive. Comparing outcomes for these groups will generate unbiased estimates of the impact of TTI on student achievement and other outcomes.

Each teacher team included the teacher who filled the vacancy as well as his or her grade-level colleagues at the same school. We expect much of the impact of TTI to operate through the teachers who filled the vacancies in the treatment and control teacher teams. We refer to them as "focal" teachers. Therefore, in addition to the team-level analysis, we are interested in the comparison between focal treatment and focal control teachers. We refer to the other teachers on a given grade-level team as "nonfocal" teachers.

Figure ES.1. Random Assignment Study Design



Data Collection

To gather data for the study, we administered surveys of teachers and principals and collected administrative records from schools and districts. We also gathered information from the program implementation process. For cohort 1, the report covers two program years, 2009–10 and 2010–11; for cohort 2, the report includes information from the first program year only, 2010–11. Data for the study are summarized below.

Candidate survey. High-performing teachers who were eligible to apply to TTI were designated as TTI transfer candidates. We surveyed candidates during the first program year. The survey asked about their background; the factors affecting their decision to apply for, interview for, and transfer into TTI positions; and their experiences in the hiring process, if applicable. The response rate was 81 percent.

Teacher background survey. All teachers on study teacher teams—focal and nonfocal—were surveyed during the first program year. The survey asked about their background, their experiences at study schools, and other factors that might affect their students' achievement.[3] The response rate was 77 percent.

Principal survey. Principals of the receiving schools were surveyed in the spring of both program years for cohort 1, and spring of the first program year only for cohort 2. The survey asked about teacher recruitment and hiring, principals' assessments of newly hired teachers, redistribution of resources across classrooms, and the school environment. The response rate was 90 percent in the first program year (cohorts 1 and 2) and 82 percent in the follow-up year (cohort 1 only).

Teacher rosters. Study schools provided teacher rosters in the fall of program years 1 and 2, as well as in the fall after the program ended for cohort 1. The rosters included information about each teacher's school and teaching assignment, and they were used to estimate teachers' retention rates. We obtained rosters from 100 percent of the schools.

Student achievement record. Districts provided student test scores linked to teachers and student demographic data in the fall after program year 1. Cohort 1 districts provided similar data in the fall after program year 2. These were used to estimate impacts on student achievement. We obtained achievement data on 100 percent of the schools in the study.

TTI program implementation records. Districts provided information related to teachers' value-added performance that was used to identify transfer candidates and retention bonuses. Some districts provided more information than others; the research team used all available data. In cases where Mathematica conducted the value-added analysis, we had detailed information on student-teacher links and student background characteristics to estimate value-added scores. In cases where value-added analysis was conducted by a third-party vendor hired by the district, we were given value-added scores directly, or, in one case, just names of highest-scoring teachers. Districts also provided data on school-level student achievement that was used to determine which schools were eligible to participate as potential receiving schools. The TTI site managers provided principal consent forms and information on the timing of teaching vacancies and when they were filled.

Study Sample

We selected school districts that were large and economically diverse. They had to have no fewer than 40 elementary schools, at least 10 of which had to be low-poverty schools and at least 15 of which had to be high-poverty schools. Low- and high-poverty schools were defined as

[3] Survey questions on teacher background information were the same for the candidate and the teacher background survey. Transfer teachers who responded to the candidate survey were not asked about their background in the teacher background survey.

having less than 40 percent or more than 70 percent of students eligible for FRL, respectively. In addition to the quantitative criteria, we selected districts according to a variety of qualitative factors related to the feasibility of implementation, including availability of test scores, data quality, hiring/transfer practices, and the local political environment. The resulting set of districts was not a random sample of a well-defined population of districts, so findings from this study cannot necessarily be generalized to other districts.

We excluded school districts in which existing or planned teacher-incentive programs would have duplicated the intervention under study, but we did come across some existing performance-incentive initiatives in some of the 10 participating school districts. In each case, we determined that the existing programs were different enough, isolated to a few schools that could be excluded from our study, or involved small enough dollar amounts that they would not interfere with the study design. Teachers and schools receiving more than $5,000 were excluded so as to avoid complicating the study by changing the effective differential in the TTI transfer incentives relative to the counterfactual. The $5,000 threshold we established was based on information in the literature on teacher responsiveness to pay (Max et al. 2007) that suggests this amount would plausibly influence teacher behavior.

Working with each district, the implementation team divided the elementary and middle schools into potential sending or potential receiving schools according to academic ranking. Schools were ranked by their students' average prior achievement level, which was determined by the previous three years of achievement data or by the past year's achievement data, depending on the district leaders' preferences.[4] The lowest-ranking schools were designated as potential receiving schools that could benefit from the intervention, and the rest were potential sending schools. We removed some schools from both pools and referred to them as exempt schools because they served a special population of students or were already implementing a program that was meant to address the problem that TTI aims to address. In the end, 21 percent of the schools were classified as potential receiving schools, 72 percent were potential sending schools, and 7 percent were exempt.

The study focused on teachers in a subset of these potential sending and receiving schools. From the potential sending schools, we surveyed the teachers who were identified as highest-performing in the district and eligible for TTI. From the potential receiving schools, we collected data on "study schools," those with teaching teams that had been randomly assigned to treatment or control status. There were 114 study schools, which represents 56 percent of the potential receiving schools.

The final sample had the following features:

Study districts. Ten school districts participated in the study, contributing both elementary and middle schools except for 3 that contributed only elementary schools or only middle schools. Six of the 10 districts were countywide, encompassing urban and nonurban areas. The districts ranged in size from just under 100 square miles to more than 1,200 square miles, which is larger than the state of Rhode Island.

[4] Achievement data from the year before the implementation of TTI were used for all but two districts, where three prior years of achievement data were used.

Study schools. Across the 10 districts, 114 of the potential receiving schools had teams that were randomly assigned to treatment or control status. The average study school was 80 percent low income (FRL).

Teacher teams. Teacher teams in the study ranged from 3rd grade through 8th grade. Some teams included more than one vacancy, and some schools included more than one team. Eighty-five teams were assigned to participate in the intervention and 80 teams were assigned to the control group. Note that we randomly assigned teacher teams within blocks and because we had some blocks with an odd number of teacher teams, the process of random assignment generated by chance an unequal number of treatment and control teams.

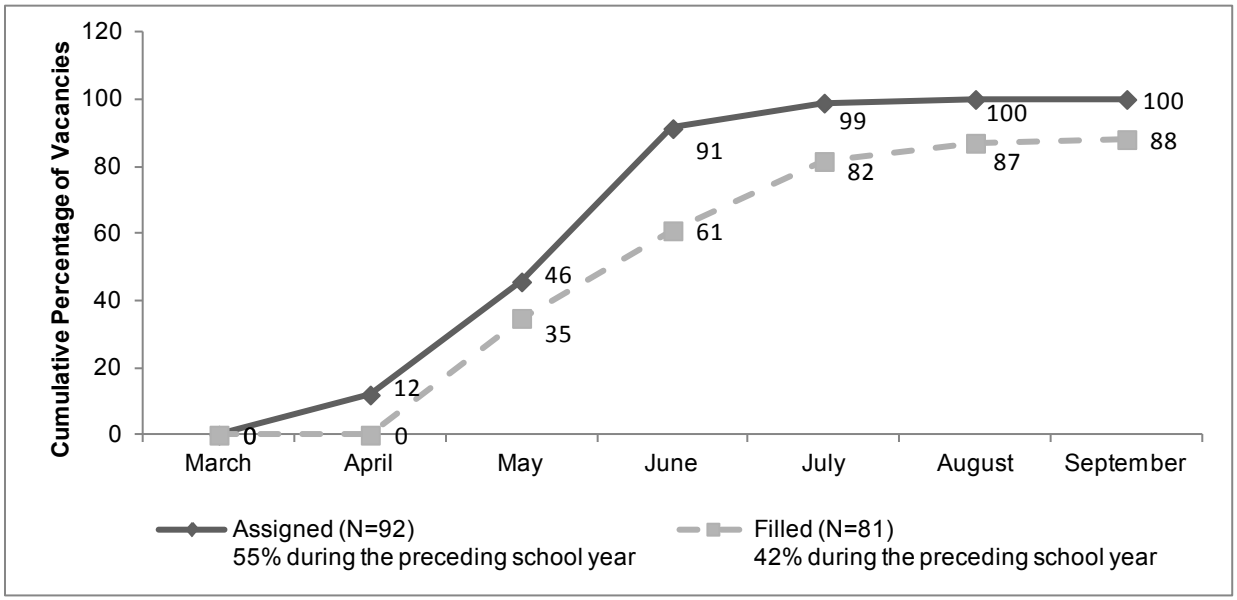
Students. The students of teacher teams in the study were low achieving and disadvantaged. In the year before implementation, the students on study teams performed at approximately the 32nd and the 33rd percentile on state standardized tests in reading and math, respectively, compared with other students in their state.

Implementation Findings: What Can We Learn from the Implementation of TTI?

The implementation of TTI can offer insights regarding the use of transfer incentives to redistribute a district's highest-performing teachers. Here, we summarize the key findings about how vacancies were filled normally and under TTI, how teachers responded to the transfer incentive, where they transferred from, and the resulting treatment contrast in teacher background. We focus especially on the teachers who filled vacancies targeted by TTI, teachers we refer to as "focal," as shown in Figure ES.1. We also consider possible effects on nonfocal teachers, defined as the focal teachers' peers in the same grade and subject.

Almost all vacancies in treatment schools were filled by TTI teachers, although a large pool of candidates was used to yield the desired number of successful TTI transfers. The implementation of TTI demonstrated that it is possible to implement a transfer-incentive program as designed for this study. The highest-performing teachers were identified in approximately the first three months of the calendar year using value-added analysis. Beginning as early as March, the implementation team, consisting of district personnel working with staff from The New Teacher Project (TNTP), conducted several months of intensive recruitment of receiving schools and transfer candidates, resulting in 88 percent of the treatment school vacancies being filled with TTI teachers. In Figure ES.2, we show that most treatment vacancies were assigned and filled in May and June. However, an initial pool of 1,514 candidates was identified to yield the 81 who ultimately transferred. Thus, an average of 5 percent of each district's highest-performing teachers in sending schools ultimately transferred to low-achieving schools. Most did not even attempt to transfer: 32 percent of eligible TTI candidates attended an information session, leaving 68 percent who did not attend; 22 percent completed an application, leaving 78 percent who did not. Fifty-five percent of the applicants interviewed for at least one vacancy, and the other 45 percent either did not follow through or were not given an opportunity to interview. Principals in treatment schools conducted an average of 3.1 interviews per vacancy, and most principals with treatment teams made an offer to only one TTI candidate.

Figure ES.2. Percentage of TTI Vacancies Assigned and Filled, by Month



Source: TTI program records.

Standard practice in the absence of treatment was to fill vacancies through a combination of new hires, transfers in, and within-school reassignments. Nineteen percent of control group vacancies were filled by teachers new to the district, 22 percent by teachers transferring from another school in the district, and 30 percent by teachers reassigned within the school. No TTI transfer candidates transferred to any of these vacancies. The average teaching experience of control focal teachers was eight years, reflecting the fact that many were experienced teachers who simply moved from elsewhere in the school or district and were not hired out of the pool of novice teachers. Only 17 percent reported being new to teaching, whereas 45 percent reported being in at least their sixth year of teaching.

The TTI teachers were more experienced than teachers who would normally fill the vacancies. The average difference in teaching experience between treatment and control focal teachers was about four years. There was also a significant difference in the percentage of teachers with National Board Certification (20 percent of focal treatment teachers compared with 9 percent of focal control teachers).

Intermediate Impacts: How Did TTI Affect School Staffing Assignments and Resource Allocation?

The opportunity for a school to fill a teaching vacancy with one of the district's highest-performing teachers could lead to changes in a range of behaviors, such as student or teaching assignment within the teaching team and the school. These changes can be considered intermediate outcomes, effects on other teachers' behavior and the allocation of resources within the teacher team or school. Understanding these intermediate impacts is important for explaining how TTI influences the internal dynamics of schools as well as for interpreting the impacts on student achievement and teacher retention.

The evidence on the strategic assignment of students to teachers as a result of TTI was mixed. We hypothesized that one way principals could react to an intervention like TTI was to

assign more challenging students to the transfer teachers on the assumption that their high performance meant transfer teachers were more capable of teaching struggling students. We did not find evidence of impacts on student assignment when we examined administrative data that described the characteristics of students assigned to treatment and control focal teachers relative to their nonfocal counterparts. We also found no evidence of impacts on student assignment when we examined principals' reports of how they assigned students to classrooms in treatment versus control teams. However, treatment focal teachers in middle schools were more likely than their peers in control teams to say they had more academically challenging students. Control focal teachers did not report differences between their students and their peer teachers' students.

We found evidence of reassignment of teachers across grades due to TTI. Another way that TTI could alter the school's internal dynamics is through resource allocation across grades. Under the status quo, principals might compensate for weak incoming teachers by moving strong peers from elsewhere in the school into their grade team. TTI may have the opposite effect: principals might move weak teachers into the grades with TTI teachers.

Using experience as an indicator of teacher quality, we found support for this hypothesis of pairing high-performing teachers with inexperienced ones. Among nonfocal teachers who had moved within their schools, the control movers were more experienced than treatment movers (by a statistically significant difference of nearly five years).

TTI teachers used less mentoring and provided more mentoring than their counterparts. We found that treatment focal teachers were less likely to receive mentoring (39 versus 59 percent had a mentor) and more likely to provide mentoring than control focal teachers (15 versus 5 percent provided mentor support). We defined a mentor as "someone who provides professional advice and direct assistance to classroom teachers." Because teachers in TTI positions were using fewer mentoring resources than the new hires on control teams, more resources were potentially available in TTI schools for supporting other teachers. However, changes in mentoring services used and provided by focal teachers were not offset by equal and opposite changes for nonfocal teachers. This opens the possibility that resource-allocation effects could spread to the larger school community, beyond just the teacher teams for which we collected data.

Test-Score Impacts: Did TTI Raise Student Achievement?

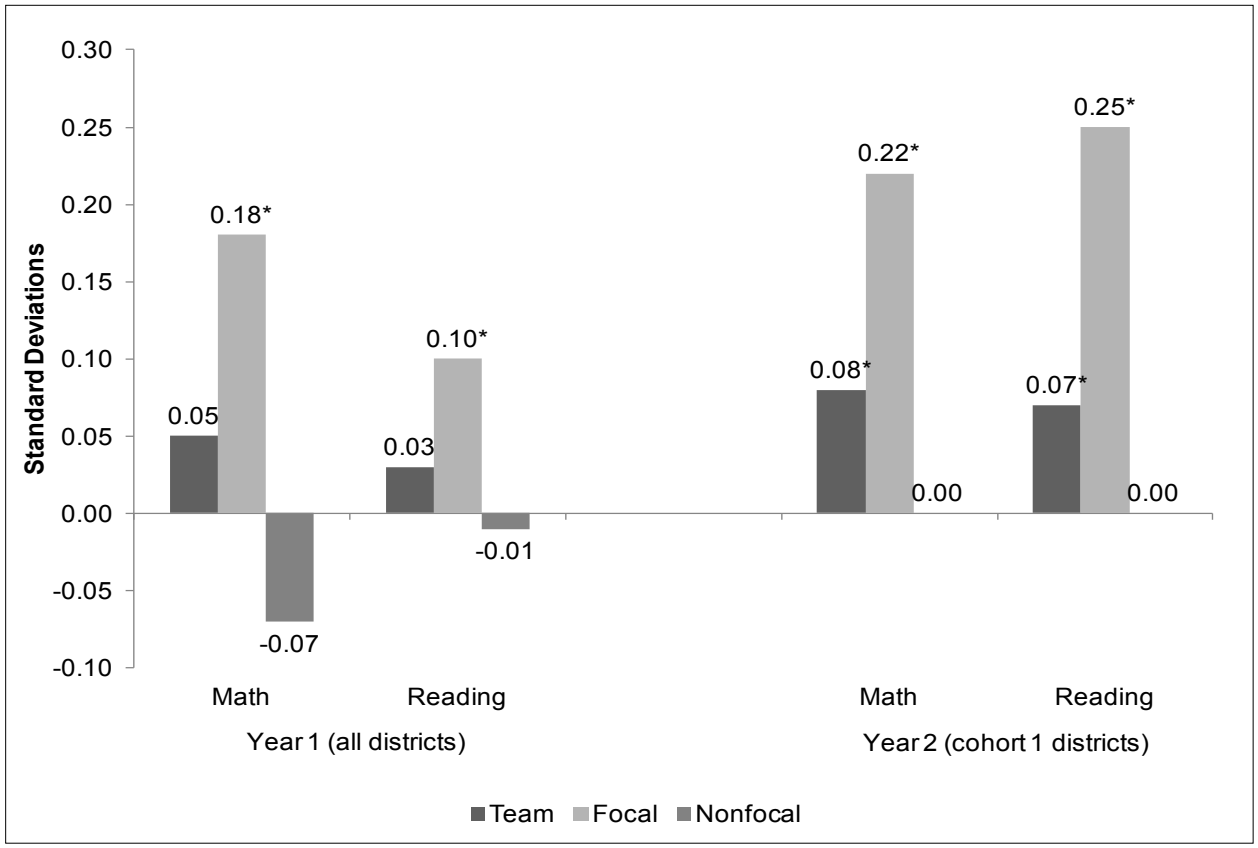
To estimate the impact of TTI on student achievement, we compared the test-score performance of students from treatment teams to the corresponding performance of students from control teams because the teacher team is the unit of random assignment. However, we expected much of the effect to be captured directly by comparing the performance of students who were taught by teachers who filled treatment vacancies (treatment focal teachers) against the performance of students who were taught by teachers who filled control vacancies (control focal teachers).

It is possible that the presence of a TTI teacher could have an effect on the team composition or the performance of other team members (nonfocal teachers). Therefore, we also report both results of the corresponding comparisons between treatment focal and control focal teachers and between treatment nonfocal and control nonfocal teachers within those teams. Program year 1 impacts were estimated using data from all 10 districts; year 2 impacts were estimated based on

cohort 1 districts because second-year follow-up data for the three cohort 2 districts were not collected.

TTI elementary school teachers had positive impacts on test scores. Comparing teams as a whole (see Figure ES.3), there were positive impact estimates for both subjects, but they were significant only in the second year of implementation: 0.08 standard deviations for math and 0.07 standard deviations for reading. The impact estimates for focal teachers ranged from one-tenth to one-quarter of a standard deviation, depending on subject and implementation year, and were positive and statistically significant for both subjects in both implementation years. The impacts on nonfocal teachers were not significantly different from zero in either year or subject. This suggests that the TTI teachers have minimal or no effect on their colleagues' performance.

Figure ES.3. Test-Score Impacts in Elementary Schools



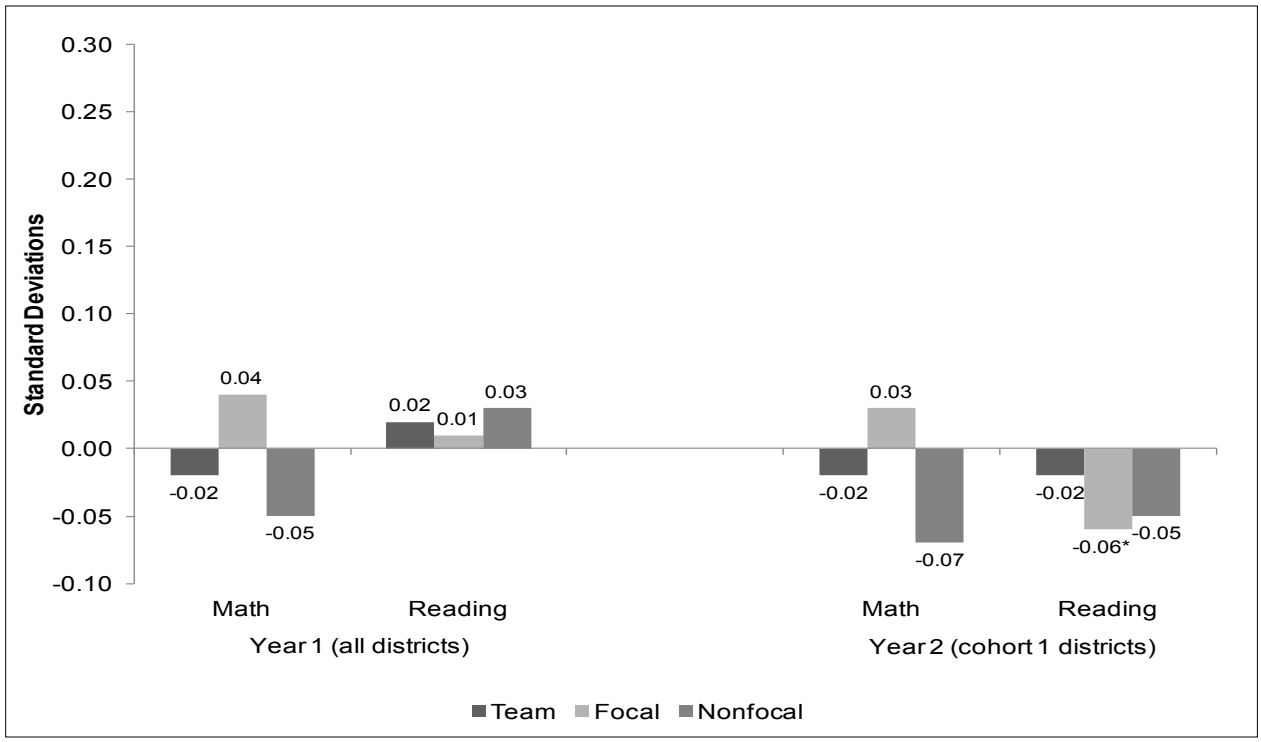
Source: District administrative data.

Note: A team consists of all classroom teachers in the grade and subject for a school. Focal teachers are those who filled study vacancies. Nonfocal teachers are the rest of the teachers on the team.

*Statistically significant at the 0.05 level, two-tailed test.

We did not find evidence that TTI was effective in middle schools. The results are shown in Figure ES.4. The impact estimates were all statistically insignificant for program years 1 and 2 except for the year 2 focal teacher impact on reading, which was negative (impact = -0.06, *p*-value = 0.031). This finding may be a middle school phenomenon or, as will be discussed next, it may be a result of the particular districts where middle schools were most heavily represented.

Figure ES.4. Test-Score Impacts in Middle Schools



Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

Impacts varied by district. The impact estimates for the team comparisons and the focal teacher comparisons varied across districts more than would be expected with sampling variation if TTI were equally effective in all sites. For example, elementary team-level district-specific impacts on math in program year 1 ranged from -0.25 to 0.48 of a standard deviation; elementary focal teacher district-specific impacts on math in program year 1 ranged from -0.15 to 0.57 of a standard deviation in program year 1. The district-specific impacts are based on small samples and most are not statistically significant, but their distribution suggests that neither the team nor focal teacher impacts are driven by results from one or two outlier districts. Also, the math and reading impacts by district are positively correlated.

These results suggest that although the variation in impacts across different grade spans may be due to real differences between elementary and middle schools, they may also be partially driven by differences in district-specific impacts because the shares of elementary and middle school teams differed across districts. One district contributed only elementary school teams to the study; two districts contributed only middle school teams; the remaining seven districts had different mixes of elementary school and middle school teams.

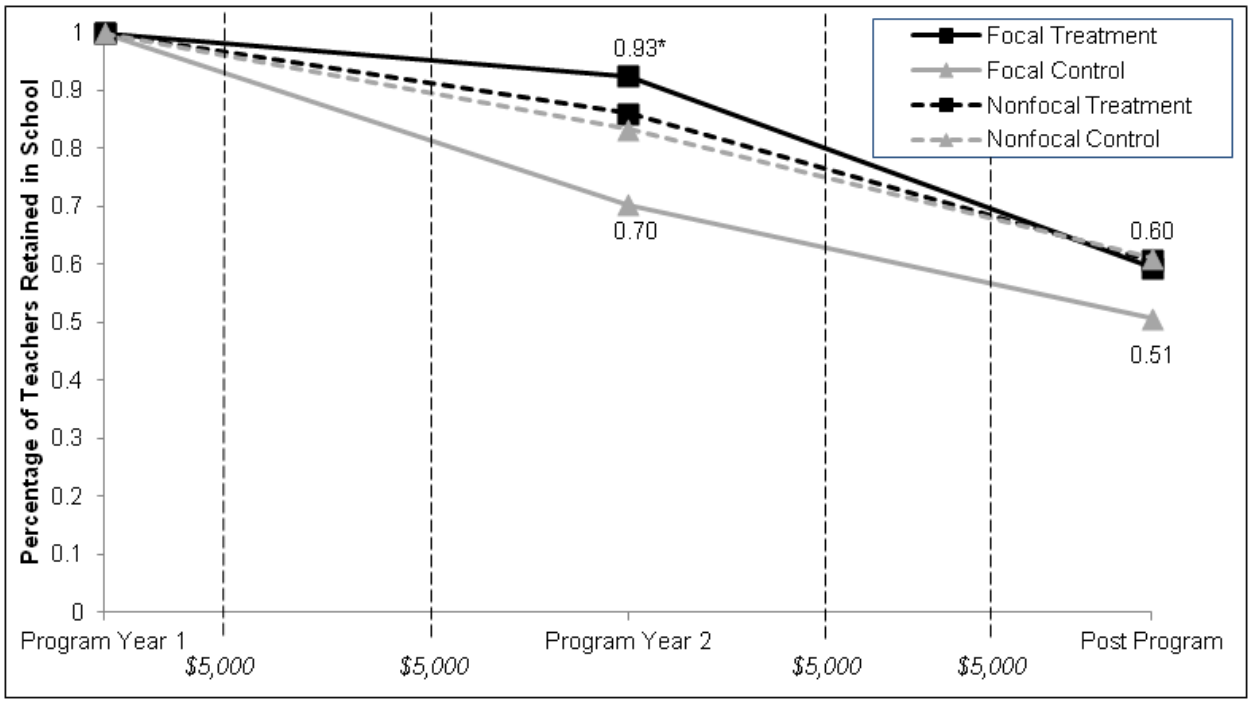
Retention Impacts: Did TTI Keep High-Performing Teachers in Their New School?

TTI teachers were offered $20,000 in five installments over a two-year period to transfer to and continue teaching in a low-achieving school in their district. Above, we discussed teachers' initial responses to the offer of a transfer incentive. Another important question for policymakers is whether the incentive would be sufficient to keep teachers at the schools into which they transferred. We examined the rates of teacher retention on treatment and control teams using teacher rosters collected during the program—while TTI teachers were still receiving incentive payments—and after the incentive payments ended.

Over the two years of the intervention, retention rates were higher for TTI teachers than for their counterparts. This finding is based on cohort 1 districts, in which impacts can be estimated after one and two years of the program (Figure ES.5). After the first year, when TTI teachers were still receiving payments for remaining at their schools, the difference in school retention between treatment and control focal teachers was 22 percentage points.[5] During the same period, the difference in school retention between treatment and control nonfocal teachers was not statistically different from zero. We observed a similar pattern for focal and nonfocal comparisons in the full sample of cohort 1 and 2 districts after the first year.

Retention rates after the TTI intervention ended were not statistically different. We used the cohort 1 districts to measure longer-term retention. After the second (last) program year, in the fall after TTI transfer teachers had received their final stipend payment, about 60 percent of the original treatment focal group returned to the same school, which was statistically indistinguishable from the control focal teachers (Figure ES.5, Post Program). The end result was that TTI transfer teachers did not leave their schools—to return to their original schools or to transfer to any destination—at a rate that was higher than any teacher hired into such a school.

Figure ES.5. Impacts on Retention in School, Cohort 1 Districts Only



Source: School rosters.

Note: Vertical dotted lines represent points at which TTI teachers received payments. Note that the first $5,000 payment was actually paid in two installments of $2,500. One installment was paid before the start of the first school year and the other was paid in the fall of the first school year.

N = 80 focal treatment teachers, 96 focal control teachers, 193 nonfocal treatment teachers, and 183 nonfocal control teachers.

*Statistically significant at the 0.05 level, two-tailed test.

[5] This impact estimate is calculated as the difference between the unrounded treatment and control means. The treatment and control means (0.93 and 0.70, respectively) presented in Figure ES.5 are rounded means.

Cost-Effectiveness

We showed that in at least some settings TTI had positive impacts on test scores, and after the two-year study period when the payments ended, treatment group teachers had not all left, but returned to their schools in year 3 at rates that were similar to their control group counterparts. Thus, the question arises as to whether the impacts are large and meaningful enough to offset the costs of generating them.

To provide a point of comparison, we compared the cost of generating the impacts of TTI with the costs of generating similar impacts if we were to implement an alternative policy, such as reduction in class size. We first estimated the incremental cost per team of implementing TTI. Then we estimated what it would have cost in terms of class-size reduction to generate those same impacts using results from the Tennessee STAR Class Size Reduction (CSR) experiment. We then subtracted the actual costs incurred per team from the estimated costs per team that would have been incurred using CSR to generate the same impacts as TTI. A positive number suggests that TTI was cheaper, and, therefore, more cost-effective.

The results of this cost-effectiveness analysis suggest that the impacts of TTI found for elementary schools would make it the cheaper alternative, compared to CSR, by $13,154 per team. Considering the long-term benefits of having high-performing teachers remain in the school could make TTI the cheaper alternative by an estimated $40,043 per team.

We found less-favorable results, however, in middle schools and in selected districts, so we cannot say that the cost comparison will always favor TTI if replicated. When we repeated the cost-effectiveness analysis based on replicating the full intervention in both elementary and middle schools, we found differences in the cost of TTI and the alternative intervention to be both negative and positive, depending on the assumptions we made. This implies that it is unclear whether TTI is more cost-effective overall for all grades and districts.

I. INTRODUCTION

In this final report, we present findings on the implementation and impacts of an intervention that identified school districts' highest-performing teachers and then used monetary incentives to encourage them to transfer into the lowest-achieving schools. The intervention, described in more detail below, was implemented in 10 school districts throughout the country. The study used random assignment to form equivalent groups that either had the chance to participate in the intervention or did not. In a report that preceded this final report (Glazerman et al. 2012), we presented implementation and intermediate impacts for 7 of the 10 districts that began implementation in 2009, which we refer to as cohort 1 districts. The other 3 districts—cohort 2—entered the study in 2010.

A. Policy Problem: Unequal Access to High-Performing Teachers

There is growing concern that the nation's most effective teachers are not working in the schools with the most disadvantaged students (Goldhaber 2008; Peske and Haycock 2006). Much of this concern is driven by research examining the disparity in such teacher characteristics as experience or certification, which are viewed as proxies for teacher effectiveness (Presley et al. 2005; Lankford et al. 2002; Education Trust 2008; Clotfelter et al. 2006; Carroll et al. 2000). These studies show that schools that serve a high proportion of low-income or minority students are more likely to employ novice teachers and teachers who lack certification in their subject area than are schools serving fewer disadvantaged students. Teacher characteristics and teacher effectiveness, however, are not equivalent. The link between teacher characteristics and student achievement has not been well established (Rivkin et al. 2005; Gordon et al. 2006; Rockoff et al. 2008; Buddin and Zamarro 2008).

More recent analysis has focused on teacher effectiveness in the classroom, as measured by student achievement growth. Specifically, concern has focused on the way that effective teachers are distributed across schools with higher and lower proportions of disadvantaged students. The measures of effectiveness are referred to as value-added estimates because they seek to describe the contribution that teachers make (the value that they add) to student achievement growth, holding constant factors that are outside the teacher's control, such as student background and prior learning (McCaffrey et al. 2004; Lipscomb et al. 2010). The question then becomes whether there are gaps in teacher effectiveness between schools serving higher- and lower-income students. In particular, the aim is to address concern about underrepresentation of high-value-added teachers in schools serving disadvantaged students, as measured in terms of poverty, low achievement, or other factors.

Evidence on teacher-effectiveness gaps, as measured by value added, is just emerging, but, with some exceptions, researchers are finding that the distribution of teacher effectiveness tends to favor schools with lower poverty levels. For example, one study that used 2005–06 data from Tennessee showed that schools with higher percentages of low-income and minority students had fewer of the most-effective teachers and more of the least-effective teachers as measured by its value added assessment system (Tennessee Department of Education 2007).

Another study, using data from North Carolina and Florida in the early 2000s, found that the average teacher effectiveness was slightly lower in high-poverty schools, though the differences were small for some subject-state combinations (Sass et al. 2012). However, the study did find that high-poverty schools had an overrepresentation of the least-effective teachers.

A third study was based on data that largely overlap with the sample used in the current study. Glazerman and Max (2011) examined the prevalence of districts' highest-performing teachers (in terms of value added) in elementary and middle schools in 10 school districts. The authors found that, on average, schools with the most disadvantaged students had a significantly lower percentage of teachers who were in the top 20 percent of the performance distribution as measured by value added.[6] This was true at the middle school level whether student disadvantage was measured by income or prior achievement; it was significant at the elementary school level only when prior achievement was used as the measure.

More recently, an analysis of the distribution of Los Angeles Unified School District (LAUSD) by the Education Trust-West found that low-income students were more likely than other students to have a low-value-added English/language arts (ELA) teacher or math teacher and less likely to have a high-value-added teacher (Hahnel and Jackson 2012).

In addition to the aforementioned research reports, journalistic accounts have compared the distribution of high-value-added teachers by school to the school or region's poverty levels. In Washington, DC, the rank ordering of the city's eight council wards was the same for presence of top-performing teachers as it was for income (Turque 2010). In New York City, *The New York Times* overlaid a map of schools by neighborhood poverty and by prevalence of high-value-added teachers and found the same relationship, suggesting a teacher-quality gap in that city as well (Fessendon 2012).

Policymakers at the federal, state, and local levels have considered a range of options to help struggling schools attract and retain highly effective teachers. One goal of such policies is to improve disadvantaged students' access to top teachers. The strategies include alternative teacher preparation and certification programs, recruitment bonuses for serving in hard-to-staff schools or subjects, intensive mentoring and professional development, and performance-based pay. The strategies have been implemented with federal funds and with funding from state, local, and nongovernment sources. Some of these policies have been implemented in the context of research studies to gauge their effectiveness, but experimental evidence to date does not suggest that any one of these strategies significantly raises student achievement in the U.S.[7]

B. One Policy Response: Selective Transfer Incentives

One strategy to address unequal access to top teachers is the use of monetary recruitment incentives that are designed specifically to move teachers with high-value-added performance to low-achieving schools. Past efforts have tried to use incentives to attract teachers with selected qualifications in Alabama, California, Tennessee, and Virginia (Max et al. 2007), but none has focused exclusively on value added. The U.S. Department of Education's Institute of Education Sciences (IES) has sponsored a study, summarized in this report, that tests the effectiveness of an

[6] Discussion of "significant differences" here and throughout refers to statistical significance. We use a 0.05 significance level, which means that a significant difference is highly unlikely (less than 5 percent of the time) to be observed in a sample if the population difference was zero. Statistical significance does not imply that the difference is necessarily meaningful to policy, nor does a lack of statistical significance imply that the difference is not meaningful for policy.

[7] See, for example, Constantine et al. (2009) on alternative teacher preparation, Glazerman et al. (2010) and Garet et al. (2008) on intensive mentoring and professional development, and Springer et al. (2010) on pay-for-performance.

intervention adopting this strategy. The intervention, known to participating school districts as the Talent Transfer Initiative (TTI), identifies a district's highest-performing teachers using value-added analysis, and it offers them an incentive of $20,000 to transfer to any one of the district's low-achieving schools targeted for the intervention.

IES contracted with Mathematica Policy Research to design an intervention that uses this strategy as well as to oversee its implementation and conduct a rigorous evaluation of its impact. The implementation was carried out in collaboration with the participating school districts by Mathematica's subcontractor, The New Teacher Project (TNTP). The program design and study design were reviewed by a technical working group of experts in the fields of teacher compensation, value-added analysis, and program evaluation. First, the broad parameters of the intervention—such as the method of identifying high-performing teachers and the size and form of the incentives—were defined. Then, TNTP developed, in consultation with participating districts, most of the operational details, including the timeline, school and teacher recruitment strategies, and communication plan. The resulting intervention design is described next.

1. Overview of the TTI

The intervention was designed to proceed as follows. The first step was to identify the highest-performing teachers, defined as the top 20 percent in each district and grade-subject pool, based on a value-added measure.[8] Highest-performing teachers were identified by their value-added scores both because some of the study districts were already using value added as a measure of teacher performance and because pay-for-performance policies like TTI are more likely to use value-added scores as performance measures. We used whatever value-added measure the district was using because that is what would have been used in the absence of the study; in cases where it was not available, we calculated it ourselves (see Appendix B for details).

We identified teachers as highest performing within three separate pools: middle school math teachers, middle school ELA teachers, and elementary multiple-subject teachers. The value-added measures relied on at least two years (and typically three years, depending on district data availability) of student achievement growth data for each teacher, where the data for each year consisted of a post-test from the end of the current school year and a pre-test from the end of the previous school year. This analysis was typically completed between January and March of the calendar year in which implementation began. This corresponds to the school year just before transfer teachers would start in their new schools. The amount of time required for the value-added analysis varied greatly, depending on the availability and quality of data on student test scores, demographics, enrollment, and student-teacher links. Value-added performance was the only criterion, although each district approved the list of highest-performing teachers by verifying that none had disciplinary actions or other serious concerns.

[8] As mentioned, value-added analysis is the statistical approach that tries to determine the unique contribution each teacher makes to student achievement, holding constant factors that are outside the teacher's control. The specific value-added model used in this study is discussed further in Appendix B.

The second step was to classify schools as "potential receiving" or "potential sending" schools. Potential receiving schools are those with the lowest achievement in the district and which the district leaders intend to help through the intervention. Low-achieving schools (those in approximately the bottom 20 percent in terms of average test scores) were targeted because it is a measure of disadvantage that high-performing teachers are in the best position to address.[9] The rest of the schools in the district, with rare exceptions for schools that were exempted because they were already receiving a comparable intervention or served a special population of students, were designated as potential sending schools.

The program's third step was to invite high-performing teachers in potential sending schools to apply to transfer; identified teachers received $20,000 over two years if they transferred and stayed in one of the targeted positions in a receiving school. Once they applied, they had to interview, receive a job offer, transfer, and remain in the position. The process began in the spring by inviting the highest-performing teachers (referred to as transfer candidates) to an information session where they were recognized for their past accomplishments and asked to consider applying to transfer as a way to help students who could benefit most from their talent. Within each district, a site manager followed up with transfer candidates individually to encourage them to apply for a transfer position as part of TTI. The site managers in the TTI study were employees of TNTP who coordinated their outreach activities with each district's human resources department.

At the same time that the site manager was recruiting transfer candidates, he or she also recruited receiving schools. The site manager worked with potential receiving-school principals by hosting an information session and conducting follow-up phone calls, inviting them to submit vacancies for consideration. Once principals volunteered a vacancy for consideration, its eligibility was confirmed by the evaluator and assigned by the evaluator to the program.[10] The site manager then performed a matchmaking function, assisting both receiving-school principals and transfer candidates by setting up interviews.

To qualify for the additional compensation, applicants had to interview with and be accepted by the principal at the receiving school and then voluntarily transfer. Ideally, the offers and acceptances were to be finalized by early summer.

Finally, the teachers who transferred were given a half-day orientation by TNTP in each district as well as the first installment of their bonus ($2,500) just before the start of the school year. Because the transfer teachers had been selected based on their strong classroom performance, they typically required no additional formal support beyond what teachers normally receive, but a TNTP staff person was assigned to provide informal support and answer questions during the first two school years. Teachers who remained in the positions into which they transferred received incentive payments at the end of each semester—in December and

[9] To define a low-achieving school, all schools were ranked by their average test scores or accountability rankings, and the TTI team worked with the district administrators to identify schools among the lowest scoring that were not already participating in a program that was similar to TTI. Targeted grades and subjects, typically multiple-subject classrooms in grades 3 through 5, and math and ELA classrooms in grades 6 through 8, were those in which standardized tests were administered in the current and prior grade.

[10] As the evaluator, we considered the grade and subject and whether the vacancy was in or adjacent to a teaching team that had already been assigned.

June—for the duration of the intervention. Only the transfer teachers themselves and their principals were part of the communication regarding their status as bonus recipients.

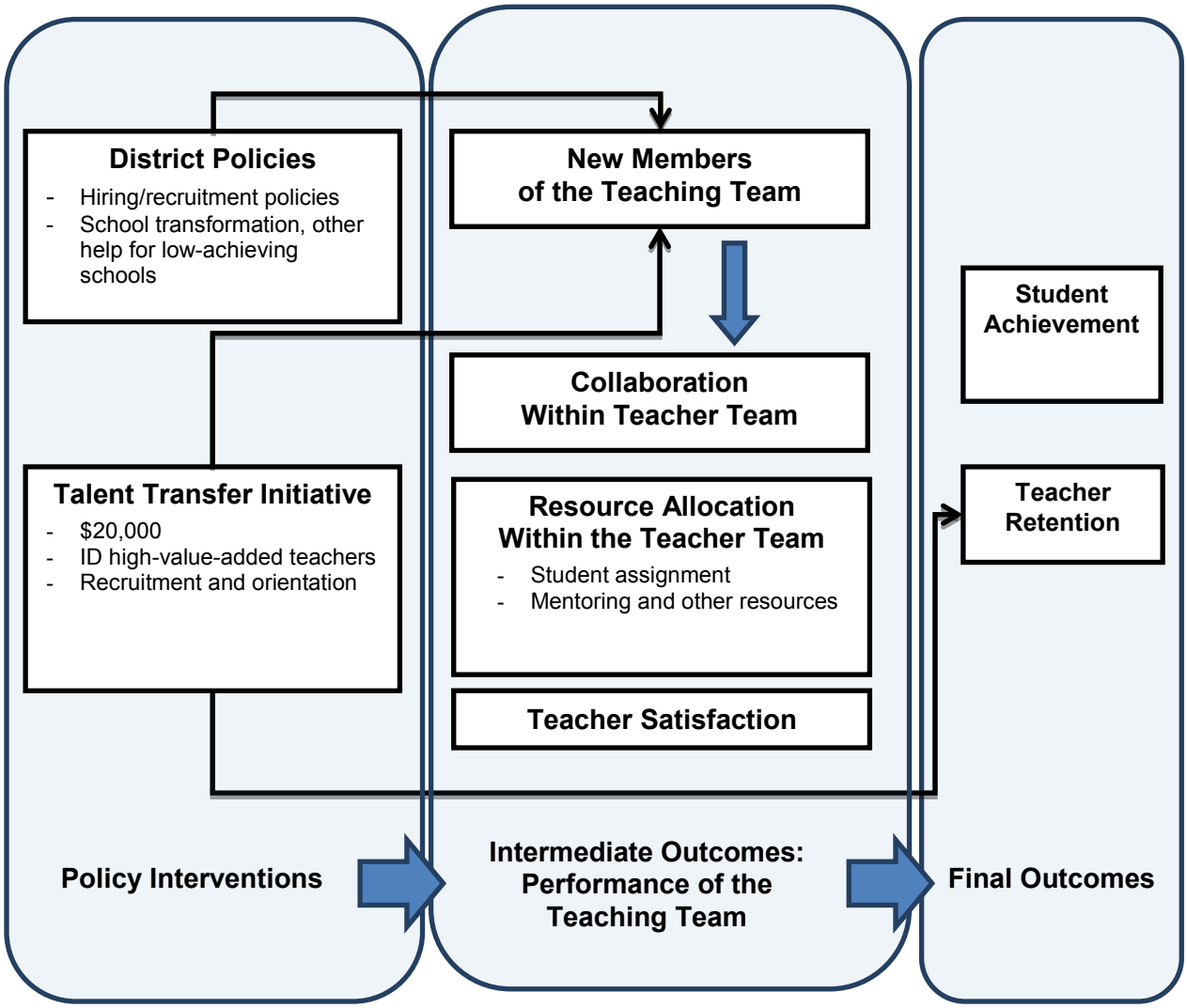
Not all of the districts' highest-performing teachers had been teaching in the potential sending schools (those that are high achieving and from which the highest-performing teachers are eligible to transfer). Although the transfer-incentive program is predicated on the idea that low-achieving schools need more high-performing teachers, it recognizes that some of the most effective teachers may, in fact, already be working in the potential receiving schools (those that are not high achieving). A critical component of the transfer-incentive program is a provision whereby teachers in the highest-performing group who were already teaching in low-achieving schools would automatically qualify for a retention bonus of $10,000. This was true regardless of whether the school or grade team was in the study or what its study status was. This bonus was also paid in installments over two years, as long as the qualifying teachers remained in their schools. Communication was private. Only bonus recipients were notified of their status.

2. Logic Model: How Can Teacher-Transfer Incentives Affect Student Achievement?

We hypothesized that TTI would improve teacher retention and the achievement outcomes of students on participating teacher teams through a series of pathways depicted in a logic model in Figure I.1. We defined the teaching team as all the teachers in the same grade and subject. In elementary schools, the team consisted of all classroom teachers in the grade. In middle schools, the teacher team consisted of all the math or ELA teachers who taught at least one class in the grade level of interest. For example, all teachers responsible for teaching 7th-grade math in the same school would make up one team. All teachers in the school who were responsible for 8th-grade ELA were considered another team.

The first component of this model suggested that TTI would operate in the context of a set of existing district policies aimed at helping raise teacher performance at low-achieving schools with teaching vacancies. The existing policies could include transfer rules, transfer incentives, signing bonuses, or other recruitment and hiring strategies. They could include a range of more general school-improvement strategies, such as school turnaround; class-size reduction; curricular changes; or changes to the working environment, such as increasing teacher induction and mentoring, or hiring a new principal.

Figure I.1. Logic Model: How Transfer Incentives Affect Teachers and Students in Receiving Schools



Note: A teacher team consists of all teachers in the same school, grade, and subject(s).

In this context, TTI represents a new intervention that uses the unique tools of identifying the highest-performing teachers using a value-added model and then offering monetary incentives ($20,000) to encourage the identified teachers to transfer to low-achieving schools. The process of identifying teachers using value added has its own pitfalls, which we consider in a separate part of the model that will be explained in the next section.

We hypothesized that several components of the transfer-incentive intervention would influence the probability that a high-performing teacher would transfer. These included the criteria used to identify transfer candidates (the value-added model and the cutoff), the size of the incentive, and how the incentive would be offered (for example, with a concerted recruiting effort that appeals to candidates' sense of duty).

All of these program factors would combine with other factors, such as the attributes that make sending schools more or less desirable as places to remain, those that make the receiving schools more or less desirable to transfer into, the match between the teachers and the principals

in the sending and receiving schools, as well as the relative desirability of commuting to the sending versus receiving school. If a high-performing teacher does not fill an identified vacant position, that vacancy may be filled by a teacher who is new to the school, the district, or the profession, or one who moves from within the school. Although unlikely, the new teacher could be highest performing. Another possibility is that the vacancy could be lost because student enrollment declines or the teacher who had planned to leave, thereby creating the vacancy, changed his or her plans.

It is important to note that hiring a TTI teacher was voluntary, not mandatory, for schools participating in the program. That means we did not seek to estimate the impact of a particular teacher forced into a particular school setting. Instead, we estimated the impact of *offering* a principal the opportunity to fill a vacancy from a pool of candidates identified in a particular way (using value-added analysis). This effort amounted to estimating the impact of whichever eligible teachers, if any, happened to apply for, successfully transfer into, and remain in targeted teaching positions. The main way that TTI is hypothesized to have an impact on final outcomes such as student achievement is through intermediate outcomes—specifically, by improving the performance and satisfaction of the teaching team that is targeted by the intervention (i.e., the team that had the opportunity to fill a vacancy with a TTI teacher). The most direct impact of the transfer incentive is the improved quality of the teacher who fills the vacancy. We refer to that teacher's impact on his or her students as the direct impact of the intervention. However, we consider all teachers in the team because whoever fills the vacancy can have indirect effects on students and teachers in the same school and grade.

One type of indirect effect is the altered collaboration that can occur within the teacher team. For instance, an experienced, higher-performing transfer teacher might help a junior colleague improve lesson planning. Research on student achievement gains in North Carolina suggests that such teacher-peer effects can be substantial and lasting (Jackson and Bruegmann 2009). Alternatively, peer teachers might resent or feel jealous of a teacher who is receiving a large bonus for doing the same job they are doing. This could have a negative impact on satisfaction, school climate, teacher retention, and student achievement.

In addition to direct and indirect effects, we hypothesize that TTI can have resource-allocation effects. In other words, altering the mix of teachers can lead to changes in the way resources—for example, mentoring and coaching—are allocated within the teaching team. If TTI were to result in an experienced, accomplished teacher filling a vacancy in a hard-to-staff school, a literacy coach in the school might have more time to spend with the existing teachers in the team than he or she would have had if the new member had been new to the profession.

Other resource-allocation effects might result if principals assign teachers to courses or students differently if they have a different type of teacher filling a vacancy on the team. Because students vary in the academic or behavioral challenges they present, and teachers vary in their ability to address different types of student needs, principals might deploy teaching resources differently in response to a transfer incentive.

The other final outcome to measure is retention of highest-performing teachers. A transfer incentive initiative like TTI would have its most direct impact on retention through the phase-in aspect of stipend payments. The $20,000 incentive paid out in installments over a two-year period encourages teachers to stay to collect payments. The other way a transfer incentive will affect teacher retention is by improving satisfaction with and attachment to the principal and

school among all teachers in the team. Therefore, it is important to measure retention both during the two-year commitment period over which the payments are made and also after that period.

In the logic model depicted in Figure I.1, we provide the overview of how TTI might influence intermediate and final outcomes, but there could be weak links in the causal chain that might hinder success. The main consideration is the quality of the teachers identified as transfer candidates. The two assumptions underlying the transfer-incentive intervention are (1) that teachers who have been identified as highest performing based on value-added analysis will continue in future years to generate large learning gains, and (2) that they can continue to be effective in their new settings after they transfer, particularly in comparison to teachers who would ordinarily be hired by low-achieving schools.

However, any measure of performance would be an estimate based on current or recent teaching wherever that teacher had been assigned. That makes it especially challenging to know whether a strong estimated performance in the years leading up to the determination of a teacher's status as a highest-performing teacher will predict performance in a subsequent year, in another setting, or with new students. Nevertheless, Xu et al. (2012) used non-experimental methods to estimate the portability of teacher skills by measuring North Carolina teachers' value added and following the movers between higher- and lower-poverty schools. They adjusted for regression to the mean expected for unusually high- or low-performing teachers and concluded that teachers were no less effective when switching to higher-poverty schools (or vice versa). The current study performs a similar analysis, except that the moves from higher- to lower-achieving schools will have been induced by the policy intervention.

The logic model presented here has several implications for studying the transfer incentives. First, the model suggests that one should pay attention to such factors as the assignment of students to classrooms and the degree to which the amount of mentoring and other supports teachers receive varies within grade-level teams. Second, the model suggests that the intervention may have an impact at two levels of aggregation. The first is the team level, which captures all of the components that feed into the total effect: direct, indirect, and resource-allocation effects. The second is the teacher level, specifically, the high-performing teacher who transferred into that teaching team.

The randomized study design, to be discussed later in this chapter, focuses on the opportunity given to some schools to use an incentive to hire a high-performing teacher. This is not the same as studying the impact of the transfer teacher because schools that have the opportunity to hire such a teacher might not necessarily do so; the transfer is voluntary on the parts of both the TTI teacher and the receiving-school principal. To capture the difference between, say, forced transfer and transfer incentives, we would study all teaching teams where a vacancy was designated as eligible for the transfer incentive, regardless of whether it was actually filled by a TTI teacher. Thus, the total impact of a transfer incentive is a weighted average of the impact of transfers plus the impact of teachers who filled vacancies outside the TTI mechanism despite the presence of the incentive.

C. Studying Teacher-Transfer Incentives

1. Research Questions and Study Design

We address the following research questions:

- What can we learn from the **implementation** of TTI? Specifically, what can we learn about timing and scale of implementation, who transfers, and from where they transfer?
- What were the **intermediate impacts** on receiving schools? Specifically, how did TTI affect the dynamics within the school, such as the allocation of resources, staffing patterns, assignment of students to teachers and courses, and school climate?
- What was TTI's **impact on student test scores** in receiving schools?
- What was TTI's **impact on teacher retention** in receiving schools?

The impact questions refer to the effect of the transfer-incentive policy relative to the absence of such a policy. In other words, we sought to measure effects relative to the outcomes that would have been realized had the school not had the opportunity to use the $20,000 incentive to fill its vacancy with a teacher designated as highest performing.

To answer the impact questions, we implemented a randomized controlled trial in which we compared outcomes within teacher teams that had the chance to fill a vacancy with a TTI teacher who would receive the $20,000 incentive for transferring to outcomes for comparable teacher teams that filled a teaching vacancy through whatever route the school would normally follow.[11] The study focused on elementary schools (which included mostly self-contained classrooms in grades 3 through 5) and middle schools (which included departmentalized subject-specific classes in grades 6 through 8). We describe the random assignment process below.

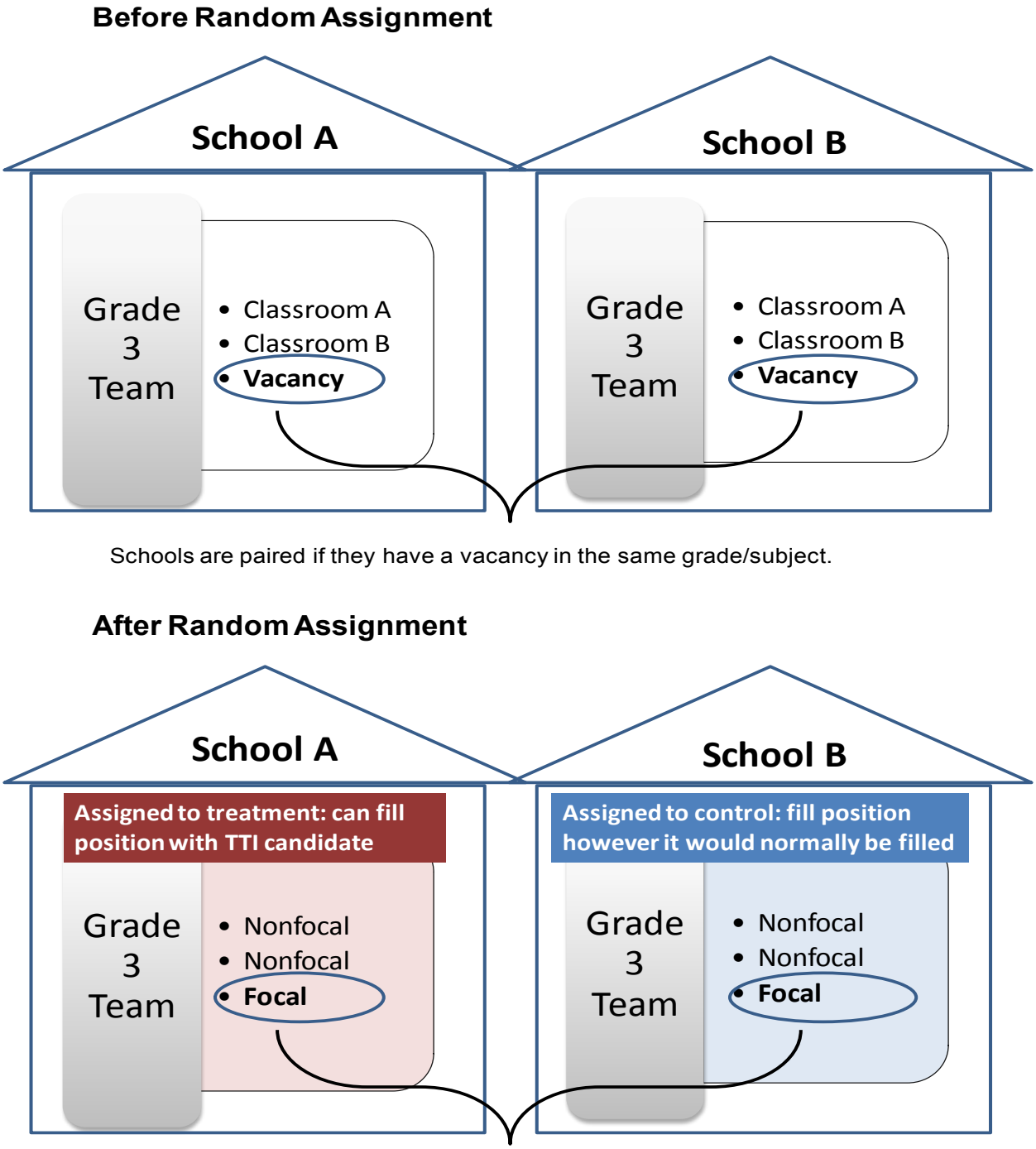
Teacher teams as the unit of analysis. The unit of analysis for the intervention and for random assignment was the teacher team (a group of teachers in the same school, grade, and subject).[12] To be eligible for the study, the teaching team had to have at least one vacancy in a tested grade.

The TTI site managers gathered information on teaching vacancies from potential receiving-school principals and provided lists of teacher teams with vacancies to Mathematica for random assignment. In Figure I.2, we provide a simplified example: two participating receiving schools each have a teaching vacancy in grade 3 (top panel). We assigned the 3rd-grade team in School A to either the treatment (TTI) or control group based on its random number and assigned the 3rd-grade team in School B to the opposite status (see grade 3 in the bottom panel). It was possible for teams to have more than one vacancy. Of the 165 teams we assigned, 151 (92 percent) had a single study vacancy, and the rest had two or three.

[11] The $10,000 retention incentives were offered to all qualifying teachers whether they were in a team assigned to treatment, to control, or not assigned to the study at all. The experimental comparison focuses on the impact of the transfer incentives only.

[12] Teachers who taught students in more than one grade could be on more than one team. For example, a middle school math teacher teaching both 6th and 8th grades could be on both 6th- and 8th-grade teams.

Figure I.2. Random Assignment Study Design



Within study teams, "focal" teachers are those who filled the vacancies. Although teams were the unit of random assignment, we hypothesized that the impact of TTI operates primarily through the teacher who fills the vacancy—termed the focal teacher. In the treatment group, the focal teacher was typically a TTI transfer. In the control group, the focal teacher was whoever the principal found to fill the position. In Chapter III, we will describe in detail the teachers who actually filled the positions. The remaining teachers in the team are called "nonfocal."

Stratified randomization in batches. The study team conducted random assignment within districts in batches because it was necessary to assign vacancies as soon as possible—so principals could begin filling vacancies as soon as they opened. Specifically, we grouped teacher teams that had vacancies in the same grade levels but were in different schools within the same district. The teams were groups into blocks (typically pairs) that were matched according to subject, and, where possible, by such school characteristics as the student achievement ranking and the percent of students eligible for free or reduced-priced lunch (FRL). This matching, which took place before random assignment, was done to improve statistical efficiency.

Assigning teams in the same school. To further strengthen the study design, we took advantage of the possibility that some pairs of schools in the same randomization batch would have eligible teams at more than one grade level. In such schools, we assigned teams in such a way as to ensure that each school in the pair had one treatment and one control team. For example, two participating receiving schools each could have a teaching vacancy in both grades 3 and 5. In such a configuration, we assigned the grade 5 teams to be the mirror image of the grade 3 random assignment status, so each school had both a treatment team and a control team. An illustration of this procedure along with more details on random assignment can be found in Appendix A.

For the vacancies in teams assigned to the treatment group, the principals were offered the opportunity to interview and hire teachers who we previously identified as highest performing (TTI candidates). The teachers who were offered jobs through TTI and accepted them were eligible to receive a $20,000 stipend over the course of two years. For the vacancies on teams assigned to the control group, the principals filled the vacancies as they normally would.

This random assignment process created two groups that, on average, should have the same student characteristics and school contexts. They differ only in the opportunity to participate in TTI. Comparing outcomes for these groups generates unbiased estimates of the impact of TTI on student achievement and teacher retention.

Control group contamination, sometimes called "spillover," is a possible concern with any study design in which treatment and control teams coexist in the same school. This can occur if a student has one teacher in the treatment group (for reading, for example) and another in the control group (for math, for example). In such cases, the math scores could be higher than we would have observed in the absence of treatment because a very strong reading teacher might help the student do well in all subjects. In other words, the treatment effect could spill over to the control group and contaminate the experiment. It can also occur if teachers collaborate across grades and/or subjects.

We sought to limit both forms of contamination by imposing the following rules on the random assignment process. In elementary schools, treatment and control teams in the same school had to be separated by at least two grade levels. This ensured that no elementary student had a teacher from both a treatment and a control team over the study's two-year period. In middle schools, teachers were sometimes responsible for classes in more than one grade level, so we required that treatment and control teams in the same school be in different subjects (math or ELA) and also be separated by at least one grade. This ensured that no student was taught the same subject by a teacher from both a treatment and control team. There was, however, the possibility that a student had a teacher from a treatment team for one subject and a teacher from a control team in the other subject because of cross-grade teaching that we discovered during the

study. This same-grade, opposite-subject overlap was possible only in 5 out of 114 schools in the study, and was likely to have occurred, based on known teacher assignments, in only 3 out of those 5 schools. In terms of students, fewer than 2 percent of cases were affected in the middle school analysis. No case was affected in the elementary school analysis.

Another way to avoid contamination was to force teams into the same treatment status by assigning them together if they did not meet the previously described adjacency rule. For example, if there were vacancies in grades 3 and 4 in an elementary school, both vacancies were assigned to the same treatment status. If there were vacancies in 6th-grade math and 8th-grade math in a receiving middle school, both were assigned to the same status. Some teams had more than one vacancy in a single study team. In such cases (14 out of 165 teams), all vacancies within the team were assigned to a common study status because teams were the unit of random assignment.

We assigned a total of 165 teams in 114 schools. There were 68 schools with one team, 42 with two teams, and 4 with 3 or more teams.

2. Survey Data

We have addressed the research questions through analysis of survey and administrative records data as well as program-implementation records. In each of the 10 districts participating in the study, surveys were conducted with teachers who were transfer candidates regardless of whether they actually transferred; teachers in teams with vacancies, including both TTI ("treatment") teams and control teams of teachers; and their principals.[13]

Candidate Survey. In fall/winter 2009, the Candidate Survey was administered to the teachers eligible for the study.[14] The survey helped us characterize the background of teachers identified as highest performing and provided information about the factors affecting teachers' willingness to apply to the TTI, to interview at low-achieving schools, and, ultimately, to transfer. The survey also served as a way to gather information about teachers' experiences during the hiring process. The Candidate Survey was repeated in fall 2010 for the three cohort 2 districts where candidates were identified in 2010.

Teacher Background Survey. This survey was administered in late winter/early spring of the first program year (2009–10 for cohort 1; 2010–11 for cohort 2) to all teachers in the study who filled one of the vacancies in treatment or control teams and to their colleagues in the same teaching team. It collected information on teachers' experiences at the study schools, along with information about their educational and professional backgrounds and other factors that could affect their students' achievement.

Principal Survey. The Principal Survey was administered in spring of each of the two program years to obtain data on teacher recruitment and hiring, principals' assessments of the teachers hired in the study's target grades, and any redistribution of resources across classrooms

[13] The survey instruments are available at http://edicsweb.ed.gov/browse/downldatt.cfm?pkg_serial_num=4024.

[14] Teachers eligible for the study were the highest-performing teachers in each of the 10 districts who were not already teaching in low-achieving schools or in schools that were exempted from the program by the district. Candidates who left the district before being notified of the program opportunity were also excluded.

(including those related to the arrival of the new hire). The first-year survey collected information about hiring during the period of the TTI transfers. The emphasis in the follow-up survey was on teacher performance and school environment.

We obtained response rates of 81, 77, 90, and 82 percent on the Candidate Survey, Teacher Background Survey, Principal Survey (program year 1), and Principal Survey (program year 2), respectively. The response rates for all surveys were similar for treatment and control groups (see below). For each of the surveys, we conducted nonresponse analysis to describe the respondent samples and the degree to which each resembled the full population of respondents and nonrespondents. The response rates are listed in Appendix A, Table A.3. More information on the characteristics of respondents and nonrespondents for each of the surveys is provided in Appendix A, Tables A.4–A.8. To account for the possibility of nonresponse bias resulting from an overrepresentation or underrepresentation of certain groups, we controlled for group characteristics in all analyses using these data.

3. Administrative Data

Teacher roster data. To compare retention rates of teachers in their new schools to those of existing teachers in the same schools, we collected teacher rosters for all study schools in the fall of each of the program's two school years, and, for cohort 1 districts, did so again in the fall of the third year, after incentive payments were no longer being made.

Student test score and background data. Districts provided student-level administrative data that capture school enrollment, student test scores, student demographics, course scheduling, and student-teacher links. The test-score outcomes are grade-specific state assessments in math and reading.[15] The 10 study districts were distributed across seven states. All pre-test and post-test scores were converted into standard deviation units, or z-scores, that express student achievement relative to the average performance for the student's own grade statewide.[16] The districts provided demographic information on students' race/ethnicity, gender, ELL status, special education status, FRL, gifted status, and age. We recoded the demographic information so they were coded consistently across districts.

Program implementation data. Districts provided information on teachers' value added-performance that was used to identify transfer candidates and retention bonuses. Some districts provided more information than others; the research team used all available data. In cases where Mathematica conducted the value-added analysis, we had detailed information. In cases where value-added analysis was conducted by a third-party vendor hired by the district, we were given scores, or, in one case, just names of highest-scoring teachers. Districts also provided data on school-level student achievement that was used to determine which schools were eligible to

[15] One of the 10 study districts administered different tests within the same grade and subject (algebra and pre-algebra in grade 8). For that district only, we used a linking procedure to convert pre-algebra scores to predicted equivalent scores on the algebra test. See the "Test Score Scaling Issues" section in Appendix F for details.

[16] The z-scores are calculated by subtracting the statewide mean scaled score for all students in that year and grade from a student's scaled score and dividing that by the statewide standard deviation of scaled scores for that same group. Z-scores greater than 4 or less than -4, which make up less than 0.1 percent of the sample, are assumed to be data errors and are set to missing. Z-scores are not equivalent to development scale scores.

participate as potential receiving schools. The TTI site managers provided principal consent forms and information on the timing of teaching vacancies and when they were filled.

In the remainder of this report, we describe the implementation, impacts, and cost-effectiveness of TTI. An interim report (Glazerman et al. 2012) focused on implementation and intermediate impacts in the 7 districts that began implementation of TTI in 2009. This report includes the experience of all 10 districts, follows the first 7 districts into the second year of implementation, and includes estimates of impacts on student test scores and teacher retention.

II. THE STUDY SAMPLE

In this chapter, we describe the study sample, including how the school districts and schools were selected and recruited into the study. We also describe the policy environment in which the study operated and the process of selecting and randomly assigning schools to either a treatment or control group. Finally, we discuss the teachers and students who were included in the study. It is important to note that we often refer to teacher teams. These are defined as groups of teachers responsible for the same subject in the same grade in a given school. For example, all 3rd-grade teachers in a given elementary school are on the same team. All 7th-grade math teachers in a middle school might be considered to be on the same teaching team.

A. School Districts

1. How Districts Were Selected

We selected 10 large and economically diverse school districts in seven states. Large districts were necessary because (1) the intervention and the study were more likely to be feasible with a large pool of sending and receiving schools to consider, and (2) the intervention required a large pool of transfer candidates. The study required a sufficient number of schools to not only implement the intervention but also to form a control group. To be "large," districts had to have at least 40 elementary schools. We identified 59 districts that met this initial screening criterion. Economic diversity is also important, because the Talent Transfer Initiative (TTI) encourages the transfer of teachers from high-achieving to low-achieving schools and we hypothesized that these gaps would be most stark when there were large income disparities within the school system. Using free or reduced-price lunch (FRL) eligibility as a proxy for determining income status, we used the number of high- and low-poverty schools to determine whether a district had a sufficient mix of schools at different achievement levels.[17] Districts were deemed to be economically diverse if they had at least 10 low-poverty elementary schools (defined as having less than 40 percent of students eligible for FRL) and at least 15 high-poverty elementary schools (defined as having more than 70 percent of students eligible for FRL). Of the 59 districts that met the size criterion, 51 were deemed sufficiently economically diverse.

In addition to the quantitative criteria, we created a prioritized list of the 51 districts based on a variety of factors, including availability of test scores, data quality, hiring/transfer practices, and the local political environment[18] These factors would affect the feasibility of conducting the program in each district. Information used in this process was based on data gathered by the three organizations represented on the recruitment team: Mathematica Policy Research, The New Teacher Project (TNTP), and Optimal Solutions Group.

[17] In 9 of the 10 study districts, the correlation between school-level achievement rank and FRL rates ranged from -0.65 to -0.91. The correlation in the 10th district was -0.38.

[18] The local political environment could affect the feasibility of implementing the TTI program in a school district. We considered such factors as whether senior district leadership supported implementation of the program, whether the district could likely reach agreement with the local teachers' union or teachers' association, if needed, to implement the program, and whether the district had any budget issues that might affect the number of vacancies.

Of the 51 prioritized-district candidates, we recruited 10: 7 districts for cohort 1 (beginning in the 2009–10 school year) and another 3 for cohort 2 (beginning in the 2010–11 school year). We chose these particular districts by starting with the largest districts and attempting to contact them in the order they appeared on the priority list, excluding any that were unwilling to participate. We prioritized 19 districts to arrive at the sample of 10.[19]

2. Description of the Districts and the Study Context

To interpret the findings of this study, one must understand the context in which it operated. This includes the overall characteristics of the specific districts participating, the economic and geographic conditions facing teachers in those districts, and the compensation policies affecting the teachers.

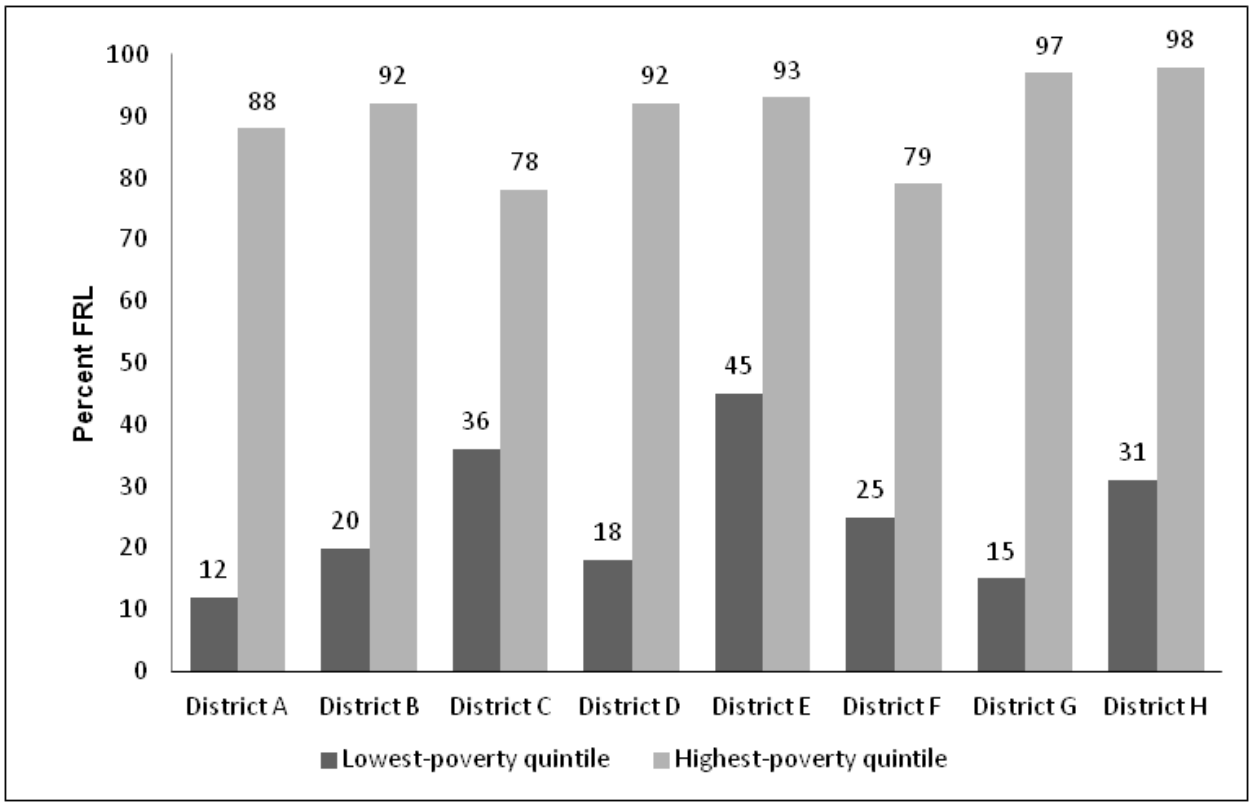
Diversity among and within participating districts. Although all of the study districts were large (with at least 40 elementary schools), they varied in size and student characteristics. Seven of the 10 had both elementary and middle schools in the study. One had only elementary schools participating, and 2 had only middle schools participating. Of the 8 study districts with elementary schools in the study, the largest in terms of elementary school enrollment had more than four times as many schools and four times as many students as the smallest. Across the 8 districts, 46 to 72 percent of elementary students were FRL eligible; the proportion who were African American or Hispanic ranged from 18 to 89 percent. The variation in district size is greater when we look at the 9 districts with middle schools in the study. The largest of the 9 in terms of middle school enrollment had more than 6 times as many middle schools and 12 times as many middle school students as the smallest. Across the 9 districts, 38 to 78 percent of middle school students were FRL eligible, and the proportion who were African American or Hispanic ranged from 20 to 92 percent.

It is also important to consider the variability in FRL rates among schools within each district, because economic diversity of schools was a district selection criterion. Within each district, we ranked all of the schools by the percentage of students FRL eligible, and divided the list into five equal-sized groups (quintiles) separately for elementary and middle schools. In Figure II.1, we show the spread of low-income students between the lowest- and highest-poverty quintile of elementary schools. For example, in District A, there was a 76-percentage-point difference in the FRL rate between the average schools in the top and bottom quintile of elementary schools. Other districts had gaps that ranged between 42 and 82 percentage points. In Figure II.2, we show the corresponding results for the 9 districts that had at least one participating middle school.

Employment landscape and geography of participating districts. Labor market conditions for teachers as well as each district's physical geography are important factors that could affect the implementation of a transfer-incentive intervention. To set the context for the current study, we examined these conditions in the 10 participating districts.

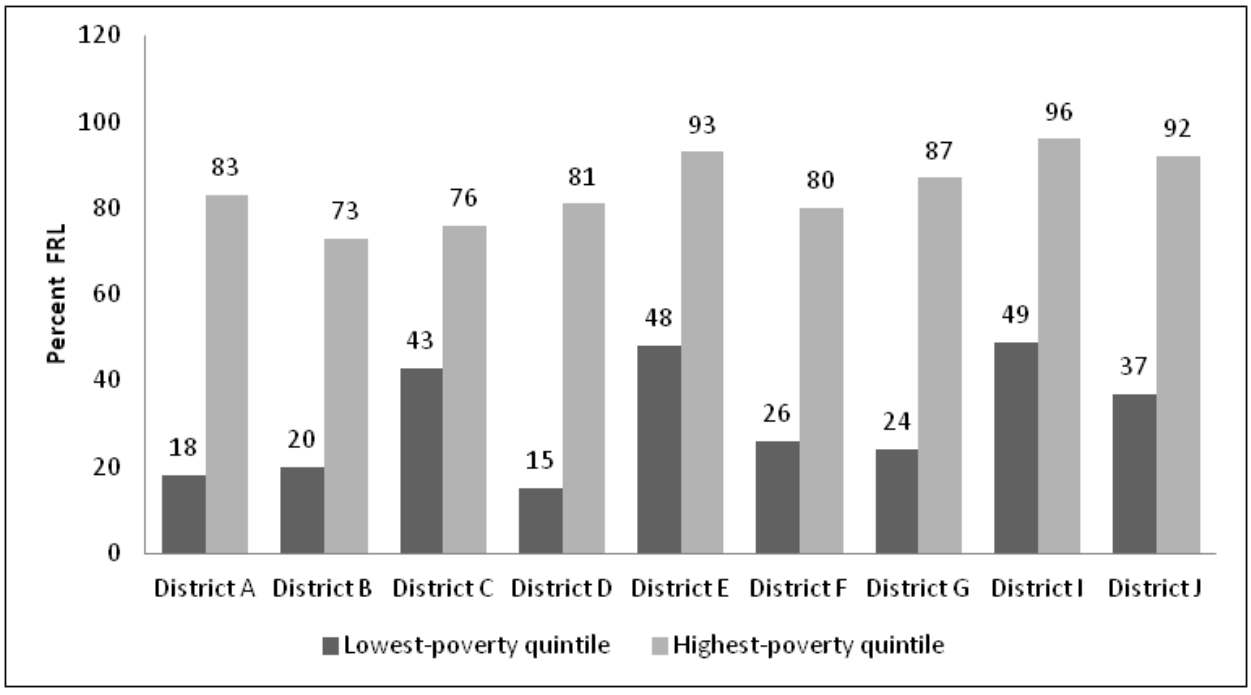
[19] Because the districts volunteered to participate, they should not be considered a statistically representative sample of those identified by the initial, quantitative selection criteria.

Figure II.1. Percentage of Students in Lowest- and Highest-Poverty Elementary Schools Who Are Low Income (FRL), by District



Source: National Center for Education Statistics Common Core of Data (http://nces.ed.gov/ccd/).

Figure II.2. Percentage of Students in Lowest- and Highest-Poverty Middle Schools Who Are Low-Income (FRL), by District



Source: National Center for Education Statistics Common Core of Data (http://nces.ed.gov/ccd/).

Eight of the 10 TTI districts are located in states with "right-to-work" laws, where each teacher must decide whether to join or pay dues to a union. Union strength is one aspect of the teacher labor market; another is the general unemployment rate. In 2008, before we began the study, the national unemployment rate was 5.8 percent. By 2009, the year when cohort 1 districts started recruiting their highest-performing teachers into TTI, the rate had risen to 9.3 percent. By 2010, the year when cohort 2 districts began full-scale implementation, the rate had risen to 9.6 percent. The local unemployment rates for each district followed a similar pattern.[20]

The geographic features of the districts are also relevant to understanding teachers' willingness to transfer, in part because of commute times. The land area of the largest of the 10 TTI districts is more than 1,900 square miles, which is larger than the state of Rhode Island (1,045 square miles).[21] The other 9 districts range in size from slightly less than 100 square miles to more than 1,200 square miles. Respondents to the Teacher Background Survey reported their average commute to school was 13.4 miles (21.3 minutes) each way.

Existing incentive programs in participating districts. We excluded school districts where existing or planned teacher-incentive programs would have duplicated the intervention under study, but we did encounter some existing policy initiatives in schools in some of the 10 participating school districts. These programs included performance incentives and signing bonuses for teachers. In each case, we determined that the existing programs were different enough, isolated to a few schools that were not in the study or that could be excluded from it, or involved small enough dollar amounts that they would not interfere with the study design. These incentive programs were funded by a variety of federal, state, and district sources. Teachers and schools receiving more than $5,000, an arbitrary threshold used to identify substantial bonus programs, were excluded from the study so as to reduce the likelihood of complicating the study by changing the effective incentive offered by the TTI intervention and the counterfactual.[22] We established the $5,000 threshold based on information in the literature on teacher responsiveness to pay (Max et al. 2007) that suggests this amount would plausibly influence teacher behavior. Only one district had an intervention very similar to TTI. It should be noted that a competing program has the potential to shrink the pool of transfer candidates, so the results of this study should be interpreted in light of the existence of these programs.

[20] From U.S. Bureau of Labor Statistics (http://bls.gov/lau/#tables) data, we found that the unemployment rates in TTI study districts ranged from 4.8 to 6.7 percent in 2008 and rose in every district in 2009, ranging from 7.2 to 13.2 percent. In 2010, the unemployment rates rose further in 9 out of 10 TTI study districts—to between 8.2 and 14.9 percent. Only one district experienced a fall in the unemployment rate: it dropped from 8.0 percent to 7.6 percent. For county districts, we used county-level unemployment rates. For all other districts, we used city-level unemployment rates.

[21] Source: U.S. Census Bureau, Census 2000 Summary File 1.

[22] We excluded such cases when we were aware of them at the outset of the study. However, six teachers (a mix of treatment nonfocal and control nonfocal) in two districts that were included in the study sample reported in the Teacher Background Survey that they were eligible for hiring or transfer bonuses greater than $5,000 in program year 1. Because they did not provide additional details on the nature of these bonuses, we could not determine if they were similar to TTI. We kept these six teachers in the study sample.

Bonuses, stipends, and additional payments received by teachers. When interpreting the study findings, it is important to relate the $20,000 incentive ($10,000 per year) offered to TTI transfer teachers to other payments offered to teachers for hiring, retention, or performance. As part of the Teacher Background Survey, we asked all teachers in study grades about the bonuses and incentives they were offered during the first year of the study.[23] We summarize in Figures II.3 and II.4 the responses of teachers in study grades. There are four groups of teachers in study grades presented in these figures: treatment focal teachers, most of whom received TTI transfer stipends; control focal teachers, who we believe filled the vacancies on teams assigned to the control group; and treatment nonfocal and control nonfocal teachers, who taught on the same teams as focal teachers but did not fill study positions and also did not receive TTI transfer stipends.[24]

By design, other bonuses in our sample that were similar to those offered by the TTI were relatively rare. In Figure II.3, we show that, with the exception of treatment focal teachers, fewer than 10 percent of respondents in each category reported being offered a hiring or transfer bonus. Across all categories, fewer than 15 percent of respondents reported being offered a retention bonus or a bonus to teach a particular grade level or subject.[25]

The average bonus amounts are summarized in Figure II.4. These averages include all teachers in each category, including those who received no bonuses. Treatment focal teachers are the only group that reported receiving, on average, more than $5,000 for any type of bonus. For those who received any type of bonus other than hiring and transfer bonuses, the amounts (not shown in the figure) were, on average, less than $6,000. The average hiring and transfer bonus amounts for nonfocal teachers (treatment and control) were greater than $5,000, but these averages are inflated by four teachers in two districts who reported bonuses of $20,000 or more. Excluding these four outliers, the average hiring and transfer bonus for nonfocal teachers was $3,425 (also not shown in the figure).

It should be noted that although 87 percent of treatment focal teachers received TTI transfer stipends of $10,000 in program year 1, only 47 percent reported having been offered transfer bonuses that year.[26] It is possible that some of these teachers considered their TTI stipends to be performance-based bonuses rather than transfer bonuses, as their past performance was emphasized in the recruiting stage. Indeed, of the 45 treatment focal teachers who reported not receiving transfer stipends, 20 percent reported receiving performance-based bonuses. Another possibility is that they thought that their TTI stipend was offered a year earlier because they were technically offered the stipend in the prior year even though they received it in program year 1, the year covered by the survey question.

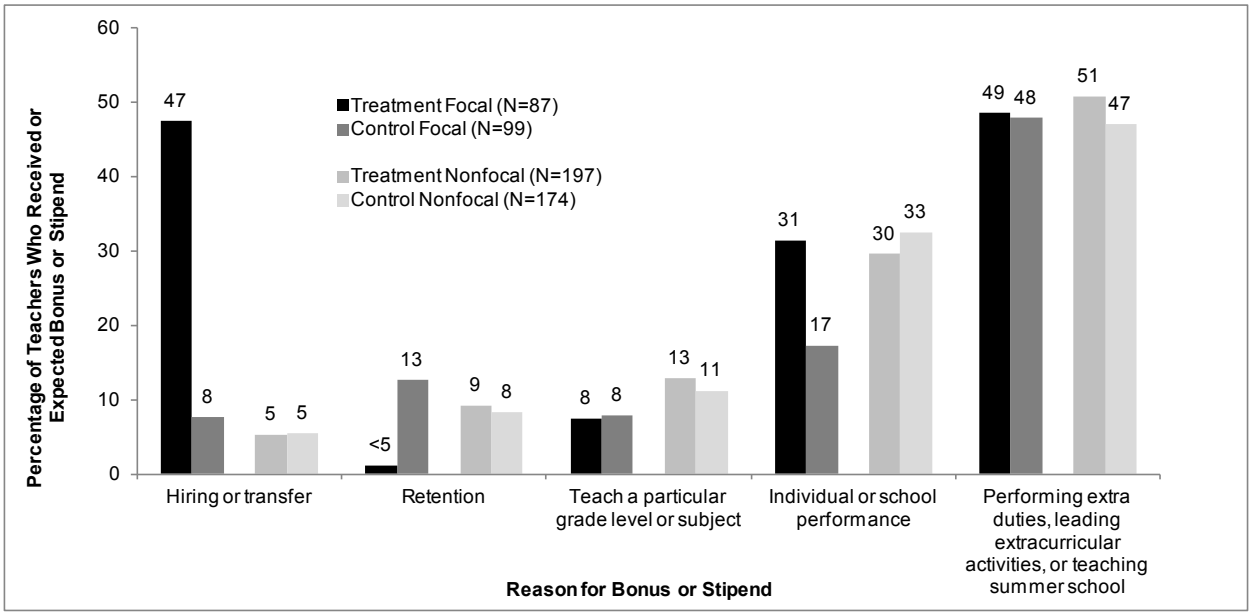
[23] The question asked teachers if they had received or had been *offered* various types of bonuses and stipends. Some teachers may have answered affirmatively that they were offered a bonus but did not actually receive it.

[24] As will be described in Chapter III, 30 nonfocal teachers on study teams were designated highest-performing at the start of the study and were eligible for $10,000 retention stipends. As with the transfer stipends, these retention stipends were paid out in installments over a two-year period.

[25] These numbers include any teachers who were receiving $5,000 per year in retention stipends through the TTI because they were deemed highest performing and were already teaching in low-achieving schools.

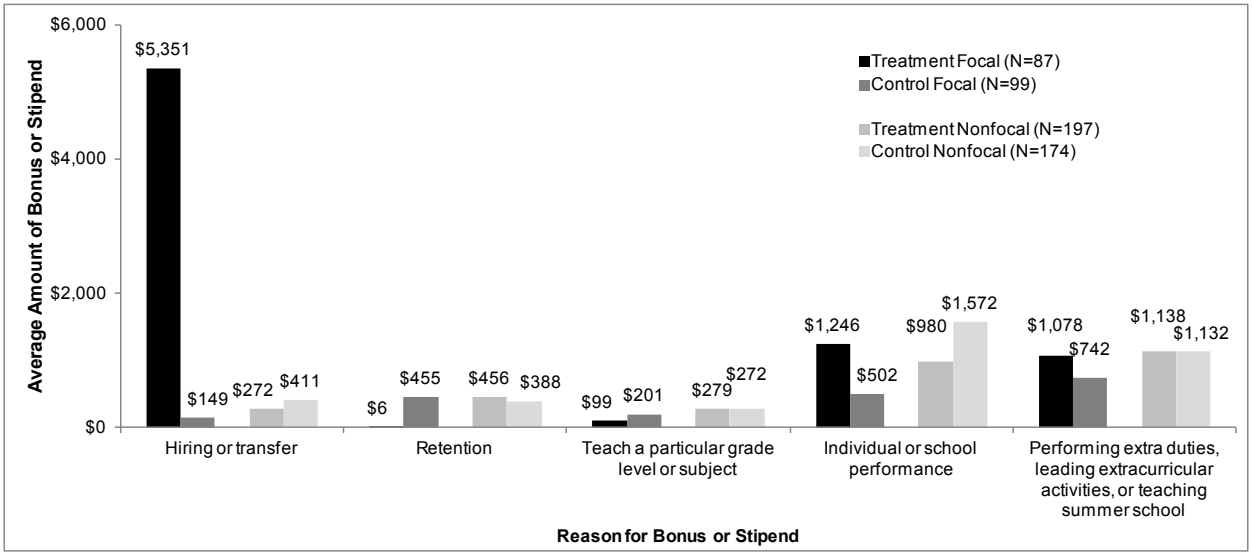
[26] Thirteen percent of treatment focal teachers either were non-TTI teachers or the teacher left before receiving the full $10,000 in the first year.

Figure II.3. Percentage of Teachers Offered Bonuses and Stipends



Source: Teacher Background Survey.

Figure II.4. Bonuses and Stipend Amounts Offered, Average Across All Teachers

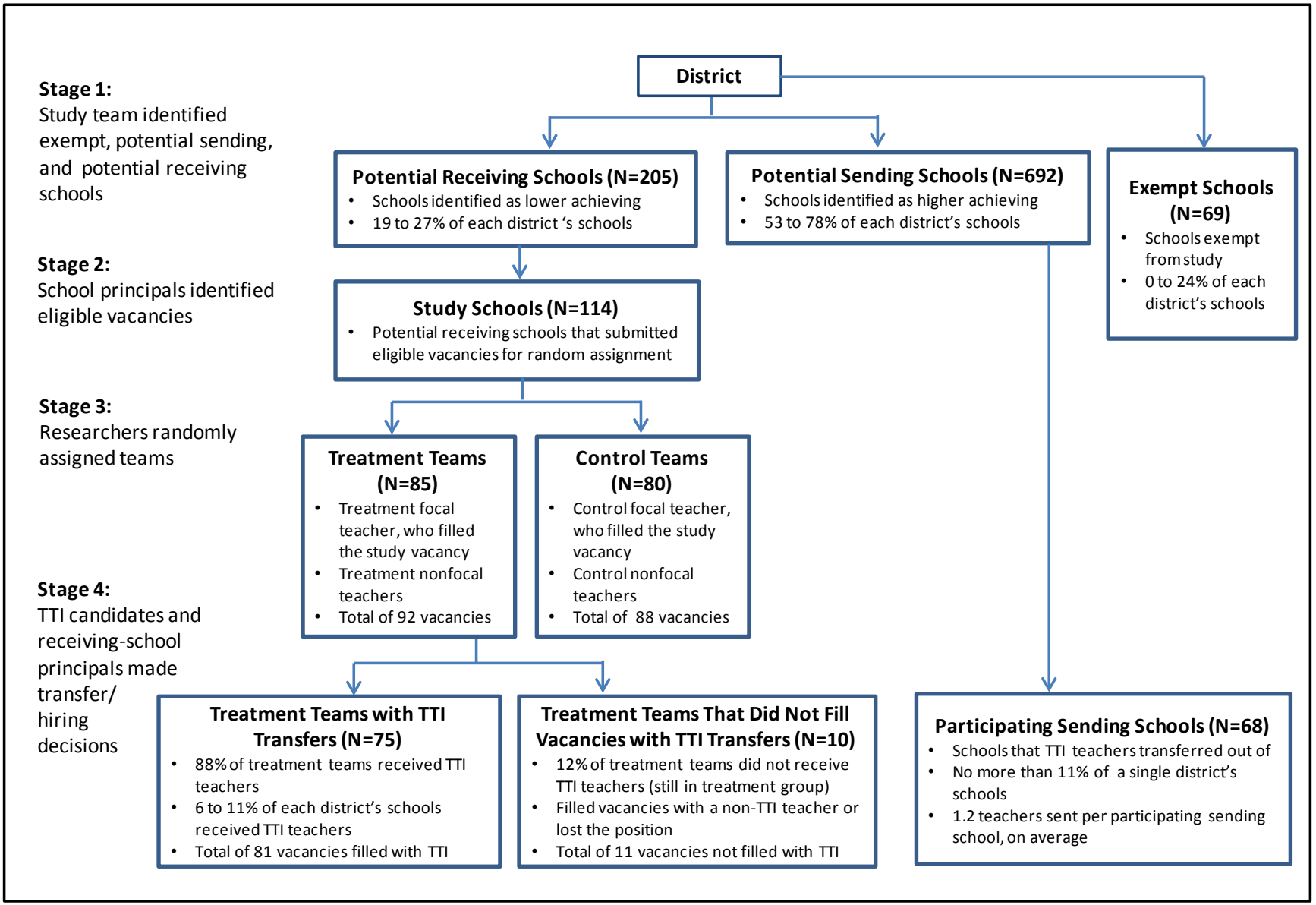


Source: Teacher Background Survey.

B. Schools

We partitioned the set of elementary and middle schools in each district into three groups: potential receiving schools, potential sending schools, and exempt schools. This was necessary because TTI was designed so that only the lowest-achieving schools (receiving schools) would be offered the opportunity to hire high-performing teachers through TTI. With the exception of schools that we exempted from TTI because of special circumstances, the remaining schools in the district formed the pool of potential sending schools. This process is shown in stage 1 in Figure II.5.

Figure II.5. Selection of Schools and Teams into the Study Sample



After describing the *potential* sending and receiving schools below, we go on to describe the random assignment process for the receiving schools with eligible vacancies. Finally, we describe the *participating* sending and receiving schools—in other words, the schools that transfer candidates actually left and the schools to which they moved or could have moved if they had been assigned to the treatment group.

1. Identifying Potential Sending and Receiving Schools

Before identifying sending and receiving schools, the districts listed schools that met our criteria for being exempt from being either a sending or a receiving school because they were already receiving a comparable intervention[27] or were primarily serving special populations.[28] In addition, one district chose to exclude low-achieving schools that had shown strong learning gains, and another district excluded schools already receiving resources through a program that targeted schools with low accountability ratings. Overall, districts participating in TTI excluded 69 schools from the program (7 percent of schools in eligible grade spans).

After exempting these schools, the primary criterion for potential receiving schools was that they must be low achieving (in approximately the bottom 20 percent). All but one district identified low-achieving schools based, at least in part, on average student test scores in the grades and subjects targeted by the program (math and reading in grades 3 through 8).[29] The district that did not use average student test scores used school accountability ratings instead, selecting all schools in the district below a cutoff, as well as the lowest-performing school in each local district within the larger district. Two other districts used accountability ratings in addition to average student test scores to identify low-achieving schools. The rankings based on average student achievement differed slightly from accountability ratings because the accountability systems included information on average achievement for student subgroups and on achievement in social studies and science. As a result, some schools in these districts were classified as potential sending schools based on their accountability rating even though they had lower overall average achievement than some potential receiving schools.[30]

[27] Adding the TTI bonus to existing bonuses would not only risk duplicating services, it would also complicate the study. It would alter the treatment and control conditions, making the treatment a larger incentive amount than in other districts and the counterfactual a different recruitment incentive, rather than no incentive. Consequently, schools already offering bonuses of $5,000 or greater were excluded from the sending and receiving pools.

[28] Special populations include primarily students who are blind or deaf, or students with severe learning disabilities. These schools often required teachers to have special training or certification and were not appropriate for TTI transfers.

[29] Achievement data from the year before implementation of TTI was used for all but two districts, where three prior years of achievement data were used.

[30] It would be theoretically possible in districts that use criteria other than overall achievement for a TTI teacher to transfer to a school with higher overall average achievement than at his or her sending school. In fact, it happened to three TTI teachers in districts that used accountability ratings instead of or in addition to average student achievement. This might serve to shrink the overall achievement contrast between sending and receiving schools, but these TTI teachers are still transferring to schools that were identified by the district as in need of high quality teachers. Such transfers are still in the spirit of the intervention and do not compromise the validity of impact estimates. These three teachers represent less than 4 percent of the assigned treatment vacancies and thus are unlikely to influence the overall impact estimates.

Working closely with the districts, the project team used the previously described criteria to rank schools in each district in terms of achievement. This was done separately for elementary and middle schools. To ensure consistency with the local school accountability systems, student achievement data came from the annual assessments used by these systems.

After ranking the schools according to achievement, the team set a cutoff in each district to identify low-achieving schools—which could receive TTI transfer teachers—and higher-achieving schools—which could send TTI transfer teachers. The cutoff between sending and receiving schools had to be carefully set in order to obtain both a sufficient number of vacancies in receiving schools and an adequate pool of eligible highest-performing teachers for transfer from sending schools. To achieve this trade-off, the project team drew upon the experience gained from a pilot study to obtain a sufficient number of vacancies and an adequate pool of eligible highest-performing teachers for transfer. Overall, 21 percent of schools (205 schools) across all 10 districts were identified as potential receiving schools, 72 percent (692 schools) were identified as potential sending schools, and the remaining 7 percent (69 schools) were deemed exempt.

Given the well-established correlation between family income and achievement (Reardon 2011; U.S. Department of Education 2013), we would expect that the lower-achieving potential receiving schools would be more disadvantaged than the higher-achieving potential sending schools. We found this to be true, using the percentage of students eligible for FRL as a measure of disadvantage. In elementary and middle schools, potential receiving schools had, on average, significantly more students eligible for FRL than potential sending schools. Race is also correlated with achievement outcomes (U.S. Department of Education 2013), and both elementary and middle receiving schools had, on average, significantly fewer white students and more African American students than sending schools (Table II.1).[31]

Table II.1. Demographic Characteristics of Potential Sending and Receiving Schools (percentages)

	Potential Receiving Schools	Potential Sending Schools	Difference
FRL			
Elementary	80	55	25*
Middle	79	61	18*
White			
Elementary	10	34	-24*
Middle	7	24	-17*
African American			
Elementary	46	26	20*
Middle	46	19	21*
Sample Size (schools)			
Elementary	130	440	
Middle	75	252	

Source: Administrative data.

*Difference is statistically significant at the 0.05 level using a two-sided test.

[31] In Chapter III, we include more detail on the achievement and poverty contrast between sending and receiving schools as well as the characteristics of TTI teachers' students at sending and receiving schools.

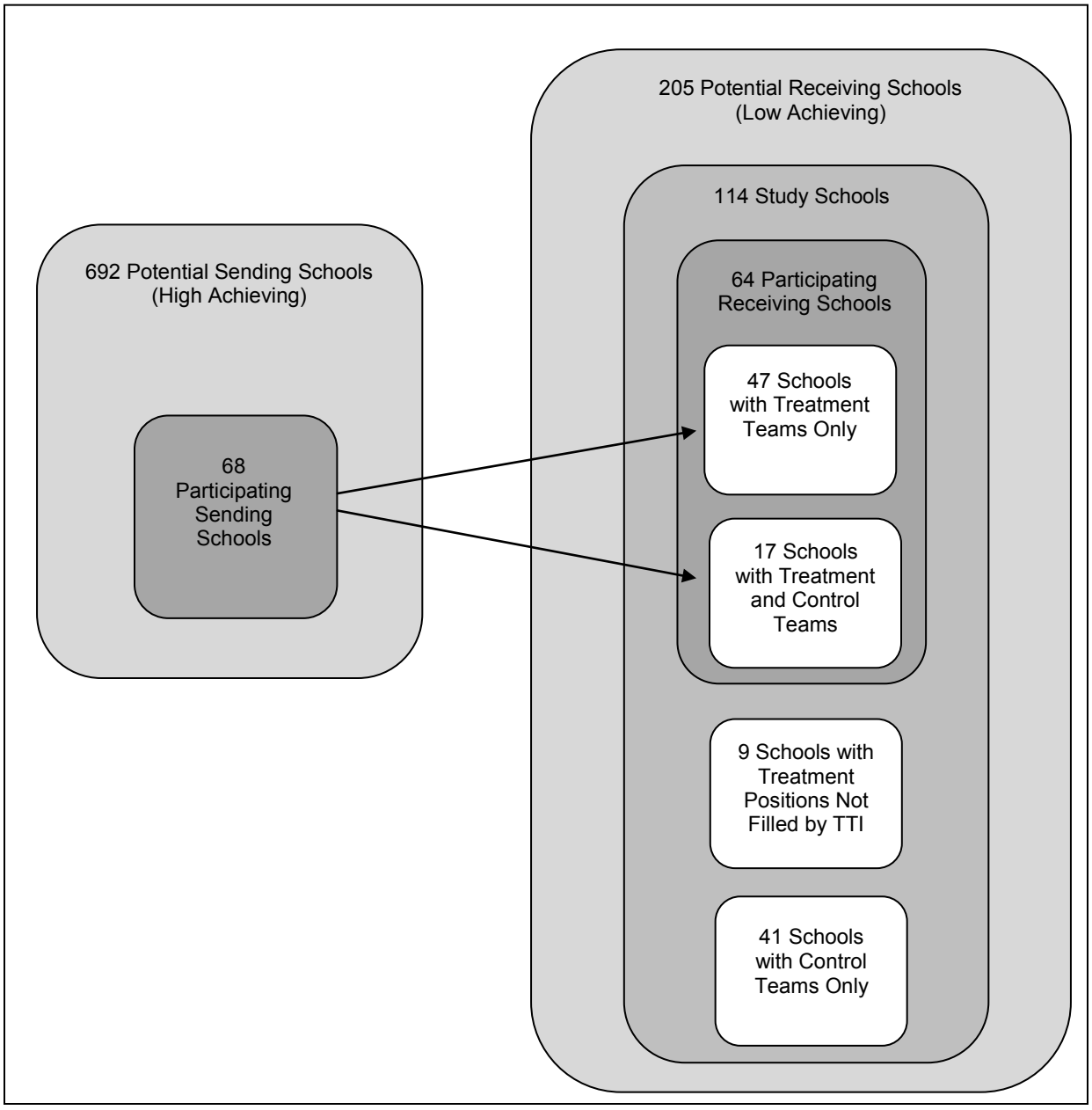
2. Defining "Participating" Sending and Receiving Schools

Not every potential sending school had a teacher transfer out through TTI. In Figure II.6, we show that, across the 10 districts, TTI transfer candidates came from 68 out of 692 (9.8 percent) of the potential sending schools (see stage 4 in Figure II.5). The percentage of affected potential sending schools is relevant for districts concerned that a transfer program like TTI might be disruptive to many of its higher-performing schools. In fact, more than 90 percent of schools identified as potential sending schools did not lose any teachers through the transfer program. This would obviously change if a similar intervention were implemented on a larger scale, but it does indicate that not every potential sending school has a qualifying transfer candidate, and that in schools that do, the candidates might not wish to leave.

Not every potential receiving school had a teacher transfer into it (see stages 2, 3, and 4 in Figure II.5). In Figure II.6, we show that the TTI teachers transferred into 64 of the potential receiving schools. Those 64 schools represent 88 percent of the 73 schools with a vacancy on a team that had been assigned to the treatment group. The 9 schools that had vacancies assigned to the treatment group but that did not receive any transfer teachers chose to fill their vacancies outside of the TTI pool or did not fill the positions with any teacher, due to a reduction in staffing requirements. The remaining 132 of the 205 potential receiving schools included both those with no vacancies submitted for random assignment (91 schools) and those where all of the vacancies submitted to the study had been assigned to the control group (41 schools).

In the rest of this report, we focus on "study teams," teacher teams within potential receiving schools that we randomly assigned to the treatment or control group. This study sample has teams from 114 schools (56 percent of the 205 potential receiving schools), as previously mentioned, including the 64 with TTI transfers, 9 with teams assigned to TTI but with no transfers, and 41 with only control teams. Treatment vacancies that were not filled through TTI are still considered to be in the treatment group in this study, following an intent-to-treat analysis approach. In other words, the treatment is defined as the *option* of hiring through TTI, whether or not a TTI candidate actually transferred in.

Figure II.6. Sending and Receiving Schools



C. Students

In random assignment studies, the treatment and control groups are expected to have similar characteristics at baseline, on average and as the sample size becomes larger, because the two groups are constructed using a lottery. In this study, baseline student characteristics in treatment and control teams are expected to be equivalent, on average, because students in both groups are from low-achieving receiving schools.

We examine baseline student characteristics at the team level because teacher teams were the units we randomly assigned. We examine them separately by subject (math and reading) but elementary school teams appear in both the math and reading analysis, while middle school teams appear in one or the other, depending on the team's subject. Therefore, we examine these

characteristics separately by grade span (elementary and middle school) as well as by subject (math and reading).

We report many results throughout this report separately by grade span because of the qualitative differences in their circumstances. TTI teachers in elementary schools were responsible for both math and reading instruction, whereas middle school teachers were hired to fill vacancies in either math or ELA, but not both. Middle schools are larger and more geographically dispersed than elementary schools. Also, the skills and certification needed to teach elementary versus middle school students may be different.

The study was originally designed to have samples large enough to detect meaningful impacts for the whole sample, not necessarily for each grade span separately, because we did not know how the sample would be allocated across grade spans. In our final analysis sample, the minimum detectable impact (MDI) on math scores for program year 1, for example, was 0.09 standard deviations. This means that we would be very likely to conclude the impact was significant if the true impact were at least this large (0.09 standard deviations, equivalent to moving a student from the 37th percentile in the state, where the control students began, to the 40th percentile).[32] If the true impact were smaller than that, it would be less likely that we could detect it. The corresponding MDI was 0.10 standard deviations for elementary school math scores and 0.16 standard deviations for middle school math scores. The larger MDI for middle school implies that we are less likely, all things being equal, to find significant impacts in middle schools. As a result, when we present findings by grade span, especially middle school, for which we have a smaller sample, a lack of statistical significance is not the same as a lack of impact. It means only that we had insufficient statistical power to detect the impact.[33]

We compared the pre-test scores and demographic characteristics of students from treatment and control teams for the math and reading analysis samples. All test scores in this report are based on standardized state assessments, and we express them in z-score units relative to the rest of the state.[34] In this way, a score of 0.0 represents a student who is average for his or her grade for the entire state in that year. A score of 1.0 represents a full standard deviation above the state mean, equivalent to the 86th percentile, and -1.0 would be a standard deviation below the state mean, equivalent to the 14th percentile.

In the elementary school math analysis sample, both groups of students had scores from their prior year that were below the state average by approximately one-third of a standard deviation in math and two-fifths of a standard deviation in reading. These scores are equivalent to the 37th percentile in math and the 35th percentile in reading. Differences between the treatment and control students' average prior achievement were not statistically significant (see Table II.2). However, racial composition was significantly different—treatment teams had higher percentages of Hispanic students (50 versus 44 percent) and lower percentages of African

[32] We used a significance level of 5 percent and a statistical power of 80 percent. This means that 5 percent of the time we would falsely reject the hypothesis of no impact and 20 percent of the time we would falsely fail to reject the hypothesis that there is no impact when, in fact, there was a real impact.

[33] For most impact estimates, we present the standard errors. These numbers provide a measure of statistical precision and can be used to calculate confidence intervals or MDIs. For example, the standard error can be multiplied by 2.80 to obtain the MDI with 80 percent statistical power at a 5 percent significance level.

[34] The study sample of 10 districts is drawn from seven states.

American students (37 versus 44 percent) than control teams. Treatment teams also had significantly higher percentages of female students, students with English language learner (ELL) status and high poverty status, as proxied by FRL.

Table II.2. Student Characteristics for Elementary School Math Analysis Sample, by Treatment Status

Student Characteristic	Treatment Mean[a]	Control Mean[a]	Difference	*p*-Value[b]
Prior Achievement[c]				
Math pre-test score	-0.33	-0.36	0.03	0.452
Reading pre-test score	-0.40	-0.39	-0.01	0.664
Other Characteristics (percentages)				
Male	49.5	51.4	-1.9	0.039*
Race/ethnicity				
White	7.2	5.9	1.3	0.201
African American	36.8	43.7	-7.0	0.009*
Hispanic	50.4	44.0	6.4	0.014*
ELL	35.6	30.9	4.7	0.026*
Special education	6.4	7.3	-0.9	0.102
FRL	77.4	74.7	2.6	0.029*
Sample Size (students)	4,236	3,941		

Source: Administrative data.

[a]Treatment and control means are within-block means. Random assignment was carried out by blocks.

[b]Standard errors of treatment-control differences are adjusted for clustering at the team level.

[c]Test scores are reported in standard deviations relative to the state average.

*Difference is statistically significant at the 0.05 level using a two-sided test.

In Table II.3, we show the same treatment-control comparison of student characteristics for the elementary school reading analysis sample. Treatment-control differences follow the same pattern as those presented in Table II.2 except gender composition is not significant (*p*-value = 0.070) in this case. For elementary schools, the math analysis sample (presented in Table II.2) and the reading analysis sample (presented in Table II.3) are nearly the same samples of students.

The middle school results are shown in Tables II.4 and II.5 for the math and reading analysis samples, respectively. The middle school math analysis sample (summarized in Table II.4) yields the same conclusions as the elementary samples—no significant difference in test scores, but statistically significant differences in demographic characteristics. The reading-analysis sample is balanced in terms of demographic characteristics, but students in the treatment group have statistically significantly higher math pre-test scores on average (Table II.5).

Table II.3. Student Characteristics for Elementary School Reading Analysis Sample, by Treatment Status

Student Characteristic	Treatment Mean[a]	Control Mean[a]	Difference	*p*-Value[b]
Prior Achievement[c]				
Reading pre-test score	-0.39	-0.38	-0.01	0.712
Math pre-test score	-0.33	-0.35	0.03	0.451
Other Characteristics (percentages)				
Male	49.5	51.1	-1.7	0.070
Race/ethnicity				
White	7.2	5.9	1.3	0.202
African American	36.9	43.8	-6.9	0.010*
Hispanic	50.3	44.2	6.1	0.020*
ELL	35.4	30.8	4.6	0.029*
Special education	6.1	6.8	-0.7	0.209
FRL	77.4	74.6	2.8	0.019*
Sample Size (students)	4,215	3,882		

Source: Administrative data.

[a]Treatment and control means are within-block means. Random assignment was carried out by blocks.

[b]Standard errors of treatment-control differences are adjusted for clustering at the team level.

[c]Test scores are reported in standard deviations relative to the state average.

*Difference is statistically significant at the 0.05 level using a two-sided test.

Table II.4. Student Characteristics for Middle School Math Analysis Sample, by Treatment Status

Student Characteristic	Treatment Mean[a]	Control Mean[a]	Difference	*p*-Value[b]
Prior Achievement[c]				
Math pre-test score	-0.46	-0.49	0.03	0.505
Reading pre-test score	-0.62	-0.59	-0.03	0.704
Other Characteristics (percentages)				
Male	49.7	50.3	-0.6	0.387
Race/ethnicity				
White	3.7	3.2	0.5	0.315
African American	19.9	45.2	-25.3	0.000*
Hispanic	74.5	47.0	27.4	0.000*
ELL	25.7	16.7	8.9	0.001*
Special education	7.6	7.7	-0.1	0.933
FRL	89.8	84.8	5.0	0.033*
Sample Size (students)	5,433	3,442		

Source: Administrative data.

[a]Treatment and control means are within-block means. Random assignment was carried out by blocks.

[b]Standard errors of treatment-control differences are adjusted for clustering at the team level.

[c]Test scores are reported in standard deviations relative to the state average.

*Difference is statistically significant at the 0.05 level using a two-sided test.

Table II.5. Student Characteristics for Middle School Reading Analysis Sample, by Treatment Status

Student Characteristic	Treatment Mean[a]	Control Mean[a]	Difference	*p*-Value[b]
Prior Achievement[c]				
Reading pre-test score	-0.49	-0.62	0.13	0.065
Math pre-test score	-0.39	-0.57	0.19	0.019*
Other Characteristics (percentages)				
Male	52.0	51.0	1.0	0.378
Race/ethnicity				
White	5.9	7.2	-1.3	0.442
African American	32.6	33.5	-1.0	0.825
Hispanic	57.8	53.7	4.1	0.244
ELL	13.5	16.6	-3.1	0.173
Special education	6.9	6.8	0.1	0.902
FRL	84.9	86.5	-1.7	0.238
Sample Size (students)	4,205	3,607		

Source: Administrative data.

[a]Treatment and control means are within-block means. Random assignment was carried out by blocks.

[b]Standard errors of treatment-control differences are adjusted for clustering at the team level.

[c]Test scores are reported in standard deviations relative to the state average.

*Difference is statistically significant at the 0.05 level using a two-sided test.

It is possible that treatment-control differences in baseline characteristics could signal assignment that was not truly random. However, the study design of random assignment across schools, as detailed in Appendix A, would necessitate students systematically changing schools because of treatment status. Specifically, African American students would have to have left TTI schools or Hispanic students would have to have moved into treatment teams in disproportionate numbers.

Instead, the more likely and plausible explanation is that differences emerged by chance. For the elementary and middle school math samples, this means that more schools with a high percentage of Hispanic students ended up in the treatment group and more of those with a low percentage of Hispanic students ended up in the control group. This makes sense when we consider that these traits (race/ethnicity and the highly correlated English language proficiency) vary considerably between schools and do not vary much within schools. In other words, schools tended to be either mostly African American or mostly Hispanic.[35] Therefore, the sample for any given district may seem very imbalanced. For example, if we randomized eight schools within a district and the Hispanic students were concentrated in one of those schools, even the best randomization would produce a treatment-control difference of 25 percentage points (0 percent versus 25 percent). To provide a sense of the prevalence of these issues in our data, we calculated the percentages of schools with more than two-thirds Hispanic students and counted how many times an odd number of such schools were randomized together. At the elementary school level, two of the three districts with such schools had an odd number of them. At the middle school level, all three districts with these two-thirds Hispanic schools had an odd number of them.

[35] For example, the intraclass correlation coefficient for Hispanic was 0.43 in the elementary math sample and 0.52 in the middle school math sample.

The impact estimates presented in Chapter V adjust for these chance differences using multiple regression. In addition, we conducted sensitivity analyses that consider alternate specifications of pretest variables and student background covariates. In Appendix F, Tables F.18–F.21 contain the results of some of these sensitivity analyses, which show that the study's findings are robust.

III. IMPLEMENTATION PROCESS AND PLACEMENT RESULTS

In this chapter, we describe in step-by-step detail the process of implementing the Talent Transfer Initiative (TTI)—from identifying the highest-performing teachers through their hiring and acceptance of positions in low-achieving schools. To receive the $20,000 transfer incentive, a teacher had to be identified as one of the highest-performing teachers in the district; be teaching in one of the designated potential sending schools; apply to the TTI; interview for, be offered, and accept a position in a receiving school; transfer; and finally, remain in his or her new school and grade-subject team for the next two years. Because this process was both voluntary and competitive, the placement results depended on the behavior of the teachers identified as highest performing (transfer candidates) and the principals in the low-achieving (potential receiving) schools. In an interim report for this study (Glazerman et al. 2012), we presented details on the implementation process and placement results for the 7 cohort 1 districts. This chapter presents the same results updated for all 10 study districts in cohort 1 and cohort 2.

A. How Were the Highest-Performing Teachers Identified and Recruited?

1. Value-Added Analysis to Identify the Highest-Performing Teachers

The first step in identifying TTI transfer candidates was to use value-added analysis to distinguish each district's highest-performing teachers. We considered three separate pools of teachers in each district—elementary school multiple-subject, middle school English/language arts (ELA), and middle school math teachers. Teachers were eligible to be considered as highest performing if they had two or more years of value-added data. This meant that teachers who were new to the district were not eligible. Teachers who taught reading or math in a tested grade in only one of the four years before the implementation of TTI, or who left before the current academic year began, were also ineligible. Among the eligible teachers, those whose value-added scores placed them in the top 20 percent in their district and pool—elementary school multiple-subject, middle school ELA, and middle school math teachers—were identified as highest performing and were designated as TTI transfer candidates.[36] As discussed in Chapter II, this cutoff was chosen so as to be selective while still providing a sufficient number of transfer candidates to yield adequate numbers of program applicants to fill all of the vacancies identified for the receiving schools. Across the 10 districts, 1,514 candidates were identified as eligible for the 81 positions that were ultimately filled, a ratio of almost 19 candidates per position.[37] Details on the process of identifying the highest-performing teachers, including the value-added models, data used, and characteristics of identified highest-performing teachers, are discussed in Appendix B.

Highest-performing teachers already in low-achieving schools. As mentioned in Chapter I, the TTI anticipated the possibility that some of the district's highest-performing

[36] The 20 percent cutoff was raised or lowered by as much as 5 points in some pools/districts. See Appendix B for details.

[37] We initially identified 2,332 teachers as highest performing, but some of them were no longer teaching or were not planning to teach during the year when the program sought to have them transfer. Counting the teachers who turned out to be ineligible for the TTI, the ratio of candidates to filled vacancies was almost 29 to one. The numbers are approximate because we had sufficient data to count the initially identified teachers for only 9 of the 10 districts. We used the ratio for the 9 districts (2,221 total to 1,442 eligible) and multiplied it by 1,514, the number of eligible teachers in all 10 districts, to extrapolate the estimated total for all 10 districts.

teachers might already be serving in low-achieving schools, and, therefore, be ineligible for the transfer incentive because they were not allowed to transfer from one low-achieving school to another via TTI. To recognize these teachers as highest performing and retain them in the lowest-achieving schools, TTI had provisions whereby these teachers were eligible for a retention stipend of $10,000 over the same two-year period, without having to apply, change schools, or be accepted by a new principal. Retention teachers were eligible for the incentive in any low-achieving schools, whether they were treatment, control, or not part of the study.

In the 10 cohort 1 and cohort 2 districts, retention teachers represented 12 percent of the highest-performing teachers identified by the program, or one retention candidate for every seven transfer candidates. The proportion of the highest-performing teachers already teaching in low-achieving schools ranged across the 10 districts from 1 to 18 percent. This variation reflects differences in the underlying distribution of these teachers.[38] Fifty-five percent of all potential receiving schools had at least one retention teacher. The within-district percentages ranged from 7 to 79 percent of potential receiving schools that had at least one retention teacher.

Focusing on just the teams of teachers in the study (treatment or control), 30 high-performing teachers were already in these low-achieving schools. These 30 teachers were distributed across treatment groups as follows: of 378 teachers in treatment teams, 17 (7.5 percent) were highest performing; of 322 teachers in control teams, 13 (4.0 percent) were highest performing. The difference is not statistically significant.

2. Identifying and Filling Teaching Vacancies

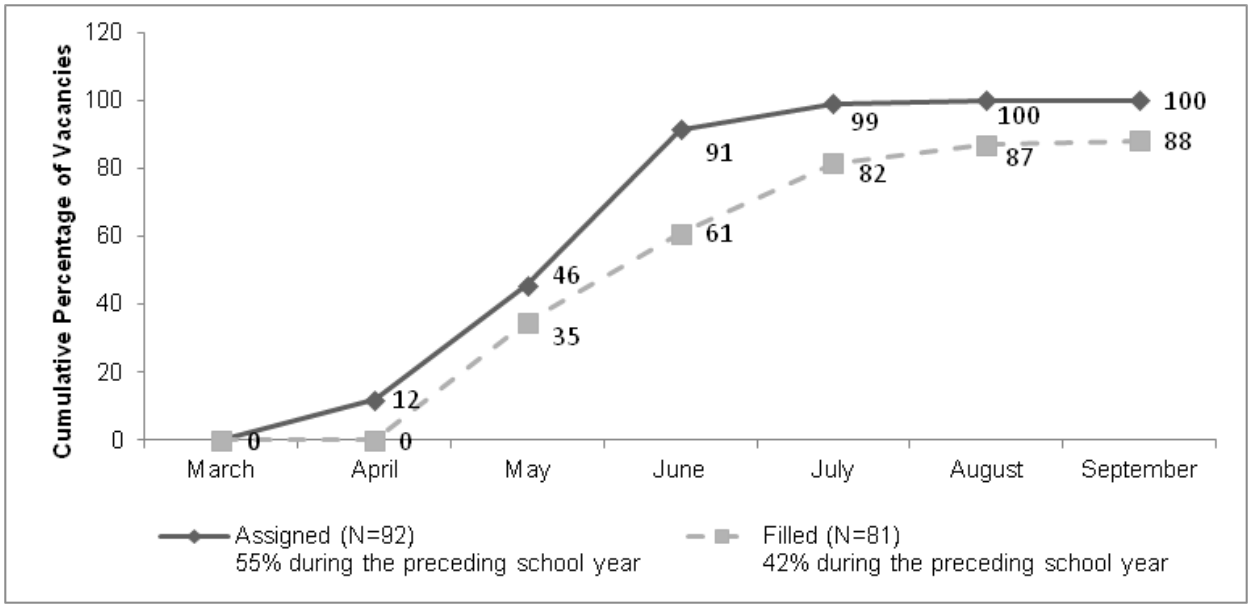
For the TTI study, a site manager designated by The New Teacher Project (TNTP) was the primary person responsible for all school and teacher recruitment. The site manager worked on both sides of the match process, with both the transfer candidates and the receiving-school principals, to fill a set number of vacancies for each pool (elementary, middle school math, and middle school ELA), determined at the district level. The TTI program relied on extensive outreach by the site managers, who served as the point of contact for teacher candidates in each district and conducted three main recruitment activities: sending invitation letters, organizing a reception that also served as an information session, and maintaining frequent communication with teacher candidates to solicit their participation and invite them to apply and interview for specific openings. In an interim report for this study (Glazerman et al. 2012), we present details on the outreach activities for the transfer candidates and receiving school principals in the cohort 1 districts, which was repeated for the cohort 2 districts.

The site managers identified teams with eligible vacancies, and Mathematica matched the teams into small groups for random assignment as they became available. Each of these groups was regarded as a randomization block, made up of teams with at least one vacancy on each team. Most blocks contained two teams, but some blocks had more than two teams, and some had just one team if no other teams were available for matching at the same time. There were teams with more than one vacancy. A total of 180 vacancies on 165 teaching teams in 89 blocks was identified between April and August. Random assignment was carried out within blocks, as discussed in Chapter II, resulting in a total of 92 vacancies on teaching teams assigned to

[38] A more complete analysis of the distribution of highest-performing teachers is presented by Glazerman and Max (2011), who relied on data from many of the same school districts in the current study.

treatment and 88 to control teams. Most vacancies were assigned and filled in May and June. In Figure III.1, we show that 79 percent of vacancies were assigned and 61 percent filled in these two months. The time between randomly assigning and filling a vacancy identified for TTI was generally short: site managers reported that vacancies were filled in as few as two days of being assigned. This general pattern of assigning and filling vacancies did not differ between elementary and middle school (see Appendix C, Figures C.1 and C.2). However, elementary school vacancies were filled at a faster rate than middle school vacancies—68 percent of the elementary school vacancies were filled by June compared with 48 percent of middle school vacancies. Among the 92 vacancies assigned to treatment, 81 (88 percent) were filled with a TTI candidate by the end of the recruitment season before the beginning of the next school year. Again, this rate was higher for elementary school (90 percent) than for middle school (85 percent).

Figure III.1. Percentage of TTI Vacancies Assigned and Filled, by Month



Source: TTI program records.

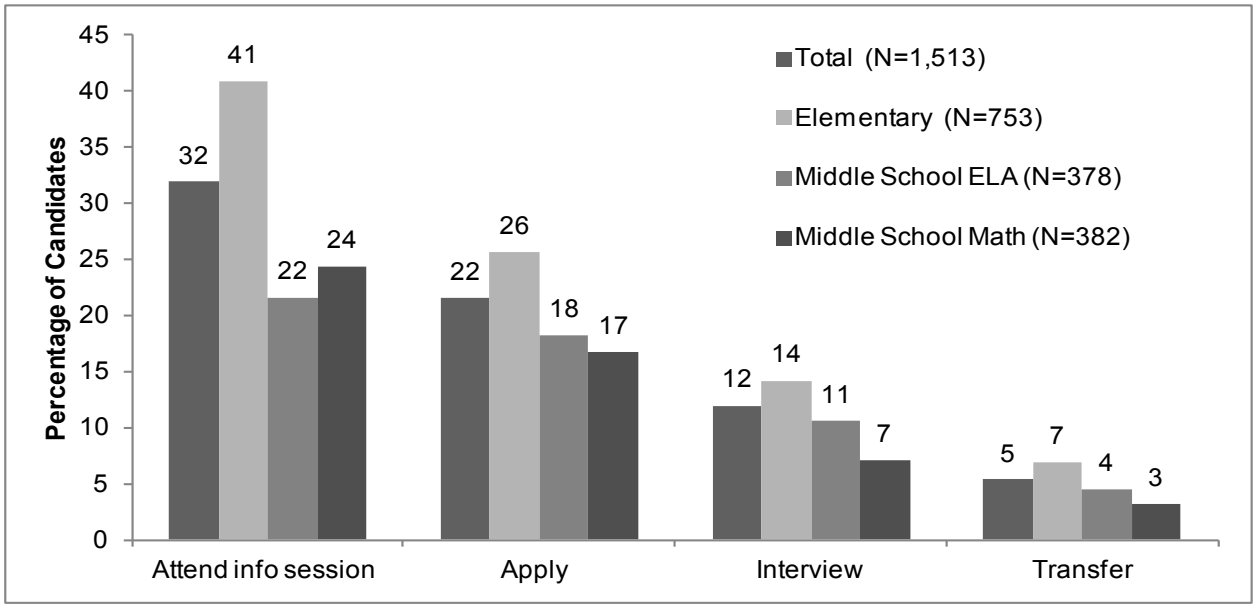
The particular month of vacancy assignment and recruitment may be less important than whether assignment and recruitment took place before or after the end of the school year. Fifty-five percent of the vacancies were assigned during the school year preceding the one when TTI teachers would begin in their new schools and 42 percent of the total filled vacancies were filled before that preceding school year ended. One district had 15 vacancies randomly assigned one week before school ended and filled all of these positions the week after school ended. In addition, many of the other vacancies were assigned with fewer than two weeks remaining in the school year. There were exceptions, however. Three of the 10 districts assigned and filled all their vacancies after the school year ended.

B. How Did Teachers React to the Transfer Incentive?

1. Take-Up Rates

To gauge the response of the candidates in the TTI, we examined the rates at which candidates took part in various phases of the process ("take-up rates"), from attending information sessions to completing an application, interviewing, and ultimately transferring. In Figure III.2, we provide a breakdown of the take-up rates by grade span and subject using TTI program records.

Figure III.2. Take-Up Rates Among TTI Transfer Candidates, by Grade Span and Subject



Source: TTI program records.

Notes Transfer candidates are the highest-performing teachers, in the top 20 percent of value-added ranking in their pool within their district. We considered three pools: elementary, middle school ELA, and middle school math.

Most teachers who were offered the transfer incentive did not apply for it or even attend an information session. Specifically, 68 percent did not attend and information session and 78 percent did not complete an online application. Forty-one percent of the elementary school candidates attended the information session/reception. For middle school, the attendance rates were 22 percent of eligible ELA teachers and 24 percent of math teachers. Almost one-third of the candidates who attended the information session did not complete an application for TTI. Application rates as a percentage of all candidates (including those who did not attend the information session) were 26 percent for elementary school, 18 percent for middle school ELA, and 17 percent for middle school math. In the Candidate Survey, we asked the candidates if they used the information session as one of the sources to obtain information about the TTI, and, if so, whether they found it useful.[39] Thirty-eight percent of those who responded used the information session as a source, and most of them (98 percent) found it useful, whether they eventually applied or not.[40]

Of those who expressed initial interest, a majority followed through to the interview stage. Fifty-five percent of applicants (12 percent of all candidates) interviewed for at least one vacancy. The other 45 percent who applied for a TTI position either did not follow through or were not given a chance to interview. Candidate Survey data suggest that of those who did interview, 104 teachers (60 percent) interviewed at one school, 40 (23 percent) interviewed at two schools, and the remaining 30 (17 percent) interviewed at three or more schools.

[39] The survey item was worded as follows: "Indicate which of the following sources you used to obtain information about the TTI:

a. Information session.
b. Internet/website.
c. Printed materials.
d. Telephone information line.
e. Email contact with program staff.
f. Other sources (please specify)."

For each of the above sources, the next survey item asked about the usefulness of the source, if used: "For each source used, rate how useful the source was in providing information you needed to make a decision about the program: 1. Not useful; 2. Somewhat useful; 3. Very useful." Sources were rated as useful if the candidate found it either "Very useful" or "Somewhat useful."

[40] The other sources used most by the candidates were Internet/website, printed materials, and phone/email contacts with TTI program staff (53, 55, and 75 percent, respectively). More than 90 percent of those who used these sources found them useful irrespective of their application status.

Data on interviews, offers, and acceptances that we obtained from the site managers provide some insight into the selectivity of the hiring process. Among the 174 candidates (12 percent of 1,514) who interviewed for at least one of the 92 available TTI vacancies, there were a total of 288 interviews because some candidates interviewed for more than one position. Thus, the average number of interviews per TTI vacancy was about 3.1. Offers were made to 97 candidates; of them, 10 received two or more offers.[41] In total, 109 offers were made by TTI schools to fill the 92 TTI vacancies, or 1.2 offers per TTI vacancy and 0.4 per candidate interviewed. This means that most principals with treatment teams made offers to only one TTI candidate and to one of three they interviewed. However, principals in TTI schools could have made offers to candidates who were not TTI candidates.[42] More details on the hiring process from the candidates' perspective are presented in Appendix C, Tables C.2 and C.3.

Candidates found most of the interviews to be informative and said they provided opportunities to communicate their strengths and receive answers to their questions. Candidates also found most of the interviews were conducted with genuine interest by the principal, reported that the interviewer was someone they could work with, and indicated the process increased their desire to teach at the school. Most of the interviewees (63 percent) met one-on-one with the principal or assistant principal; 59 percent reported that they interviewed with other school staff.[43] It was less common for candidates to give a teaching demonstration (11 percent), receive a school tour (32 percent), or meet students at the school (8 percent).

We also examined the characteristics of teachers who applied to transfer and successfully transferred, and compared them with teachers who did not apply. These characteristics are presented in Appendix C, Table C.6. In addition to tabulating the characteristics of these teachers by how far in the TTI process they went, we conducted a multivariate regression to identify factors associated with TTI take-up behavior. All of the data are in Appendix C. We summarize the key findings here.

Candidates who applied to TTI were different in some consistent ways from candidates who did not apply. They were more likely, holding other variables constant, to have lower income, be African American, be unmarried, or be less satisfied with policies at their current school. Also, more likely to apply were married candidates with children under 5 years old living with them.

Candidates who went through the entire process and transferred to a TTI position, 5 percent of the original pool, were more likely than those who did not transfer to have lower income, be African American, and be married and have children under 5 years old living with them. Contrary to the hypothesized direction of the effect, transfer candidates who were satisfied with their salary and their students during the application period were two times *more* likely to transfer than those who were not satisfied with their salary and their students, a statistically significant relationship.

[41] The number of offers received is self-reported by candidates in the Candidate Survey.

[42] We also used the Principal Survey to examine hiring rates as reported by the principals for both treatment and control teams and found similar results (see Appendix C, Table C.1). Differences in the number of applicants considered and interviewed between the treatment and the control teams were not statistically significant.

[43] The two types of interviews are not mutually exclusive.

2. Retention of Transfer and Retention-Stipend Teachers Over Two Years

Looking beyond the transfer process, we also examined whether TTI teachers stayed at study schools throughout the duration of the program. Based on program records of incentive payments, 71 of the 81 teachers who transferred through TTI (88 percent) stayed into the fall of the second program year. Of the 10 teachers who left, 9 served for the entire school year before leaving. In Table III.1, we present the percentage of teachers, by cohort, who were present in study schools and eligible for incentive payments in the fall and spring of the second program year. Nearly all teachers were present during the first year.

We do not have program records on incentive payments for the fall after program completion, but the rosters collected for cohort 1 schools indicate that 36 out of 63 (57 percent) cohort 1 TTI teachers were still in study schools at that time. Limiting the sample to only schools that provided valid rosters in all three years and did not close or reconstitute, 59 percent of cohort 1 TTI teachers returned in the fall after program completion.[44]

We also examined the retention rates of the high-performing teachers already in low-achieving schools who were eligible for retention stipends. In Table III.1, we show that, according to payment records, 84 percent of cohort 1 retention-stipend teachers and 89 percent of cohort 2 retention-stipend teachers returned in the fall of year 2. Rosters collected from cohort 1 schools in the fall after program completion indicate that 62 percent of cohort 1 retention-stipend teachers returned to their schools after incentive payments ended.

Table III.1. Percentage of Teachers Receiving Study Payments After the First Program Year

		Percentage Receiving Payments	
	Number of Teachers	Fall Year 2	Spring Year 2
Transfer Teachers			
Cohort 1	63	90.5	88.9
Cohort 2	18	77.8	72.2
Retention-Stipend Teachers			
Cohort 1	119	84.0	83.2
Cohort 2	62	88.7	87.1

Source: Program records.

C. Where Did TTI Transfer Teachers Come From?

The goal of any intervention similar to TTI is to help low-achieving schools by recruiting strong teachers from schools that are not low achieving. Some types of transfers may achieve that goal better than others. In designing the intervention, it was necessary to designate all schools in the district as being in one of two groups: low achieving (potential receiving schools) or not low achieving (potential sending schools). As described in Chapter II, school percentile ranks were calculated for elementary and middle schools separately within each district from

[44] Here, we report unadjusted retention rates of TTI teachers. These numbers differ from those presented in Chapter VI, which are adjusted retention rates estimated in an impact analysis. The impact analysis includes all treatment focal teachers even if they were not TTI transfers, and it compares the retention rates of these teachers to control focal teachers.

school composite scores in reading and math in the year or years before the school selection. For the analysis below, we used ranking in school average test scores in 2008–09 for cohort 1 districts and in 2009–10 for cohort 2 districts, replacing school average scores with school proficiency rates for one of the districts.

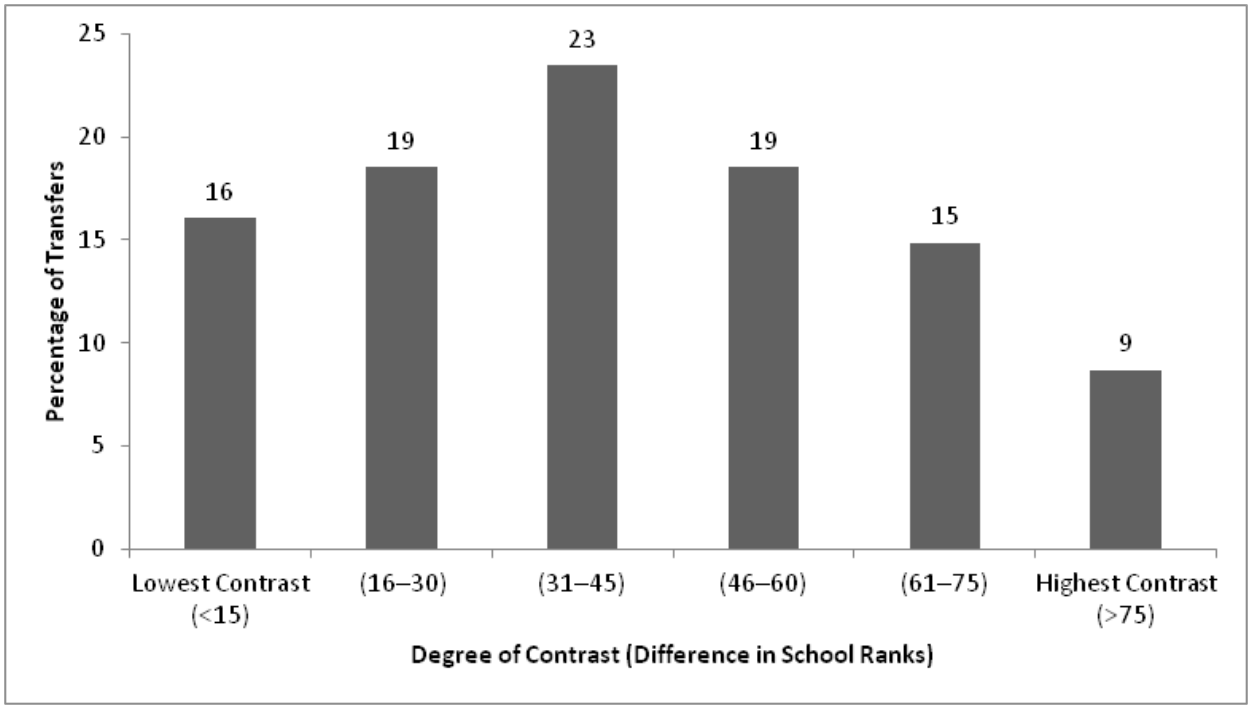
Establishing a discrete and somewhat arbitrary dividing line in the distribution between the groups means that it is possible for a teacher to transfer from a school just above the threshold to one just below it. We call these moves lower-contrast transfers because the difference in achievement ranking between the sending and receiving schools is comparatively small. The transfer incentive might be counterproductive in the case of lower-contrast transfers if the sending school is itself in need of strong teachers and has difficulty filling the vacancy created by the transfer.

We grouped the transfers by the degree of contrast, measured as the difference in the rank between the sending school and the receiving school for a given transfer. The maximum degree of contrast would be a transfer from the highest-achieving school in the district to the lowest-achieving school, a difference of 100 percentile points.

On average, transfer teachers did not move from schools just above the threshold of low achieving defined for this study. If they had, we would find that the schools differed by just a few percentile points in terms of their achievement ranking. However, transfer teachers also did not typically move from the highest-ranked group of schools, which would result in ranking differences near the maximum of 99 points (highest- to lowest-ranked in the district). The average contrast in school achievement rank is about 41 percentile points, with a median contrast of 42 percentile points. Specifically, the average sending school was ranked in the 57th percentile, where the 100th percentile is the highest-achieving school in the district, and the average receiving school was ranked in the 17th percentile.[45] Thirty-five percent of the transfers involved highest-performing teachers moving between schools that were ranked within 30 percentile points of each other in the rank distribution. On the other hand, 25 percent of the transfers involved highest-performing teachers moving between schools that were more than 60 percentile points apart. In Figure III.3, we summarize the contrasts for the 81 teachers who successfully transferred to low-achieving schools. We found similar patterns when we looked at the contrast in achievement ranks for elementary and middle school transfers separately: in elementary schools, 34 percent of the transfers were between schools that were ranked within 30 percentile points of each other; the corresponding number for middle school transfers was 36 percent (Appendix C, Figures C.3 and C.4). However, 13 percent of the transfers at the elementary school level involved moves between schools that were less than 15 percentile ranks apart. The corresponding percentage for middle school transfers was 24.

[45] The achievement rank of the highest-achieving potential receiving school ranged across districts and pools from 20th to 42nd. Achievement rankings used for the analysis in this section include schools exempt from the study, some of which are ranked in the bottom 20 percent. Also, as discussed in Chapter II, some districts used accountability ratings in addition to achievement ranks to identify potential receiving schools. As a result, some schools in these districts were classified as potential receiving schools based on their accountability rating even though they had higher average achievement than some potential sending schools.

Figure III.3. Contrast in TTI Transfer Teachers' Sending- and Receiving-School Achievement Rank



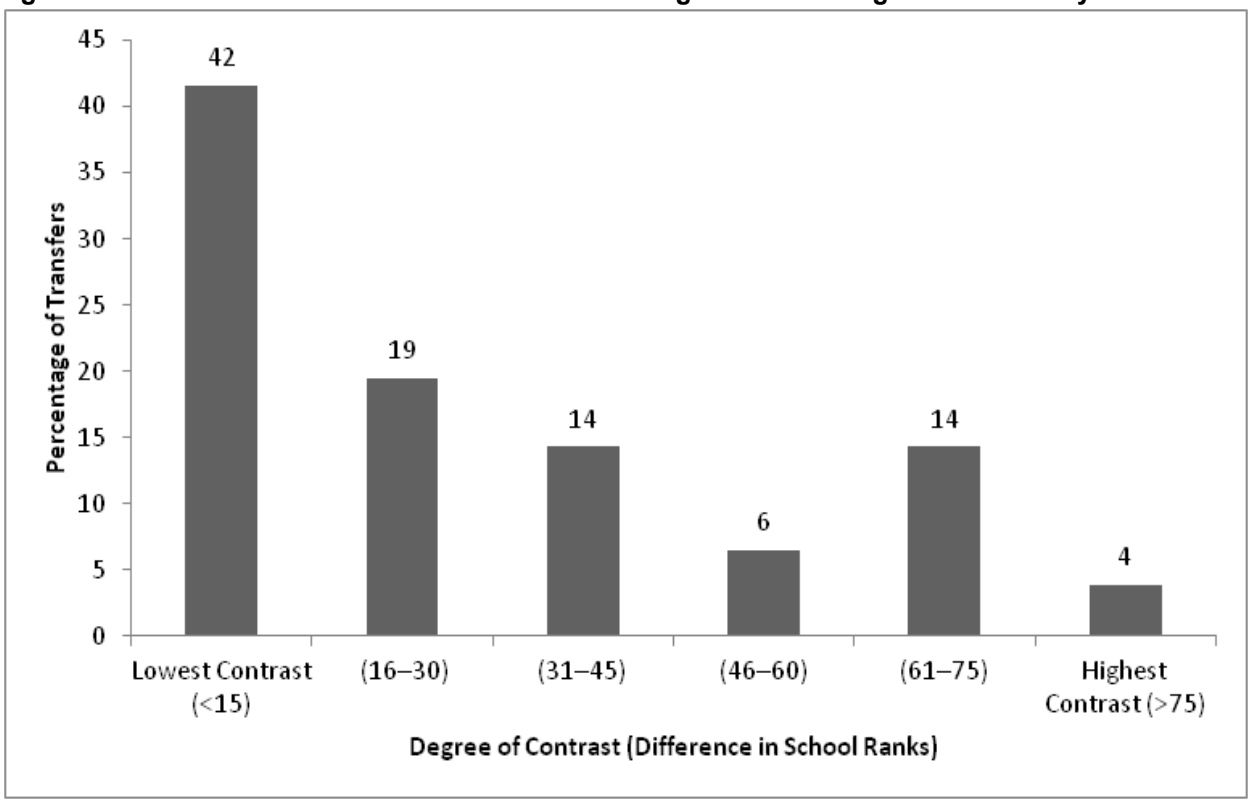
Source: Administrative data and TTI program records (N = 81).

Although the poverty status of schools was not taken into account when identifying potential receiving schools, it is still informative to examine the contrast in the percentile rank positions based on poverty status (as measured by percentage of students eligible for free or reduced-price lunch [FRL]) of schools that transfer teachers left, compared with the ones to which they moved, because FRL is used as a measure of student disadvantage in many federal programs. If we found only low-contrast transfers, based on FRL ranks, it would suggest that TTI is redistributing teachers within the same group of disadvantaged students.

The average contrast in poverty rank was about 23 percentile points, with an average sending school percentile rank of 45 and an average receiving school rank of 21.[46] Eighteen percent of transfer teachers moved between schools whose ranks were more than 60 percentile points apart based on poverty status (Figure III.4). However, 42 percent of the transfer teachers moved between schools that were fewer than 15 percentile ranks apart from each other. This further validates the results of the multivariate analysis above: candidates are more likely to transfer if they have a higher percentage of disadvantaged students (as measured by FRL) in their original school or if the student population in the school into which they transfer is similar to the one in their original school. We found similar patterns when we looked at the contrast in poverty ranks for elementary and middle school transfers (Appendix C, Figure C.5).

[46] We ranked schools in descending order of poverty status, as measured by the percentage of students eligible for FRL. Therefore, schools with higher poverty status or higher percentage of students eligible for FRL have a lower rank.

Figure III.4. Contrast in TTI Transfer Teachers' Sending- and Receiving-School Poverty Rank



Source: Administrative data and TTI program records (N = 81).

As one additional way of comparing the circumstances of the transfer candidates' original schools to those to which they transferred, we compared the candidates' students before and after the transfer. This is relevant to the question of whether teacher-student match effects plausibly play a role in determining the impacts of TTI.

Because this analysis uses detailed data from before the candidates transferred, we focus on the subset of seven districts that provided such data.[47] In Table III.2, we show the average student characteristics before and after transfer for the 52 transfer teachers who came from these seven districts. The transfer teachers moved to locations where they would be teaching a significantly lower percentage of white students and a significantly higher percentage of FRL-eligible students. They would also be teaching a higher percentage of minority students—a difference of 9 percentage points for African American students and 6 percentage points for Hispanic students—but these differences are not statistically significant. Consistent with the findings in the contrast and multivariate analyses, transfer teachers were already teaching classes with an average share of FRL-eligible students of 68 percent in their sending schools, but it rose to 93 percent in the receiving schools, a significant difference of 24 percentage points. Given that the teachers transferred from higher-achieving and lower-achieving schools, this shift in student

[47] Administrative data on the background characteristics of students in TTI transfer teachers' original schools were available only for the seven districts for which the study team conducted value-added analysis to identify highest-performing teachers. The three other districts were already using value-added systems developed by an external vendor and gave us teacher-level value-added data or a list of highest-performing teachers. See Appendix B for details on the highest-performing-teacher identification process.

characteristics is consistent with prior research on the correlation between family income and achievement and between race and achievement (Reardon 2011; U.S. Department of Education 2013).

The differences in test scores between transfer teachers' students in sending and receiving schools were statistically significant for this group. The average student in the transfer teachers' classrooms scored 0.06 standard deviations below the district average in reading, placing the students in the 48th percentile. The same teachers' students in the schools to which they transferred had scored 0.48 standard deviations below the district average, which placed them in the 32nd percentile. For math, the differences were -0.09 standard deviations (47th percentile) and -0.44 standard deviations (33rd percentile).

Table III.2. Characteristics of TTI Transfer Teachers' Students Before and After Transferring

Characteristic of Average Student (percentages unless noted)	In Sending-School Classrooms	In Receiving-School Classrooms	Difference (Receiving Minus Sending)	*p*-Value
Demographic				
White	23.7	8.7	-15.0*	0.000
African American	27.8	36.6	8.8	0.138
Hispanic	40.9	46.9	6.0	0.330
Economic				
FRL	68.1	92.6	24.4*	0.000
Academic				
Special education (SPED)[a]	10.8	11.0	0.2	0.913
Limited English proficient	15.9	15.1	-0.7	0.859
Average reading score[b]	-0.06	-0.39	-0.32*	0.001
Average math score[b]	-0.09	-0.38	-0.29*	0.000

Source: Administrative data.

Note: Data pertain to a subgroup consisting of seven districts that provided student-level data (N = 52 teachers who transferred in the seven districts and for whom detailed student data were available). Due to missing data, the sample size was 38 teachers for SPED and 45 teachers for FRL and ELL. The sample sizes were 43 for reading scores and 42 for math scores because not all teachers taught both math and reading.

[a]The SPED category in two of the seven districts includes gifted students. These two districts are not included in the table. One teacher who taught 100 percent SPED students in the sending school is also not included.

[b]Average reading and math scores are given in fraction of a standard deviation computed within district and grade.

*Difference is statistically significant at the 0.05 level using a two-sided test based on the teacher sample.

D. Who Filled the Study Vacancies?

One might expect that the treatment-group vacancies would be filled by teachers identified through the TTI process as described throughout this chapter, and that control-group vacancies would be filled by hiring new teachers into the profession. In reality, vacancies can be and were filled through a variety of means. Although most vacancies in the control group were filled through new hires or transfers from other schools, several were filled by moving teachers from another grade or subject within the school—and in some cases, the position was lost altogether

because of declining enrollment. Furthermore, although most vacancies in the treatment group were filled through TTI, some were filled outside the program. The details are discussed next.

First, however, we note that identifying the teacher who filled the study vacancy was potentially problematic, mainly in the control group, due to data limitations. For nearly one-quarter of the control teams (23 percent) there was some uncertainty as to the identity of the focal teachers because of teachers who did not respond to the survey and principals who could not be reached to verify. We faced the same problem for 5 percent of the treatment teams. For such cases, we included all possible focal teachers in the analyses throughout this report and assigned each one a weight equal to the number of vacancies on the team divided by the number of uncertain but possible focal teachers on the team. We refer to this as the inclusive definition of focal teachers. It differs from the definition used in an interim report for this study (Glazerman et al. 2012), which was more selective. The selective definition used only focal teachers who were known with certainty to have filled a specific vacancy. Use of the selective definition requires discarding substantial numbers of cases based primarily on survey nonresponse, raising concerns of selection bias. For the analysis in the interim report, it was more important to describe the teachers who were hired than to compare treatment and control characteristics, so the possibility of selection bias was not the overriding concern. For this report, we considered the selective definition of focal teachers as a sensitivity test, but we relied on the inclusive definition for the main analysis to avoid the risk of selection bias in impact estimates. In Appendix D, we provide a full discussion of this issue with comparisons of selective and inclusive focal teacher definitions.

1. Control-Group Vacancies

A transfer-incentive strategy like TTI is intended to improve the quality of teachers filling new vacancies in low-achieving schools. It is, therefore, important to understand how those vacancies would have been filled if there had been no TTI intervention. Before conducting the study, we hypothesized that many such vacancies would be filled by individuals who are just now entering the teaching profession but some could be filled in other ways. Teachers might transfer from other schools, be hired from other districts, or move from another grade within the same school.

We identified 114 potential receiving schools whose principals indicated they had at least one vacancy eligible for the study, for a total of 180 vacancies.[48] After randomly assigning teacher teams with those vacancies to either a treatment group that could hire through the TTI or a control group that could not, we followed the control group to learn more about the business-as-usual condition. We assigned 80 teams with 88 vacancies to the control group and 85 teams with 92 vacancies to the treatment group.

In Table III.3, we show that the control-group vacancies were filled by a combination of new hires, transfers in, and within-school reassignments. The number of positions that were not filled is also shown. This breakdown illustrates what would have happened to teaching vacancies in low-achieving schools in the absence of TTI. Seventeen (19 percent) of the 88 control positions were filled with new hires, and most of those (11 of the 17) were new to the profession. Another 19 (22 percent) were transfers from other schools, none of them TTI candidates, and

[48] Sixty-three schools had one vacancy, 41 schools had two, and 10 schools had three, four, or five.

26 (30 percent) moved from another position within the school. Most of these within-school moves were the result of moving the vacancy to another grade, so there were more teachers who might have been new to the school but who were not included in the study because we focused only on the grades we randomly assigned. Another 9 positions were filled by individuals whose backgrounds we could not ascertain because they did not respond to the Teacher Background Survey. An additional 8 positions (9 percent) were lost because of a drop in student enrollment, an increase in class size, or because the teacher who was leaving to create the vacancy returned to his or her position. Finally, there were 9 vacancies (10 percent) whose ultimate status could not be determined unambiguously. In these cases, there were so many teachers within a grade who did not complete a survey that we could not determine which teacher filled the vacancy, and the principal provide insufficient information for us to determine the hiring outcomes.

Table III.3. How Study Schools Filled Their Vacancies in the Absence of a Transfer Program (Focal Teachers, Control Group Only)

Final Status of the Vacancy	Number	Percentage
Positions Filled		
New to teaching	11	12.5
New to the district	6	6.8
Transfer from other school	19	21.6
Transfer from another grade	26	29.5
Unknown origin/uncertain	9	10.2
Position Lost, Transfer Canceled, or Layoff Rescinded[a]	8	9.1
Unknown Status[b]	9	10.2
All Vacancies	88	100.0

Source: Teacher Background Survey, Principal Survey, principal follow-up interviews.

[a]Teachers whose transfers out of the study school were canceled or whose layoffs were rescinded were treated as the focal teacher for this study.

[b]These are teaching teams whose vacancy was not filled, or where the focal teacher was not identifiable.

The characteristics of control focal teachers are summarized in Table III.4. Although most survey respondents (54 percent) designated as focal teachers in the control group were new to the school, not all were new to the profession. In fact, 17 percent reported being new to teaching, whereas 45 percent reported being in at least their sixth year of teaching. The average experience of control focal teachers was eight years. Thirty-six percent held at least a master's degree; 9 percent held National Board Certification, an advanced teaching credential that requires a lengthy application process and demonstration of mastery through a portfolio and other materials; and 24 percent had undergraduate degrees from institutions rated very competitive or higher. Their experience profiles show that many (see Table III.4) were experienced teachers who simply moved from elsewhere in the school or district and were not hired from the pool of beginning teachers.

The demographic characteristics of control focal teachers are also shown in Table III.4. Eighty-two percent of control focal teachers were female, 42 percent were white, and 54 percent were married. The average age of the control teachers was 37 years old.

Table III.4. Characteristics of Control Focal Teachers (percentages)

Characteristic	Average[a]
Professional Background	
Years of Experience in Teaching (average years)	8.2
Years of Experience in Teaching (percentages by category)	
1 (first year teaching)	17.2
2–5 years	38.0
6–10 years	20.0
11+ years	24.7
First Year in the School	54.4
First Year in the Grade	58.3
Has Regular Certification for Grade/Subject Taught	94.1
Has a Master's or Doctorate Degree	36.1
Has National Board Certification	9.0
Has Undergraduate Degree from Institution Rated Very Competitive or Higher by Barron's	23.9
Personal Background	
Female	82.1
Race/Ethnicity	
White, non-Hispanic	42.1
African American, non-Hispanic	33.0
Hispanic or Latino	15.5
Average Age (years)	36.5
Married or Living with a Partner	54.4
Home Owner	48.0
Sample Size (number of respondents)	99

Source: Teacher Background Survey.

[a]Results are weighted to account for the possibility that more focal teachers were identified in a grade team than there were vacancies identified. The sum of weights equals the number of vacancies being described.

2. Treatment-Group Vacancies

Most vacancies assigned to the intervention were filled with TTI candidates. Of 92 positions, 80 (87 percent) had a successful transfer and stayed beyond the start of the school year. Three others were subsequently moved to a different grade within the same school and are not included in our analysis of treatment focal teachers below. Of the 7 positions assigned to the treatment group that were not filled by a TTI teacher, 3 were lost because of enrollment declines or because teachers were recalled from a layoff notice that had created the vacancy; the remaining 4 positions were unable or unwilling to select a match with TTI candidates and the principal hired outside the pool. In Table III.5, we summarize the status of the vacancies in teacher teams assigned to the treatment group and whether or not they were filled by a TTI transfer. In Table III.6, we present the characteristics of the teachers who filled those vacancies and responded to the Teacher Background Survey. We repeated the data from Table III.4 in Table III.6 in order to facilitate the treatment-control comparison. The teachers in the treatment group all had at least 2 years of experience and a large proportion (44 percent) had at least 11 years of experience. The difference in experience level between treatment focal and control focal teachers (about 4 years) was statistically significant. Other treatment-control differences are pointed out in Table III.6.

Table III.5. How Study Schools Filled Their Vacancies Using Transfer Program (Focal Teachers, Treatment Group Only)

How Vacancy Was Filled	Number of Vacancies	Percentage
Filled with TTI candidate	81[a]	88.0
Filled outside TTI	7	7.6
Position lost, transfer canceled, or layoff rescinded	4	4.3
All vacancies	92	100.0

Source: Teacher Background Survey, Principal Survey, principal follow-up interviews.

[a]Three TTI candidates were placed in a different grade in the school than the originally designated vacancy.

Table III.6. Characteristics of Treatment and Control Focal Teachers (percentages)

Characteristic	TTI Transfers	All Treatment Focal	All Control Focal[a]	Treatment-Control Difference	*p*-Value
Professional Background					
Years of Experience in Teaching (average years)	12.5	11.8	8.2	3.6*	0.002
Years of Experience in Teaching (percentages by category)					
1 (first year teaching)	0.0	0.0	17.2	-17.2*	0.000
2–5 years	6.3	12.3	38.0	-25.7*	0.000
6–10 years	44.7	43.6	20.0	23.6*	0.001
11+ years	49.0	44.1	24.7	19.3*	0.007
Has a Master's or Doctorate Degree	46.7	46.2	36.1	10.1	0.188
Has National Board Certification	20.8	20.2	9.0	11.2*	0.035
Has Undergraduate Degree from Institution Rated Very Competitive or Higher by Barron's	15.2	16.3	23.9	-7.6	0.229
Transferred via TTI	100.0	75.5	0.0	75.5*	0.000
Personal Background					
Female	82.7	83.3	82.1	1.2	0.847
Race/Ethnicity					
White, non-Hispanic	48.3	49.5	42.1	7.4	0.350
African American, non-Hispanic	27.9	27.7	33.0	-5.3	0.465
Hispanic or Latino	17.1	16.6	15.5	1.1	0.848
Age (years)	42.6	41.8	36.5	5.3*	0.001
Married or Living with a Partner	64.6	62.1	54.4	7.7	0.323
Home Owner	82.0	77.4	48.0	29.3*	0.000
Sample Size (number of teachers)	78	89	99		

Source: Teacher Background Survey.

[a]Results are weighted to account for the possibility that more focal teachers were identified in a grade team than there were vacancies identified. The sum of weights equals the number of vacancies being described.

*Difference between treatment focal mean and control focal mean is statistically significant at the 0.05 level using a two-sided test. Test was conducted only for the "All Treatment" group.

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

IV. INTERMEDIATE IMPACTS

The logic model for the Talent Transfer Initiative (TTI), presented in Chapter I, suggests that the opportunity to fill a teaching vacancy with one of the district's highest-performing teachers could alter the school's internal dynamics, inducing certain changes within the teaching team. In this chapter, we document some of the changes, which can include indirect impacts and resource-allocation effects. Indirect impacts are those that affect nonfocal teachers, that is, teachers and students other than the transfer teachers and their own students. They can result from teacher collaboration and sharing of ideas. The changes in the teaching team also include resource-allocation effects, where the overall teaching workload may be unchanged but merely shifted from one teacher to another within the team, or sometimes within the school.

A. Assignment of Teachers to Students and Grades

An important type of resource-allocation effect is the impact of a transfer incentive on the assignment of students to teachers and teachers to grades. Under normal circumstances, when filling a teaching vacancy in a hard-to-staff school, the principal may strategically assign students such that the teacher who fills the vacancy works with the less-challenging students. Alternatively, the new teacher may be assigned more challenging students. Similarly, a principal may seek to pair weaker teachers with stronger colleagues and vice versa.

Each of these effects may be attempts by principals to offset the perceived disadvantage of hiring a new teacher in the absence of TTI, or responses of principals to the perceived advantage of hiring a high-performing teacher through TTI.

1. Assignment of Students to Teachers

The evidence regarding whether students were assigned differently to teachers as a result of TTI was mixed. Using administrative data, we did not find evidence of impacts on student assignment. If focal teachers were assigned different types of students (based on prior test scores or student demographics) than nonfocal teachers, the differences were statistically indistinguishable between the treatment and control teams. Detailed evidence on this point was first provided in the interim report based on the 7 cohort 1 districts (Glazerman et al. 2012). In this report, we provide the updated version of these analyses for all 10 districts in Appendix E, Figures E.1–E.8. We did not reach different conclusions when we examined these results separately by grade span (elementary and middle schools separately), nor did we find evidence of principals reporting a different method of assigning students to classrooms in treatment versus control teams. We asked principals directly how students were assigned to classrooms in the specified grade teams.[49] The results (Table IV.1) suggest that the most common assignment mechanism was random or balanced assignments: principals reported that students in 57 percent of control teams and 44 percent of treatment teams were assigned this way. The difference was

[49] The survey item was worded as follows: "Which ONE of the following statements best describes how students were assigned to classrooms/teachers in [the given grade] for 2009–2010. Students were assigned:

a. At random (or similar method to ensure balance of academic level, gender, and/or behavioral problems).
b. By matching student needs to teachers' specific abilities.
c. By creating homogeneous groups based on ability or course difficulty.
d. By 'looping' or a related approach to keep previous year student rosters mostly intact.
e. Other (please specify)."

not statistically significant (*p*-value = 0.111). Approximately one-quarter of treatment teams and 20 percent of control teams formed their student rosters by matching students to teachers, according to principals. Ability grouping was reported by principals to be used in 19 and 17 percent of treatment and control teams, respectively. None of the treatment-control differences was statistically significant.

Table IV.1. How Students Were Assigned to Classrooms, Principal Report (percentages)

Method of Assignment[a]	Treatment	Control	Impact	*p*-Value
All Schools				
Random (or similar method to balance academic level, gender, and/or behavioral problems)	43.8	56.6	-12.8	0.111
Matching student needs to teachers' specific abilities	23.8	19.7	4.0	0.547
Homogeneous groups based on ability or course difficulty	18.8	17.1	1.6	0.791
Sample Size (teams)	80	76		
Elementary Schools				
Random	49.1	56.6	-7.5	0.441
Matching	20.8	20.8	0.0	1.000
Ability grouping	17.0	13.2	3.8	0.592
Sample Size (teams)	53	53		
Middle Schools				
Random	33.3	56.5	-23.2	0.104
Matching	29.6	17.4	12.2	0.322
Ability grouping	22.2	26.1	-3.9	0.756
Sample Size (teams)	27	23		

Source: Principal Survey.

Note: None of the treatment-control differences is statistically significant at the 0.05 level.

[a]Percentages do not add up to 100 because some principals had only one teacher teaching the subject in that team, students "looped" with their teachers (the student roster moved intact from one grade to the next, keeping the same teacher), or some other circumstance.

Evidence supporting the resource-allocation hypothesis did emerge, however, from another source—the data from subjective teacher reports. Focal teachers were more likely to report having more academically challenging students than their peers in treatment teams than in control teams. All teachers in the study teams, focal and nonfocal, were asked if they felt that their students were more challenging, less challenging, or equally challenging to teach, in terms of academic ability, compared with students of other teachers in the same school, grade, and subjects.[50] The results are summarized in Figure IV.1, which shows the percentages of treatment and control teachers divided into focal and nonfocal teachers.

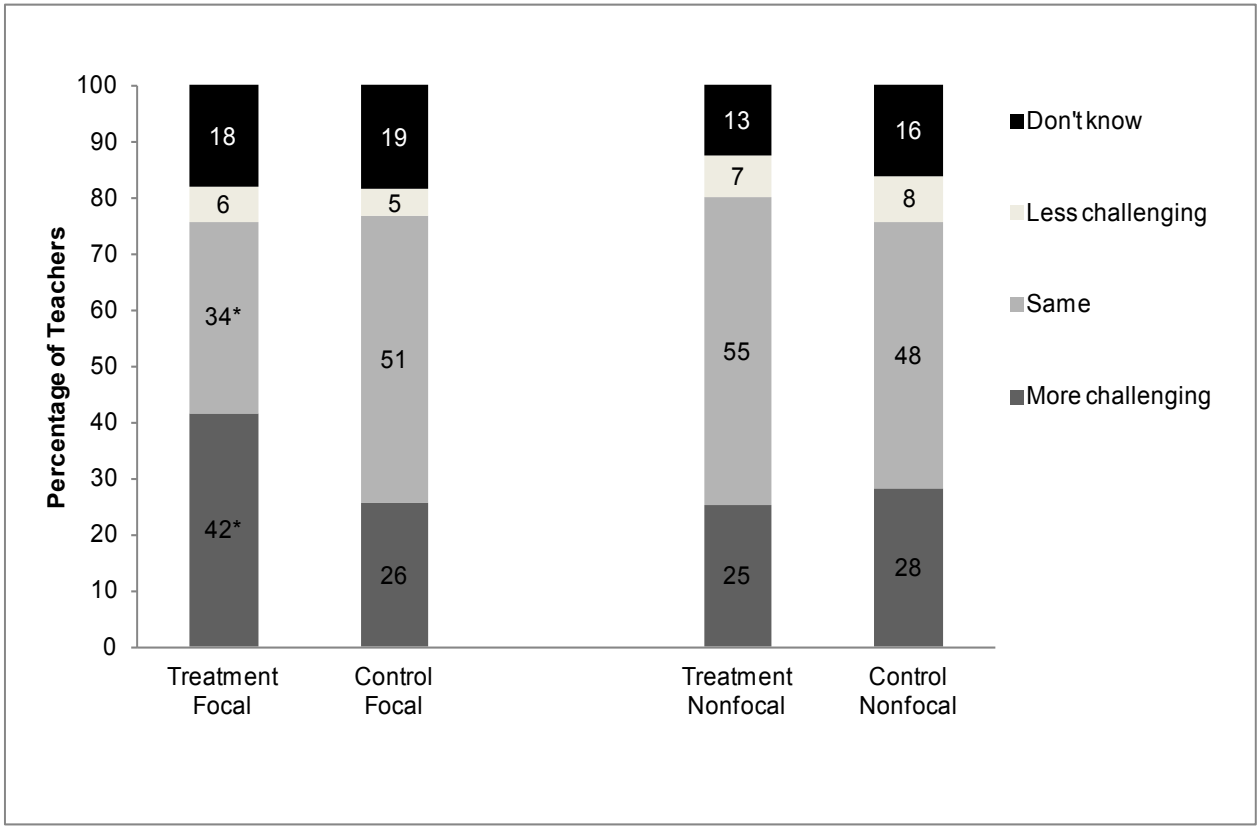
In Figure IV.1, we show the percentages of teachers in each group (treatment/control and focal/nonfocal) who found their own students more challenging, the same, or less challenging in

[50] The survey item was worded as follows: "Think about the ABILITY LEVELS of the students assigned to your class(es) this year compared with those of student assigned to your colleague(s) teaching the same grade level or subjects in your school. Would you say the students in YOUR class(es) are…

a. More challenging in ability.
b. About the same level of ability.
c. Less challenging in terms of ability.
d. Cannot judge. I am unfamiliar with the ability levels of the students in the other class(es)."

terms of ability than the students of other teachers on their team. The most common response for all teachers except treatment focal was that their own students were "about the same" as the other teachers' students in terms of academic challenges, although between 13 and 19 percent of teachers said they did not know whose students were more or less academically challenging.

Figure IV.1. Classroom Assignment of Students Who Are More or Less Academically Challenging, Teachers' Perceptions, by Their Treatment and Focal Status



Source: Teacher Background Survey.

Note: N = 86 treatment focal, 97 control focal, 194 treatment nonfocal, and 172 control nonfocal respondents.

*Treatment-control differences are statistically significant at the 0.05 level using a two-sided test.

The treatment-control differences in response to this question were statistically significant for focal teachers. Treatment focal teachers were more likely to report having more academically challenging students than were control teachers (42 versus 26 percent), and they were less likely to report having students with the same degree of challenge (34 versus 51 percent). We analyzed the same data separately for elementary and middle school teachers (not shown) and found that the significant differences arose from middle school teachers' experiences. In middle schools, where ability grouping is more prevalent, 52 percent of treatment focal teachers versus 9 percent of control focal teachers reported that their students were more challenging in terms of academic ability. We repeated this exercise using a question about students' behavioral challenges but did not find statistically significant differences. Those results are shown in Appendix E, Figure E.9.

2. Assignment of Teachers to Grades

Another possible way that TTI could alter the school's internal dynamics is through resource allocation across grades. Specifically, principals under the status quo, represented by control school teams, may compensate for weak incoming teachers by moving strong peers—veterans who can mentor their less-experienced colleagues—from elsewhere in the school into their grade team. Using similar logic, principals with treatment teams may move weaker or less-experienced teachers into teams with an incoming TTI teacher in order to take advantage of pairing weak teachers with strong ones. Comparing the characteristics of the nonfocal teachers in the treatment and control teams is one way to explore this hypothesis of strategic staffing.

Table IV.2 presents characteristics of the nonfocal teachers by treatment status. It compares the number of years of teaching experience for all nonfocal teachers (treatment versus control). It also compares these same characteristics for nonfocal mover and stayer teacher subgroups (treatment versus control). This subgroup comparison is important because the most logical place to find differences, if there were strategic reassignment, would be in the teachers who are new to the teaching team (the movers).

The results presented in Table IV.2 show significant treatment-control differences in experience among nonfocal movers. Teachers who were new to the treatment teams had almost five years less experience than teachers who were new to the control teams (a pattern of disproportionately pairing the high-performing TTI teachers with less experienced teachers). Thus treatment-control differences in nonfocal teacher characteristics are consistent with a hypothesis of strategic reassignment of teachers.

Table IV.2. Characteristics of Nonfocal Stayers and Movers

Characteristic	Treatment Mean	Control Mean	Difference	*p*-Value	Sample Size (Teachers)	
					Treatment	Control
Years of Classroom Experience						
All teachers	10.5	11.5	-1.0	0.292	198	176
Stayers	11.6	11.3	0.3	0.782	124	129
Movers	10.8	15.4	-4.6*	0.044	51	33
Years of Teaching Experience in the District						
All teachers	8.6	9.6	-1.0	0.234	198	176
Stayers	9.7	9.4	0.3	0.748	124	129
Movers	8.5	13.1	-4.6*	0.033	51	33
National Board Certified (percentage)						
All teachers	13.4	17.5	-4.1	0.296	192	168
Stayers	18.5	18.8	-0.3	0.961	118	123
Movers	6.3	20.1	-13.8	0.077	51	31

Source: Teacher Background Survey.

*Differences are statistically significant at the 0.05 level using a two-sided test. "Stayers" refers to teachers who were in the same school and taught the same grade/subject in the previous year. "Movers" are teachers who were in the same school the previous year but who taught a different grade or subject.

B. Teachers' Mentoring and Leadership Roles

Strategic assignment of students is one way in which principals might adapt to the opportunity to fill vacancies with transfer teachers in a program like TTI. Another way is to shift mentoring resources that would have gone toward supporting a new, presumably inexperienced teacher to other teachers, or to reduce the level of mentoring in that grade altogether, freeing more resources for teams in other grades in the school or for other types of interventions in the same grade.

1. Mentoring Received

We did find evidence of a resource-allocation effect on mentoring resources. Treatment focal teachers reported having a mentor at a significantly lower rate than control focal teachers (39 versus 59 percent, *p*-value = 0.009). The results are shown in Table IV.3.[51] We report results for the entire sample, not separately by grade span, because the findings were qualitatively similar and the sample sizes for middle schools (29 treatment focal and 35 control focal) were too small to support detailed analysis.

Table IV.3. Mentoring Received by Focal and Nonfocal Teachers (percentages)

	Focal Teachers			Nonfocal Teachers		
Outcome[a]	Treatment	Control	Difference	Treatment	Control	Difference
Had a mentor	38.6	58.7	-20.0*	45.0	47.0	-2.1
Was mentored by another teacher	20.6	35.6	-15.0*	19.0	20.1	-1.1
Was mentored by a coach or facilitator	16.3	15.6	0.7	20.4	17.5	2.9
Time with mentor (minutes per week)	35.1	49.3	-14.1	46.3	50.4	-4.1
Sample Size[b]	87	99		199	175	

Source: Teacher Background Survey.

[a]Units are percentages unless otherwise indicated.

[b]Sample size is number of teachers.

*Differences are statistically significant at the 0.05 level using a two-sided test.

The source of mentoring differed significantly when we asked whether another teacher in the school was providing the assistance: 21 percent of treatment focal teachers versus 36 percent of control focal teachers reported receiving mentoring support from a fellow teacher. Reports of assistance from literacy coaches, math coaches, and similar types of school supports did not statistically differ for treatment and control focal teachers. We asked about mentoring support from principals, but the numbers of respondents were too small to report, and the treatment-control differences were not statistically significant.

[51] The survey item measuring "Had a Mentor" was worded as follows: "Are you assigned to, and currently working with, a person (or persons), such as a mentor, coach, lead teacher, or other school or district leader, who provides professional advice and direct assistance to you in your teaching duties?"

Differences for nonfocal teachers (right side of Table IV.3) were not statistically significant. This suggests that when there were differences in focal teachers' mentor receipt, they did not come at the expense of nonfocal teachers.

2. Mentoring Provided and Other Leadership Roles

Yet another way to take advantage of TTI might be to assign additional duties or responsibilities to TTI transfer teachers themselves. The design of the intervention did not require principals to create or require any special duties or roles for teachers as a condition of being hired or receiving the TTI bonus, but there was no restriction against a principal imposing such a condition or simply assigning the teacher or requesting that the teacher fill such a role.

In Table IV.4, we show the percentages of teachers by treatment and focal status who reported playing such roles. The evidence indicates that treatment focal teachers provided more mentoring to their peers than did control focal teachers (15 percent versus 5 percent who provided mentoring). Although the difference was not statistically significant (p-value = 0.058), the time spent mentoring other teachers went in the same direction: 24 minutes provided by treatment focal teachers versus 6 minutes per week provided by control focal teachers. This positive impact on mentoring support provided by treatment focal teachers was not offset by a significant negative impact for nonfocal teachers.

We also examined the rates at which focal and nonfocal teachers reported playing leadership roles in their schools and found that none of the treatment–control differences was statistically significant. The survey, administered in the spring of the first program year, asked teachers about several activities that they could have been involved in, such as serving as a grade-level or subject chair or serving on a committee.[52] Results are shown in the bottom of Table IV.4.

The weight of evidence on mentoring and leadership suggests that the intervention did alter the mentoring relationships within the teaching team. However, the lack of impact on nonfocal teachers suggests that changes in mentoring services used and provided by focal teachers were not offset by equal and opposite changes for nonfocal teachers. Because we did not collect data on teachers in other grades, we cannot measure the extent to which the extra resources were allocated elsewhere in the school.

[52] The exact question was asked: "In which of the following activities are you currently involved at your school?

(a) Serving as a grade-level or subject area chair
(b) Serving on a school improvement committee
(c) Working to obtain external funding for my school (i.e., grants or funding from external organizations for projects/supplies/materials
(d) Leading or promoting teacher collaboration
(e) Observing or providing feedback to other teachers
(f) Other (Please specify)."

Table IV.4. Mentoring and Other Leadership Provided by Focal and Nonfocal Teachers (percentages unless indicated otherwise)

	Focal Teachers			Nonfocal Teachers		
Outcome[a]	Treatment	Control	Difference	Treatment	Control	Difference
Support Provided						
Provides mentor support	15.0	5.4	9.6*	16.8	16.3	0.5
Number of teachers mentored	0.3	0.1	0.2	1.1	0.5	0.6
Average minutes per week of mentoring provided	24.1	6.3	17.9	22.4	28.3	-5.9
Leadership Roles						
Serves as grade-level or subject-area chair	15.1	21.9	-6.8	29.1	33.9	-4.8
Serves on school-improvement committee	26.5	32.8	-6.3	36.1	41.2	-5.1
Works to obtain external funding for the school	16.2	14.9	1.3	14.8	15.0	-0.2
Leads or promotes teacher collaboration	52.5	51.0	1.5	40.4	48.4	-8.0
Observes or provides feedback to other teachers	32.9	29.2	3.7	39.4	36.9	2.5
Involved in other activities	30.3	30.8	-0.5	33.0	25.2	7.8
Sample Size[b]	87	99		199	175	

Source: Teacher Background Survey.

[a]Units are percentages unless otherwise indicated.

[b]Sample size is number of teachers.

*Differences are statistically significant at the 0.05 level using a two-sided test.

C. Teacher Attitudes

Evidence on teachers' attitudes reinforces the idea that TTI had an impact on study teams. We asked teachers toward the end of the first program year to indicate the extent to which various aspects of their jobs were challenging during their first year at the school, and we also asked them to rate their satisfaction with different aspects of their jobs.

We found that treatment focal teachers were less likely to find their work to be a challenge in two of the five areas we asked about.[53] The results, in Table IV.5, show that 77 percent of treatment focal teachers said that teaching low-performing or disadvantaged students was a challenge, compared with nearly all (95 percent) of the control focal teachers. The corresponding percentages for nonfocal teachers were 87 percent for both treatment and control. (Inasmuch as the question asked about respondents' first year in the school, nonfocal teachers are more likely to be describing experiences from an earlier period). The pattern for student discipline and classroom management, as well as for interaction with parents, were similar (control focal percentages higher than treatment focal, with nonfocal in between).

[53] The survey question asked, "To what extent did you find each of the following a challenge during your first year at this school?" and each possible challenge (listed in the table verbatim) could be marked "not a challenge," "minor challenge," or "major challenge." We computed the percentage who said minor or major challenge, but the results were robust to an alternative definition that used major challenge only.

One area in which there was a statistically significant difference in the opposite direction was in gaining support from fellow teachers. More than half of treatment focal teachers found this a challenge, compared with 34 percent of control focals. There were no significant differences for nonfocal teachers, nor were there differences between focal or nonfocal treatment and control teachers in whether it was a challenge to gain support from the principal.

Table IV.5. Teacher-Reported Challenges Associated with First Year at the School, by Focal and Nonfocal Teachers

	Focal Teachers			Nonfocal Teachers		
Outcome[a]	Treatment	Control	Difference	Treatment	Control	Difference
Teaching low-performing or disadvantaged students	77.3	95.0	-17.7*	87.3	86.5	0.8
Student discipline and classroom management	72.1	84.8	-12.7*	78.8	78.1	0.7
Interacting with parents	57.6	64.6	-7.0	62.4	63.7	-1.4
Gaining support from fellow teachers	50.7	33.9	16.9*	40.7	42.9	-2.2
Gaining support of the principal/administration	44.5	42.9	1.6	40.3	31.8	8.5
Sample Size (teachers)	87	99		199	175	

Source: Teacher Background Survey.

[a]Outcomes are percentages who said each item was a "major challenge" or "minor challenge."

*Differences are statistically significant at the 0.05 level using a two-sided test.

We also asked teachers whether they felt supported during their first year in the school. We tabulated results from these questions and compared treatment to control teacher responses in Table IV.6. For each question, we asked teachers if they strongly disagree, somewhat disagree, somewhat agree, or strongly agree with statements about the support they received in their first year at the school.[54] The table reports the percentages who agreed, either "somewhat" or "strongly" with each statement.

In the survey, which was administered in the second half of the year and asked questions about the respondents' first year of teaching in the school, focal treatment teachers were more likely to report that they felt like outsiders in their school than control focal teachers did (38 versus 23 percent, a statistically significant difference of 15 percentage points). The other four measures of whether teachers felt supported and integrated into the team went in the same direction, but were not statistically significant.

[54] The statements were: "My orientation to this school was useful," "I received material supports needed to integrate into this school (e.g., classroom equipment)," "I received social support needed to integrate into this school," "Other teachers here made me feel welcome," and "I often felt like an outsider at this school." For the last item, we reversed the coding of the variable so a higher percentage agreeing would be a positive result, consistent with the other statements.

Table IV.6. Teacher Reports on Supportive Environment, by Focal and Nonfocal Teachers

	Focal Teachers			Nonfocal Teachers		
Outcome[a]	Treatment	Control	Difference	Treatment	Control	Difference
Orientation was useful	66.5	75.8	-9.3	78.8	80.6	-1.8
Received material support	73.4	73.4	-0.1	83.9	82.5	1.4
Received social support	68.0	75.1	-7.1	84.3	82.8	1.4
Welcomed by other teachers	86.8	91.0	-4.2	91.7	86.1	5.6
Did not feel like an outsider	61.9	77.1	-15.2*	81.5	74.6	6.9
Sample Size (teachers)	87	99		199	175	

Source: Teacher Background Survey.

[a]Outcomes are percentages who "somewhat" or "strongly" agreed with the statement.

*Differences are statistically significant at the 0.05 level using a two-sided test.

In terms of satisfaction, treatment focal teachers were more satisfied with their compensation than were control focals (78 versus 62 percent), but all other differences were not statistically significant. We summarized satisfaction measures using four index variables in addition to compensation: leadership/policies, professional environment, school environment and facilities, and students and their families. The results are summarized in Appendix E, Table E.1.

Taken together, the data suggest that TTI transfers may have felt more like outsiders and felt challenged to gain the support of their colleagues relative to their control-group counterparts, but they were equally satisfied with their new schools and less likely to find their students' behavior or low achievement to be a challenge.

D. Principal Reports on School Climate and Teacher Contributions

When policymakers introduce monetary incentives for selected teachers to transfer to a new school, there could be positive or negative impacts on the new school's climate. Positive impacts could come from the fresh ideas and insights that a high-performing transfer teacher might bring. Negative impacts could come from resentment or morale problems due to differentiated pay, which can lead to a breakdown in trust and collaboration. The most cost-effective way to learn about such issues was to survey principals separately about each of their teaching teams that were included in the study. We asked about three dimensions of school climate: (1) degree of collaboration, (2) trust and mutual respect, and (3) sharing of ideas within grade teams.

On average, principals reported levels of collaboration, trust, and sharing of ideas after the intervention began that were between 3 and 4 on a 5-point scale in both treatment and control teams. We also asked principals to rate these aspects of school climate for the year before the transfers (if they had been working in the school), and we followed up with the principals in cohort 1 districts to ask about the second program year. Differences in all program years were not statistically significant. The results are shown in Appendix E, Table E.3.

We also asked principals to rate the contributions of each of their teachers to the school, and found no statistically significant differences between their ratings of treatment and control teachers. We asked about teachers' contribution to three areas: (1) leadership, (2) activities

outside the classroom, and (3) the school in general. The findings are shown in detail in Appendix E, Table E.4.

E. Summary of Intermediate Impact Findings

Taken together, the findings from this chapter suggest that TTI did have impacts on the internal dynamics of the low-achieving schools targeted by the intervention, although not along every dimension where we hypothesized that impacts could be realized. In Table IV.7, we summarize the findings.

Table IV.7. Summary of Findings on Intermediate Impacts

Question	Evidence[a]	Data Source[b]
Were students assigned differently to teachers?	Mixed	A, P, T
Were teachers assigned differently to grades?	Positive	T
Was more mentoring received by study teachers?	Insufficient	T
Was more mentoring provided by study teachers?	Positive	T
Were there differences in other leadership roles played?	Insufficient	T
Did TTI change teacher attitudes?	Positive in some areas	T
Did TTI change school climate and teacher contribution to the school?	Insufficient	P

[a]Evidence refers to results from hypotheses tests related to treatment-control differences. "Positive" refers to positive statistical difference between treatment and control teams. "Mixed" refers to results that differ in sign or significance. "Insufficient" refers to differences that were not statistically different.

[b]A = administrative data; P = Principal Survey; T = Teacher Survey.

V. IMPACTS ON STUDENT ACHIEVEMENT

To estimate the impact of the Talent Transfer Initiative (TTI) on student achievement, we compared the test-score performance of students from treatment teams to the corresponding performance of students from control teams. According to the logic model presented in Chapter I, however, we expect much of the effect to operate directly through the teacher who filled the designated vacancy, also known as the focal teacher. Therefore, we report not only the results from the team analysis but also the results of the corresponding comparisons between treatment focal and control focal teachers and between treatment nonfocal and control nonfocal teachers within those teams.

We begin this chapter with a brief description of the data and our main approach to estimating impacts, which we refer to as the benchmark model. In the rest of the chapter, we discuss findings by grade span, district, and overall.

A. Data and Methods

For the test-score impact analyses, we used student-level administrative data that capture school enrollment, test scores, student demographics, course scheduling, and student-teacher links. The test-score outcomes are grade-specific state assessments in math and reading.[55] All pre-test and post-test scores were converted into standard-deviation units, or z-scores, that express student achievement relative to the average performance for the student's own grade statewide.[56] The districts provided demographic information on students' race/ethnicity, gender, English language learner (ELL) status, special education (SPED) status, free or reduced-price lunch (FRL) eligibility, gifted status, and age.

By comparing the performance of students on state assessments for treatment teams to the corresponding performance of students in control teams, we obtain an unbiased estimate of the total impact of TTI. Because we randomly assigned teams to treatment status (as described in Chapter II), the treatment and control groups should be equivalent in every way, on average, except for treatment status itself. The only differences are those that arise by chance. We reported some of those chance differences in terms of observable characteristics in Chapter II. We controlled for such differences using a regression model that includes students' background characteristics.[57]

[55] One of the 10 study districts administered different tests within the same grade and subject (algebra and pre-algebra in grade 8). For that district only, we used a linking procedure to convert pre-algebra scores to predicted equivalent scores on the algebra test. See Appendix F, "Test Score Scaling Issues," for details.

[56] The z-scores are calculated by subtracting the statewide mean scaled score for all students in that year and grade from a student's scaled score and then dividing that by the statewide standard deviation of scaled scores for that same group. Z-scores greater than 4 or less than -4, which make up less than 0.1 percent of the sample, are assumed to be data errors and are set to missing.

[57] The control variables include the student's same-subject pre-test z-score, race/ethnicity, gender, ELL status, special education status, and FRL status, as well as over-age-for-grade status, an indicator of whether the student belonged to a study team that has at least one retention-stipend teacher, grade dummies, block dummies, and imputation dummies. For more details, see Appendix F, "Benchmark Regression Model."

We conducted impact analyses separately by elementary and middle schools for math and reading in three ways: a team analysis that consists of all students in the study teams, a focal-teacher analysis that consists only of students in study teams taught by focal teachers, and a nonfocal-teacher analysis that consists only of students in study teams taught by nonfocal teachers.[58] To be included in the analysis samples for math or reading, a student must have a valid and nonmissing post-test score in that subject in addition to enrolling in a study team.

We tested the robustness of the test-score impact findings to a wide variety of alternative methods. We augmented the benchmark model with additional covariates, adjusted modeling assumptions, and used alternative sample inclusion rules. For example, we adjusted the specification of the pre-test covariate in the benchmark model, and corrected the pre-test covariate in the benchmark model for measurement error. The impact estimates fluctuated more when the sample changed (for example, when we used the alternate definition of focal teacher in Tables F.14 and F.15), but the results also showed a pattern of heterogeneous impacts like the one presented in this chapter (see Appendix F, Table F.13 for a list of all the sensitivity analyses we implemented). We also addressed concerns regarding the possibility of downward-biased impact estimates due to the presence of both treatment and control teams within some schools. These results are found in Appendix F, Tables F.9–F.12.

B. Impacts in Elementary and Middle Schools

1. Main Findings

In Table V.1, we show the benchmark impact estimates by grade span and program year based on three types of treatment-control comparisons: team, focal, and nonfocal. Following the logic model introduced in Chapter I, the team impact represents the overall impact of TTI on all students in study teams, regardless of whether they were taught by focal or nonfocal teachers. Teams were the units of randomization. Program year 1 impacts were estimated using data from all 10 districts; program year 2 impacts were estimated using data from cohort 1 districts only because the study plan did not include data collection from the 3 cohort 2 districts for 2011–12.[59]

[58] Throughout this chapter, we report results from the benchmark model that uses an inclusive definition of both focal and nonfocal teachers. That is, on teams where it was unclear which teacher filled the vacant position or where no teacher filled the vacant position (because it was lost due to declining enrollment, for example), we used a weighted average of all the possible focal teachers for the focal-teacher analysis and a weighted average of the possible nonfocal teachers for the nonfocal analysis. Students of teachers whose status was truly unknown appear in both analyses. We repeated the focal and nonfocal analyses using a selective definition as well, and discuss them in Appendix F, Tables F.11 and F.12. Refer to Appendix D for details on how weights were assigned to each teacher.

[59] Refer to Appendix F, Tables F.1 and F.2, for program year 1 impacts estimated using data from cohort 1 districts only. A similar story emerges, so we present the full sample results for program year 1 here and throughout this chapter.

Findings in Table V.1 indicate that TTI elementary school focal teachers had positive impacts on their students in both program years and both subjects.[60] The impacts were 0.18 standard deviations for math and 0.10 standard deviations for reading in the first year, equivalent to increases of 7 and 4 percentile points, respectively, assuming students start at the 40th percentile relative to all other students in their state. In the second year, these impacts were 0.22 and 0.25 for math and reading, respectively, equivalent to increases of 9 and 10 percentile points. When we consider team comparisons, the impact estimates were positive in both subjects and program years but significant only in the second year of implementation. None of the impacts for nonfocal teachers was significantly different from zero.

Table V.1. Test-Score Impacts in Elementary Schools

Program Year, Subject, and Comparison Type	Impact	Standard Error	*p*-Value	Sample Size[a]
Year 1 (cohorts 1 and 2)				
Math				
Team	0.05	0.04	0.204	8,177
Focal teacher	0.18*	0.05	0.000	3,751
Nonfocal teacher	-0.07	0.04	0.104	6,516
Reading				
Team	0.03	0.03	0.380	8,097
Focal teacher	0.10*	0.05	0.043	3,804
Nonfocal teacher	-0.01	0.04	0.849	6,642
Year 2 (cohort 1 only)				
Math				
Team	0.08*	0.04	0.049	7,565
Focal teacher	0.22*	0.06	0.001	3,327
Nonfocal teacher	0.00	0.04	0.951	6,240
Reading				
Team	0.07*	0.03	0.020	7,481
Focal teacher	0.25*	0.05	0.000	3,201
Nonfocal teacher	0.00	0.03	0.958	7,024

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the focal- and nonfocal-teacher comparisons, sample size refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in the analyses of focal and nonfocal teachers because they can be linked to more than one teacher. Students in the analyses of focal and nonfocal teachers are weighted proportionately to the probability that the teacher to whom they are linked is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples. Information on the number of teams, teachers, and students that are included in each analysis can be in found in Appendix F, Table F.28.

When assessing the impacts of program years 1 and 2, it is important to keep in mind two intrinsic differences in these sets of impacts. One difference between the two program years is that the second-year impacts are based on a sample that does not include cohort 2 districts. To remove this difference, we also compared first-year and second-year impacts for the same (smaller) sample of cohort 1 districts only. The results, presented in Appendix F, Table F.1, fit into the same pattern of significant impacts in year 1.

[60] The estimate for focal teachers in reading was not significant at the 0.05 level (p-value = 0.087), but otherwise fits into a pattern of positive impact estimates.

A second difference is that by the second program year, the transfer teachers and principals had had more time to respond to and adjust to TTI. In the second year, the composition of teachers in the study grades changed somewhat: some TTI transfer teachers left, and principals had opportunities to assign teachers strategically to grades and subjects after having observed a year of TTI. Any one or a combination of these factors can explain the observed results.

In middle schools (Table V.2), we did not find evidence that TTI was effective. The impact estimates were all statistically insignificant for program years 1 and 2, except for the year 2 focal-teacher impact on reading, which was negative (impact = -0.06, *p*-value = 0.031).

Table V.2. Test-Score Impacts in Middle Schools

Program Year, Subject, and Comparison Type	Impact	Standard Error	*p*-Value	Sample Size[a]
Year 1 (cohorts 1 and 2)				
Math				
Team	-0.02	0.06	0.694	8,875
Focal teacher	0.04	0.09	0.633	2,827
Nonfocal teacher	-0.05	0.05	0.381	8,549
Reading				
Team	0.02	0.03	0.442	7,812
Focal teacher	0.01	0.05	0.831	3,261
Nonfocal teacher	0.03	0.03	0.311	7,224
Year 2 (cohort 1 only)				
Math				
Team	-0.02	0.06	0.726	2,627
Focal teacher	0.03	0.06	0.654	1,575
Nonfocal teacher	-0.07	0.07	0.345	1,788
Reading				
Team	-0.02	0.02	0.317	3,488
Focal teacher	-0.06*	0.02	0.031	2,090
Nonfocal teacher	-0.05	0.03	0.068	2,898

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the focal- and nonfocal-teacher comparisons, sample size refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples. Information on the number of teams, teachers, and students that are included in each analysis can be in found in Appendix F, Table F.29.

2. Is Middle School Really Different?

Because we reach different conclusions by looking at elementary schools versus middle schools, one might conclude that the effectiveness of TTI depends on grade span. However, one should exercise caution when generalizing the findings based on the difference in impacts in elementary and middle schools. The difference may be because of real differences in TTI effectiveness in elementary versus middle schools, but they may also be partially driven by heterogeneous district impacts because the shares of elementary and middle school teams differ across districts. One district contributed only elementary school teams to the study; two districts contributed only middle school teams.

When we formally tested whether the elementary and middle school impacts were statistically different from each other, results were mixed.[61] For the first program year, which included the full sample (cohorts 1 and 2), the difference between the elementary and middle school impacts were not statistically significant in either math or reading. For program year 2, which included only the cohort 1 districts, we did find that the difference in elementary versus middle school impacts were statistically significant in both tested subjects. To understand whether the differing hypothesis test results were because of the inclusion of cohort 2 or because of the difference in program year, we conducted one additional test. We restricted the sample to cohort 1 and repeated the elementary-middle difference test for program year 1. The result was that difference in impacts remained statistically insignificant, suggesting that the real difference by grade span was specific to program year 2. To further explore the question of differential impacts, we next look at impacts by district.

C. Impacts by District

The distribution of district-level TTI impacts provides evidence on the degree of consistency of impacts across different settings. We examined the distribution of district-level TTI impacts across all possible combinations of grade spans, subjects, comparisons and program years. In all but two cases,[62] we found that impact estimates varied by district more than we would expect under normal sampling variation if TTI were equally effective in all sites and all of the variation were due to sampling error.[63] This means that future implementers should plan for the possibility of results that differ from the averages presented in this report.

In Figures V.1 and V.2, we illustrate the variation in impact estimates across districts. These figures show the focal-teacher impacts in elementary schools by district for math in program years 1 and 2, respectively. In both figures, the district impacts are sorted by the size of the impacts, starting from the left with the district with the lowest impact estimate. The horizontal line in the figures represents the size of the overall focal-teacher impact for math in elementary schools. Focal-teacher impacts on math scores in program year 1 in elementary schools range from -0.15 to 0.57 of a standard deviation (Figure V.1); the corresponding program year 2 impacts range from -0.24 to 0.62 of a standard deviation (Figure V.2).[64] Elementary school TTI impacts do not appear to be driven by results from one or two outlier districts. Focal-teacher district-specific math impacts tended to be greater in districts that also had greater reading impacts (program year 1 correlation = 0.79; program year 2 correlation = 0.66).

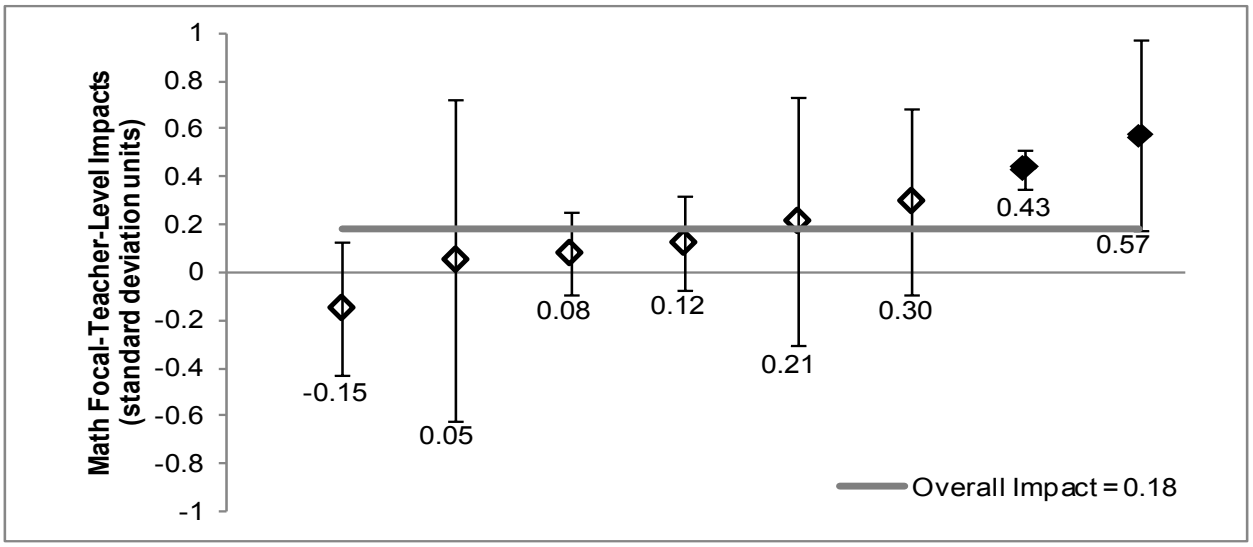
[61] The test was conducted by including an interaction between treatment status and grade span and testing whether the coefficient on the interaction term was significantly different from zero with a traditional t-test.

[62] The two cases are program year 2 impacts for elementary focal teachers in reading (Appendix F, Figure F.4) and program year 2 impacts for elementary nonfocal teachers in math (Appendix F, Figure F.8).

[63] For each subject, we conducted an F-test to determine if the district-specific impacts are jointly equal to one another at the 5 percent level.

[64] District-level TTI impacts are based on smaller sample sizes than average TTI impacts. Thus, the aim of this section is to examine the distribution of the *magnitudes* of district-level impacts, rather than to examine the prevalence of *statistical significant* district-level impacts. The results for reading are presented in Appendix F, Section F.4.

Figure V.1. Year 1 Impacts on Math Scores, Elementary Focal Teachers, by District (cohorts 1 and 2)

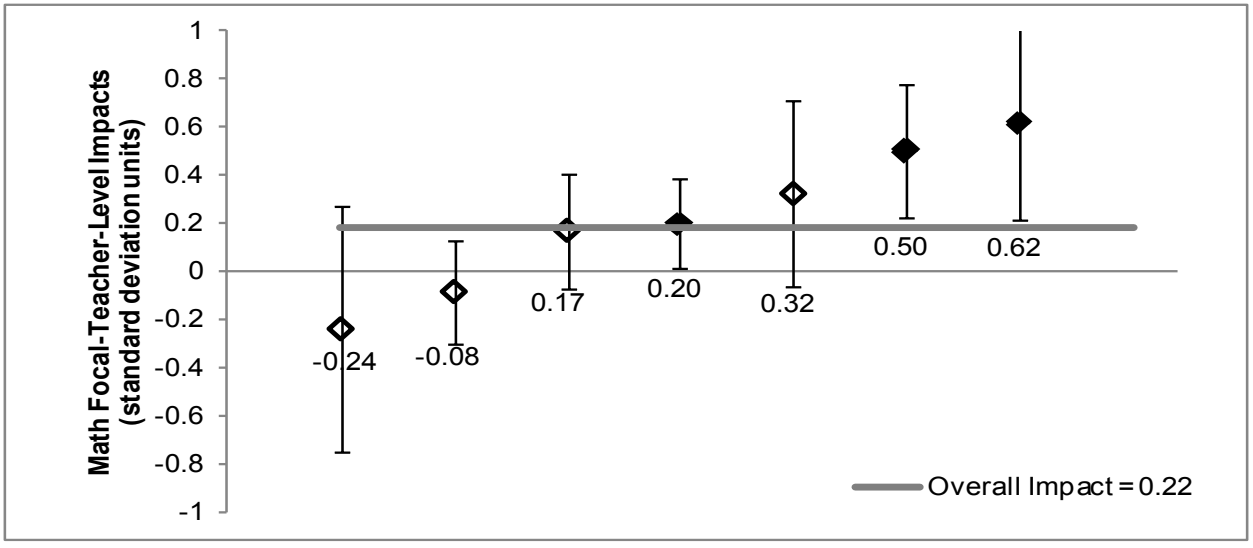


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a solid horizontal line denotes a statistically significant overall impact. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

Two study districts do not have elementary school teams. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 99 to 1,395. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure V.2. Year 2 Impacts on Math Scores, Elementary Focal Teachers, by District (cohort 1)



Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a solid horizontal line denotes a statistically significant overall impact. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 97 to 1,446. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Elementary school team and nonfocal-teacher district impacts in both subjects, elementary school focal-teacher district reading impacts, and middle school impacts in both subjects can be found in Appendix F, Figures F.1 to F.22.

D. Combined Elementary and Middle School Impacts

The variation in TTI impacts by district, and the fact that elementary-middle school differences in impacts were not statistically significant for program year 1, suggests an explanation more nuanced than simply that TTI has positive impacts in elementary schools but not in middle schools. We are unable to disentangle whether TTI was effective in certain grade spans, in certain districts, or a combination of these two factors. Therefore, we computed the impact for the entire sample combined, which we use as part of the cost-benefit analysis in the last chapter of this report. The results are shown in Table V.3. As one would expect, the combined elementary and middle school estimates lie in between the elementary and middle school results shown above in Tables V.1 and V.2.

There are two equally valid ways to construct the overall mean impacts, each of which involves different weights applied to the observations in our study sample. The benchmark model places equal importance on each occurrence of a student's observation, which effectively allocates more weight to teams and teachers with more students and teachers. This approach is attractive because it measures the impact of TTI on the average *student* and gives more weight to teams that contribute more evidence. An equally plausible alternate approach is to give each *vacancy* equal weight. For the team analysis, this means that each TTI vacancy is given equal weight, regardless of team size (number of teachers or students). For the focal-teacher and nonfocal-teacher analyses, this gives equal weight to each teacher, regardless of the number of students he or she teaches. The benchmark approach answers the policy question, "What is the impact of TTI on the average student?" The alternate approach answers a different policy question: "What is the average impact of TTI on each team or classroom?" In Table V.3, we show the benchmark and vacancy-weighted estimates for program years 1 and 2, pooling together study teams from elementary and middle schools. Both are considered in our cost-benefit analysis presented in the final chapter of this report.

The combined benchmark-team impacts were not statistically significant, except in program year 2 for reading, and all focal-teacher impacts were statistically significant except in program year 1 for reading. Also, the nonfocal impact for program year 1 math was negative and statistically significant (impact = -0.06 standard deviations).

The vacancy-weighted impact estimates are greater than or equal to the student-weighted benchmark estimates: TTI had a positive significant impact on the team's math and reading scores in both program years. The effect sizes of the team impacts were 0.05, 0.06, 0.10, and 0.09 for program year 1 math, year 1 reading, year 2 math, and year 2 reading, respectively. The focal-teacher impact estimates were at least 0.10 in year 1, and 0.21 in year 2. Only the nonfocal-math impact estimate in year 1, which was -0.06 and statistically insignificant under the benchmark model, became lower (-0.07) and statistically significant for the alternative (vacancy-weighted) model.

We interpret the results using the vacancy weights as another indication of heterogeneous TTI impacts for different subgroups. By placing an equal weight on each classroom, the

vacancy-weighting approach gives smaller classrooms more weight as compared with the benchmark model.

Table V.3. Combined Test-Score Impacts, Benchmark and Vacancy Weighted

Program Year, Subject, and Comparison Type	Benchmark Impact	Benchmark *p*-Value	Vacancy-Weighted Impact	Vacancy-Weighted *p*-Value	Sample Size[a]
Year 1 (all districts)					
Math					
Team	0.00	0.901	0.05	0.091	17,052
Focal teacher	0.10*	0.027	0.14*	0.002	6,578
Nonfocal teacher	-0.06*	0.047	-0.07*	0.022	15,065
Reading					
Team	0.03	0.105	0.06*	0.008	15,909
Focal teacher	0.07	0.055	0.10*	0.006	7,065
Nonfocal teacher	0.02	0.545	0.02	0.499	13,866
Year 2 (cohort 1 only)					
Math					
Team	0.05	0.141	0.10*	0.010	10,192
Focal teacher	0.12*	0.021	0.21*	0.000	4,902
Nonfocal teacher	-0.01	0.873	0.03	0.504	8,028
Reading					
Team	0.05*	0.039	0.09*	0.000	10,969
Focal teacher	0.13*	0.001	0.21*	0.000	5,291
Nonfocal teacher	-0.01	0.811	0.03	0.377	9,922

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the focal- and nonfocal-teacher comparisons, sample size refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in the analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples. Information on the number of teams, teachers and students that are included in each analysis can be in found in Appendix F, Table F.30.

Another explanation for the increase in team math and reading impacts is the dilution factor. A transfer-incentive intervention has greater potential for impact on the team if there are more vacancies as a fraction of all positions on the team. This is because the smaller the team, the greater the influence of any one transfer teacher on that team. By applying vacancy weights, we generate team-impact estimates that are pulled toward the impact of the average-sized team instead of giving extra influence to larger teams with more students.

E. Interpreting the Impact Estimates

1. TTI Impacts Versus Effectiveness of Transfer Teachers

The impact estimates presented in this chapter represent the effect of having the opportunity to hire a high-performing teacher identified by TTI on the teacher team regardless of whether or not the vacancy was filled with a TTI candidate. This impact of the "opportunity to hire" is

different from the effectiveness of the actual transfer teacher. The average impact of the transfer teachers themselves may be slightly larger than the estimated impact of TTI we reported above once we account for the treatment vacancies not filled through TTI.

The most common way to account for unfilled vacancies on treatment teams is to assume that the impact of TTI was zero for all such teams, so all of the observed impacts are attributable to the teams that had a TTI transfer teacher.[65] Following Bloom (1984), we can obtain estimates of transfer-teacher effectiveness by dividing the team and focal-teacher impacts by an adjustment factor equivalent to the proportion of students in treatment teams taught by TTI transfers.

Applying this adjustment, we calculated that the program year 1 impacts of TTI could be 14 to 38 percent higher than the estimates presented above suggest. For math, the resulting elementary school team and focal adjusted estimates in program year 1 are 0.06 (adjustment factor = 0.83) and 0.24 (adjustment factor = 0.77) of a standard deviation respectively; for reading, the elementary school team and focal adjusted estimates in program year 1 are 0.03 (adjustment factor = 0.86) and 0.13 (adjustment factor = 0.80) of a standard deviation, respectively. In other words, the magnitudes of the impact estimates shown in Table V.1 (and in other tables and figures throughout this chapter) would be slightly larger if we were to take into account the fact that not every treatment-team vacancy was filled with a TTI teacher. Throughout this report, we present estimates of the opportunity to hire through TTI, which can be converted to impacts of transfer teachers themselves by inflating them by approximately 5 to 50 percent, depending on the comparison, subject, grade span, and program year. The specific adjustment factors corresponding to any given estimate can be requested from the authors.

2. Resource-Allocation Effects

We noted that the focal and nonfocal treatment-control comparisons can be regarded as estimates of the direct and indirect effects, respectively, of TTI as long as we assume no resource-allocation effects (redistribution of resources) within the team. Furthermore, the team-impact estimates fully capture the impact of TTI if there is no resource allocation across teams in the same school. In reality, there are likely to be at least some strategic reallocations of resources in response to TTI, as documented in Chapter IV and Appendix E.

If treatment focal teachers were given fewer resources and more challenging workloads than their nonfocal peers relative to what occurred in control teams, the direct impacts understate the true (direct) effect of TTI teachers and the nonfocal-impact estimates overstate the indirect effects. If the opposite is true—that the net resource allocation favors treatment teachers—then the focal-teacher impacts will overstate the direct effects and the nonfocal estimates will understate the indirect effects.

The data presented in Chapter IV were mixed regarding the possibility of resource-allocation effects within study grades. There were no detectable differences in terms of the prior test scores of treatment and control focal teachers' students relative to their nonfocal peers. Similar numbers of teams had unusually high- or unusually low-performing students. When teachers themselves reported having more challenging students, it was treatment focal teachers

[65] We consider the opportunity to hire a TTI teacher as the "intent to treat" and the impact of the transfer teacher as the effect of treatment on the treated.

whose reported rates of "academically more challenging" students were highest, but there was no corresponding offset among their nonfocal peers. Evidence on mentoring also suggests that if there was a resource-allocation effect, it was for treatment focal teachers to help their peers, so it should have equalized the differences in focal and nonfocal effects, not magnified them. Focal teachers used fewer and provided more mentoring resources rather than the other way around.

We also documented evidence of resource-allocation effects across grades in a direction that could lead us to understate the true impact of transfer teachers. That is, principals under the status quo, represented by control school teams, may compensate for weak incoming teachers by moving strong peers—veterans who can mentor their less-experienced colleagues—from elsewhere in the school into their grade team. Equivalently, when TTI is introduced, principals may move weak teachers into the grades with TTI teachers, using the same strategy of pairing weak teachers with strong ones, or at least more-experienced ones with less-experienced ones. If principals do pair strong teachers with weak ones, that compensating behavior would bias impact estimates downward.

Data on the average level of experience, presented in Chapter IV, showed that control nonfocal movers had almost five years more experience, on average, than treatment nonfocal movers. The implication is that our estimates of the impact of TTI may understate the full benefit of the intervention for participating schools due to potential resource-allocation effects.

VI. IMPACTS ON TEACHER RETENTION

Talent Transfer Initiative (TTI) teachers were offered $20,000 in five installments over a two-year period to transfer to and continue teaching in a low-achieving school in their district. In Chapter III, we examined whether the promise of this monetary incentive was sufficient to attract teachers to apply and transfer to study schools. By analyzing the rates of teacher retention on treatment and control teams, we can more fully understand teachers' responses to this incentive after the transfers. We examine whether the incentive was sufficient to keep teachers at their new schools both during the intervention—while TTI teachers were still receiving incentive payments—and after the incentive payments ended.

Teacher retention is also relevant for understanding the impact of the intervention on student achievement. In Chapter V, we presented evidence that students of treatment focal teachers in elementary schools performed significantly higher on math and reading standardized tests than their control peers in the first year of the study, and these differences were larger in the second year. If treatment focal teachers positively affect their students' performance, retaining those teachers is necessary for the effects to persist. In this chapter, we focus on retention relative to the control group.

A. Data and Methods

Information on teacher retention is based on teacher rosters we collected from districts and schools in the fall of program years 1 and 2 for all 10 districts. We also collected teacher rosters in the fall after the completion of the intervention for the 7 cohort 1 districts.

We used the teacher rosters to identify the teachers on study teams in the fall of program year 1, immediately after TTI teachers transferred into treatment positions. The teacher sample for the team-level retention analysis is based on teachers who are on the study teams, including focal and nonfocal teachers.[66]

We used the background survey and information collected from principals and The New Teacher Project (TNTP) implementation team to identify the focal teachers on study teams.[67]

The retention analysis focused on within-school retention over three years. We measured (1) one-year school retention by tracking whether study teachers taught in the same school in the fall of program years 1 and 2 and (2) two-year school retention based on their teaching in the same school in the fall of program year 1 and the fall of the year after TTI payments ended. Two-year retention was measured for the seven cohort 1 districts only.

[66] Thirty teachers on study teams were designated at the start of the study as highest performing and were eligible for $10,000 retention stipends paid out in installments over two years. These retention-stipend teachers taught on both treatment and control study teams but were not eligible to be focal teachers because they were already in study positions before program year 1. We have no way to estimate the counterfactual for these teachers, but we examine their retention rates in a non-experimental analysis in Appendix G.

[67] In Appendix D, we describe the process of identifying focal teachers.

We adjusted the retention rates using a linear regression that accounts for the block randomized design of the study. The regression model also accounts for whether the nontransfer teachers were receiving a retention stipend—which was not part of the treatment—because a retention stipend could have affected teachers' decisions to stay in their positions throughout the study.[68]

We estimated the intervention's impact on teacher retention for the full sample, including all teachers on study teams. We also compared teacher retention within the focal sample, including only treatment and control focal teachers.[69] We can think of these focal-teacher comparisons as unbiased estimates of impact under the assumption of zero resource-allocation effects. Examining the intervention's impact on retention of focal teachers is a more direct test of the impact of the incentive payments and transfer program; the team-level impact allows us to test how the intervention affects the retention of the team as a whole, including nonfocal peer teachers. We also estimated retention for elementary and middle school teachers separately to examine whether the intervention's impacts on retention differed by grade span.[70]

B. Retention Impacts

Our primary discussion of retention impacts focuses on cohort 1 districts, for which the study collected data for years 1 and 2. TTI had a significant positive impact on one-year retention of teachers in their schools, but the positive impact did not persist in cohort 1 after they completed the intervention. In Figure VI.1, we show the one- and two-year retention rates for cohort 1 teachers.[71] The figure shows that 93 percent (regression adjusted) of treatment focal teachers remained in their schools while they were receiving incentive payments, but ultimately left at rates similar to other teachers on study teams once they were no longer eligible for payments at the end of the second year. Sixty percent of treatment focal, treatment nonfocal, and control nonfocal teachers returned to their schools in the fall after the intervention.[72]

[68] We repeated this analysis, referred to as the "benchmark," in several ways to check the robustness of the findings to different methods and samples. For example, we used a logistic regression instead of a linear regression, a random-effects model instead of a fixed-effects model, and alternative sets of covariates. The benchmark results are robust across most sensitivity tests, although the impact estimates are smaller when controlling for team-level student characteristics.

[69] Throughout this chapter, we report impacts based on the inclusive definition of focal teachers. This means that all study teams contribute to the estimates, even if we are uncertain about which teacher filled the study vacancy, or if the vacancy was lost. For teams for which there is uncertainty about the identity of the focal teacher, we identified all potential focal teachers and assigned each an equal weight that sums to the number of vacancies originally identified in the random assignment process. In Appendix G, we compare the retention impacts using both the inclusive and selective definitions of focal teachers. The impact estimates remain statistically significant when using the selective definition.

[70] Retention was also estimated by district. District-specific results show there is variation, but the impacts are not significantly different across districts. See Appendix G for the results of the district-specific analyses.

[71] As described above, we did not have data to measure the two-year retention rate for cohort 2 teachers.

[72] The retention rates presented in Figure VI.1 can be compared with one-year retention rates reported in the 2008–09 Teacher Follow-Up Survey. In this national sample, 84.5 percent of all teachers stayed in the same school between 2007–08 and 2008–09; during these years, 77.3 percent of teachers in their first three years of teaching stayed in the same school (http://nces.ed.gov/pubs2010/2010353.pdf).

In Figure VI.1, the solid lines represent the retention rates of focal teachers on study teams, and the dotted lines represent nonfocal teachers. Program year 1 represents the first roster-collection time point, immediately after TTI teachers transferred into study schools. The data points at year 2 show the percentages of teachers in each group that remained in their schools between years 1 and 2 (one-year retention). The data points post-program show the percentages of teachers in each group who remained in their schools between years 1 and 3 (two-year retention).[73]

In the graph, we illustrate the statistically significant 22.3-percentage-point[74] difference between treatment and control focal teachers for one-year retention, and the statistically insignificant 9-percentage-point difference for two-year retention. Although TTI did not result in a significant impact on retention after the program's conclusion, treatment focal teachers did not leave *en masse* immediately after the incentive payments ended. In fact, they stayed in the study schools at the same rate as treatment and control nonfocal teachers.

This figure also clearly demonstrates that the presence of a TTI teacher on a team had no impact on the retention of other teachers on the team. The retention patterns of treatment and control nonfocal teachers are very similar, and not statistically distinguishable, throughout the study period.

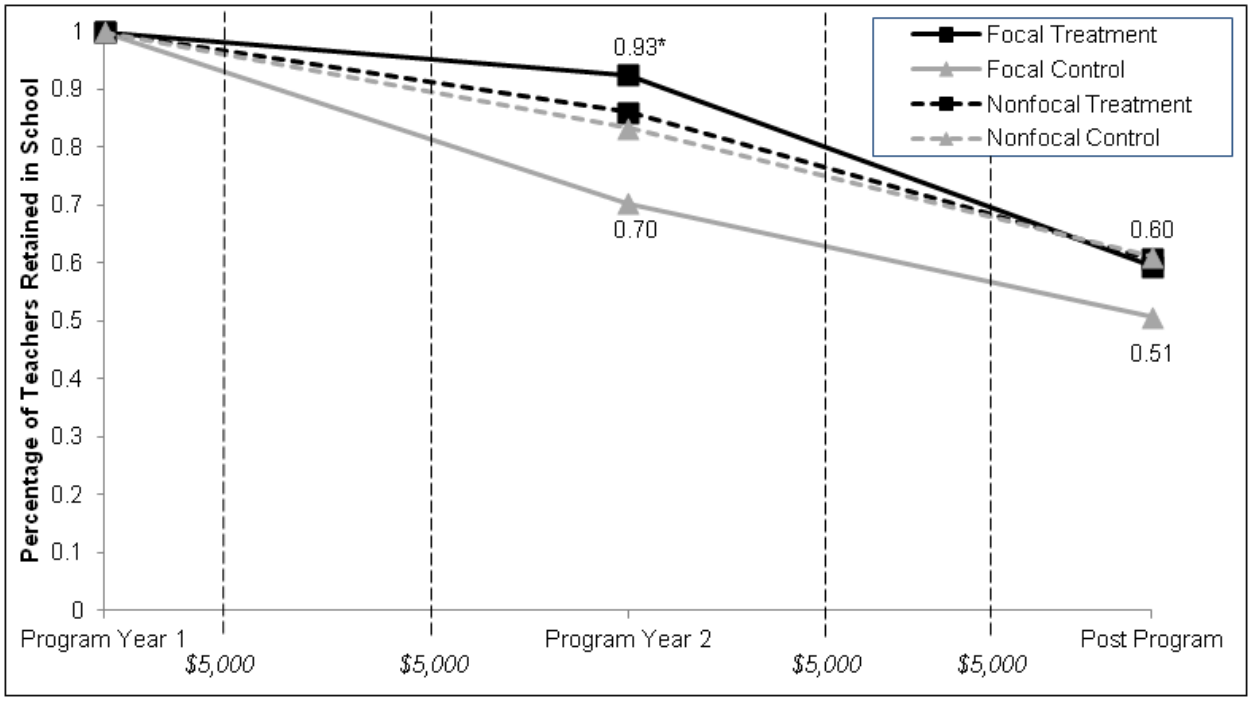
Figure VI.1 presents data only from cohort 1, but we found a similar one-year retention pattern in the full sample, which includes both cohorts. In Table VI.1, we show that in both the cohort 1 sample and the full sample, there is a statistically significant 7-percentage-point difference in retention rates between treatment and control teachers.

The team-level impacts appear to be driven almost entirely by differences in retention for focal teachers, rather than by the effect of focal teachers on their peers. When limiting the sample to focal teachers, we found a 22-percentage-point impact in the cohort 1 sample and a 23-percentage-point impact in the full sample. In contrast, the retention rates of nonfocal teachers on treatment and control teams were statistically indistinguishable (in the full sample, 79 percent on treatment teams, compared with 77 percent on control teams).

[73] See Table VI.1 for point estimates, impact estimates, p-values, and sample sizes that correspond to the one-year impacts in Figure VI.1. See Appendix G, Table G.6, for the two-year impacts in Figure VI.1.

[74] This impact estimate is calculated as the difference between the unrounded treatment and control means. The treatment and control means (0.93 and 0.70, respectively) presented in Figure VI.1 and Table VI.1 are rounded means.

Figure VI.1. Impacts on Retention in School, Cohort 1 Districts Only



Source: School rosters.

Note: N = 80 focal treatment teachers, 96 focal control teachers, 193 nonfocal treatment teachers, and 183 nonfocal control teachers.

Vertical dotted lines represent points at which TTI teachers received incentive payments. Note that the first $5,000 payment was paid in two installments of $2,500. One installment was paid before the start of the first school year, and the other was paid in the fall of the first school year.

*Statistically significant at the 0.05 level, two-tailed test.

Table VI.1. One-Year Impacts on Retention in School

Sample	Treatment	Control	Impact	*p*-Value	N
Cohort 1					
All teachers on study teams	0.87	0.80	0.07*	0.021	498
Focal teachers	0.93	0.70	0.22*	0.002	176
Nonfocal teachers	0.86	0.83	0.03	0.455	374
Cohorts 1 and 2					
All teachers on study teams	0.81	0.74	0.07*	0.024	725
Focal teachers	0.89	0.66	0.23*	0.000	230
Nonfocal teachers	0.79	0.77	0.02	0.570	559

Source: School rosters.

Note: Data from cohort 1 districts are from 2010–11 and 2011–12; data from cohort 2 districts are from 2011–12. The samples of focal and nonfocal teachers are not mutually exclusive because they are based on the "inclusive" definition of focal and nonfocal (see Appendix D for details).

*Statistically significant at the 0.05 level, two-tailed test.

We also found one-year retention impacts for focal teachers in the elementary and middle school subgroups (Table VI.2). At the elementary level, the one-year retention rate for treatment focal teachers was 22 percentage points higher than for control focal teachers. Middle school teachers on both treatment and control teams had somewhat lower retention levels than elementary school teachers, but the impact of TTI is positive and significant for middle school teachers, with a 25-percentage-point difference between treatment and control.[75]

A closer look at the middle school results shows that one-year retention impacts were concentrated in middle school math. Although the sample for middle school math includes only 42 focal teachers, the intervention had a statistically significant impact of 41 percentage points on one-year retention of math teachers. The impact for middle school English/language arts (ELA) teachers was 11 percentage points and was not statistically significant.

Table VI.2. One-Year Impacts on Retention in School for Focal Teachers, by Grade Span

	Treatment	Control	Impact	*p*-Value	N
Elementary	0.92	0.70	0.22*	0.007	154
Middle School	0.83	0.59	0.25*	0.015	76
Middle school math	0.93	0.52	0.41*	0.001	42
Middle school ELA	0.75	0.64	0.11	0.493	34

Source: School rosters.

*Statistically significant at the 0.05 level, two-tailed test.

In Appendix G, Tables G.6 and G.7, we present one- and two-year retention impacts by grade span and subject at the team, focal, and nonfocal levels. The one-year impact estimates at the team and nonfocal levels are not statistically significant; they follow a pattern similar to the impacts estimated on the full sample. The team-level impact estimates are smaller than the focal-teacher impacts, and the nonfocal impact estimates are close to zero. None of the two-year retention impacts is statistically significant.

The findings from this chapter suggest that TTI had a significant impact on teacher retention, primarily through the focal teachers. Treatment focal teachers had significantly higher one-year retention rates than control focal teachers, who represent "business as usual" in the absence of TTI. We observe this trend in both elementary and middle schools, but the impact is largest for middle school math teachers. Although we did not find a significant impact of TTI on two-year retention, treatment focal teachers did not leave their schools at higher rates than other teachers after they stopped receiving incentive payments.

[75] Retention rates for elementary and middle school teachers were estimated by splitting the sample into subgroups. An alternate model with grade span-treatment interaction variables estimated on the full sample yields the same estimates, and an F-test confirms that there is no significant difference between the elementary and middle school impacts ($p = 0.812$).

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

VII. COST-EFFECTIVENESS

This report has provided evidence on the implementation experiences and impacts of the Talent Transfer Initiative (TTI). We showed that filling teaching vacancies by using transfer incentives, as designed for this study, is feasible. Nearly all vacancies (88 percent) on teams assigned to TTI were filled with an eligible high-performing teacher. However, the conditions that made this possible are important, and they have cost implications. For each of the 81 positions ultimately filled, 19 transfer candidates had to be identified. A dedicated site manager in each district devoted the equivalent of one-third to one-half of his or her work time for about five months in each district to facilitate the transfer process. Retention rates near 90 percent meant that nearly all of the promised incentive payments were claimed.

In terms of impacts on test scores, TTI elementary school teachers had positive and statistically significant impacts for both reading and math in both implementation years. Focal-teacher impacts were 0.18 and 0.10 standard deviations in math and reading, respectively, in program year 1, and 0.22 and 0.25 standard deviations in year 2. When we consider teams, which include focal teachers and nonfocal teachers, we see that impacts were positive and significant for both subjects only in year 2. We did not find any statistically significant effects at the middle school level in program years 1 or 2, except for a focal-teacher impact on reading in year 2, which was negative (impact = –0.06) and statistically significant.

In terms of impacts on teacher retention, we found that TTI payments, spread over two years, caused many teachers to delay their exits from their new schools: 89 percent of the treatment focal teachers returned after the first year, compared with 66 percent in the control schools. After the second program year, when TTI transfer teachers had already received their final payments, 60 percent of the treatment focal teachers were still in their schools. This was not statistically different from the retention rates of control focal teachers. Thus, TTI transfer teachers had not all exited by the time payments ended, so the transfers were permanent in the sense that the teachers on treatment teams were no more likely to exit low-achieving schools than teachers who had not been part of TTI.

A key question is whether the findings were sufficiently beneficial to warrant the expense of the intervention. In the remainder of this chapter, we compare the cost of implementing TTI to the estimated costs of generating impacts as large as those of TTI from an alternative intervention—class-size reduction (CSR)—whose costs are easy to determine and whose impacts have been estimated from a randomized experiment.

A. Cost-Effectiveness Methods

Our approach to TTI cost estimation was to identify the categories of cost that would be part of a future implementation of an intervention like TTI, calculate the cost within each category, and divide the sum of those costs by the number of teaching teams exposed to the intervention. To provide a point of comparison, we estimated the cost of generating similar-sized test-score impacts using CSR. This method assumes that we can generate a given (known) increment in test-score impact by spending a specific amount on CSR. This may be a strong assumption, but it provides a useful benchmark.

Calculating the cost of student achievement gains generated from CSR. We used the Tennessee class-size experiment documented by Mosteller (1995) to estimate the cost of generating a specific size of impact on student achievement from an alternative intervention. The Tennessee study was a randomized controlled trial in which 6,500 students in 330 classrooms in approximately 80 schools were randomly assigned either to small classrooms (13 to 17 students per teacher) or to larger classrooms (22 to 25 students). Using the national average annual teacher salary of $55,000 (U.S. Department of Education 2012), we compared the resulting cost estimate per student to the reported impact per student on standardized math and reading tests (0.27 and 0.23 standard deviations, respectively). This yielded a cost of $4,724 and $5,545 for each standard-deviation increase in math and reading test scores.[76] Using the average of these two numbers, we assumed that it costs $5,134 per student to raise student test scores by one standard deviation based on the CSR intervention.

Although the Tennessee study included a long-term follow-up that showed lasting benefits for students exposed to several years of reduced class sizes, the impact of a single year of intervention was not found to persist (Finn et al. 2001), so this calculation reflects the one-time impact associated with a one-time cost, which is appropriate to the current evaluation.

Units. We expressed all costs in dollars per team instead of dollars per vacancy or dollars per transfer teacher in order to make them comparable to the impacts of TTI, which are calculated at the team level. In the current study, there were 1.08 targeted vacancies per team and 0.95 filled vacancies (transfers) per team, so one can adjust slightly upward or downward in accordance with these values to obtain costs per vacancy or per transfer.

Discounting. We accounted for the fact that costs incurred later in time are worth less than costs incurred in the present. We converted all costs incurred after the first year to their present value using an annual discount rate of 1.6 percent. This means, for example, that $1,016 in costs incurred in program year 2 is equivalent to $1,000 incurred in program year 1.[77]

Combining math and reading impacts. We took the simplest approach and assumed that policymakers give a standard-deviation change in math the same value as an equal change in reading. We computed an average impact and average cost-per-unit increase in student achievement. Estimates based on each subject individually, holding the other one constant, are also available from the authors.

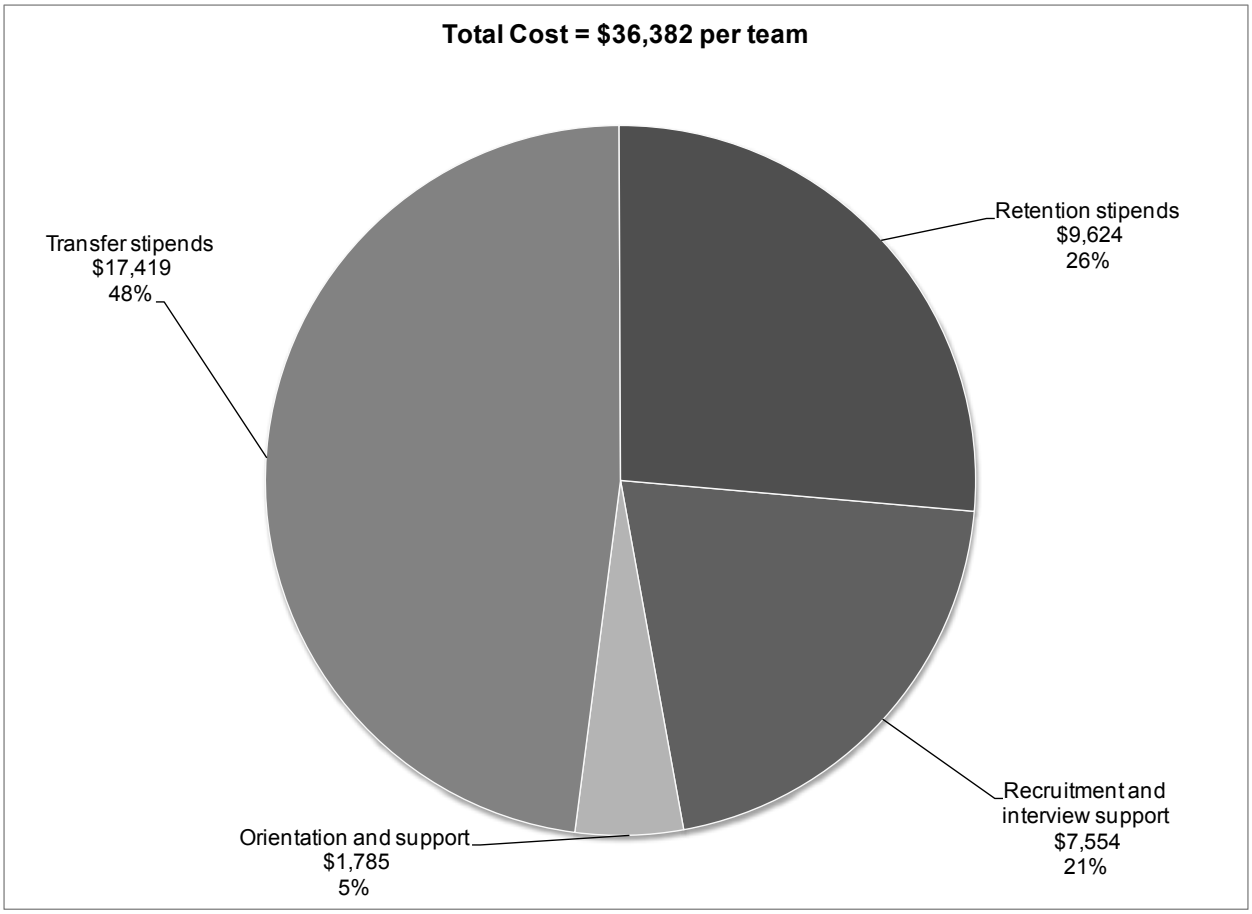
[76] The cost is equal to the change in the average number of teachers per student multiplied by the average teacher's salary, divided by the test-score impact. For math, this is equal to (1/15–1/23) x $55,000/0.27.

[77] The rate of 1.6 is the nominal discount rate recommended for short-term investments by the U.S. Office of Management and Budget, Circular A-94 Appendix C, revised December 2011, available at [http://www.whitehouse.gov/omb/circulars_a094/a94_appx-c]. Accessed June 3, 2012.

B. Costs of TTI

We estimated that the average cost of implementing TTI was $36,382 per targeted teaching team. The breakdown by category of cost is shown in Figure VII.1. The majority of costs took the form of stipend payments made to teachers: 48 percent was spent on transfer stipends, and 26 percent was spent on retention stipends. Another 21 percent was spent on recruiting teachers and schools, including information sessions and personal contact, and the remaining 5 percent was spent on the time of site managers and others who provided half-day orientations and were available to support teachers during the first two years after their transfers.[78]

Figure VII.1. Costs of TTI per Team



Source: TTI expenditure data.

Transfer stipends. The largest cost component was payments to the teachers for transferring, which averaged $17,419 per team in this study. Calculating the cost of transfer incentives is straightforward: it is $20,000 per transfer, minus the value of payments forfeited by teachers who left before fulfilling their two-year commitment. Policymakers can set a cap on the number of slots to fill with transfer incentives.

[78] The number of retention stipends required per team may vary depending on the scale and context of implementation. See the implementation report (Glazerman et al. 2012) for further information.

Retention stipends. In addition to transfer incentives, however, future implementers must consider the cost of retention stipends for eligible highest-performing teachers who are already in low-achieving schools, estimated in this study at $9,624 per team. Unlike transfer teachers, these teachers do not have to apply and interview to be eligible for payments, so their costs cannot be capped ahead of time. The number of such teachers is known only after the value-added analysis is completed and the teachers' current (or expected) school assignments for the coming year are finalized. In the 10 districts participating in TTI, 181 teachers received retention stipends, compared with 81 TTI transfer teachers.[79] One reason for such a high ratio (more than two to one) is that the study required identification of approximately double the number of teacher teams in order to form a control group. Additionally, there were low-achieving schools not participating in the study that had high-performing teachers who were eligible for retention stipends. We assume that the districts, in the absence of a study, would have paid about half of the retention stipends that were actually paid out. That is, the treatment group and half of the nonstudy teams would still have been part of the intervention.

Recruitment and interview support. Recruitment and interview costs were estimated at $7,554 per team. The effort associated with recruiting teachers (transfer candidates) and receiving-school principals—from the initial information sessions to the final matchups and signing of transfer agreements—also varied by district and local circumstances, but typically took the equivalent of about one-third to one-half of the site managers' time for five months. This component includes the cost of developing and maintaining a TTI website with secure access points for applicants. Website costs were averaged over the 10 districts.

Orientation and support. The final component measured was the cost of ongoing support for TTI transfer teachers. Each district had a half-day orientation with costs of $1785 per team, including materials development, facilitator travel, and facilitator time averaged over the 10 sites. Site managers responsible for the recruitment and placement of TTI transfer teachers also provided ongoing support and helped verify employment for stipend-eligible teachers. The implementation team incurred costs for this employment verification, but we did not include them because we believe that in future implementations such verification would not have to be completed by an external party, as the district will already have the information. The costs that we did include were mostly for time spent responding to queries, specifically about continued eligibility for TTI stipends.

Other costs. The cost of identifying the highest-performing teachers (value-added analysis) was intentionally omitted from our calculation. We assumed that such costs are not specific to this intervention and that a district or state might already have invested in such performance measures. For those districts, the marginal cost of identifying the highest-performing teachers based on their value added is likely to be negligible. However, to inform future implementers who might not have invested already in a performance measurement system, we describe the process briefly here.

[79] We initially identified for the study 208 highest-performing teachers teaching in lowest-achieving schools. Seventeen of those who were eligible for retention stipends were on teaching teams assigned to treatment. The other 191 teachers eligible for retention stipends were in control teams, nonstudy teams in study schools, or nonstudy schools.

The amount of effort required to identify candidates (to conduct value-added analysis and verify teacher eligibility) depends on the quality and availability of the data, which varied considerably by district. In some cases, almost three months of intensive effort was needed to clean data, merge dozens of files, reconcile anomalous data, match ID codes, account for multiple courses and multiple tests, perform checks, and repeat the process for updated data when errors or omissions were found. The final step of verifying teacher eligibility took one to five days.

We assumed that the value of time invested by teachers and principals in the interviewing process, not counting the TTI site managers' effort, was the same for treatment and control groups. In other words, the cost would be incurred even in the absence of a transfer incentive, so we did not include it in our calculations.

C. Cost Comparison

The primary impacts of TTI were the increases in student test scores, expressed in standard-deviation units. To place the magnitude of these impacts into context, we estimated what it would cost in each year to generate them through CSR.[80] This calculation allowed us to compare the cost of TTI with the cost of CSR to generate the same test-score impact as TTI and to find out whether TTI was cost-effective. If the CSR cost was higher, we could conclude that TTI is more cost-effective. As we change assumptions about how to estimate the effectiveness of TTI, the impact estimate changes, and so does the cost of producing that same impact through CSR. Thus, we have different comparisons, depending on the assumptions, which are discussed below.

We focused on the combined elementary and middle school results because we wanted to use all the evidence from this study. Although elementary and middle school results were different, impact estimates also varied by district, and district and grade span were confounded. Because it might have been difficult to choose which estimates were relevant to future implementation, we used the full sample average. Given the variability in impacts across districts noted in Chapter V, average findings should be considered a starting point from which individual results may vary.

For simplicity, we assumed that the impacts observed in year 2, which were estimated using the 7 cohort 1 districts only, applied to all 10 districts (cohorts 1 and 2).[81] We also assumed that the impact that TTI continued to have after the second year dropped to zero, although we included one set of calculations that assumes that the results from year 2 are repeated in year 3, with no impact beyond that point. This is another way of saying that any long-term impacts of TTI are treated as an unmeasured benefit of the intervention.

We estimated the dollar value of TTI impacts in five ways, shown as separate columns in Table VII.1. First, we used the benchmark-impact estimates, but set any statistically insignificant impact estimates to zero. Second, we used the benchmark estimates regardless of statistical significance. Third, we used the estimates for the full sample from the vacancy-weighted analyses reported in Chapter V. Fourth, we used the impact estimates for elementary schools

[80] We used the discount-rate assumptions mentioned above when combining costs across years.

[81] Alternatively, substituting the costs for cohorts 1 and 2 in year 1 and cohort 1 only in year 2 would yield similar results.

only. We excluded middle school because the middle school impact estimates are either statistically insignificant or negative, findings that clearly indicate that TTI would not be cost-effective at the middle school level. Finally, we went back to the benchmark estimates and assumed that the results from year 2 were obtained in year 3 as well.

The insight provided in Table VII.1 is that TTI may cost less than an alternative intervention to generate the same impacts. However, under some assumptions, it is not more cost-effective, based on impacts observed within two years. Specifically, if we use the benchmark model and treat nonsignificant impact estimates as equal to zero (column 1), then the intervention is a more expensive way to raise test scores than the CSR alternative used in this example.

If we assume that the true impacts would match our estimates (point estimates) regardless of their statistical significance (column 2), TTI would save more than $7,000 per team relative to CSR. If we thought that the more replicable result was the one that weighted each vacancy equally instead of each student, then the intervention would cost $30,000 less than CSR per team. Similarly, if we assume the intervention would be implemented in elementary schools only, with the expectation of repeating the elementary results, TTI would again appear to be the cheaper alternative by $13,000 per team. And finally, if we add one more year of impact in the calculation and assume the impact estimate for year 3 is the same as the impact estimate for year 2, then, even with the benchmark model, TTI costs $40,000 less per team than the alternative. We do not show in Table VII.1 the results one would obtain by using estimates from the middle school subgroup or the lower-impact districts. In those cases, TTI would not have positive impacts and could never be a preferred alternative.

Table VII.1. Cost of TTI Relative to Estimated CSR Alternative

	Benchmark, Significant Only (1)	Benchmark, Point Estimates (2)	Weighted by Vacancy (3)	Elementary Only (4)	Benchmark Plus One Year (5)
TTI Impact by Year					
Year 1	0.00	0.02	0.03	0.04	0.02
Year 2	0.03	0.05	0.10	0.08	0.05
Year 3					0.05
CSR Cost of Increasing Test Scores ($ per std dev per student)	5,134	5,134	5,134	5,134	5,134
Team Size (students per team)	131	131	104	83	131
CSR Cost of Generating TTI impact ($ per team)					
Year 1	0	10,095	16,019	17,046	10,095
Year 2	16,560	33,120	49,928	31,458	33,120
Year 3					32,599
CSR Cost of Generating Impact Due to Retention of Teachers ($ per team)	0	611	970	1,032	611
Total CSR Cost of Generating TTI Impacts ($ per team)	16,560	43,826	66,917	49,535	76,425
Cost of TTI ($ per team)	36,382	36,382	36,382	36,382	36,382
Cost Savings from TTI ($ per team)	**-19,821**	**7,445**	**30,535**	**13,154**	**40,043**

Note: Impact estimates, costs per unit of test-score increase, and team size are averages of math and reading, giving equal weight to both subjects.

Impacts on teacher retention affected the cost-effectiveness calculation in four ways. First, an impact on the retention of highest-performing teachers presumably improved average teacher performance. Such impacts—found after the first year—are already captured in the findings related to student achievement.

Second, beyond the intervention period, we found no impact on retention rates relative to control teachers. More importantly, we found that the retention rate for treatment teachers did not fall to zero. Therefore, one might expect the impacts on test scores to continue beyond the two-year observation period. For that reason, we considered an alternative cost-effectiveness calculation that made different assumptions about how to extrapolate the test-score impacts beyond year 2 (column 5 in Table VII.1). If one were to extrapolate beyond year 2, the intervention would appear even more cost-effective.

Third, impacts on the retention rate of teachers receiving a retention stipend for remaining in low-achieving schools would also presumably raise student achievement if the teachers stayed longer than they would have without the stipend. Because we include the cost of paying all of these teachers, it was important to account for any impact that they might generate. Inasmuch as they might have remained in their schools anyway without the retention stipend, we needed an estimate of the net impact of the stipend on their probability of staying. We did not have an experimental-impact estimate so we used the non-experimental estimate implied by the findings presented in Appendix G, Tables G.8 and G.9. We found that retention was 7 percentage points higher for retention teachers than for all others after the first year. After the second year, when the payments had ended, there was no difference, so we focused on only the 7 percent of high-performing retention teachers who stayed one year longer. Absent any better information, we assumed that this was the true impact of TTI, and we applied that difference to our best estimate of the value that these teachers added above and beyond the effectiveness of the average teachers who would replace them if they left. The most plausible estimate of this effectiveness increase would be the impact of treatment focal teachers. To be conservative, we used the same value as the first-year team impacts of TTI and incorporated this into the calculation of TTI benefits. These benefits are included in all the calculations shown in Table VII.1.

Finally, a benefit of increasing teacher retention is a reduction in the cost of replacing the teacher. However, because the main effect of TTI was just to delay teachers' attrition by one year, and the retention impacts we observed were defined in terms of school retention (and not retention in the district), we assumed that the benefits would be negligible and did not attempt to quantify them.

D. Unmeasured Effects of TTI

The cost comparisons above were based on replicating the impacts that TTI had on test scores in the targeted teaching teams. However, this exercise may fail to capture some unmeasured impacts of the intervention, therefore understating the cost savings associated with its implementation relative to another approach. As mentioned, the most obvious unmeasured component is the impact that TTI teachers continue to have after they stop receiving stipend payments but remain in the targeted schools. Above, we extrapolated impacts for one year beyond the study's two-year observation period, but at least some portion of the impact could persist longer. Thus, the true benefit of TTI is likely to be undervalued: TTI would appear to generate even more cost savings if its benefits were measured well beyond the two-year observation period of the study.

Another potential effect that we were not able to quantify is the positive spillover to other grades in the same school. We found suggestive evidence of this in our data. We did not have measures of teacher performance for teachers in all grades, just those grades in which there was a TTI assigned team. However, we did document a difference in the experience levels of nonfocal teacher movers: new nonfocal teachers in treatment teams were less experienced, and new nonfocal teachers in control teams were more experienced. This means that weaker teachers might have been moved into the treatment teams to benefit from stronger peers, only to swap positions with stronger teachers who had originally been in the teams where the TTI teacher was placed. Thus, students in the other grades would, presumably, have benefited. Our estimates do not capture the benefits that may have spilled over beyond the treatment teams. If the opposite occurred in control teams—stronger teachers were moved into study teams and their places taken by relatively weaker teachers—this underestimation of TTI benefits would be magnified further.

Two other unmeasured effects have to do with the redistribution of highest-performing teachers. We did not attempt to quantify the social value associated with making the distribution of highest-performing teachers more equitable, but one might wish to consider as a positive effect of the intervention the progress toward closing the teacher-effectiveness gap between higher- and lower-income students as a benefit of the intervention.

On the other hand, we also did not quantify the possibility of negative impacts on sending schools that lost their highest-performing teachers. Such schools may have had to replace a highest-performing teacher with an average or novice teacher who, at least initially, would be lower performing than the transfer teachers who left. The schools would have had to expend resources interviewing and hiring replacement teachers, the less-experienced replacement teachers could require extra mentoring support, and the loss of high-performing teachers could harm morale. These outcomes could burden sending schools in ways that offset the previously mentioned benefits to society. This should be considered when weighing the full effect of selective transfer incentives. A working assumption behind the design of TTI was that the schools with the highest-performing teachers were desirable workplaces, and it would therefore not be difficult to fill their vacancies with good teachers. This assumption, however, cannot be tested with the data available to this study.

REFERENCES

Bloom, Howard. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review,* vol. 8, 1984, pp. 225–246.

Buddin, Richard, and Gema Zamarro. "Teacher Quality, Teacher Licensure Tests, and Student Achievement." Santa Monica, CA: RAND, May 2008.

Carroll, Stephen, Robert Reichardt, and Cassandra Guarino. "The Distribution of Teachers Among California's Districts and Schools." Santa Monica, CA: RAND, October 2000.

Clotfelter, Charles, Helen Ladd, Jacob Vigdor, and Justin Wheeler. "High Poverty Schools and the Distribution of Teachers and Principals." Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, December 2006.

Constantine, Jill M., Daniel W. Player, Timothy W. Silva, Kristin Hallgren, Mary Grider, and John G. Deke. "An Evaluation of Teachers Trained Through Different Routes to Certification." Final report submitted to the U.S. Department of Education, Institute of Education Sciences. Princeton, NJ: Mathematica Policy Research, February 2009.

Education Trust. "Their FAIR Share: How Texas-Sized Gaps in Teacher Quality Shortchange Low-Income and Minority Students." Washington, DC: The Education Trust, February 2008.

Fessendon, Ford. "Schools With the Most Top-Rated Teachers." *New York Times.* February 25, 2012, p. A19.

Finn, Jeremy D., Susan B. Gerber, Charles M. Achilles, and Jayne Boyd-Zaharias. "The Enduring Effects of Small Classes." *Teachers College Record*, vol. 103, no. 2, April 2001, pp. 45–83.

Garet, Michael, Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Jones, Audrey Falk, Howard Bloom, Fred Doolittle, Pei Zhu, and Laura Sztejnberg. "The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, September 2008.

Glazerman, Steven, Ali Protik, Bing-ru Teh, Julie Bruch, and Neil Seftor. "Moving High-Performing Teachers: Implementation of Transfer Incentives in Seven Districts." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, April 2012.

Glazerman, Steven, Eric Isenberg, Sarah Dolfin, Martha Bleeker, Amy Johnson, Mary Grider, and Matthew Jacobus. "Impacts of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, June 2010.

Glazerman, Steven, and Jeffrey Max. "Do Low-Income Students Have Equal Access to the Highest-Performing Teachers?" *NCEE Evaluation Brief*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, April 2011.

Goldhaber, Dan. "Addressing the Teacher Qualifications Gap: Exploring the Use and Efficacy of Incentives to Reward Teachers for Tough Assignments." Washington, DC: Center for American Progress, November 2008.

Gordon, Robert, Thomas Kane, and Douglas Staiger. "Identifying Effective Teachers Using Performance on the Job." Washington, DC: Hamilton Project, Brookings Institution, April 2006.

Hahnel, Carrie, and Orville Jackson. "Learning Denied: The Case for Equitable Access to Effective Teaching in California's Largest School District." Oakland, CA: The Education Trust-West, January 2012.

Huber, Peter J. "The Behavior of Maximum Likelihood Estimation under Nonstandard Conditions." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1*, edited by L.M. LeCam and J. Neyman. Berkeley, CA: University of California Press, 1967.

Isenberg, Eric, and Heinrich Hock. "Design of Value-Added Models for IMPACT and TEAM in DC Public Schools, 2010–2011 School Year." Washington, DC: Mathematica Policy Research, May 2011.

Jackson, C. Kirabo, and Elias Bruegmann. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers." Working Paper No. 15202. Cambridge, MA: National Bureau of Economic Research, July 2009.

Lankford, Hamilton, Susanna Loeb, and James H. Wyckoff. "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis." *Educational Evaluation and Policy Analysis*, vol. 24, no. 1, 2002, pp. 37–62.

Lipscomb, Stephen, Bing-ru Teh, Brian Gill, Hanley Chiang, and Antoniya Owens. "Teacher and Principal Value-Added: Research Findings and Implementation Practices." Cambridge, MA: Mathematica Policy Research, September 2010.

Max, Jeffrey, Allison McKie, and Steven Glazerman. "Feasibility of a Star Teacher Demonstration." Washington, DC: Mathematica Policy Research, February 2007.

McCaffrey, Dan F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton. "Models for Value-Added Modeling of Teacher Effects." *Journal of Educational and Behavioral Statistics*, vol. 29, no. 1, 2004, pp. 67–101.

Morris, Carl N. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of American Statistical Association*, vol. 78, no. 381, 1983, pp. 47–55.

Mosteller, Frederick. "The Tennessee Study of Class Size in the Early Grades." *The Future of Children: Critical Issues for Children and Youths*, vol. 5, no. 2, summer/fall 1995, pp. 113-127.

Peske, Heather, and Kati Haycock. "Teaching Inequality: How Poor and Minority Students Are Shortchanged on Teacher Quality." Washington, DC: The Education Trust, June 2006.

Presley, Jennifer, Bradford White, and Yuqin Gong. "Examining the Distribution and Impact of Teacher Quality in Illinois." Edwardsville, IL: Illinois Education Research Council, 2005.

Puma, Michael J., Robert B. Olsen, Stephen H. Bell, and Cristofer Price. "What to Do When Data Are Missing in Group Randomized Controlled Trials" (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. 2009.

Reardon, Sean. "The Widening Academic-Achievement Gap between the Rich and the Poor: New Evidence and Possible Explanations." Stanford, CA: Stanford University, July 2011.

Rivkin, Steven, Erick Hanushek, and John Kain. "Teachers, Schools, and Academic Achievement." *Econometrica*, vol. 73, no. 2, March 2005, pp. 417–458.

Rockoff, Jonah, Brian Jacob, Thomas Kane, and Douglas Staiger. "Can You Recognize an Effective Teacher When You Recruit One?" Cambridge, MA: National Bureau of Economic Research, November 2008.

Sass, Tim, Jane Hannaway, Zeyu Xu, David Figlio, and Li Feng. "Value Added of Teachers in High-Poverty Schools and Lower-Poverty Schools." *Journal of Urban Economics*, vol. 72, nos. 2-3, September-November 2012, pp. 104-122.

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." Nashville, TN: National Center on Performance Incentives, Vanderbilt University, 2010.

Tennessee Department of Education. "Tennessee's Most Effective Teachers: Are They Assigned to the Schools That Need Them Most?" Nashville, TN: Tennessee Department of Education, 2007.

Turque, Bill. "Top Teachers Have Uneven Reach in District." *The Washington Post.* November 14, 2010. Available at [http://www.washingtonpost.com/wp-dyn/content/article/2010/11/13/AR2010111303782.html]. Accessed June 7, 2012.

U.S. Department of Education, National Center for Education Statistics. *Digest of Education Statistics*. Available at [http://nces.ed.gov/programs/digest/d10/tables/dt10_083.asp]. Accessed June 3, 2012.

U.S. Department of Education. "NAEP Data Explorer." Available at [http://nces.ed.gov/nationsreportcard/naepdata/]. Accessed April 16, 2013.

White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, vol. 48, no. 4, 1980, pp. 817–830.

Xu, Zeyu, Umut Ozek, and Matthew Corritore. "Portability of Teacher Effectiveness Across School Settings." CALDER Working Paper, Washington, DC: American Institutes for Research, June 2012.

APPENDIX A

SUPPLEMENTAL MATERIALS FOR CHAPTERS I AND II

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

A. Random Assignment Procedures

In this section, we provide more details about the study's random assignment process described in Chapter 1. We begin by describing each step of the process.

Step 1. Identify batches of teacher teams with vacancies. This step involved gathering information from potential receiving schools about teacher teams with vacancies. We conducted random assignment in batches so principals could begin filling vacancies as soon as they opened. As TTI managers provided information about teacher teams with vacancies, we waited approximately two weeks before conducting random assignment for a batch of vacancies. The number of batches per district ranged from one to eight; the average was three to four batches per district.

Step 2. Group teacher teams in the same grade and subject into blocks. After identifying a batch of vacancies, we matched teacher teams in different schools but in the same grade and subject into blocks. The number of blocks per district ranged from 3 to 24; the average district had 9 blocks. Although we attempted to match teacher teams on the basis of such school characteristics as the student achievement ranking and the percentage of students eligible for FRL, this was feasible in only about 17 percent of batches where we had four or more schools with a vacancy in the same grade and subject (Table A.1). For 83 percent of batches, we had three or fewer schools available for matching.

Table A.1. Maximum Number of Schools per Batch with Vacancies in the Same Grade and Subject

Number of Schools with Matching Vacancies	Number of Batches	Percentage of Batches
1	10	28
2	17	47
3	3	8
4 or more	6	17
Total	36	100

Step 3. Randomly assign the teacher teams within a block. Once teacher teams were matched into blocks, we randomly assigned the teams in each block to treatment or control. Although we assigned pairs of teams in 65 percent of blocks, the remaining blocks had an odd number of teacher teams (Table A.2). We randomly assigned a single teacher team in 25 percent of blocks and we randomly assigned three teacher teams in 10 percent of the blocks. We conducted random assignment between April and August: 39 percent of the blocks were assigned in April or May, 49 percent in June, and 11 percent in July and August.

Table A.2. Number of Teacher Teams per Block

Teams per Block	Number of Blocks	Percentage of Blocks
1	22	25
2	58	65
3	9	10
Total	89	100

As described in Chapter 1, when two schools in the same randomization batch had eligible teams at more than one grade level, we assigned teams in the same school. In these schools, shown in Figure A.1, we assigned teams so that each school in the pair had one treatment and one control team. In this example, two participating receiving schools each have a teaching vacancy in grades 3 and 5 (top panel). In such a configuration, we assigned the grade-5 teams to be the mirror image of the grade-3 random assignment status, so each school had both a treatment team and a control team. In the example in the bottom panel of the figure, we show the result where grade 3 in School A and grade 5 in School B were assigned to the treatment group, and grade 3 in School B and grade 5 in School A were assigned to the control group.

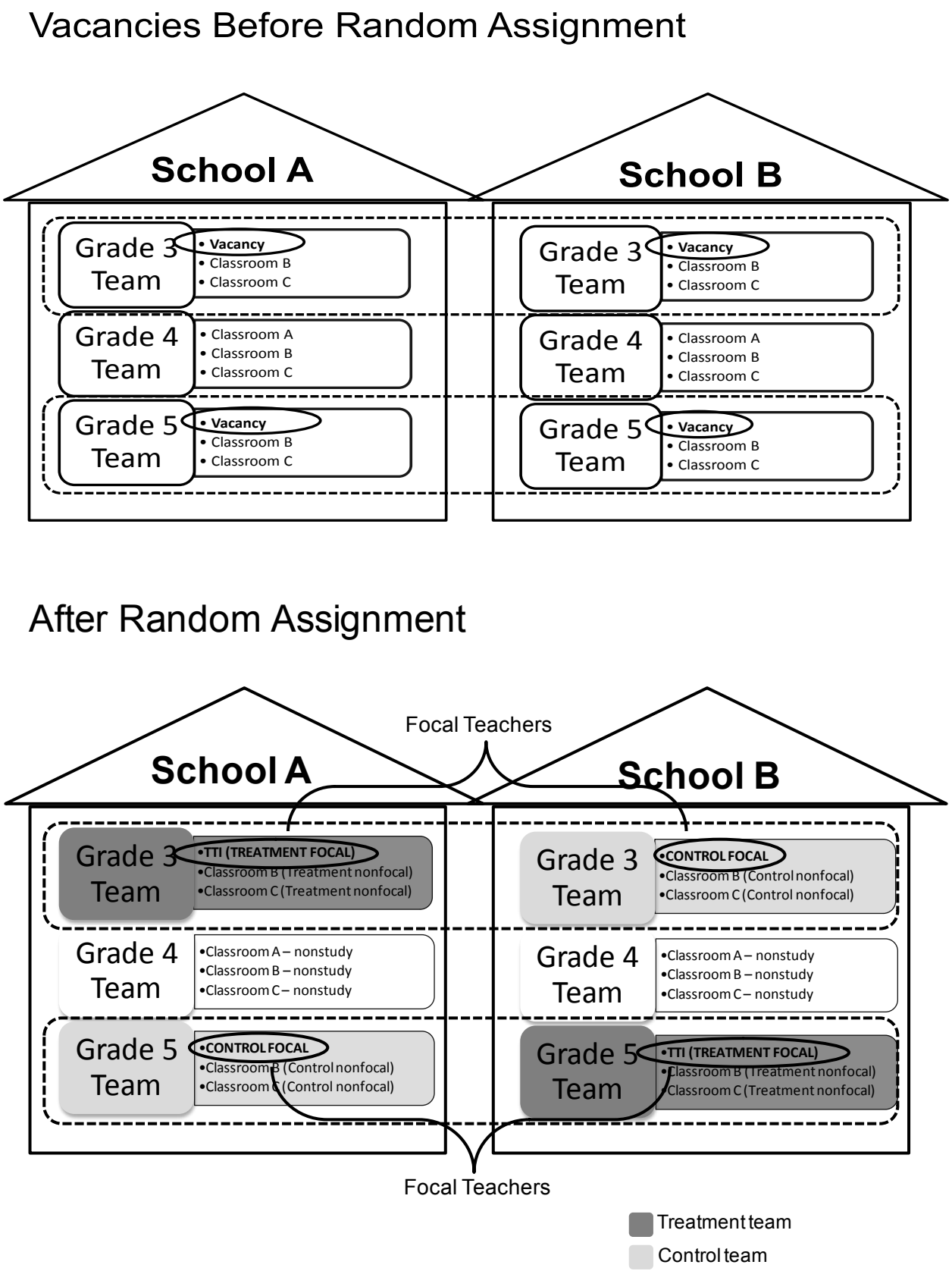
We used the following rules when randomly assigning teacher teams in the same school to avoid contamination.

- In elementary schools, treatment and control teams in the same school had to be separated by at least two grade levels. This ensured that no elementary student had a teacher from both a treatment and a control team during the study's two-year period.
- In middle schools, teachers were sometimes responsible for classes in more than one grade level, so we required that treatment and control teams in the same school be in different subjects (math or ELA) and also be separated by at least one grade. This ensured that no student was taught the same subject by a teacher from both a treatment and control team.

There was, however, the possibility that a student had a teacher from a treatment team for one subject and a teacher from a control team in the other subject, because of cross-grade teaching that we discovered during the study. This same-grade, opposite-subject overlap was possible in only 5 out of 114 schools in the study and is likely to have occurred, based on known teacher assignments, in only 3 out of those 5 schools. In terms of students, fewer than 2 percent of cases were affected in the analysis of middle schools. No case was affected in the analysis of elementary schools.

Another way to avoid contamination was to force teams into the same treatment status by assigning them together if they did not meet the previously described adjacency rule. For example, if there were vacancies in grades 3 and 4 in an elementary school, both vacancies were assigned to the same treatment status. If there were vacancies in 6th-grade math and 8th-grade math in a participating receiving middle school, both were assigned to the same status. Some teams had more than one vacancy in a single study team. In such cases (14 out of 165 teams), all vacancies within the team were assigned to a common study status because teams were the unit of random assignment.

Figure A.1. Random Assignment Study Design



B. Survey Response

We analyzed survey response rates and examined the degree to which each sample of respondents resembles the full population of respondents and nonrespondents in terms of characteristics we can measure for all sample members. In the following tables, we present the results of this analysis for the surveys of transfer candidates, teachers in treatment and control teams, and their principals. In Table A.3, we summarize response rates by instrument and treatment status, and in Table A.4, we examine the distribution of candidate survey respondents as well as the full sample of respondents and nonrespondents across districts, grade-subject pools, the top 10 percent based on a value-added ranking in their grade-subject pool, and their application status.

We examined survey response rates by cohort, district, grade, and school characteristics by treatment and control status (Table A.5 for the teacher survey and Table A.6 for the principal survey). We also compared the respondents to the full sample of respondents and nonrespondents by cohort, district, grade, and school characteristics (Table A.7) to gauge the extent to which our analysis sample reflects the intended sample of teachers in study grades.

Table A.3. Summary of Response Rates, by Instrument and Treatment Status

		Response Rate[a]		
Survey Instrument	Number Eligible	All	Treatment	Control
Candidate Survey				
Cohort 1 districts	1,012	82.5	n.a.	n.a.
Cohort 2 districts	502	77.7	n.a.	n.a.
All	1,514	80.9	n.a.	n.a.
Teacher Background Survey				
Cohort 1 districts	469	76.5	76.0	77.0
Cohort 2 districts	220	76.9	76.0	78.2
All	689	76.6	76.0	77.4
Principal Survey, Program Year 1				
Cohort 1 districts	124	90.3	90.6	90.0
Cohort 2 districts	41	90.2	95.2	85.0
All	165	90.3	91.8	88.8
Principal Survey, Program Year 2				
Cohort 1 districts only	123[b]	82.1	81.0	83.3

[a]Response rates account for the fact that some respondents, when contacted and surveyed, turned out to be ineligible. The rates are calculated as the number of eligible completes divided by the estimated number of eligible cases attempted. We assume that the fraction of noncompletes that would have been determined ineligible is equal to the fraction of completes that were determined eligible.

[b]One school in a cohort 1 district closed and was not included in the program year 2 survey.

n.a. = not applicable

Table A.4. Respondents Versus Full Sample of Respondents and Nonrespondents (percentages)

	Candidate Survey	
Subgroup	Respondents	Full Sample
District		
A	12.2	12.0
B	6.0	7.8
C	21.5	22.0
D	5.1	4.8
E	6.1	5.9
F	9.5	8.9
G	7.8	7.8
H	6.3	5.9
I	13.8	13.9
J	11.8	13.4
Total	100.0	100.0
Pool		
Elementary	51.4	49.5
Middle school English/language arts (ELA)	24.1	25.2
Middle school math	24.5	25.2
Total	100.0	100.0
In Top 10% of Value-Added Distribution		
Elementary pool	46.9	47.3
Middle school ELA pool	52.1	52.9
Middle school math pool	46.8	46.4
Application Status*		
Did not apply	75.2	78.5
Applied but did not transfer	18.2	16.2
Transferred	6.6	5.4
Total	100.0	100.0
Sample Size[a]	**1,225**	**1,514**

[a]Sample size for comparing whether candidates are in the top 10 percent of value-added ranking in their pool between respondent and the full sample is different because we have value-added ranking for candidates only in the seven districts where we conducted value-added analysis ourselves.

*Difference in distributions by application status is statistically significant at the 0.05 level using chi-square test of independence.

Table A.5. Survey Response Rates by Subgroup, Teacher Survey (percentages)

	Teacher Background Survey		
Subgroup	Treatment	Control	Difference
All	**76**	**77**	**1**
Cohort			
1	76	77	-1
2	76	78	-2
District			
A	85	73	12
B	91	89	2
C	69	79	-10
D	64	62	2
E	67	95	-28*
F	91	64	27*
G	71	78	-7
H	88	79	9
I	73	74	-1
J	79	90	-11
Grade			
3	76	81	-5
4	82	80	2
5	78	78	1
6	87	86	0
7	63	67	-3
8	67	69	-3
School Poverty			
Lower poverty (≤ 80% free or reduced-price lunch [FRL])	74	71	3
Higher poverty (> 80% FRL)	77	85	-8
School Race/Ethnicity			
Majority African American	76	82	-6
Majority Hispanic	74	76	-2
Majority white	45	51	-6
No majority	91	69	23*
School Size			
Smaller (≤ 700 students)	80	75	5
Larger (> 700 students)	72	80	-9
Sample Size	**374**	**315**	

*Difference is statistically significant at the 0.05 level using a two-sided test.

Table A.6. Survey Completion Rates by Subgroup, Principal Survey (percentages)

	Year 1 Principal Survey (one response per team)			Year 2 Principal Survey (one response per team)		
Subgroup	Treatment	Control	Difference	Treatment	Control	Difference
All	**92**	**89**	**3**	**81**	**83**	**-2**
Cohort						
1	91	90	1	81	83	-2
2	95	85	10	n.a.	n.a.	n.a.
District						
A	100	100	0	89	89	0
B	83	100	-17	83	100	-17
C	95	82	13	65	68	-3
D	100	67	33	75	100	-25
E	86	100	-14	71	80	-9
F	83	100	-17	100	100	0
G	83	80	3	100	80	20
H	100	100	0	n.a.	n.a.	n.a.
I	90	67	23	n.a.	n.a.	n.a.
J	100	100	0	n.a.	n.a.	n.a.
Grade						
3	82	100	-18	79	79	0
4	100	100	0	94	92	2
5	100	89	11	87	78	9
6	92	82	10	67	75	-8
7	91	80	11	80	80	0
8	78	70	8	50	100	-50*
School Poverty						
Lower poverty (≤ 80% FRL)	93	86	7	76	80	-4
Higher poverty (> 80% FRL)	90	92	-1	90	89	1
School Race/Ethnicity						
Majority African American	90	89	-3	83	78	5
Majority Hispanic	95	93	2	83	90	-7
Majority white	100	50	50	50	100	-50
No majority	85	82	3	75	83	-8
School Size						
Smaller (≤ 700 students)	92	91	1	86	85	0
Larger (> 700 students)	91	83	8	71	75	-4
Sample Size	**85**	**80**		**63**	**60**	

*Difference is statistically significant at the 0.05 level using a two-sided test.

Table A.7. Respondents Versus Full Sample of Respondents and Nonrespondents, Teacher Survey (percentages)

	Teacher Background Survey	
Subgroup	Respondents	Full Sample
Cohort		
1	67.9	68.1
2	32.1	31.9
Total	100.0	100.0
District		
A	11.1	10.7
B	6.5	5.5
C	24.4	25.4
D	3.8	4.6
E	6.7	6.4
F	9.5	9.3
G	5.9	6.1
H	4.8	4.4
I	19.7	20.6
J	7.6	7.0
Total	100.0	100.0
Grade		
3	22.3	21.9*
4	19.1	18.0
5	18.3	18.0
6	16.8	14.8
7	13.7	16.3
8	9.7	11.0
Total	100.0	100.0
School Poverty		
Lower poverty (≤ 80% FRL)	47.0	49.5*
Higher poverty (> 80% FRL)	53.1	50.5
Total	100.0	100.0
School Race/Ethnicity		
Majority African American	39.9	38.8*
Majority Hispanic	47.3	48.3
Majority white	1.7	2.8
No majority	11.1	10.2
Total	100.0	100.0
School Size		
Smaller (≤ 700 students)	53.2	52.5
Larger (> 700 students)	46.8	47.5
Total	100.0	100.0
Sample Size	**524**	**689**

*Difference in Teacher Survey respondents versus full sample distributions for grade, school poverty, and school race/ethnicity are statistically significant at the 0.05 level using chi-square test of independence.

Table A.8. Respondents Versus Full Sample of Respondents and Nonrespondents, Principal Survey (percentages)

	Year 1 Principal Survey (one response per team)		Year 2 Principal Survey (one response per team)	
Subgroup	Respondents	Full Sample	Respondents	Full Sample
Cohort				
1	75.2	75.2	100.0	100.0
2	24.8	24.9	n.a.	n.a.
Total	100.0	100.0	100.0	100.0
District				
A	12.1	10.9	15.8	14.6*
B	6.7	6.7	9.9	8.9
C	24.8	25.5	27.7	34.2
D	4.0	4.2	5.9	5.7
E	7.4	7.3	8.9	9.8
F	14.1	13.9	21.8	17.9
G	6.0	6.7	9.9	8.9
H	8.7	7.9	n.a.	n.a.
I	10.1	11.5	n.a.	n.a.
J	6.0	5.5	n.a.	n.a.
Total	100.0	100.0	100.0	100.0
Grade				
3	20.8	20.6	21.8	22.8
4	20.8	18.8	26.7	23.6
5	22.8	21.8	26.7	26.8
6	14.1	14.6	6.9	8.1
7	12.1	12.7	7.9	8.1
8	9.4	11.5	9.9	10.6
Total	100.0	100.0	100.0	100.0
School Poverty				
Lower poverty (≤ 80% FRL)	51.7	52.1	64.4	67.5
Higher poverty (> 80% FRL)	48.3	47.9	35.6	32.5
Total	100.0	100.0	100.0	100.0
School Race/Ethnicity				
Majority African American	40.9	41.2	44.6	45.5
Majority Hispanic	43.6	41.8	41.6	39.8
Majority white	2.0	2.4	3.0	3.3
No majority	13.4	14.6	10.9	11.4
Total	100.0	100.0	100.0	100.0
1. ..... chool Size				
Smaller (≤ 700 students)	67.8	66.7	76.2	73.2
Larger (> 700 students)	32.2	33.3	23.8	26.8
Total	100.0	100.0	100.0	100.0
Sample Size	**149**	**165**	**101**	**123**

*Difference in Principal Survey respondents versus full sample distributions for district in year 2 is statistically significant at the 0.05 level using chi-square test of independence.

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

APPENDIX B

VALUE-ADDED ANALYSIS TO IDENTIFY HIGHEST-PERFORMING TEACHERS

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

The first step for the Talent Transfer Initiative (TTI) was to identify the highest-performing teachers in each study district.[82] To do this, the districts estimated teachers' value added to student achievement based on two or more years of test-score data from state assessments. Value added represents the amount of learning growth that can be attributed to the teacher, holding constant the factors outside the teacher's control. It can be estimated by measuring growth in student achievement over time and comparing the actual scores of each teacher's students to the predicted scores, given the prior achievement and possibly other characteristics of that teacher's students. It requires longitudinal data and a reliable student-teacher link. Using more than one year of data is meant to increase the statistical precision and stability of the estimates, identifying teachers with high persistent performance—in other words, a strong track record.

The value-added estimates were prepared by either the participating districts, working with an outside vendor, or the study team, who performed this analysis for the district as a condition of their participation in the study. We used whatever value-added measure the district was using because that is what would have been used in the absence of the study; in cases where it was not available, we calculated it ourselves. One of the study districts had its vendor conduct the data analysis and then supplied the TTI team with a list of teachers the district identified as being highest performing based on these pre-existing measures of teacher effectiveness. Two other districts gave the TTI team information from their vendor on teachers' value-added estimates, which the study team then combined across years and used to identify the top performers.[83] For the remaining seven districts, Mathematica used raw data on student achievement, demographics, and enrollment to link students to teachers, and then computed teachers' value added, which took place between January and March 2009 for four districts in cohort 1, and between January and May 2010 for three districts in cohort 2. The approach used by Mathematica for its seven districts is described below, but the other three districts followed a similar approach in their estimation of teacher effectiveness.

Mathematica did not attempt to duplicate the methods used by the other districts. Instead, the goal was to estimate a model that could plausibly have been adopted by the district in regular implementation of an intervention such as TTI. As a result, the study's impacts should be interpreted as the result of whatever process the district might have used to identify top teachers using value-added methods. We examined the main impacts separately for the districts that used Mathematica's estimates and those that used SAS Institute estimates, and we did not find differences in the results. What follows is a discussion of procedures that Mathematica used in estimating teacher value added.

[82] Value-added estimation described here is purely a program-implementation function. It was not used to estimate the impacts of TTI.

[83] The same vendor, the SAS Institute, conducted the value-added analysis for each of the three districts. The methods used by the SAS Institute are described at http://www.sas.com/govedu/edu/k12/evaas/index.html.

A. Estimation Equation

We estimated a value-added model separately for three pools of teachers: elementary school teachers, middle school math teachers, and middle school English/language arts (ELA) teachers. Elementary school included grades 3 to 5 or 4 to 5, and middle school included grades 6 to 8. We used up to three waves of student achievement growth data to identify highest-performing teachers.

All of a teacher's student observations for a particular year were dropped from the estimation sample if the teacher was linked to fewer than five students' test scores in that year. Students who spent less than 20 percent of the school year with a teacher were also excluded from the estimation sample for that teacher.

The estimation equation is:

$$(1)\quad Y_{ijt} = \lambda_{t-1} * Y_{ij,t-1} + \alpha_1 * X_{ijt} + \alpha_2 * Z_{jt} + \beta_j * D_{ijt} + e_{ijt}$$

where Y_{ijt} is the post-test score for student i who is taught by teacher j in year t; $Y_{ij,t-1}$ is the pre-test score for that same student, which is assumed to capture previous inputs into student achievement; and e_{ijt} is the error term. X_{ijt} is a vector of control variables that includes the following student-level variables: indicators for gender, race/ethnicity, free or reduced-price lunch (FRL) status,[84] English language learner (ELL) status, special education (SPED) status, disability type, grade repetition status, and overage-for-grade status.[85] Z_{jt} includes the following teacher-level variables: the percentage of a teacher's students who were mobile, the percentage of a teacher's students who were grade repeaters, and class size. Grade-by-year dummies are also included in Z_{jt} to eliminate any mean differences between grade levels and years. Dosage (D_{ijt}) is a variable that equals the percentage of the year student i in year t was taught by teacher j, or zero if student i was not taught by teacher j in year t. D_{ijt} is expressed as a vector of such dosage variables that includes separate values for each teacher-year. The coefficients λ_{t-1}, α_1, α_2, and β_j are parameters to be estimated. The performance measures ("teacher effects") are contained in the vector β_j, which is the set of coefficients of the dosage variables.

After initial estimation of the teacher effects, we standardized subject-specific performance measures (one for math and one for English/language arts [ELA], if applicable) within each grade level.[86] We then excluded from the rankings any teachers who had fewer than two years of subject-specific performance measures. Although some elementary schools are departmentalized, the majority of elementary school teachers taught in self-contained classrooms. For these teachers, performance measures were calculated by taking the average of their math and ELA performance measures. The top 20 to 25 percent of teachers in each of the three pools—elementary school teachers, middle school math teachers, and middle school ELA teachers—were identified as being the highest-performing teachers in their respective districts.

[84] One district did not provide data on FRL.

[85] Missing values in $Y_{ij,t-1,}$ and X_{ijt} were imputed with predicted values from a regression model.

[86] This assumes that the distribution of teacher effectiveness is the same in each grade within a district, but has the benefit of removing any artificial differences associated, for example, with the properties of the assessment instrument and the ways such properties vary by grade.

B. Controlling for Measurement Error in the Pre-Test

Before estimating equation (1), we corrected for measurement error in the pre-test by fitting an errors-in-variables regression model.[87] We obtained the reliability for each test, when available, from either the test publisher or the school district. We then employed a two-stage procedure. In the first stage, we estimated the following errors-in-variables regression model by using the average published reliability of the test across grades and years to remove the bias caused by the measurement error in the pre-test:[88]

$$(2)\quad Y_{ijt} = \lambda_{t-1} * Y_{ij,t-1} + \alpha_1 * X_{ijt} + \beta_j * D_{ijt} + e_{ijt}$$

The control variables for student background characteristics in equation (2) are identical to those used in equation (1). Using $\hat{\lambda}_{t-1}$, the estimated value for the coefficient of the pre-test from equation (2), we calculated the estimated adjusted gain for each student in each year:

$$(3)\quad \hat{G}_{ijt} = Y_{ijt} - \hat{\lambda}_{t-1} * Y_{ij,t-1}$$

The second-stage regression model pools the data from all years and uses the adjusted gain as the dependent variable:

$$(4)\quad \hat{G}_{ijt} = \alpha_1 * X_{ijt} + \alpha_2 * Z_{jt} + \beta_j * D_{ijt} + e_{ijt}$$

In equation (4), we accounted for the correlation in outcomes for students in different years by using robust standard errors (Huber 1967; White 1980). This errors-in-variables measurement-error correction method underestimates the standard errors of β_j because it treats $\hat{\lambda}_{t-1}$ as identical to its true value, λ_{t-1}; if $\hat{\lambda}_{t-1}$ is estimated precisely, it will be negligible. By substituting equation (3) into (4), rearranging terms, and treating $\hat{\lambda}_{t-1}$ as λ_{t-1}, we arrived at equation (1).

C. Shrinkage Estimator

After estimating equation (1) to obtain performance measures from the β_j coefficients, we applied a shrinkage procedure outlined in Morris (1983) to calculate empirical Bayes performance measures and standard errors. Using this procedure, the empirical Bayes estimate of each performance measure is approximately the precision-weighted average of the original performance measure (an individual element of the β_j vector) and the mean of all the point estimates (all the elements of β_j):

[87] We implemented this model by using the eivreg command in Stata.

[88] The errors-in-variables correction works by subtracting the reliability from the diagonal terms of the regression crossproduct matrix. The resulting parameters are consistent for the normal distribution. See Isenberg and Hock (2011) for a recent application.

$$(5)\ \beta_j^{EB} \approx \left(\frac{\frac{1}{\sigma_j^2}}{\frac{1}{\sigma_j^2}+\frac{1}{\sigma_\beta^2}} \right) \beta_j + \left(\frac{\frac{1}{\sigma_\beta^2}}{\frac{1}{\sigma_j^2}+\frac{1}{\sigma_\beta^2}} \right) \mu_\beta ,$$

where β_j^{EB} is the empirical Bayes estimate of an element of the β_j vector, β_j is the original point estimate, σ_j is the standard error of the original point estimate, μ_β is the mean of all the point estimates, and σ_β is the standard deviation of all the point estimates.

Due to the precision weighting of the original estimate and the mean of all the point estimates, the empirical Bayes performance measure is designed to place relatively more weight on the mean when the original estimate has a high standard error. This is especially important for an intervention like TTI, because the focus is on the upper tail of the teacher-performance distribution. Random estimation error will vary across teachers when we try to estimate their value added, because they have different numbers of students, their students can be more or less homogeneous, and their students' characteristics can be more or less similar to the population average. Each of these factors influences the precision of the individual teacher's value-added estimate. Most important, if that precision does vary, the most imprecisely estimated teacher effects will be overrepresented in both tails of the distribution (because the variance in the effect estimates will contain true variation in teacher quality plus a larger error variance). As a result, an intervention like TTI would identify an artificially high number of teachers with small classes or outlier students unless the estimates were corrected. The empirical Bayes shrinkage adjusts the estimates to account for this phenomenon.

D. Diagnostics

We also conducted a series of robustness checks to ensure the stability of the rankings generated for the model described above: excluding all control variables except for year and grade dummies, estimating the model without controlling for measurement error in the pre-test, including higher-order terms of the pre-test variables, and estimating the model separately by each of the three school years.

E. Rules for Identifying Highest-Performing Teachers

Teachers were eligible to be considered as highest performing if they had two or more years of value-added data.[89] For the seven districts for which we estimated value-added scores, we used three years of student growth, demographic characteristics, and enrollment data from school years 2005–06 to 2007–08 for cohort 1 districts, and from 2006–07 to 2008–09 for cohort 2 districts. Individual teachers who had taught two of the three years could still qualify. In the two districts that provided teacher value-added scores directly from the external partner, one provided these scores for school years 2005–06 to 2007–08, and the other provided scores for two of these three years.

[89] The percentage of teachers eligible to be considered highest performing ranged from 35 to 65 percent for elementary school teachers across districts and from 22 to 70 percent for middle school teachers. In other words, there was one pool of teachers in which 78 percent of teachers we identified who had ever taught a student in that pool did not teach a sufficient number of students in the same pool consistently for three years in a row.

Eligible teachers were identified as highest performing if their value-added scores placed them in the top 20 percent in their district and pool (pools were defined as multiple subjects in elementary school, middle school ELA, and middle school math). The choice of 20 percent as the arbitrary cutoff usually generated a large enough pool to fill the target number of vacancies. (As shown in Chapter III, 88 percent of the vacancies were filled). We adjusted the cutoff for some pools in some of the districts either to be more selective or to slightly enlarge the pool of candidates. Specifically, for elementary teachers it was lowered to 15 percent in one district and 18 percent in another, and raised to 25 percent in one district; for middle school math teachers it was raised to 23 percent in one district and to 25 percent in three districts; and for middle school ELA teachers it was lowered to 15 percent in one district and raised to 25 percent in two districts.

F. Characteristics of Highest-Performing Teachers

Across the 10 districts, 1,514 candidates were identified as eligible for the 81 positions that were ultimately filled, a ratio of almost 19 candidates per position.[90] In Table B.1, we compare value-added scores of the highest-performing teachers to the other eligible teachers in the 7 districts in which we estimated value-added scores. Our value-added model estimated teachers' contribution to student achievement growth—value added—in terms of standardized student test scores, scaled so that a score of 1.0 represents one standard deviation above the mean for the distribution of test-takers (students) districtwide in each respective district.[91] By construction, this scaling results in a value-added score of zero for the average teacher in the analysis sample for a given pool within a district. Also, the value added by any given teacher is the amount of extra progress (if positive) that the teacher's students made with him or her relative to the average teacher in terms of district-level student standard-deviation units.

By definition, average value-added scores for the highest-performing teachers are higher than those of the other eligible teachers. The mean value-added score for the highest-performing teachers of all grades together (grades 3 to 8) for ELA was 0.13 standard deviations above the value-added score of the average teacher; for math it was 0.23 standard deviations above. The mean value-added score for all other eligible teachers was significantly lower: 0.15 standard deviations below the score of the average highest-performing teacher for ELA, and 0.27 standard deviations below that of the average highest-performing teacher for math.

[90] We initially identified 2,332 teachers as highest performing, but some of them were no longer teaching or were not planning to teach in the year during which the program sought to have them transfer. Counting the teachers who turned out to be ineligible for TTI, the ratio of candidates to filled vacancies was almost 29. These numbers are approximate because we had sufficient data to count initially identified teachers for only 9 of the 10 districts. We used the ratio for the 9 districts (2,221 total to 1,442 eligible) and multiplied it by 1,514, the number of eligible teachers in all 10 districts to extrapolate the estimated total for all 10 districts.

[91] In Chapter V, we examine impacts of TTI on student test scores in later years (after possible exposure to TTI). Those scores are scaled relative to the *state* distribution in each state. We use the district as the reference group for value-added measures because state test norms were not available in every district for every year that contributed to the value-added analysis.

Translating standard deviations to percentiles, the average highest-performing teacher would move his or her students up by an average of 5.9 percentile points for ELA and 10.6 percentile points for math in a school year, compared with the average non-highest-performing teacher in the district.[92] In Table B.1, we present the mean value-added scores for these groups separately as well as by school type (elementary or middle). As mentioned, all of these results are based on 7 of the 10 districts that provided detailed data. The magnitude of these differences in average value-added scores may differ for the other 3 districts.

One question about value-added measures is whether those with high scores have more advantaged students even after controlling for the influences of student background on test-score growth. Value-added measures computed by the SAS Institute, for example, take into account prior achievement, and those computed by Mathematica account for prior achievement as well as demographic characteristics of students. We wanted to tabulate the characteristics of students of highest-performing and all other teachers to assess how much their students differed.

Table B.1. Value-Added Scores: Highest-Performing Versus Other Eligible Teachers

	Highest-Performing Teachers[a]		Other Teachers[a]		Difference
	Mean	Sample Size	Mean	Sample Size	Mean
All Grades					
ELA	0.13	1,070	-0.02	4,267	0.15*
Math	0.23	1,153	-0.04	3,952	0.27*
Elementary (grades 3–5)					
ELA	0.14	571	-0.03	2,095	0.17*
Math	0.24	571	-0.04	2,102	0.28*
Middle School (grades 6–8)					
ELA	0.12	499	-0.02	2,172	0.14*
Math	0.22	582	-0.04	1,850	0.26*

Source: Estimation by study team from administrative data.

Notes: Data pertain to a subgroup consisting of the seven districts whose value-added estimates were calculated by the study team.

[a]Value-added scores are in student-level standard-deviation units standardized at the district level.

*Difference is statistically significant at the 0.05 level using a two-sided test.

In Table B.2, we describe the students of the highest-performing teachers compared with those of other eligible teachers in the seven districts for which we estimated value-added scores.[93] The data describe students during the 2005–06 through 2007–08 school years for cohort 1 districts and 2006–07 through 2008–09 school years for cohort 2 districts, the period to which the value-added analysis pertains. Highest-performing teachers had, on average, significantly higher proportions of white students and significantly lower proportions of African American and Hispanic students compared with other eligible teachers. Average prior achievement of students of the highest-performing teachers was significantly higher in both math and reading, by 0.31 and 0.22 standard deviations, respectively. Also, they had a significantly lower proportion of economically disadvantaged students measured by FRL status, and significantly lower proportions of students who had SPED or ELL status.

[92] We translated the student-level standard deviations into percentiles assuming that student test scores are normally distributed.

[93] Administrative data on student background characteristics were not available for the other three districts.

Table B.2. Student Characteristics: Highest-Performing Versus Other Eligible Teachers (percentages)

Student Characteristic	All	Highest-Performing Teachers	Other Teachers	Difference
Demographic				
Male	52.4	51.2	52.7	-1.5*
White	23.1	28.2	21.7	6.5*
African American	24.8	20.5	26.0	-5.5*
Hispanic	44.9	43.0	45.5	-2.5*
Economic				
FRL status	67.2	61.1	68.8	-7.7*
Academic				
Math pre-test score[a]	-0.06	0.18	-0.13	0.31*
ELA pre-test score[a]	-0.06	0.12	-0.10	0.22*
SPED status	18.4	14.0	19.6	-5.5*
ELL status	16.5	13.5	17.3	-3.8*
Sample Size (teachers)	7,776	1,652	6,124	

Source: Administrative data.

Notes: Data pertain to the subgroup of seven districts whose value-added estimates were calculated by the study team. Sample size for FRL status, pre-test scores and limited English proficiency status is lower because of missing data at the student level.

[a]Test scores are in student-level standard-deviation units standardized at the district level.

*Difference is statistically significant at the 0.05 level using a two-sided test.

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

APPENDIX C

SUPPLEMENTAL MATERIALS FOR CHAPTER III

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

In this appendix, we provide supplemental tables for the analysis of the transfer process described in Chapter III.

A. Self-Reported Experiences Related to the Process of Transferring to a School

In Tables C.1 and C.2, we present data on the transfer and hiring process from the perspectives of the receiving-school principals and transfer applicants, respectively. Tables C.3 and C.4 contain additional information on the interview process for candidates who interviewed for TTI vacancies; for those who did not, the reasons for not applying are included.

Table C.1. Hiring Rates in the Treatment and Control Teacher Teams

	Treatment	Control	Difference	*p*-Value
Applicants considered per vacancy	4.16	4.59	-0.43	0.658
Applicants considered per vacancy (including teams where no applicant was considered)	3.20	3.30	-0.10	0.900
Applicants interviewed per vacancy	3.10	3.22	-0.12	0.761
Applicants interviewed per vacancy (including teams where no applicant was interviewed)	2.39	2.31	0.08	0.834
Offers made per applicant interviewed	0.42	0.33	0.08	0.126
Offers accepted per offer made	0.93	0.94	-0.01	0.831

Source: Principal Survey.

Notes: Analysis conducted at the teacher-team level. Sample sizes are 59 treatment teams and 41 control teams.

Table C.2. Candidate Interview Process and Perceptions, by Transfer Status (percentages)

Perception of Interview	Interviewed But Did Not Transfer	Transferred
Interview was informative	74.8	90.1
Had opportunities to communicate strengths during the interview	87.4	95.4
Principal/interviewer was genuinely interested	71.7	86.7
Principal/interviewer responded to candidate's questions about the position	88.9	92.6
Principal seemed like someone the candidate could work with	73.8	85.8
Interview increased candidate's desire to teach at the school	50.0	73.7
Sample Size		
Number of interviews	127	135
Number of candidates	93	81

Source: Candidate Survey.

Table C.3. Structure of Candidate Interview, by Transfer Status (percentages)

Interview Structure	Interviewed But Did Not Transfer	Transferred
Had a one-on-one interview with the principal or assistant principal	63.8	61.7
Interviewed with other school staff	58.3	61.2
Asked to give a teaching demonstration	11.0	11.2
Given a tour of the school	22.1	41.5
Met students at the school	3.9	10.5
Sample Size		
Number of interviews	127	135
Number of candidates	93	81

Source: Candidate Survey.

Table C.4. Top Self-Reported Reasons for Not Applying to TTI (percentages)

Factor	All Reasons (marked all that applied)	Most Important Reason
Happy at old school	87.1	33.1
Commuting issues	54.6	6.9
Not interested in starting at a new school	53.7	6.3
Concern about not being able to return to current school	53.3	6.7
Not confident about self-effectiveness in a TTI school	47.9	2.5
Concerns about TTI school/neighborhood safety	39.6	3.2
Concerns about being unwelcome and not receiving enough support at the TTI school	38.3	6.8
Child care or family-related issues	27.1	8.4
Committed to another school and did not want to go back on word	26.5	3.4
Stipend not big enough	26.5	4.5
Students in receiving schools too challenging	25.0	5.0
Timing of application did not work out	19.1	2.1
Grade level/subject area of vacancies not ideal	17.0	2.7
Does not support the philosophy of the program	11.3	2.7
Application process too difficult or too time-consuming	8.3	< 0.5
Did not like the principals at the TTI schools	4.6	0.9
Other	6.0	4.5
Sample Size	1,015	1,139

Source: Candidate Survey.

B. Types of Teachers Who Applied to Transfer and Successfully Transferred

We examined the demographic, residential, and professional characteristics of candidates by their application status to better understand the influence of their observable background characteristics in their career choices. In Table C.5, we present a comparison of characteristics between those who did not apply, those who applied but did not transfer, and those who transferred. The pattern of teacher characteristics is similar when we looked at them separately for elementary and middle school teachers.

Table C.5. Characteristics of Candidates, by Application Status (percentages unless otherwise noted)

Characteristic	Did Not Apply	Applied But Did Not Transfer	Transferred	All Candidates
Demographic				
Age				
25–35	20.6	31.1	27.8	23.0
36–45	25.8	25.7	39.2	26.7
46–55	30.0	25.7	24.1	28.8
55+	23.6	17.6	8.9	21.5
Male	17.8	19.3	17.7	18.0
Black	12.5	22.1	27.2	15.0
Hispanic	15.0	13.0	14.8	14.7
Married	71.3	60.1	61.7	68.6
Have Co-Residing Children	41.8	44.4	54.3	43.1
Residential				
Own Home	87.5	79.8	81.5	85.7
Average Commute Time[a]				
Less than 10 minutes	17.4	18.5	8.9	17.0
10–25 minutes	52.8	47.7	51.9	51.8
25+ minutes	29.8	33.8	39.2	31.2
Professional				
Base Salary (dollars)	51,856	47,699	46,604	50,740
Other Compensation (dollars)	3,825	3,551	3,565	3,757
Years of Experience in Teaching				
0 (first year teaching)	0.0	0.0	0.0	0.0
2–5 years	06.3	10.3	07.4	07.1
6–10 years	25.1	40.4	43.2	29.1
11+ years	68.6	49.3	49.4	63.8
Competitiveness of Undergraduate Institution				
Very competitive	19.3	18.4	14.8	18.9
Competitive	35.7	43.3	49.4	37.6
Has a Master's or Doctorate Degree	45.2	51.4	45.7	46.4
Has National Board Certification	11.7	7.5	11.1	11.1
Sample Size	920	223	81	1224

Source: Candidate Survey.

[a]Average commute time is for school year 2008–09 for candidates in cohort 1 districts and 2009–10 for candidates in cohort 2 districts.

Because many of the candidate characteristics listed above are related to one another, it may be difficult to isolate specific factors that best explain a candidate applying and/or transferring. Therefore, we performed multivariate analyses to understand which factors correlate with candidates' decisions to apply and transfer. We used a logistic regression with application status (whether a candidate applied or not) as the outcome, then repeated the analysis using transfer status (whether a candidate transferred or not) as the outcome.

The explanatory variables included in the regressions are the following: (1) a measure of income of the candidate, which is the base salary plus any additional compensation; (2) a set of personal characteristics of the candidate, including gender, race, marital status, and an indicator for whether the candidate has co-residing children under age 5; (3) a set of residential characteristics, including whether the candidate owns a home, and his or her average commute time from home to current school; (4) a set of professional characteristics, including two indicators denoting the competitiveness of the candidate's undergraduate institution,[94] candidate's degree, and National Board Certification status; and (5) a set of indicator variables summarizing the candidate's satisfaction with different aspects of his or her current school, including school leadership/policy, payments and benefits, professional environment, school environment and facility, and students.[95] We also accounted for any unobserved (by the researcher) effects at the district level that influenced decisions of all candidates within districts similarly, such as district union policies or district labor-market conditions. Standard errors of estimated logit coefficients account for clustering at the school level to account for the possibility of unmeasured factors common to the same potential sending school at which multiple candidates may teach.

Relative to the teachers who did not apply, teachers who applied to TTI were different in some consistent ways (Table C.6). They were more likely, holding other variables constant, to have lower income, be African American, be unmarried, or be less satisfied with their current school policy.[96] Married candidates with co-residing children under age 5 were also more likely to apply. To illustrate the magnitude of some of these likelihoods, at the average level of income of $54,568, African American teachers were 14 percentage points more likely than white teachers to apply for a TTI position, all other things also held at their mean values. Also, at the average level of income of $54,568, unmarried candidates were 10 percentage points more likely to apply for a TTI position. None of the other personal, professional, or residential characteristics was a significant predictor of application.

[94] Competitiveness of undergraduate institution is based on Barron's Profiles of American Colleges (2003). We used two indicators—the very competitive indicator takes a value of one if the candidate's undergraduate institution is one of those listed as "most" or "highly" or "very competitive" and the competitive indicator takes a value of one if the candidate's undergraduate institution is "competitive." According to Barron's profiles, a "very competitive" undergraduate institution is one that admits less than 75 percent of applicants and whose students were ranked at least in the top 35 to 50 percent in high school.

[95] The indicator variables for satisfaction were constructed from a series of aspects of a candidate's current school for which the candidate chose his or her satisfaction level on a four-point Likert-type scale: very dissatisfied, somewhat dissatisfied, somewhat satisfied, and very satisfied. A candidate was assumed satisfied for an aspect if he or she was somewhat or very satisfied. Aspects were pre-grouped in the survey questionnaire to reflect satisfaction with school leadership/policies, compensation, professional environment, school environment and facility, and students and their families. We also conducted factor analysis to confirm that the items (aspects) loaded into the five pre-defined categories. The dummy variables summarizing satisfaction with the five categories were constructed as follows: a candidate was defined as satisfied for a category and was given a value of one if he or she was satisfied with more than 50 percent of the aspects within that category, or given a zero otherwise. We also constructed an alternative set of dummy variables using a more restrictive definition, where a candidate was defined as satisfied for a category if he or she was satisfied with all the aspects within a category. However, using this alternative set of dummy variables did not change the regression results.

[96] For all of these variables and any other variables that are reported to be significant in this section, the *p*-value in the regression was less than 0.05. In a different specification, we used base salary instead of income, and the results were the same. The correlation between base salary and income is 0.89.

Table C.6. Factors Related to the Probability of Applying

Factor[a]	In All 10 Districts		In 7 Districts with Student Data[b]	
	Odds Ratio	Standard Error	Odds Ratio	Standard Error
Dependent Variable: Probability of Applying to TTI				
Income (thousands of dollars)	0.97*	(0.01)	0.98	(0.01)
Demographic Variables				
Male	1.08	(0.24)	1.16	(0.35)
African American	2.11*	(0.48)	1.69	(0.57)
Hispanic	1.03	(0.26)	0.52	(0.18)
Married	0.59*	(0.10)	0.75	(0.18)
Married with co-residing children under 5	1.64*	(0.37)	1.48	(0.45)
Residential Variables				
Owns home	0.73	(0.16)	0.48	(0.23)
Travel time in application school year (hours)	0.93	(0.33)	1.08	(0.50)
Professional Variables				
Attended a very competitive undergraduate institution	0.77	(0.18)	1.00	(0.33)
Attended a competitive undergraduate institution	0.68	(0.13)	0.85	(0.26)
Has master's degree or higher	1.28	(0.21)	1.44	(0.33)
Candidate for certification or certified	1.15	(0.23)	0.57	(0.20)
Satisfaction Indicators				
Satisfied with school policy	0.54*	(0.11)	0.65	(0.18)
Satisfied with salary	1.20	(0.20)	0.90	(0.21)
Satisfied with professional environment	0.67	(0.15)	0.68	(0.22)
Satisfied with facilities	0.98	(0.22)	0.92	(0.28)
Satisfied with students	1.04	(0.18)	1.15	(0.27)
District Indicators				
B	2.60*	(1.00)		
C	1.38	(0.45)		
D	2.32*	(0.93)		
E	0.83	(0.32)	0.30*	(0.13)
F	4.13*	(1.37)	2.44*	(0.84)
G	2.13	(0.83)		
H	1.74	(0.85)	0.60	(0.33)
I	1.70	(0.68)	0.58	(0.27)
J	0.58	(0.25)	0.33*	(0.15)
In Top 10% of Value-Added Ranking			1.40	(0.29)
Percentage of Free or Reduced-Price Lunch (FRL) Current Students			1.03*	(0.01)
Constant	2.79	(1.50)	0.61	(0.43)
Sample Size	1,127		672	
Log-Likelihood	-566.89		-300.05	
Likelihood Ratio (LR) Chi-Squared	106.05		94.34	
p-Value of LR Chi-Squared	0.00		0.00	

[a]Factors included are dummy variables unless otherwise noted.

[b]The seven districts are those where we estimated value added and had student-level data.

*Coefficient is statistically significant at the 0.05 level, two-sided test.

Next, we examined the probability of transferring for all candidates irrespective of their application status. Relative to the teachers who did not do so, teachers who went through the entire process and transferred to a TTI position were more likely to have lower income, be African American, and be married with co-residing children under 5 years old (see Table C.7).

Contrary to the hypothesized direction of the effect, transfer candidates who were satisfied with their salary and their students during the application period were two times *more* likely to transfer than those who were not satisfied with their salary and their students, a statistically significant relationship.

Table C.7. Factors Related to the Probability of Transferring

	In All 10 Districts		In 7 Districts with Student Data[b]	
Factor[a]	Odds Ratio	Standard Error	Odds Ratio	Standard Error
Dependent Variable: Probability of Transferring to a Low-Achieving School				
Income (thousands of dollars)	0.97*	(0.01)	0.97	(0.02)
Demographic Variables				
Male	1.03	(0.37)	1.35	(0.66)
African American	2.81*	(1.03)	2.39	(1.11)
Hispanic	1.45	(0.56)	0.97	(0.50)
Married	0.62	(0.21)	0.67	(0.30)
Married with co-residing children under 5	2.18*	(0.79)	1.24	(0.62)
Residential Variables				
Owns home	0.94	(0.37)	1.20	(0.56)
Travel time in application school year (hours)	1.58	(0.82)	1.64	(1.26)
Professional Variables				
Attended a very competitive undergraduate institution	0.72	(0.29)	0.84	(0.49)
Attended a competitive undergraduate institution	0.82	(0.26)	1.18	(0.58)
Has master's degree or higher	0.95	(0.26)	1.33	(0.47)
Candidate for certification or certified	1.05	(0.37)	1.01	(0.56)
Satisfaction Indicators				
Satisfied with school policy	0.65	(0.23)	0.56	(0.26)
Satisfied with salary	1.92*	(0.54)	1.49	(0.49)
Satisfied with professional environment	0.53	(0.21)	0.47	(0.24)
Satisfied with facilities	0.84	(0.33)	1.94	(1.04)
Satisfied with students	2.09*	(0.72)	2.98	(0.97)
District Indicators				
B	1.40	(0.85)		
C	0.73	(0.36)		
D	0.89	(0.75)		
E	0.58	(0.39)	0.26	(0.19)
F	2.20	(1.19)	1.41	(0.80)
G	1.21	(0.77)		
H	1.12	(0.75)	0.51	(0.36)
I	1.17	(0.67)	0.63	(0.39)
J	0.39	(0.32)	0.28	(0.24)
In Top 10% of Value-Added Ranking			0.90	(0.29)
Percentage of FRL Current Students			1.02*	(0.01)
Constant	0.28	(0.25)	0.05*	(0.06)
Sample Size	1,127		672	
Log-Likelihood	-244.78		-138.16	
Likelihood Ratio (LR) Chi-Squared	71.07		57.82	
p-Value of LR Chi-Squared	0.00		0.00	

[a]Factors included are dummy variables unless otherwise noted.

[b]The seven districts are those where we estimated value added and had student-level data.

*Coefficient is statistically significant at the 0.05 level, two-sided test.

In addition to the information on candidates' personal, professional, and residential characteristics from the Candidate Survey, we also examined if student characteristics and the value-added scores of candidates were related to their decisions to apply and/or transfer. Because students in the potential receiving schools are perceived as more disadvantaged, we hypothesized that candidates who have a higher percentage of disadvantaged students before transferring might be more willing to apply for and/or transfer to TTI positions.

However, data on student characteristics and value-added scores measured before the candidate transfer were available only for the seven districts in which we estimated value-added scores and had student-level data. In Table C.8, we present the value-added scores and student characteristics by candidates' application status for these districts.

For the seven districts where we estimated value-added scores, we examined the probability of candidates applying, using the same multivariate approach discussed above. In addition to the explanatory variables already included, we added an indicator variable indicating whether a candidate was in the top 10 percent of the value-added ranking. We also included the percentage of current students who were FRL eligible.[97] As hypothesized, candidates with a higher percentage of disadvantaged current students were more likely to apply. Teachers who applied were not significantly different in any of their personal, residential, or professional background characteristics than those who did not apply.

Table C.8. Value-Added Scores and Student Characteristics of Candidates, by Application Status (percentages except for value-added scores)

Characteristic	Did Not Apply	Applied But Did Not Transfer	Transferred	All Candidates
Teacher Value-Added				
Math (score)[a]	0.22	0.23	0.21	0.22
Reading (score)[a]	0.13	0.14	0.14	0.13
Percentage in top 10	48.3	48.9	44.8	48.2
Student Demographics				
Male	50.6	51.5	50.8	50.8
White	36.4	24.8	23.7	34.0
African American	16.5	25.6	27.4	18.4
Hispanic	38.4	41.2	41.3	39.0
Student Economic Status				
FRL	52.1	67.3	67.5	55.1
Student Academic Status				
Special education (SPED) status	15.0	13.4	15.7	14.8
English-language learners (ELLs)	11.5	15.9	15.5	12.3
Sample Size	815	158	53	1,026

Source: Administrative data and Candidate Survey. Data pertain to a subgroup consisting of seven districts that provided student-level data.

[a]Value-added scores are in student-level standard-deviation units standardized at the district level.

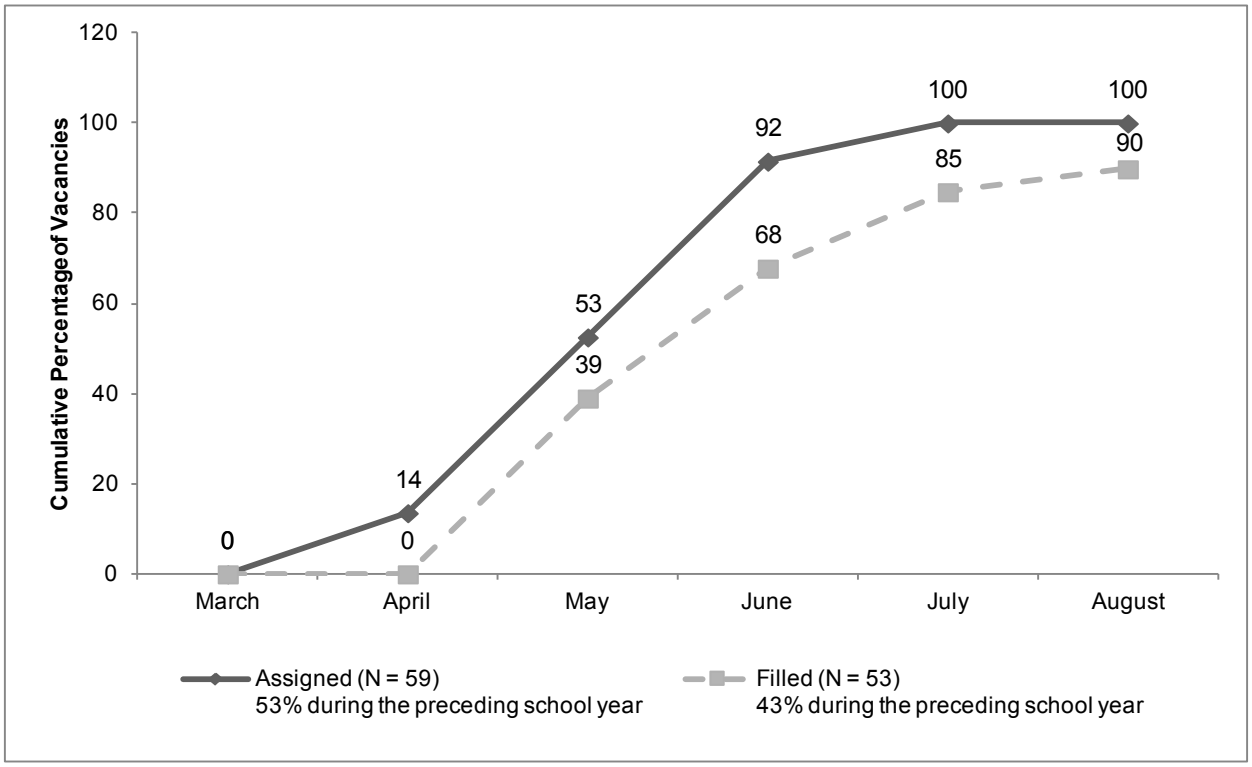
[97] Percentage African American, percentage Hispanic, and percentage ELLs are three other measures of student disadvantage that we did not include in the regression because of their high correlation with percentage of FRL status: 0.30, 0.45, and 0.45, respectively.

When examining transfer behavior irrespective of application status and focusing on these seven districts, none of the background characteristics of the candidates was a statistically significant predictor of transfers. Similar to application, candidates with a higher percentage of disadvantaged current students measured by FRL status were more likely to transfer.

C. Identifying and Filling Vacancies at Different Grade Spans

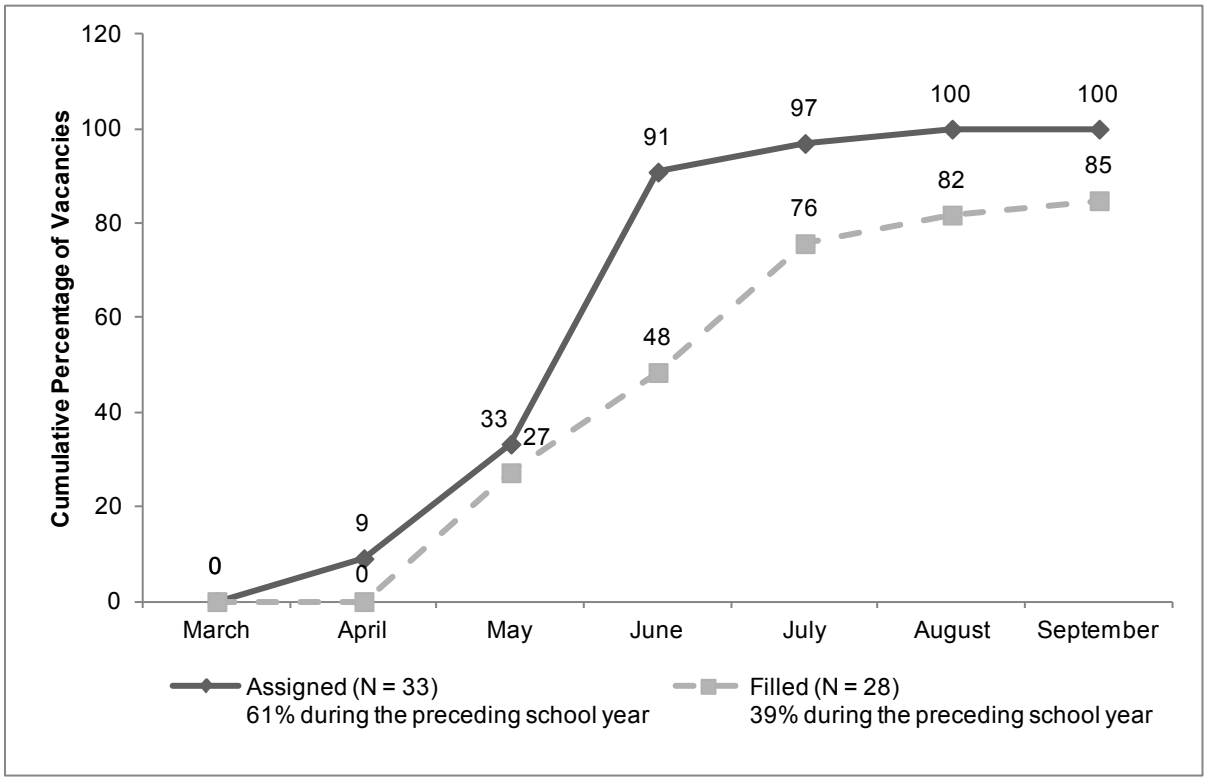
When we examined the timing of identifying and filling vacancies by grade span, we found that for both elementary and middle schools, all vacancies were identified between April and August—59 for elementary and 33 for middle. Most vacancies were assigned and filled in May and June. In Figures C.1 and C.2, we show that 78 percent of elementary school vacancies and 82 percent of middle school vacancies were assigned in these two months. Sixty-one percent of elementary school vacancies were also filled in these two months. Middle school vacancies took longer, but 76 percent were filled by between May and July. By the end of the recruitment season before the beginning of the next school year, 90 percent of elementary school vacancies assigned to treatment were filled with a TTI candidate, which was higher than the middle school rate of 85 percent.

Figure C.1. Percentage of Elementary-Level TTI Vacancies Assigned and Filled, by Month



Source: TTI program records.

Figure C.2. Percentage of Middle School-Level TTI Vacancies Assigned and Filled, by Month



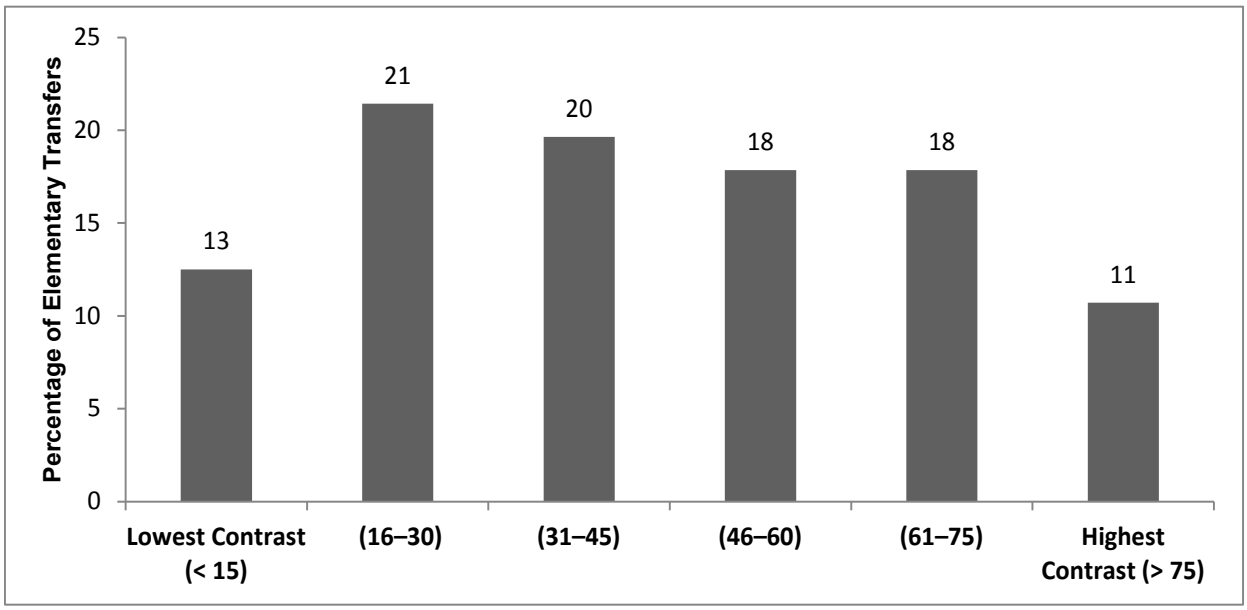
Source: TTI program records.

D. Types of Transfer by Grade Span

We examined whether TTI transfer teachers moved between schools that are close to each other in achievement ranks by grade span. Such lower-contrast moves are possible because the dividing line between potential sending and receiving schools was somewhat arbitrary. We grouped the transfers by the degree of contrast, measured as the difference in the rank between the sending school and the receiving school for a given transfer. The maximum degree of contrast would be a transfer from the highest-achieving school in the district to the lowest-achieving school, a difference of 100 percentile points.

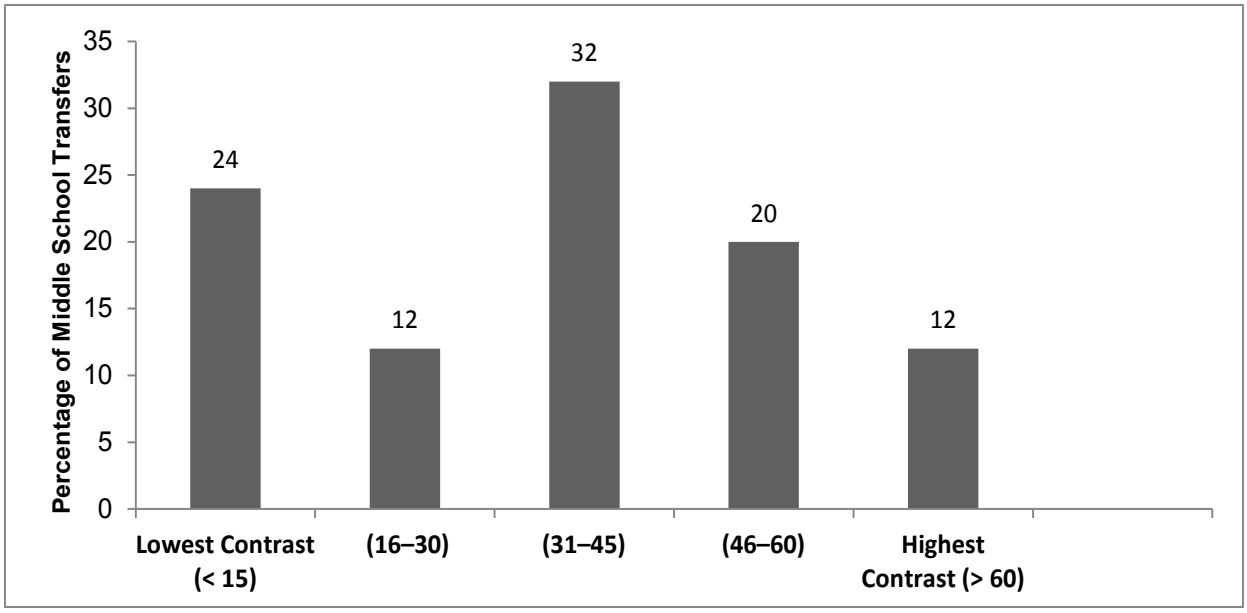
In elementary schools, 34 percent of the transfers were between schools that were ranked within 30 percentile points of each other; the corresponding number for middle school transfers was 36 percent (Figures C.3 and C.4). However, 13 percent of the transfers at the elementary school level were between schools fewer than 15 percentile ranks apart. The corresponding percentage for middle school transfers was 24.

Figure C.3. Types of Elementary-Level Transfers, by Achievement Rank



Source: Administrative data and TTI program records (N = 52)

Figure C.4. Types of Middle School-Level Transfers, by Achievement Rank

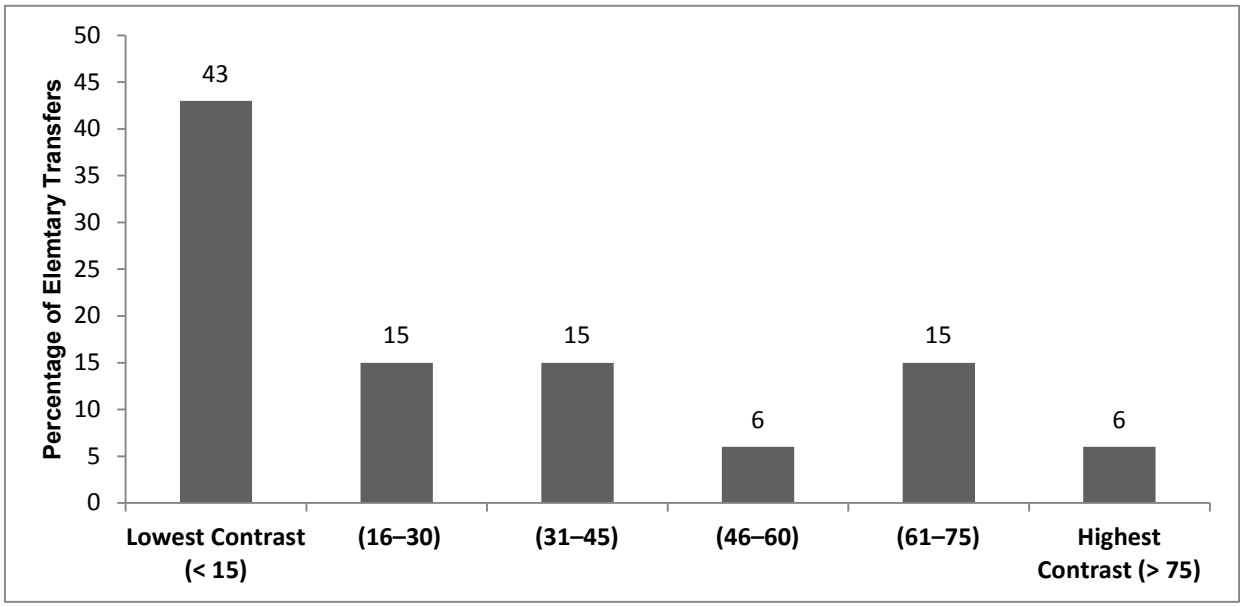


Source: Administrative data and TTI program records (N = 29).

Although the poverty status of schools was not taken into account when identifying potential receiving schools, it is still informative to examine the contrast in the percentile-rank positions based on poverty status (as measured by percentage of students eligible for FRL) of schools that transfer teachers left, compared with the ones to which they moved, because FRL is used as a measure of student disadvantage in many federal programs. If we found only low-contrast transfers, based on FRL ranks, it would suggest that TTI is redistributing teachers within the same group of disadvantaged students.

We also examined the contrast in school ranks between the sending school and the receiving school for a given transfer by the percentile rank positions based on poverty status (as measured by percentage of students eligible for FRL).[98] Nineteen percent of elementary school teachers and 13 percent of middle school teachers who transferred moved between schools whose ranks were more than 60 percentile points apart based on poverty status. However, 43 percent of the elementary school teachers and 38 percent of the middle school teachers who transferred moved between schools that were fewer than 15 percentile ranks apart from each other. Figure C.5 shows the distribution by poverty ranks for elementary schools. Middle schools, for which the number of transfers was too small to display in detail, followed a similar pattern.

Figure C.5. Types of Elementary-Level Transfers, by Poverty Ranks



Source: Administrative data and TTI program records (N = 52).

[98] We ranked schools in descending order of poverty status as measured by the percentage of students eligible for FRL. Therefore, schools with higher poverty status or higher percentage of students eligible for FRL have a lower rank.

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

APPENDIX D

IDENTIFICATION OF FOCAL TEACHERS

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

A. The Importance of Identifying Focal Teachers

Following the logic model for studying the transfer incentive intervention laid out in Chapter I and the experimental study design described in Chapter II, it is critical to identify the "focal" and "nonfocal" teachers on both the treatment and control teams. The focal teacher is defined as the person who fills the vacancy that was identified at the time of random assignment. All other teachers on the team are considered nonfocal.

The distinction between focal and nonfocal teachers is important for understanding the counterfactual outcomes—those that would have been realized in the absence of the Talent Transfer Initiative (TTI). To understand the counterfactual, it helps to think about who would have taught the students who were taught by the TTI transfer teacher had the TTI not been made available to their school. They might have been taught by a teacher who the principal hired or who moved into that grade and/or subject under the school's normal processes for filling a vacancy. Alternatively, the students might have been assigned to classes differently and, therefore, taught by some combination of the teachers who were already in that teaching team (grade and subject). Focal teachers were defined as those who were new to the grade, meaning that they could be new hires, new transfers from another school, transfers from another position within the school, or substitute teachers who were made permanent.

B. Challenges in Identifying Focal Teachers

Identifying the focal teacher was not always possible, especially on control teams. For 85 percent of treatment teams, the vacant positions were filled by someone who transferred via the TTI. In such cases, it was easy to determine the identity of the focal teacher because the TTI transfers were tracked carefully to ensure that stipends were paid to the correct people. For treatment vacancies that were not filled by TTI teachers, we typically still had information from the TTI implementation team about the newly hired teacher because the site manager was in regular contact with the principal trying to fill that position. On control teams, however, principals were not required to communicate with the TTI implementation team or the study team, so we had to rely on the Teacher Background Survey to identify teachers who were new to the grade in the fall after random assignment. When the survey data did not resolve a teacher's status, we attempted to contact the principal. The teacher and principal reports were not always complete and did not always agree, so we reconciled the discrepancies and used the best available information, coding uncertain cases as we encountered them. Principal reports took precedence over the responses in teacher surveys because principals were asked directly for the name of the teacher who filled a particular vacancy. However, the responses teachers gave in the survey about their own years in the school and grade were considered more accurate than principals' recollections. Therefore, we overrode the principal if he or she identified as the focal teacher someone who said on a survey that he or she was not new to the grade.

Unfortunately, however, there were several situations in which we were not able to resolve the identity of the focal teacher. In some instances, the data showed more teachers who were new to the grade than the number of vacancies identified. This could happen if vacancies opened up after random assignment. In other cases, if not everyone on the team responded to the survey, the data could show fewer teachers who were known to be new to the grade than the number of vacancies. This could also happen if there was an error or omission on the teacher roster supplied by the school because we relied on the rosters to form the sample frame for the Teacher Background Survey. Another possibility is that a vacancy was lost because of some combination

of enrollment declines, increases in class size, or current teachers taking on more classes in that grade and subject.

In all cases where the data did not single out one teacher per vacancy, there were teachers whom we refer to as having "ambiguous focal status."[99] Eighteen of 80 control teams (23 percent) and 4 of 85 treatment teams (5 percent) had at least one ambiguous focal teacher. Ambiguous focal teachers were found in every grade and in 8 of the 10 districts. However, the teams with ambiguous focal teachers should not be considered a random subsample of the full study sample. They were significantly more likely to be in the control group (percentages noted above), and the reasons for ambiguity suggest that the teachers in question are likely to be different in many ways from those with unambiguous focal teacher status. Survey nonresponse is the main reason for the ambiguity on teams for which we did not have program information. Communication or engagement with principals can also be a factor because when we called principals to resolve uncertain cases, some did not respond to our requests and others said they did not know who filled the vacancy.

C. Our Approach to Identifying the Focal Teachers

Because there were ambiguities, we used two different definitions of focal teacher: one using a selective rule and the other an inclusive rule. The selective definition classifies only teachers who can be linked to the study vacancy based on information provided by principals or teachers' reports. If more than one teacher was new to the team and the principal did not identify which teacher filled the study vacancy, we designated all new teachers as focal teachers and assigned each a proportional weight that sums to the number of vacancies on the team.[100] Under the selective rule, if neither principals nor teachers provided sufficient information about who filled the vacancy on a team, no focal teacher was identified for that team, and the team was not included in the analysis.

For the inclusive definition, on the other hand, we classified at least one teacher on every team as the focal teacher, even if there was limited evidence that the person was the true focal teacher. In cases where we could not determine which teacher was the focal teacher, we included all ambiguous focal teachers and assigned each of them a proportional weight that sums to the number of vacancies in the team. The sum of the weights represents the number of vacancies for which we could identify at least one focal teacher who responded to the survey.

[99] As discussed below, some teams had more vacancies than those known at the time of random assignment. If we knew the identity of the teachers filling the vacancies on such teams, we did not count these as ambiguous.

[100] Even though the particular teacher who filled the study vacancy was unknown or not certain, we included these new entrants to the team in the selective definition for two reasons. First, the teachers were not hired for specific vacancies or "chairs," so the distinction of who filled which position was generally meaningless. Second, the average outcome for these new entrants is a reasonable approximation to the counterfactual because in both cases the student is taught by someone new to the grade.

The nonfocal teachers were handled in the same way. On teams where we could identify the focal teacher, we also could be sure of the nonfocal teachers. On teams with ambiguous focal teachers, we could not be sure if the teachers were focal or nonfocal. As a result, ambiguous focal teachers were also considered to be nonfocal teachers for the inclusive nonfocal analysis. They were weighted according to the number of nonfocal positions expected to be on the team based on survey responses and teacher rosters. As a result, some teachers are included in both the inclusive focal and nonfocal samples. Importantly, however, under the inclusive rule, every team has at least one nonfocal teacher identified.[101]

In Table D.1, we present an illustration of identifying and weighting focal and nonfocal teachers under three different scenarios. In this example, all three teams have one study vacancy out of three positions on the team, but our knowledge about the focal teacher differs.

- **Team 1.** This is the full-information case. Based on survey responses, we know that teacher A is the focal teacher and teachers B and C are nonfocal teachers. Teacher A is identified as the focal teacher under both the selective and inclusive definitions and is weighted as 1 in both cases. Teachers B and C are nonfocal teachers under both definitions and are weighted as 1 in both cases.

- **Team 2.** This is the case of extra vacancies. Based on survey responses, we know that teachers D and E were new to the team, but we cannot link the teachers to specific vacancies. Both are identified as focal teachers under the selective and inclusive definitions. Each is weighted as 0.5 because there was only one treatment-eligible vacancy on the team. Teacher F was not new to the team and is identified as a nonfocal teacher under both the selective and inclusive definitions. Because there were two nonfocal positions on the team (three teachers minus one focal teacher), teacher F is assigned a nonfocal weight of 2.[102]

- **Team 3.** This is the case of completely missing data. We do not know which teacher filled the study vacancy because all teachers were survey nonrespondents; therefore, no teacher is identified as focal or nonfocal under the selective definition. Under the inclusive definition, all three teachers are identified as focal teachers and each is weighted as 0.33. The sum of the weights is equal to 1—the number of vacancies on the team. Likewise, all three teachers are also identified as nonfocal teachers under the inclusive definition, because we do not know if they were new to the team or not. Each is assigned a nonfocal weight of 0.67. The nonfocal weights sum to 2 because there are two nonfocal positions on the team (three teachers minus one focal teacher). The three teachers are used in both the focal and nonfocal comparisons.

[101] This approach was used to identify focal and nonfocal teachers in the Teacher Background Survey sample as well as in the test-score analysis sample. Because the survey sample was based on teacher rosters, it did not align exactly with the sample of teachers in the test-score analysis. This could have been due to roster error or to teacher transfers during the school year (rosters were collected in the fall and test-score data were collected in the spring). For this reason, although the approach used was identical, a given team might have a different sample of focal and nonfocal teachers (with different weights) in each analysis.

[102] An alternative approach would be to assign a weight of 1 to the known nonfocal teacher (Teacher F) and weights of 0.5 to each of the teachers who could have been the nonfocal new hire (Teachers D and E). Our default approach, assigning all the weight to Teacher F in this case, reduces the amount of double-counting.

Table D.1. Example of Focal Teacher Identification

	Survey Information	Selective Focal Weight	Inclusive Focal Weight	Selective Nonfocal Weight	Inclusive Nonfocal Weight
Team 1					
Teacher A	New to team	1	1		
Teacher B	Not new to team			1	1
Teacher C	Not new to team			1	1
Sum of Weights		1	1	2	2
Team 2					
Teacher D	New to team	0.5	0.5		
Teacher E	New to team	0.5	0.5		
Teacher F	Not new to team			2	2
Sum of Weights		1	1	2	2
Team 3					
Teacher G	Survey nonrespondent		0.33		0.67
Teacher H	Survey nonrespondent		0.33		0.67
Teacher I	Survey nonrespondent		0.33		0.67
Sum of Weights			1		2

For the test-score-impact analysis, there was an additional complication in identifying elementary focal teachers. Unlike at the middle school level, where all teams were defined as either reading or math, elementary teams were included in both subject analyses. This is because most elementary schools in the study had self-contained classrooms, and departmentalization was not always formalized or documented. It was possible, however, for an elementary focal teacher to teach only reading or math. In these cases, a known focal teacher could be linked in the data to students for one subject but not the other. Although the identity of the focal teacher would be known for one subject, there was no true focal teacher in the opposite subject. To avoid dropping this team from the opposite-subject analysis of focal teachers, all teachers linked to students on the team for the opposite subject were identified as focal and weighted equally (as in the example of Team 3 in Table D.1).

The analysis in this report uses the inclusive definition so as to maximize the sample size and avoid discarding any teams' data. Survey nonresponse is the main reason that many teams have an ambiguous focal teacher. In the example in Table D.1, because of survey nonresponse, team 3 has no teachers identified as focal or nonfocal under the selective definition. That team would not be included in the selective focal analysis because we cannot determine whether the nonresponding teachers are new to the school or teaching team. As previously noted, this problem is mainly an issue for the control teams; in most treatment teams the treatment focal teacher is the TTI transfer, and his or her identity is known from program records regardless of survey response. Because the selective definition affects the treatment and control teams differently, relying on the selective definition would risk introducing a severe selection bias into our impact estimates. The inclusive definition does not have this problem, but using it does mean that the control group will include some "extra" teachers who are not strictly filling the vacant position. In that case, we need to interpret the control focal condition more broadly as representing teachers of students who could have been assigned to the focal teacher had there not been an intervention. If these extra teachers are more experienced than the true focal teachers, the inclusive definition would dampen the impact estimate.

An interim report from Glazerman et al. (2012) used the selective definition of focal teachers. For that analysis, it was more important to describe the teachers who were hired rather than to compare treatment and control characteristics, so the possibility of selection bias was not the overriding concern. For this report, we considered the selective definition of focal teachers as a sensitivity test (see Chapter V and Appendix G).

D. Comparing the Inclusive and Selective Definitions

In Table D.2, we compare the average characteristics of focal teachers using the selective versus the inclusive definition. Under both definitions, the treatment focal teachers have significantly more years of teaching experience and are more likely to hold National Board Certifications than control focal teachers. Treatment focal teachers are also older, on average, and a greater percentage of them are home owners.

Table D.2. Characteristics of Focal Teachers Under Alternative Definitions

Characteristic	Control Focal (Selective Definition)	Control Focal (Inclusive Definition)[a]	Treatment Focal (Selective Definition)	Treatment Focal (Inclusive Definition)
Professional Background				
Years of Experience in Teaching (average years)	7.4	8.2	11.9*	11.8*
Years of Experience in Teaching (percentages by category)				
1 (first year teaching)	20.4	17.2	0.0*	0.0*
2–5 years	38.2	38.0	11.4*	12.3*
6–10 years	20.1	20.0	44.7*	43.6*
11+ years	21.3	24.7	43.9*	44.1*
First Year in the School	64.1	54.4	93.1*	90.9*
First Year in the Grade	65.3	58.3	40.9*	40.7*
Has Regular Certification for Grade/Subject Taught	95.6	94.1	98.7	97.5
Has a Master's or Doctorate Degree	32.7	36.1	46.9	46.2
Has National Board Certification	8.5	9.0	20.7*	20.2*
Has Undergraduate Degree from Institution Rated Very Competitive or Higher by Barron's	26.2	23.9	15.9	16.3
Personal Background				
Female	84.2	82.1	82.9	83.3
Race/Ethnicity				
White, non-Hispanic	43.5	42.1	49.9	49.5
African American, non-Hispanic	29.3	33.0	28.0	27.7
Hispanic or Latino	17.1	15.5	16.2	16.6
Average Age (years)	35.7	36.5	41.8*	41.8*
Married or Living with a Partner	54.5	54.4	62.4	62.1
Home Owner	46.1	48.0	78.0*	77.4*
Sample Size (number of respondents)	66	99	85	89

Source: Teacher Background Survey.

Results are weighted to account for the probability of survey nonresponse and the probability of the teacher being the focal or nonfocal teacher.

*Statistically significant at the 0.05 level, two-tailed test.

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

APPENDIX E

SUPPLEMENTAL MATERIALS FOR CHAPTER IV

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

In this appendix, we present supplemental information corresponding to Chapter IV, which was a discussion of intermediate impacts.

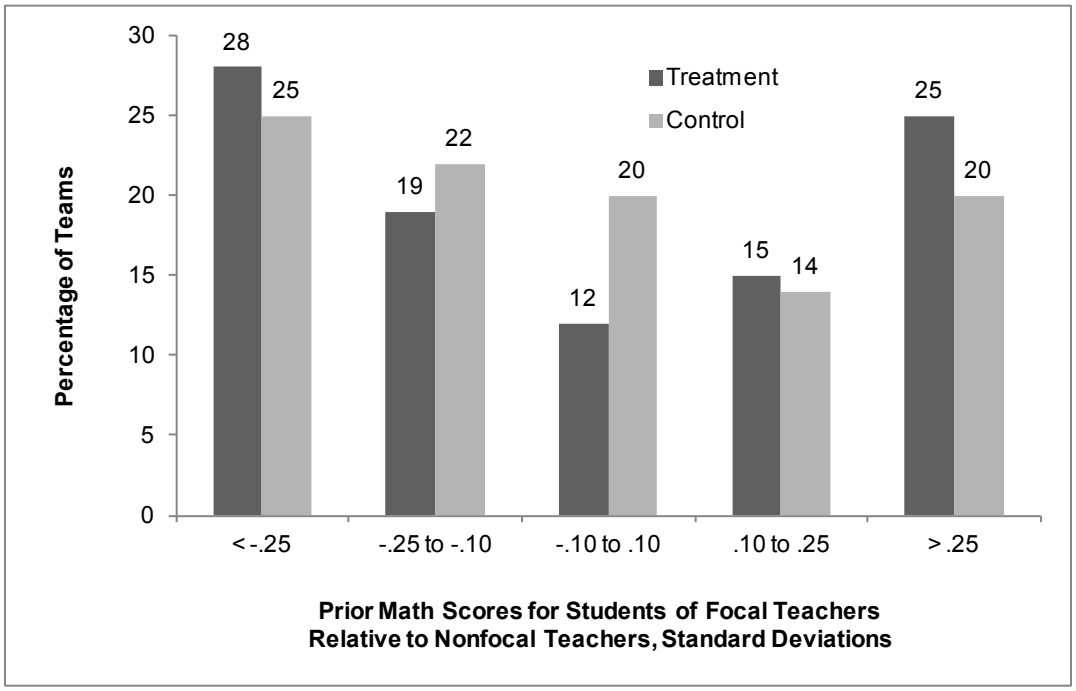
A. Assignment of Students to Teachers

Evidence from administrative data. We compared the prior achievement and demographic background characteristics of students assigned to focal versus nonfocal teachers separately for treatment and control teams. We computed focal versus nonfocal differences in average characteristics on each team and then presented their distribution in Figure E.1 for prior math scores, in Figure E.2 for prior reading scores, and in Figure E.3 for percent free or reduced-price lunch (FRL). The differences in test scores are reported in standard-deviation units, so a difference of 0.25 is one-quarter of a standard deviation. This size difference would separate, for example, a student at the 50th percentile in his or her state from one at the 40th percentile (10 percentile points).

In the figures, we illustrate the disparities within teams and the difference in disparities by treatment status. Taller bars on the left side of the graph (less than -0.10 standard deviations or greater than 5 percent FRL) imply that focal teachers on those teams were assigned more disadvantaged students than were nonfocal teachers on the team. Taller bars on the right side of the graph (greater than 0.10 standard deviations or less than -5 percent FRL) imply the opposite: that focal teachers were assigned students who were less disadvantaged. Tall bars in the middle category (-0.10 to 0.10 standard deviations or -5 to 5 percent FRL) suggest that there was no differential assignment of students.

In all three cases, the treatment-control differences were not statistically significant. Even though the test score figures (Figures E.1 and E.2) showed that focal teachers on some teams had more-disadvantaged students (taller bars on the left side of the figure) and others had less-disadvantaged students (taller bars on the right side of the figure) and the heights of those bars were not always the same for adjacent (treatment-control) pairs, the pattern is consistent with random chance. We conducted a chi-square test for the independence of treatment status and relative disadvantage and failed to find significant differences for math pre-test, reading pre-test, or FRL.

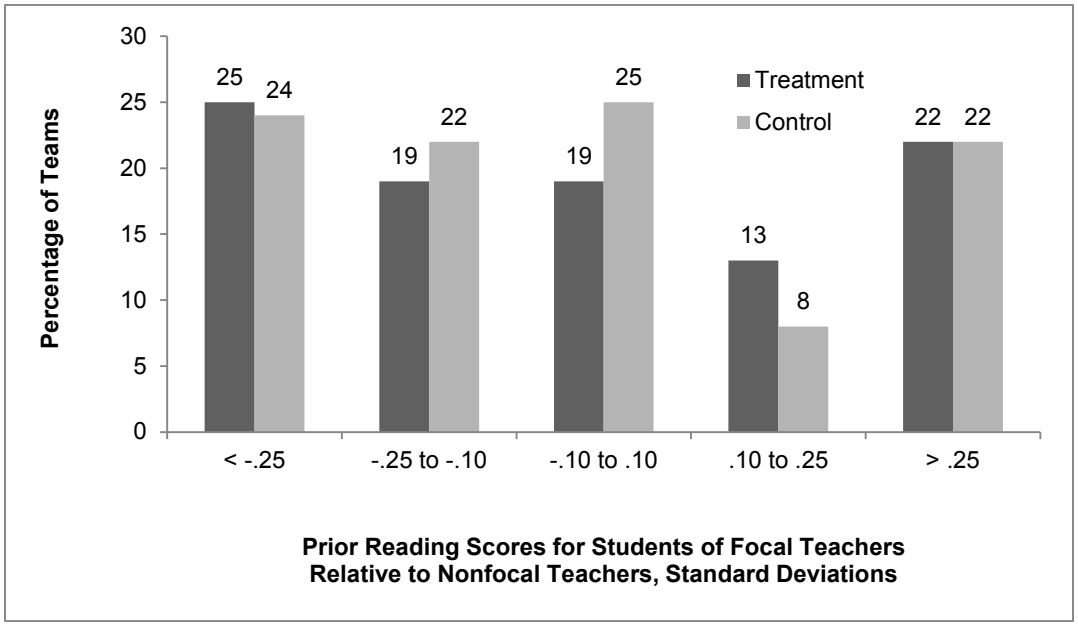
Figure E.1. Students' Prior Math Scores in Focal Teachers' Classrooms



Source: Administrative data.

Note: N = 67 teams in the treatment group and 51 teams in the control group. Distributions are not significantly different based on a Pearson's chi-square test of independence.

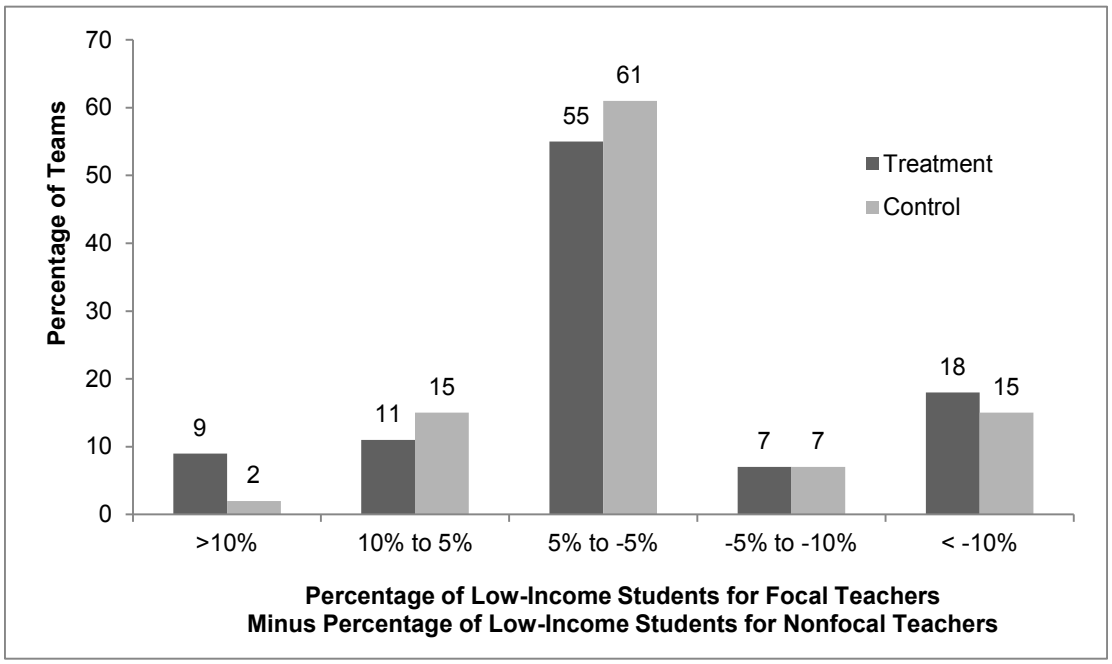
Figure E.2. Students Prior Reading Scores in Focal Teachers' Classrooms



Source: Administrative data.

Note: N = 67 teams in the treatment group and 51 teams in the control group. Distributions are not significantly different based on a Pearson's chi-square test of independence.

Figure E.3. Low-Income Students in Focal Teachers' Classrooms

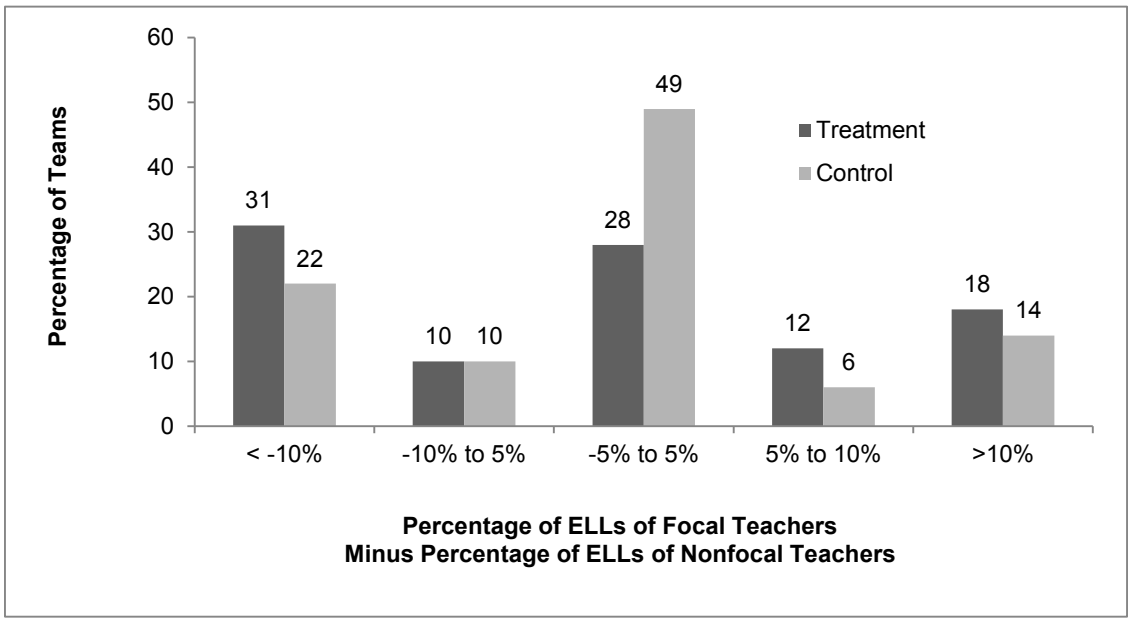


Source: Administrative data.

Note: N = 55 teams in the treatment group and 41 teams in the control group. Two districts did not provide sufficient detail on FRL. Distributions are not significantly different based on a Pearson's chi-square test of independence.

The findings were robust. We repeated this analysis for elementary and middle school teams only, and we repeated it for the second program year (cohort 1 districts only) and found no significant treatment-control differences. We constructed similar figures demonstrating the distributions for differentials in terms of percentages of students who received special education (SPED) services, and belonging to certain race/ethnic categories (see Figures E.4–E.8). As with the test score and FRL results, the differences in distributions were not statistically significant for any of these characteristics except for the percentage of white students in program year 1, which is likely to be spurious, given the large number of hypotheses being tested and the lack of any pattern.

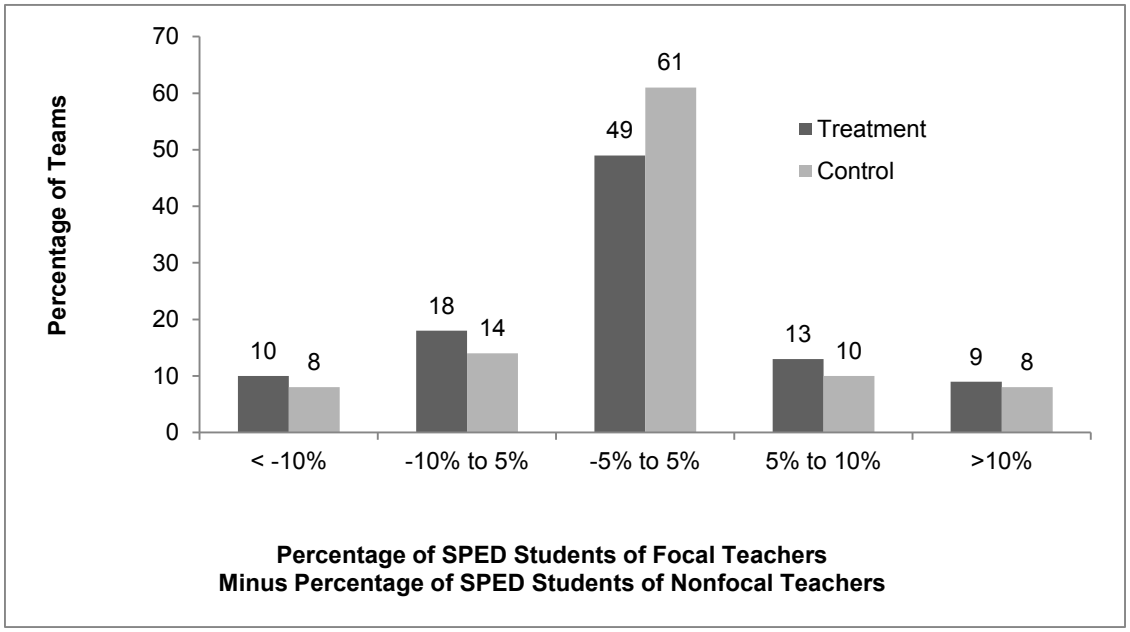
Figure E.4. ELLs, Relative Percentage in Focal Teachers' Classrooms



Source: Administrative data.

Note: N = 67 teams in the treatment group and 51 teams in the control group. Distributions are not significantly different based on a Pearson's chi-square test of independence.

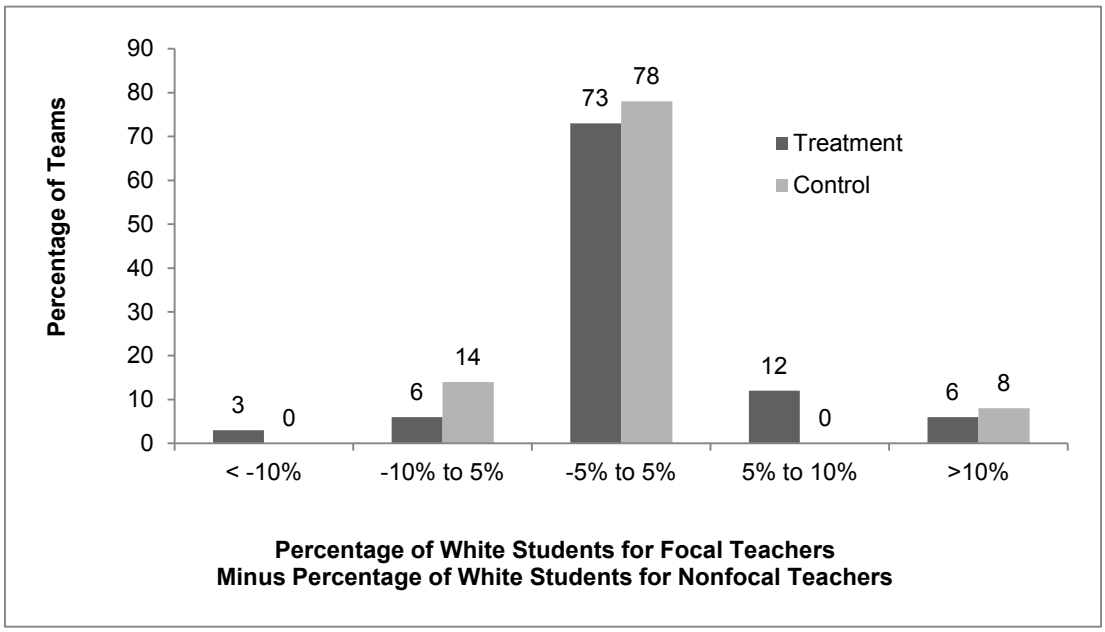
Figure E.5. SPED Students, Relative Percentage in Focal Teachers' Classrooms



Source: Administrative data.

Note: N = 67 teams in the treatment group and 51 teams in the control group. Distributions are not significantly different based on a Pearson's chi-square test of independence.

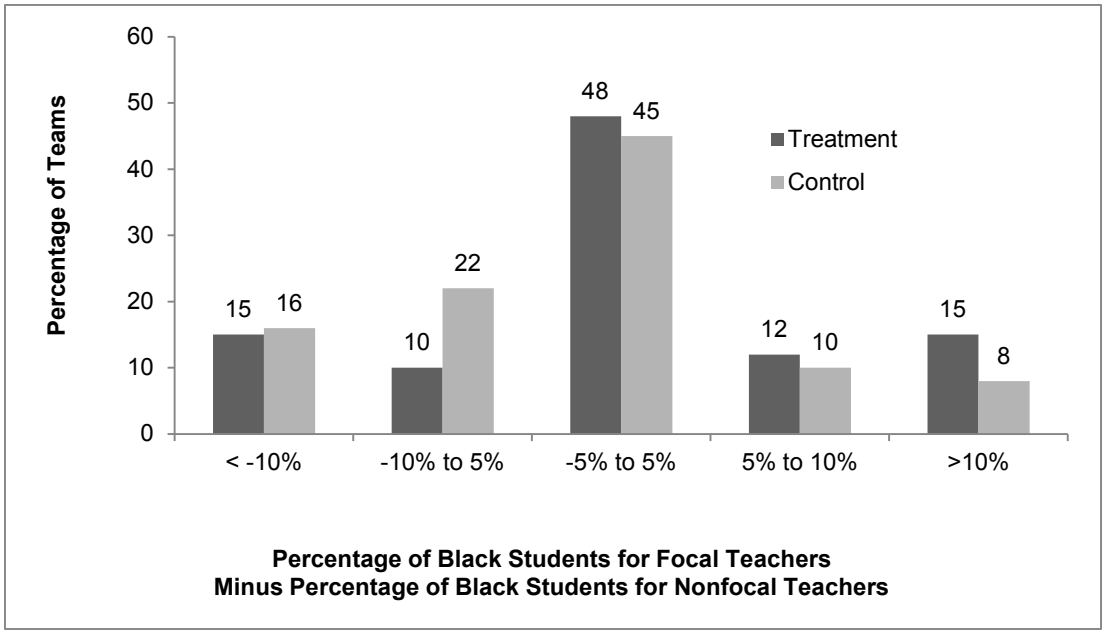
Figure E.6. White Students, Relative Percentage in Focal Teachers' Classrooms



Source: Administrative data.

Note: N = 67 teams in the treatment group and 51 teams in the control group. Pearson's chi-square test of independence is rejected at the 0.05 level, suggesting that the distributions by treatment status are not the same.

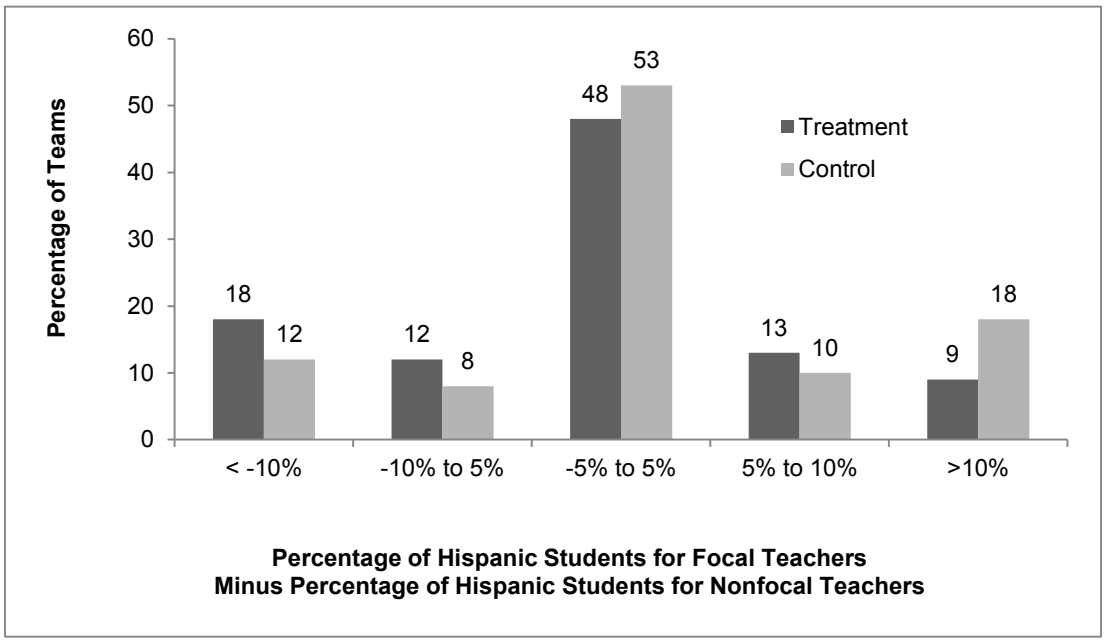
Figure E.7. Black Students, Relative Percentage in Focal Teachers' Classrooms



Source: Administrative data.

Note: N = 67 teams in the treatment group and 51 teams in the control group. Distributions are not significantly different based on a Pearson's chi-square test of independence.

Figure E.8. Hispanic Students, Relative Percentage in Focal Teachers' Classrooms



Source: Administrative data.

Note: N = 67 teams in the treatment group and 51 teams in the control group. Distributions are not significantly different based on a Pearson's chi-square test of independence.

Evidence from Teacher Survey data. In Chapter IV, we presented evidence on teacher reports of whether their own students were more, less, or equally challenging—in terms of academic abilities—than the students assigned to their peers in the same grade. Figure E.9 mirrors Figure IV.1 in Chapter IV, but shows the corresponding percentages of the same groups of teachers for the question about how challenging the students were in terms of behavior.[103] As was the case when they assessed academic challenges of dealing with their students, teachers most commonly reported having "about the same" level of behavioral challenges as their peers rather than "more" or "less." However, they reported having more-challenging students more frequently than less-challenging students in all four groups. The treatment-control differences in teacher perception of students' behavioral challenges were not statistically significant for focal or nonfocal teachers.

[103] The survey item was worded as follows: "Think about the DISCIPLINARY ISSUES of the students assigned to your class(es) this year compared with those of student assigned to your colleague(s) teaching the same grade level or subjects in your school. Would you say the students in YOUR class(es) are…

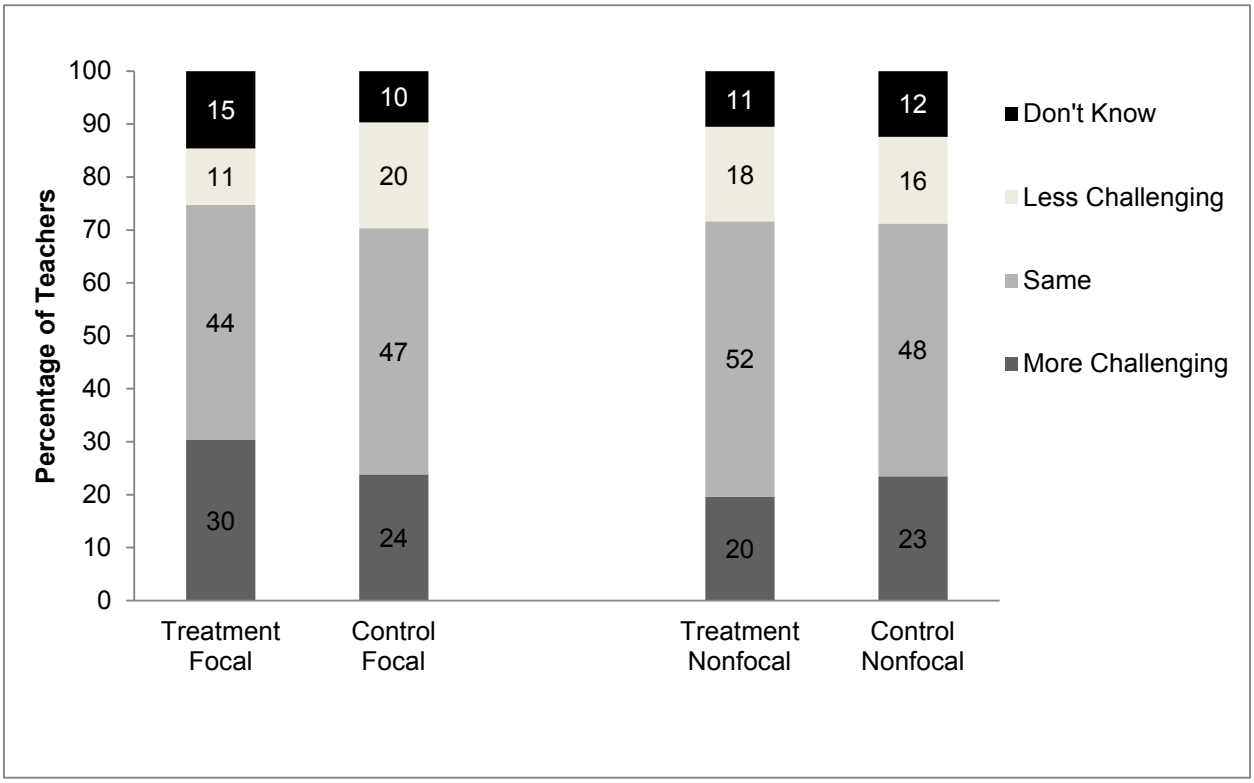
a. More challenging in terms of disciplinary issues.

b. About the same in terms of disciplinary issues.

c. Less challenging in terms of disciplinary issues.

d. Cannot judge. I am unfamiliar with the disciplinary issues of the students in the other class(es)."

Figure E.9. Classroom Assignment of Students Who Are More or Less Behaviorally Challenging, Teachers' Perceptions, by Their Treatment and Focal Status



Source: Teacher Background Survey.

Note: N = 86 treatment focal, 97 control focal, 195 treatment nonfocal, and 172 control nonfocal respondents. Treatment-control differences are not statistically significant at the 0.05 level using a two-sided test.

When we examined the same phenomenon separately for elementary and middle schools (not shown), we found that the pattern in elementary schools was basically the same as the full sample with no significant treatment-control differences. In middle schools, there was a nonsignificant difference: 30 percent of treatment focal teachers versus 11 percent of control focal teachers reported their own students were "more challenging" (p-value = 0.069).

B. Teacher Satisfaction

In terms of satisfaction, treatment focal teachers were more satisfied with their compensation than were control focals (78 versus 62 percent), but all other differences were not statistically significant. The results are summarized in Table E.1. We asked whether respondents were very dissatisfied, somewhat dissatisfied, somewhat satisfied, or very satisfied with each of 22 aspects of teaching that we grouped into five categories. Satisfaction in each category was measured as the percentage of items for which the respondent was "somewhat" or "very" satisfied. The compensation category had two items: salary and benefits. It should be noted that most treatment focal teachers (TTI transfers) had been receiving payments as part of the intervention. The one category for which less than one-half of respondents in all four groups were satisfied was "students and their families," which included the following four items: student motivation to learn, student discipline and behavior, student academic performance, and parental involvement in the school.

Table E.1. Teacher Satisfaction with School, by Focal and Nonfocal Teachers

	Focal Teachers			Nonfocal Teachers		
Outcome[a]	Treatment	Control	Difference	Treatment	Control	Difference
Leadership/policies	67.2	68.5	-1.2	70.3	68.1	2.2
Compensation	77.8	61.7	16.1*	62.0	61.2	0.8
Professional environment	72.5	70.8	1.8	73.7	71.1	2.7
School environment and facilities	73.6	69.7	3.9	71.0	70.8	0.2
Students and their families	41.9	35.7	6.2	37.0	32.6	4.4
Sample Size (teachers)	87	99		198	176	

Source: Teacher Background Survey.

[a]Units are average percentages who reported being "satisfied" or "very satisfied" on items making up each composite.

[b]Sample size is number of teachers.

*Differences are statistically significant at the 0.05 level using a two-sided test.

C. Principal Reports on Team Climate

In order to gauge the impact of TTI on teacher collaboration, trust, and sharing of ideas, we asked principals to rate the treatment and control teacher teams on each of these three dimensions.[104] The level of collaboration was measured from "highly independent" (1) to "highly collaborative" (5). The other two measures were scaled from "no extent" (1) to "a great extent" (5). The results are shown in Table E.2.

The results presented in Table E.2 were robust. We computed the changes over time for principals who had been in the same school in the prior year and found that year-to-year changes were not significantly different between treatment and control schools. We also found no significant treatment-control differences when we focused on elementary schools only or middle schools only, or when we collapsed the 5-point scale, comparing the percentages of respondents who answered either 4 or 5.

[104] The survey items were worded as follows:

"On a scale of 1 to 5, where 1 is 'Highly independent' and 5 is 'Highly collaborative,' how would you rate the level of collaboration among teachers in grade X?"

"On a scale of 1 to 5, where 1 is 'Little or no extent' and 5 is 'Great extent,' how would you rate the extent to which teachers in grade X trust and mutually respect one another?"

"On a scale of 1 to 5, where 1 is 'Little or no extent' and 5 is 'Great extent,' how would you rate the extent to which teachers in grade X seek ideas from one another?"

Table E.2. Principal Reports on Team Climate

Climate Measure (score on a 5-point scale)	Treatment	Control	Impact	Effect Size	*p*-Value
Level of Collaboration					
Pre-implementation	2.85	2.82	0.03	0.02	0.897
Program year 1	3.77	3.62	0.14	0.13	0.411
Program year 2	3.69	3.53	0.16	0.15	0.426
Degree of Trust and Mutual Respect					
Pre-implementation	2.87	3.05	-0.18	-0.15	0.410
Program year 1	3.77	3.90	-0.13	-0.13	0.424
Program year 2	3.82	3.81	0.01	0.01	0.965
Teachers Seek Ideas From One Another					
Pre-implementation	2.91	3.00	-0.09	-0.07	0.677
Program year 1	3.86	3.78	0.08	0.08	0.633
Program year 2	3.91	3.76	0.15	0.15	0.437
Sample Sizes (teams)					
Pre-implementation (cohorts 1 and 2)	67	61			
Program year 1 (cohorts 1 and 2)	80	77			
Program year 2 (cohort 1 only)	55	56			

Source: Principal Survey.

Note: Principals who were new to the school were not asked to report on pre-implementation climate. Effect size is the impact divided by the standard deviation for the pooled treatment and control groups.

D. Principal Ratings of Teacher Contributions

We also asked principals to offer their subjective ratings of their own teachers' contributions to the school outside of the classroom. We asked principals in the spring to rate three aspects of their teachers' performance: (1) leadership, (2) contribution to activities outside the classroom, and (3) whether they are assets to the school in general.[105]

The results, presented in Table E.3, show that the differences were not statistically significant. This does not necessarily mean that teachers in the treatment group were not held in higher regard along these dimensions, but that the differences were not large enough for us to conclude that they were due to the intervention and not to other factors, given the size of our sample (160 teams).

In Table E.3, we show the percentages of teams whose principals agreed or agreed strongly with each statement. We also calculated the percentage who agreed strongly (or disagreed strongly) and the average score, assuming that each increment on the 4-point scale was equal. Nearly all of the differences were statistically insignificant. The same was true when we analyzed the data separately by grade span (elementary and middle school).

[105] The question asked principals to rate on a 4-point scale whether they strongly disagree, disagree, agree, or strongly agree with the following statements: (a) This teacher is demonstrating leadership skills with peers and other school staff; (b) This teacher is contributing to school activities outside of his/her own classroom, including leading student groups or assisting in after-school student activities; and (c) This teacher is an asset to the school.

Table E.3. Principal Ratings of Teacher Contributions

Description of Focal Teacher (percentage of teams whose principal agrees or agrees strongly)	Treatment	Control	Impact	*p*-Value
Demonstrates leadership skills with peers and other school staff	70.2	68.4	1.8	0.805
Contributes to school activities outside the classroom	63.1	69.7	-6.6	0.378
Is an asset to the school	80.7	76.3	4.4	0.501
Sample Size (teams)	83	76		

Source: Principal Survey.

Note: None of the differences are statistically significant at the 0.05 level, two-sided test.

APPENDIX F

SUPPLEMENTAL MATERIALS FOR CHAPTER V

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

In this appendix, we supplement the Chapter V summary of the test-score findings in several ways: with extra detail on the methods used to estimate the impacts on student achievement, imputation methods used to handle missing data, and additional district-specific impact figures. We also describe test-score-scaling issues in two districts and the results of further analyses of the sensitivity of the benchmark results. Finally, we present sample sizes corresponding to tables and figures in Chapter V.

A. Benchmark Model

1. Benchmark Regression Model

The test-score impacts presented in Chapter V were derived from the following ordinary least squares model:

$$(5.1).\ Y_{it} = \alpha + \beta * X_{it} + \gamma * Y_{i,t-1} + \delta * T_{it} + \theta * D_i + \eta * G_i + \rho * P_i + \lambda * I_i + \varepsilon_{it}$$

where Y_{it} and $Y_{i,t-1}$ are student *i*'s post-test and pre-test z-scores respectively. X_{it} is a vector that consists of the following student background covariates: race/ethnicity, gender, English-language learner (ELL) status, special education status (SPED), free or reduced-price lunch eligibility (FRL) status, over-age-for-grade status, and an indication of whether the student belonged to a study team that has at least one retention-stipend teacher. T_{it} is the treatment indicator that equals one when student *i* belonged to a treatment study team for academic year *t*. P_i is a vector of grade 3 pre-test interactions that refer to the set of interaction terms created by multiplying student *i*'s pre-test score by an indicator for whether the student was in grade 3 and by a set of district dummies. These dummies were included because some of the study districts administered a different test for 2nd-grade students, which we used as a pre-test for 3rd-grade study teams.

We also included a vector of block dummies (D_i) to account for the block random assignment design, a vector of grade dummies (G_i), and a vector of imputation dummies (I_i), one for each of the pre-test or student background covariates that were imputed. α is the constant term; ε_{it} is the error term. To account for the possibility that students within a team are more similar to each other than to students in other teams, we clustered standard errors at the team level. δ is the coefficient of interest; depending on the sample of students included in the regression, it represents the team, focal teacher, or nonfocal teacher impact of the Talent Transfer Initiative (TTI) on test scores. Because each student observation contributes equally to the regression, teams and teachers with more students are weighted more heavily in the regression.

2. Missing Data Imputation

When particular demographic information was unavailable for the whole district,[106] we set the value of that variable to an arbitrary constant, the mode of the study population with nonmissing values for that variable. When demographic information was missing for only some of the students in the district, we imputed (filled in) missing values with team-level means. In the

[106] Four of the 10 districts provided information on gifted status; 9 provided information on age.

case of missing pre-test scores,[107] we implemented a single stochastic regression imputation strategy, estimated separately by treatment status, one of the four imputation strategies recommended by Puma et al. (2009). That is, we filled in missing values with predicted values based on a set of best-available covariates. The covariates included in this regression are race, gender, ELL status, SPED status, FRL status, gifted status, over-age status, academic-year indicators, and district-by-grade interactions. For all demographic variables and pre-test scores, we created accompanying indicator variables to indicate the presence of imputed values. We did not impute post-test scores.

3. Cohort 1 Districts Only, Impacts in Program Year 1

In Chapter V, we showed impacts for all 10 districts in program year 1 but for just the 7 cohort 1 districts in the second program year because we did not follow cohort 2 districts into a second year. However, to allow readers to compare the same sample in both program years (the 7 cohort 1 districts), we present results analogous to Tables V.1 and V.2 limiting the sample to cohort 1 districts only, shown in Table F.1.

Table F.1. Test-Score Impacts in Cohort 1 Districts Only, Program Year 1

Program Year, Subject, and Comparison Type	Treatment Mean	Control Mean	Impact	Standard Error	*p*-Value	Sample Size[a]
Elementary School						
Math						
Team	-0.28	-0.33	0.05	0.04	0.235	7,513
Focal teacher	-0.26	-0.45	0.19*	0.05	0.000	3,472
Nonfocal teacher	-0.31	-0.24	-0.07	0.04	0.119	6,109
Reading						
Team	-0.36	-0.39	0.02	0.04	0.489	7,432
Focal teacher	-0.42	-0.52	0.10	0.06	0.087	3,528
Nonfocal teacher	-0.37	-0.36	-0.01	0.05	0.832	6,231
Middle School						
Math						
Team	-0.36	-0.45	0.08	0.05	0.097	2,786
Focal teacher	-0.34	-0.45	0.11	0.10	0.296	1,415
Nonfocal teacher	-0.40	-0.47	0.07	0.04	0.099	2,109
Reading						
Team	-0.51	-0.54	0.02	0.03	0.505	3,747
Focal teacher	-0.59	-0.57	-0.02	0.05	0.748	2,381
Nonfocal teacher	-0.54	-0.54	-0.00	0.04	0.922	3,000

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the focal and nonfocal teacher, sample size refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher to whom they are linked is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

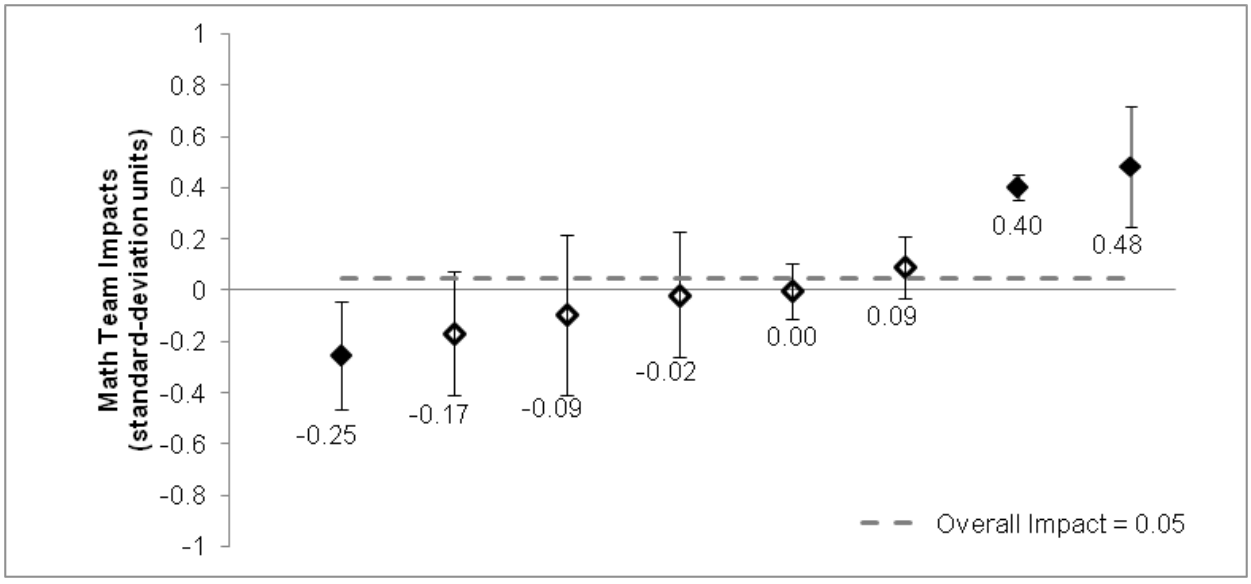
[107] In addition to students who were missing pre-test scores in the data, we also set a student's pre-test score to missing if the student had an irregular grade progression (that is, if the student repeated or skipped a grade) between the pre-test and post-test years.

4. Additional District-Specific Impact Figures

In Chapter V, we presented impact estimates for elementary school math by district for focal comparisons in program years 1 and 2. For completeness, we include additional district-level impact figures for elementary school math team and nonfocal comparisons, elementary school reading team, focal, and nonfocal comparisons, and middle school team, focal, and nonfocal comparisons in both subjects. In Figures F.1–F.4, we show the district-level elementary school math team and nonfocal comparisons; in Figures F.5–F.10, we show the district-level elementary school reading impact estimates for the team, focal, and nonfocal comparisons. In Figures F.11-F.16, we show the district-level middle school math team, focal, and nonfocal comparisons. In Figures F.17–F.22, we present the district-level middle school reading team, focal, and nonfocal comparisons. The dark, horizontal line on each graph represents the full-sample impact estimate—solid if statistically significant, dashed otherwise.

5. Math Impacts in Elementary Schools

Figure F.1. Year 1 Impacts on Math Scores, Elementary Teams, by District (cohorts 1 and 2)

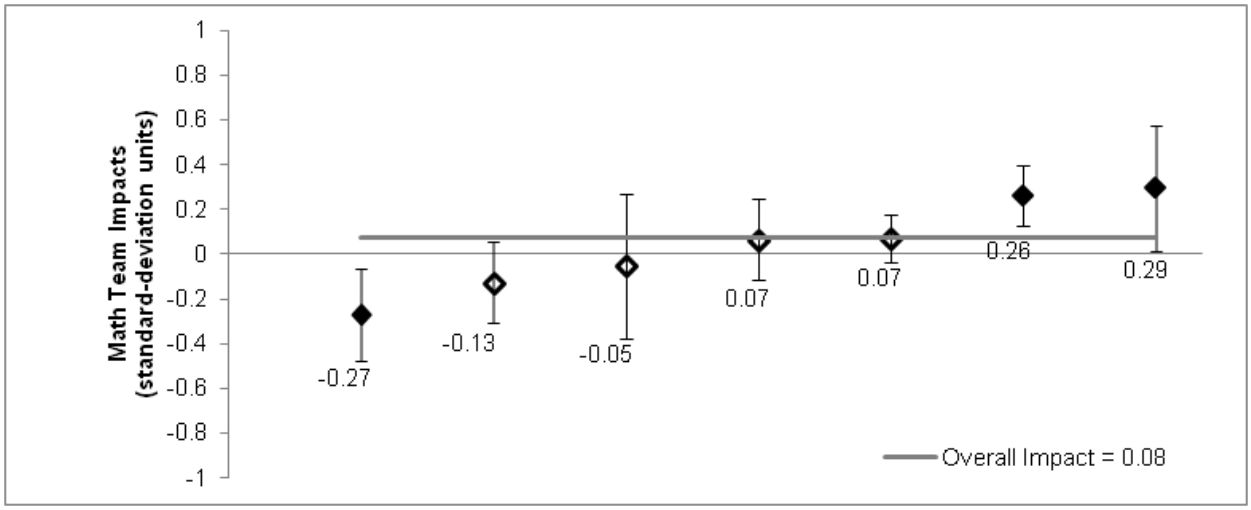


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

Two study districts do not have any elementary school teams. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of student-team combinations in each district range from 332 to 3,167. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.2. Year 2 Impacts on Math Scores, Elementary Teams, by District (cohort 1)

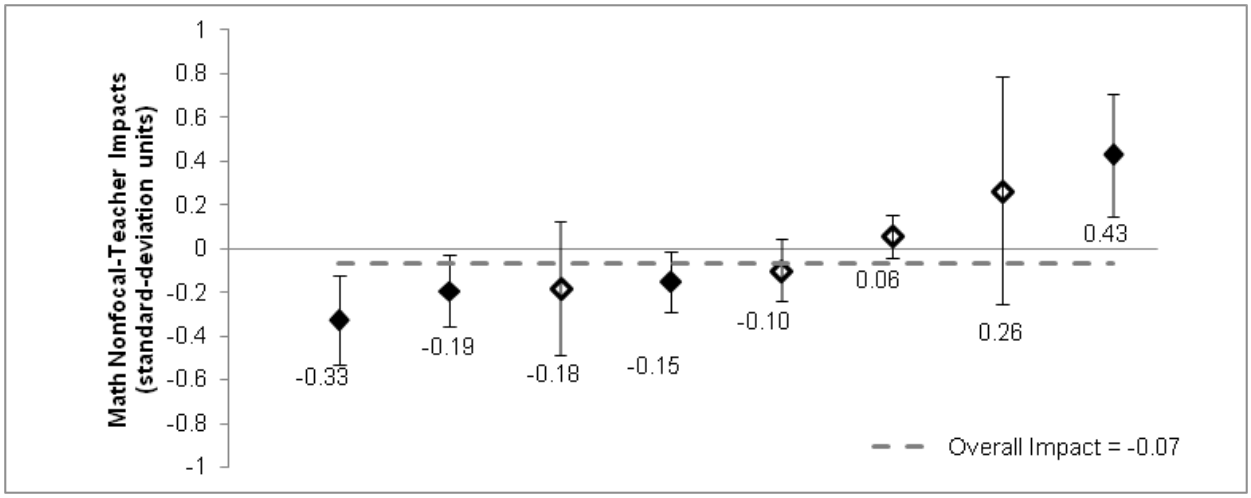


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a solid horizontal line denotes a statistically significant overall impact. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of student-team combinations in each district range from 341 to 3,353. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.3. Year 1 Impacts on Math Scores, Elementary Nonfocal Teachers, by District (cohorts 1 and 2)

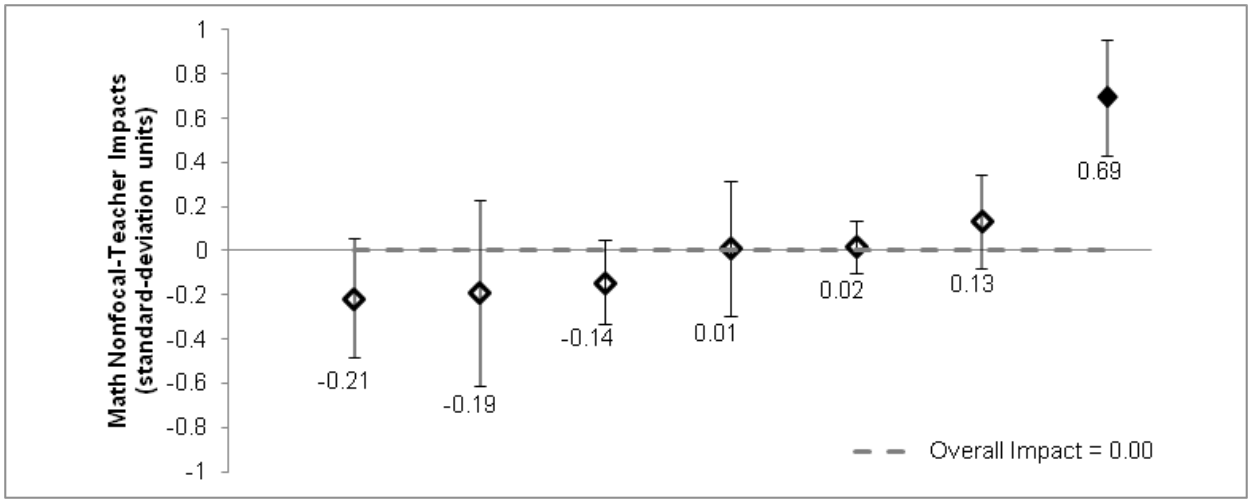


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. A solid horizontal line denotes a statistically significant overall impact. A dashed line denotes a statistically insignificant overall impact. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals

Two study districts have no elementary school team. Sample sizes of unique student-teacher combinations in each district range from 292 to 2,977. We do not report sample sizes for specific data points in the figure to avoid linking results to specific districts.

Figure F.4. Year 2 Impacts on Math Scores, Elementary Nonfocal Teachers, by District (cohort 1)



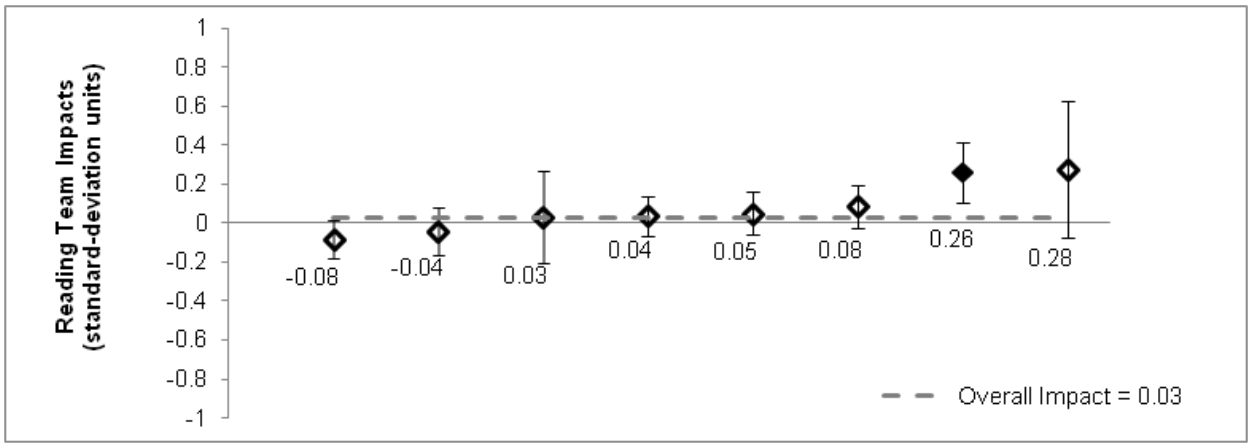
Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was not rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 287 to 3,135. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

6. Reading Impacts in Elementary Schools

Figure F.5. Year 1 Impacts on Reading Scores, Elementary Teams, by District (cohorts 1 and 2)

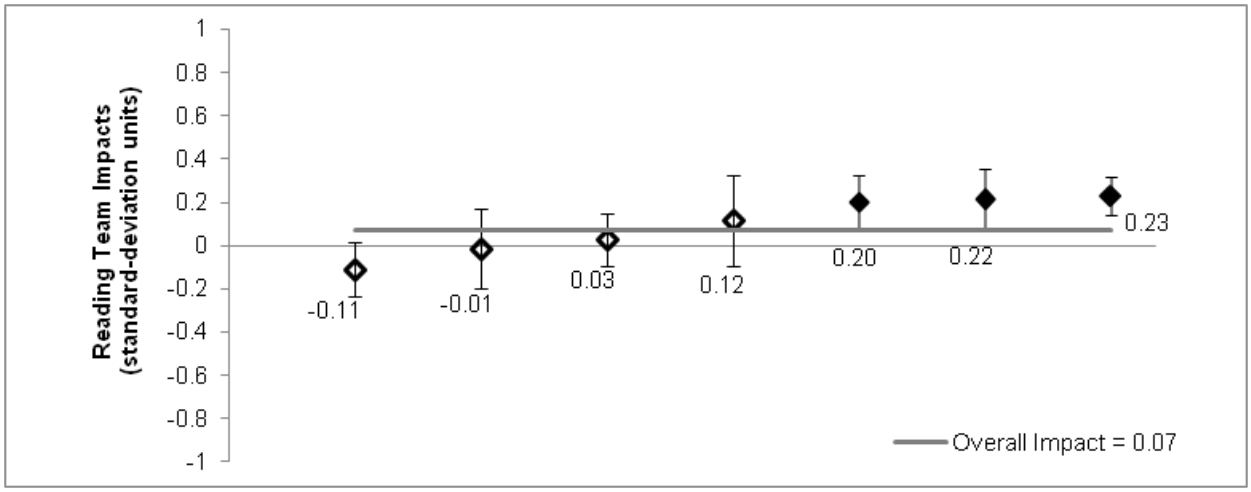


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

Two study districts do not have elementary school teams. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 332 to 3,114. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.6. Year 2 Impacts on Reading Scores, Elementary Teams, by District (cohort 1)

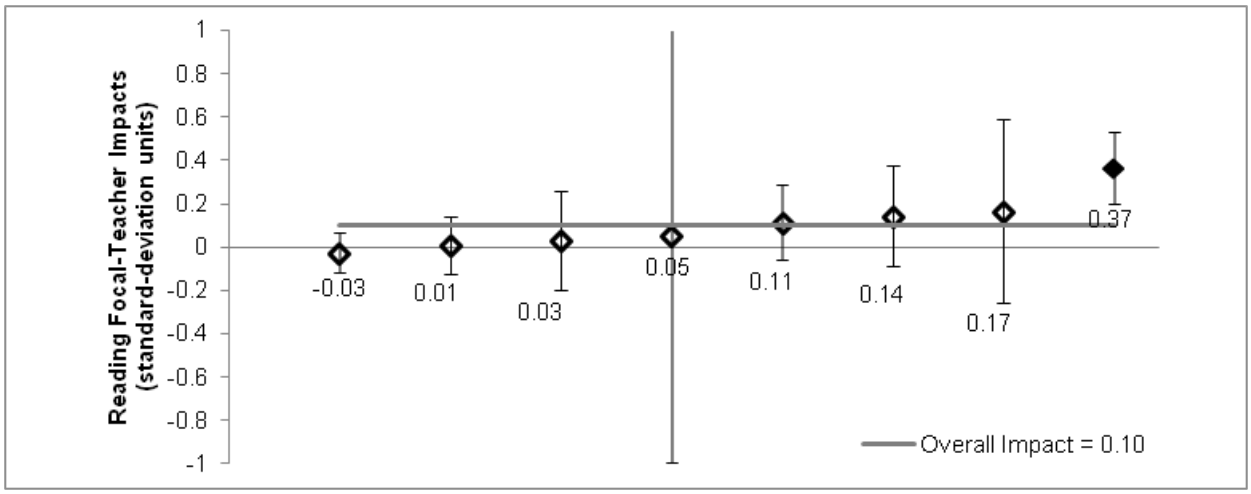


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a solid horizontal line denotes a statistically significant overall impact. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of student-team combinations in each district range from 340 to 3,304. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.7. Year 1 Impacts on Reading Scores, Elementary Focal Teachers, by District (cohorts 1 and 2)

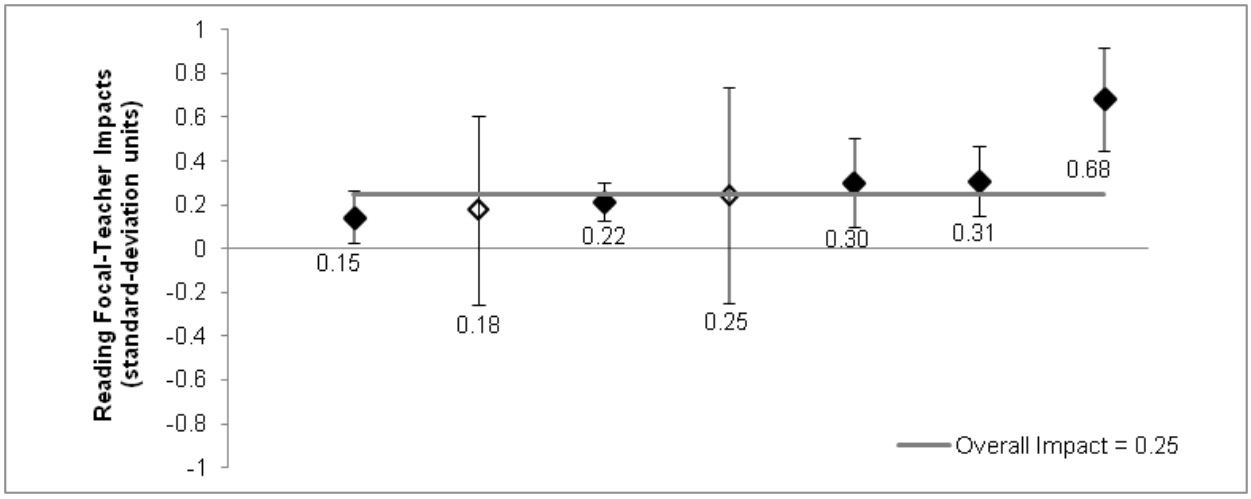


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a solid horizontal line denotes a statistically significant overall impact. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

Two study districts have no elementary school team. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 76 to 1,511. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.8. Year 2 Impacts on Reading Scores, Elementary Focal Teachers, by District (cohort 1)

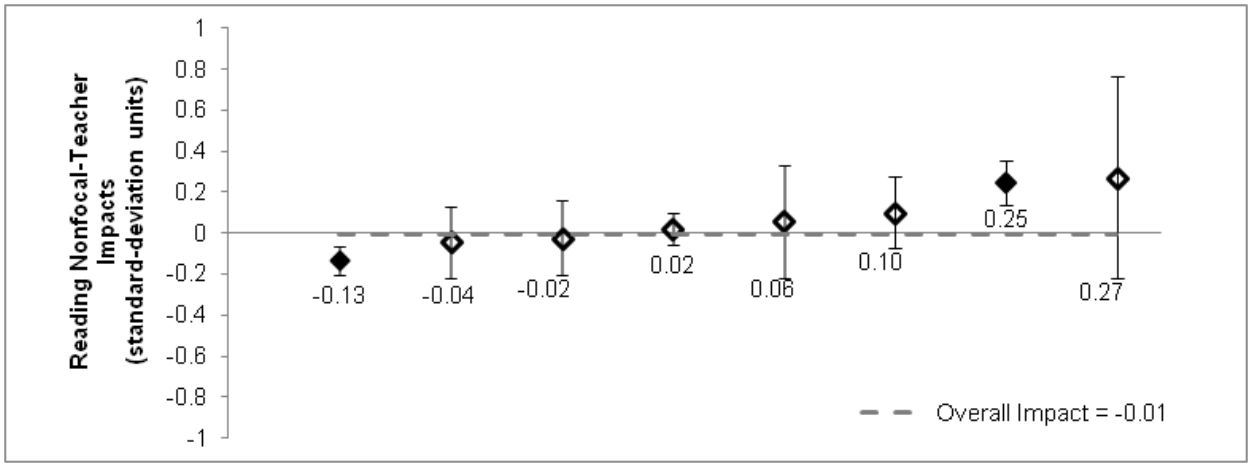


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a solid horizontal line denotes a statistically significant overall impact. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was not rejected at the 5 percent level. Sample sizes of student-teacher combinations in each district range from 91 to 1,378. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.9. Year 1 Impacts on Reading Scores, Elementary Nonfocal Teachers, by District (cohorts 1 and 2)

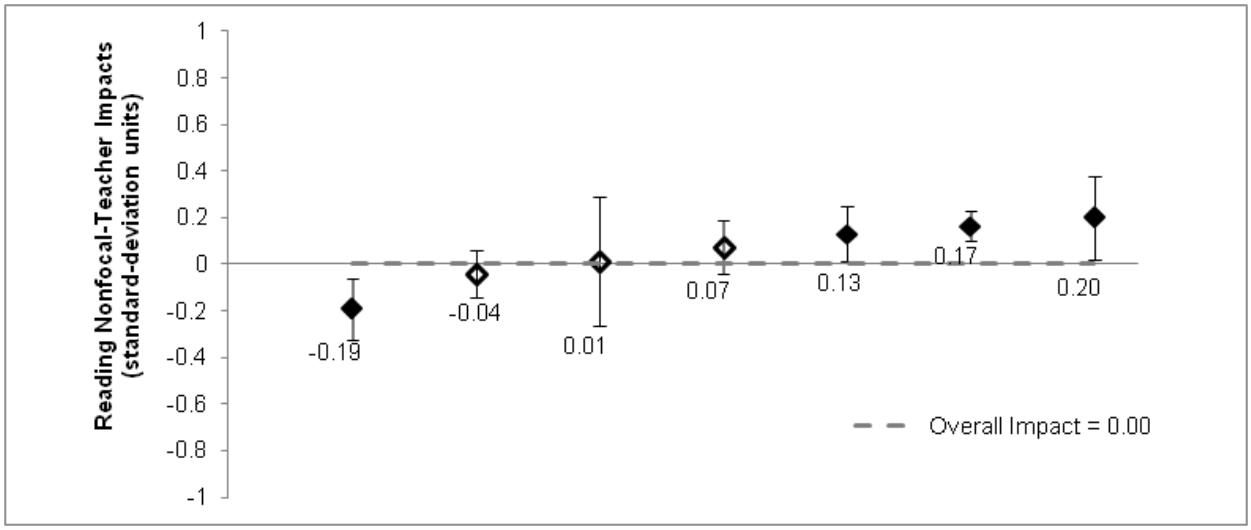


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

Two study districts have no elementary school team. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 289 to 3,063. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.10. Year 2 Impacts on Reading Scores, Elementary Nonfocal Teachers, by District (cohort 1)



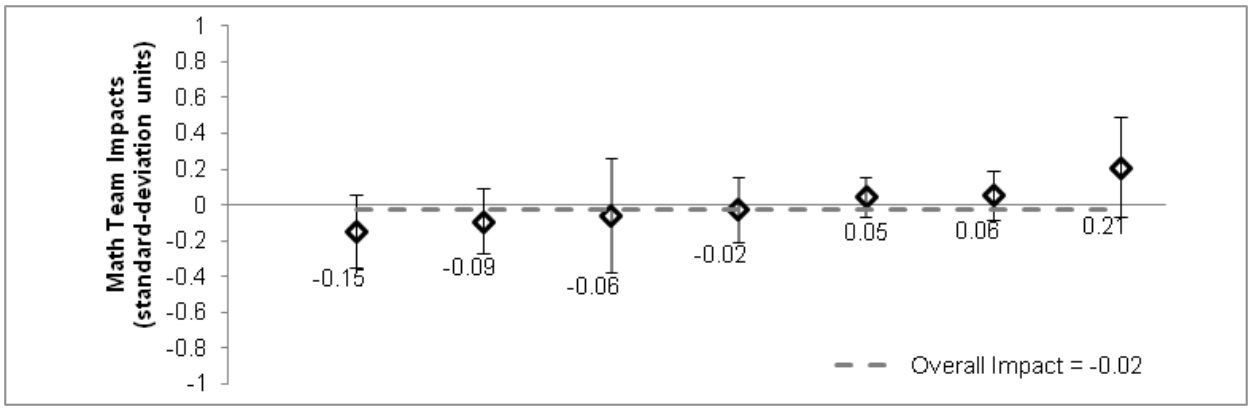
Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 287 to 3,925. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

7. Math Impacts in Middle Schools

Figure F.11. Year 1 Impacts on Math Scores, Middle Teams, by District (cohorts 1 and 2)

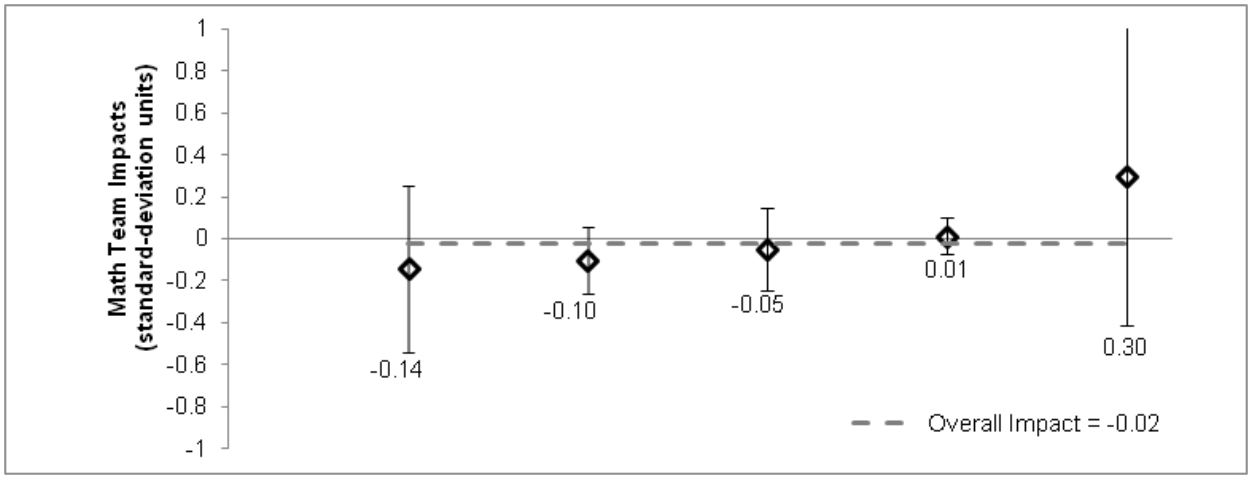


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

Two study districts do not have a middle school math team; another has no middle school team. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-team combinations in each district range from 336 to 5,010. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.12. Year 2 Impacts on Math Scores, Middle Teams, by District (cohort 1)

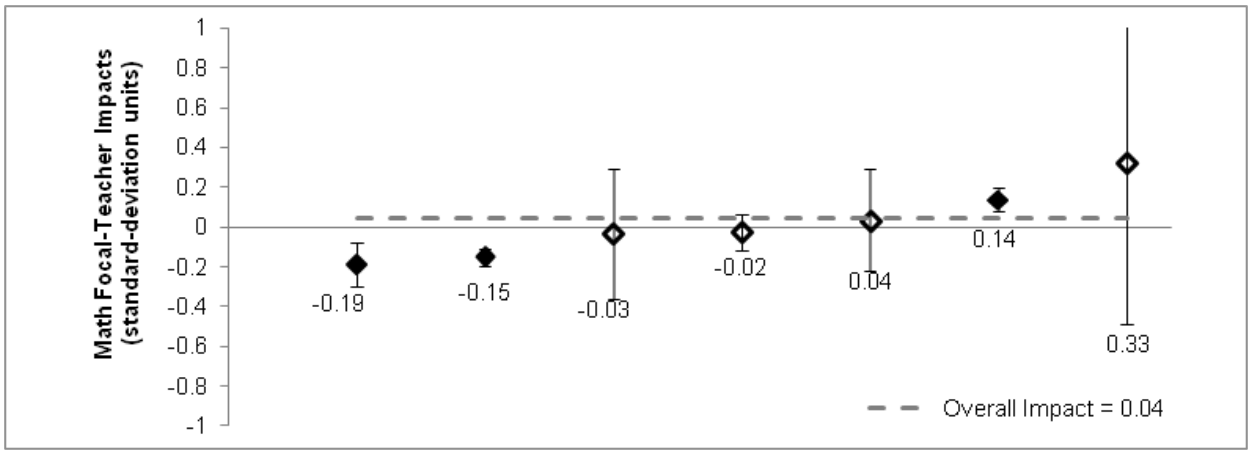


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

Two study districts have no middle school math team; another has no middle school team. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-team combinations in each district range from 320 to 1,059. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.13. Year 1 Impacts on Math Scores, Middle Focal Teachers, by District (cohorts 1 and 2)

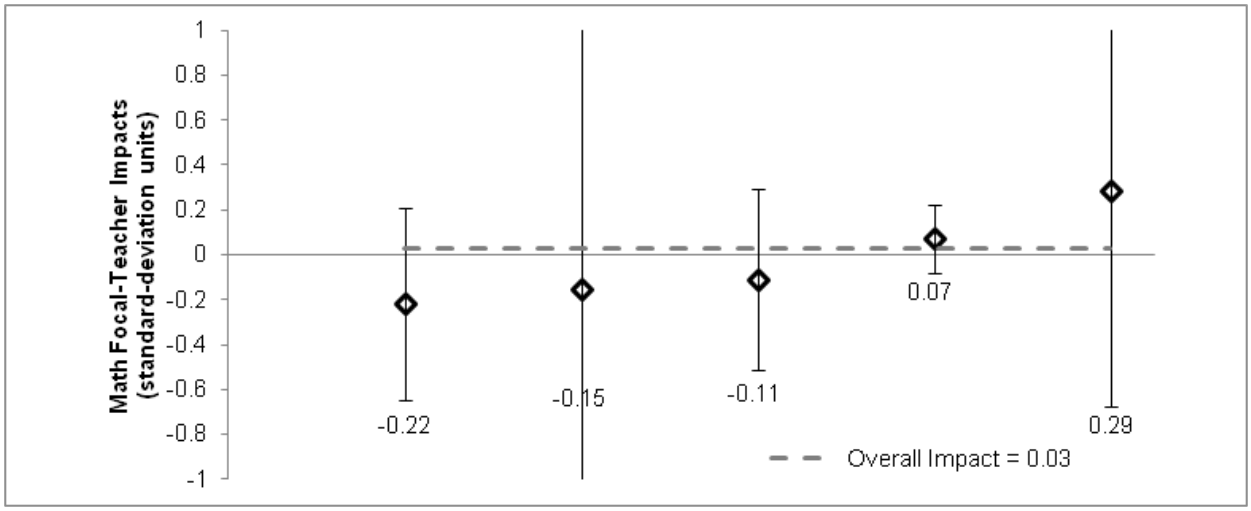


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

Two study districts have no middle school math team; another has no middle school team. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 99 to 981. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.14. Year 2 Impacts on Math Scores, Middle Focal Teachers, by District (cohort 1)

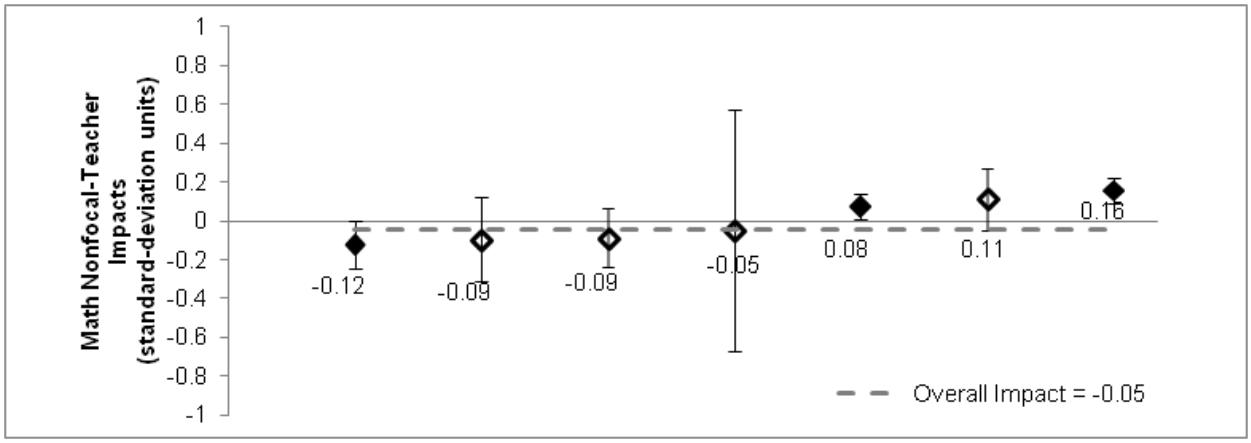


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

Two study districts have no middle school math team; another has no middle school team. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 241 to 498. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.15. Year 1 Impacts on Math Scores, Middle Nonfocal Teachers, by District (cohorts 1 and 2)

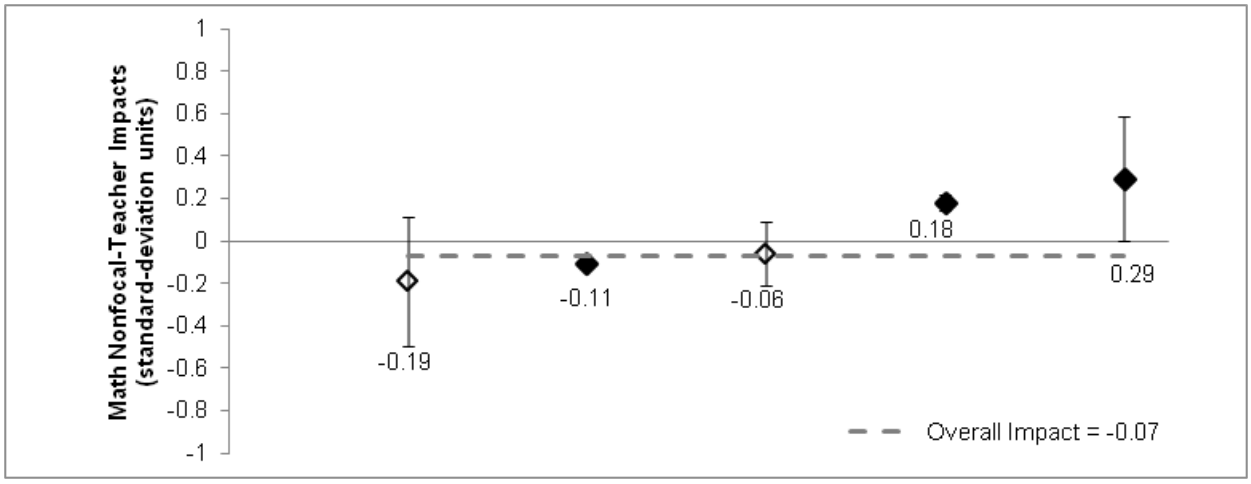


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

Two study districts have no middle school math team; another has no middle school team. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 234 to 5,341. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.16. Year 2 Impacts on Math Scores, Middle Nonfocal Teachers, by District (cohort 1)



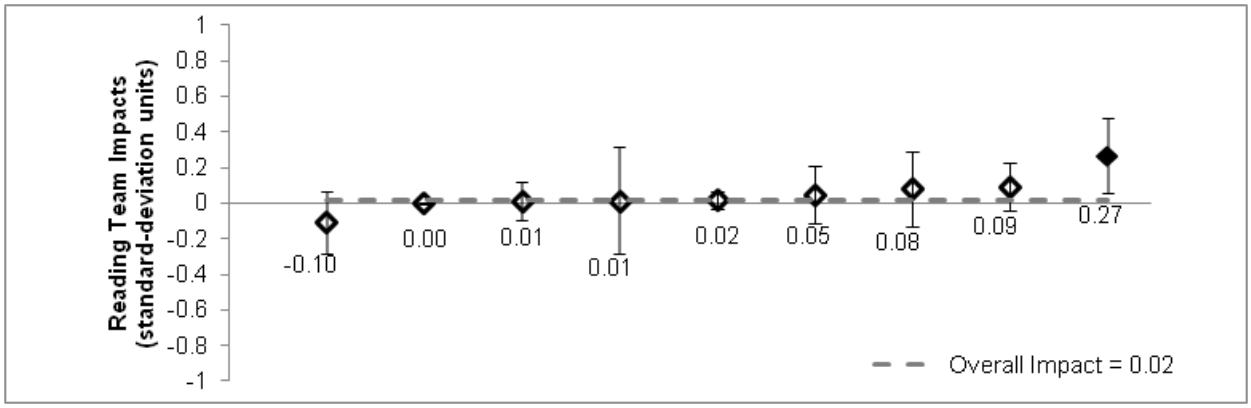
Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

Two study districts have no middle school math team; another has no middle school team. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 222 to 718. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

8. Reading Impacts in Middle Schools

Figure F.17. Year 1 Impacts on Reading Scores, Middle Teams, by District (cohorts 1 and 2)

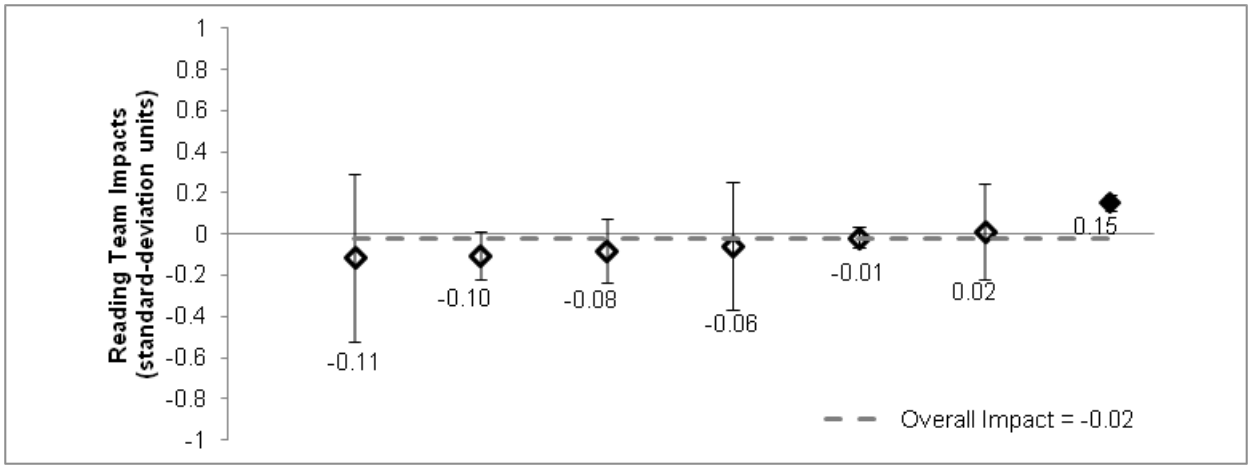


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

One study district does not have middle school teams. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-team combinations in each district range from 272 to 3,236. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.18. Year 2 Impacts on Reading Scores, Middle Teams, by District (cohort 1)

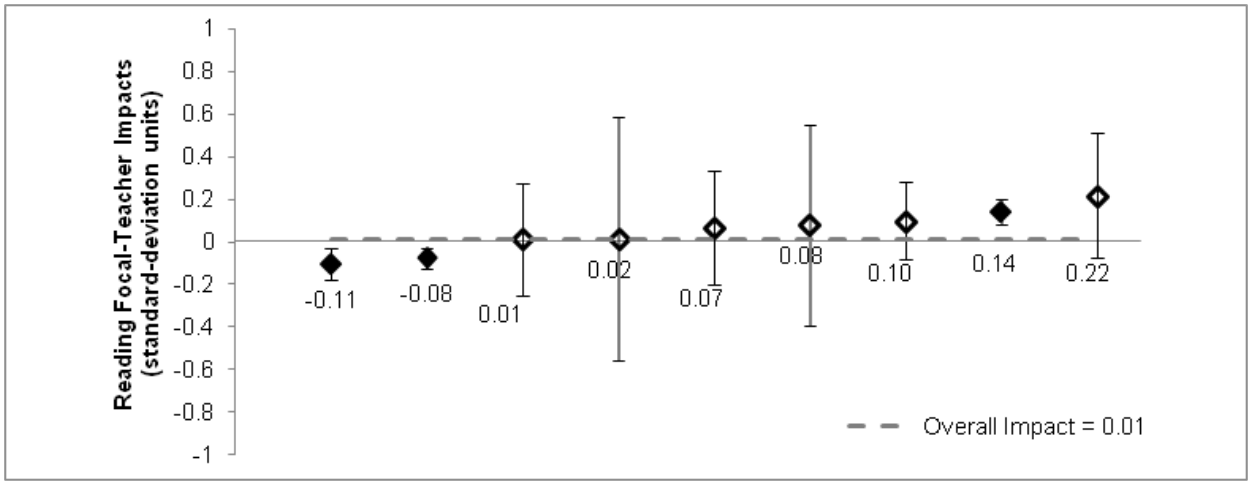


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

One study district does not have middle school teams. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-team combinations in each district range from 227 to 932. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.19. Year 1 Impacts on Reading Scores, Middle Focal Teachers, by District (cohorts 1 and 2)

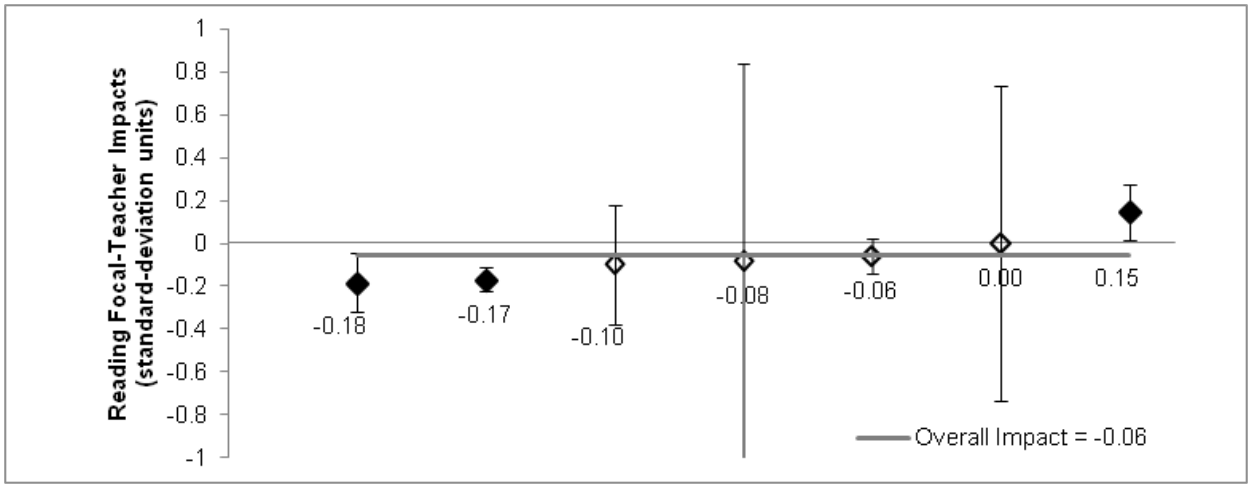


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

One study district does not have middle school teams. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 104 to 637. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.20. Year 2 Impacts on Reading Scores, Middle Focal Teachers, by District (cohort 1)

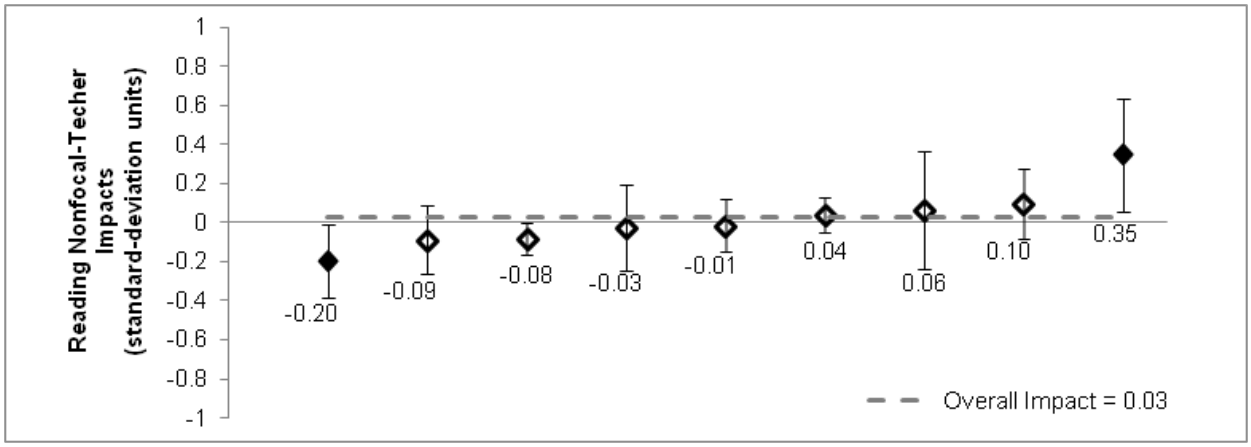


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a solid horizontal line denotes a statistically significant overall impact. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

One study district does not have middle school teams. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of student-teacher combinations in each district range from 156 to 672. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.21. Year 1 Impacts on Reading Scores, Middle Nonfocal Teachers, by District (cohorts 1 and 2)

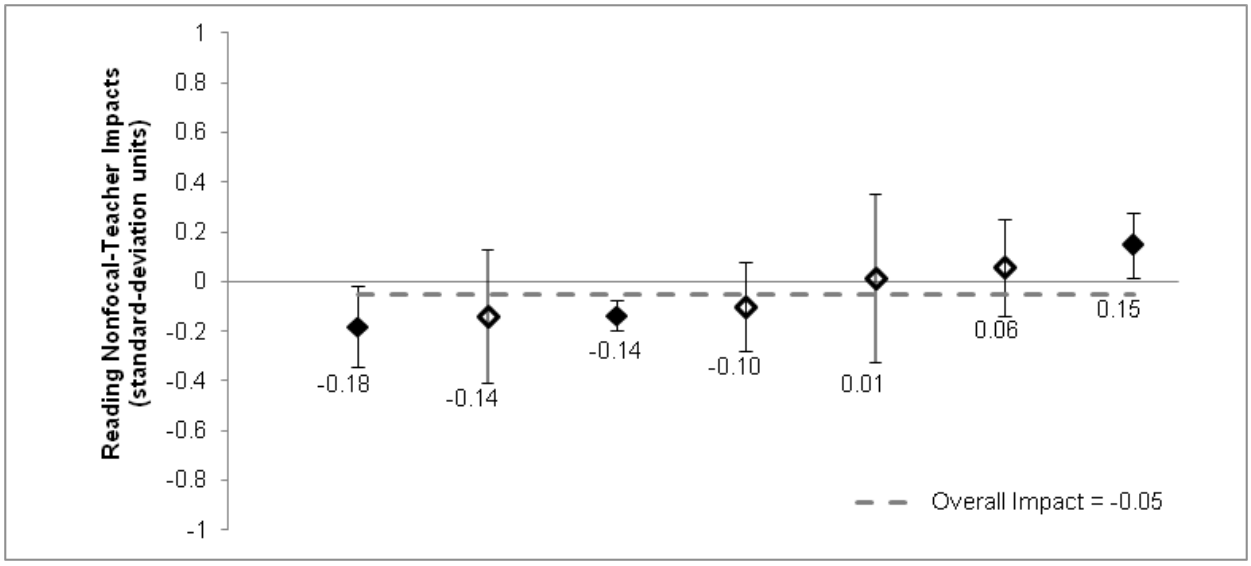


Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

One study district does not have middle school teams. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 239 to 3,222. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure F.22. Year 2 Impacts on Reading Scores, Middle Nonfocal Teachers, by District (cohort 1)



Source: District administrative data.

Notes: The horizontal line denotes the size of the overall impact. Here, a dashed horizontal line denotes an overall impact that is statistically indistinguishable from zero. Each diamond marker represents an impact estimate from one district. A black diamond marker represents an impact that is statistically significant at the 0.05 level (two-tailed test). A hollow diamond represents an impact that is statistically indistinguishable from zero. Bars denote 95 percent confidence intervals.

One study district does not have middle school teams. The null hypothesis of the F-test conducted to determine whether the district-level impacts were jointly equal to one another was rejected at the 5 percent level. Sample sizes of unique student-teacher combinations in each district range from 216 to 976. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

9. Impacts by Grade

In Chapter V, we showed that TTI appears to be more effective in elementary schools than in middle schools, although this observation may be confounded by district and cohort effects. Here, we present test-score comparisons by grade using teams (Table F.2), focal teachers (Table F.3), and nonfocal teachers (Table F.4). For all three comparisons, test-score impacts in both math and reading did not have a monotonic (steadily increasing or steadily decreasing) relationship with grade level.

Table F.2. Team-Level Test-Score Comparisons, by Grade

	Math				Reading			
Program Year and Grade	Impact	Standard Error	*p*-Value	Sample Size[a]	Impact	Standard Error	*p*-Value	Sample Size[a]
Year 1 (all districts)								
Grade 3	-0.04	0.06	0.540	2,932	0.02	0.06	0.744	2,896
Grade 4	0.19*	0.07	0.011	2,571	0.08	0.06	0.200	2,547
Grade 5	0.05	0.04	0.261	2,565	0.01	0.02	0.523	2,548
Grade 6	0.09	0.10	0.384	3,987	-0.06	0.03	0.121	1,508
Grade 7	0.07*	0.01	0.002	736	0.02	0.03	0.571	4,986
Grade 8	-0.06	0.07	0.406	4,261	0.13*	0.04	0.016	1,424
Year 2 (cohort 1 only)								
Grade 3	0.11	0.08	0.217	2,561	0.06	0.06	0.303	2,533
Grade 4	0.14*	0.06	0.036	2,394	0.17*	0.04	0.000	2,365
Grade 5	0.03	0.04	0.477	2,610	0.03	0.03	0.201	2,583
Grade 6	-0.03	0.02	0.271	600	-0.03	0.04	0.403	1,328
Grade 7	0.30*	0.03	0.003	729	0.02	0.05	0.618	771
Grade 8	-0.11*	0.03	0.025	1,298	-0.07*	0.02	0.011	1,389

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]Sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.3. Focal-Teacher Test-Score Comparisons, by Grade

	Math				Reading			
Program Year and Grade	Impact	Standard Error	*p*-Value	Sample Size[a]	Impact	Standard Error	*p*-Value	Sample Size[a]
Year 1 (all districts)								
Grade 3	0.02	0.07	0.728	1,242	-0.03	0.09	0.763	1,258
Grade 4	0.35*	0.10	0.001	1,193	0.23*	0.08	0.008	1,223
Grade 5	0.19*	0.07	0.013	1,245	0.09	0.05	0.077	1,254
Grade 6	0.42*	0.16	0.020	1,325	-0.09	0.04	0.059	824
Grade 7	-0.12*	0.02	0.006	439	0.01	0.09	0.882	1,550
Grade 8	-0.05	0.10	0.646	1,134	0.16*	0.05	0.020	956
Year 2 (cohort 1 only)								
Grade 3	0.29*	0.12	0.022	909	0.33*	0.08	0.000	848
Grade 4	0.26*	0.10	0.019	1,259	0.27*	0.08	0.003	1,155
Grade 5	0.08	0.08	0.313	1,159	0.12*	0.04	0.005	1,198
Grade 6	0.17*	0.02	0.009	264	-0.06	0.04	0.175	604
Grade 7	0.29*	0.06	0.020	529	-0.06	0.05	0.310	575
Grade 8	-0.12*	0.02	0.002	782	-0.05	0.03	0.156	911

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]The number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in the focal-teacher analyses because they can be linked to more than one teacher. Students in the focal-teacher analyses are weighted proportionately to the probability that the teacher they are linked to is the focal teacher. Due to the uncertainty involved in focal-teacher identification, some teachers and their students were included in both the focal and nonfocal samples.

Table F.4. Nonfocal-Teacher-Level Test-Score Comparisons, by Grade

	Math				Reading			
Program Year and Grade	Impact	Standard Error	*p*-Value	Sample Size[a]	Impact	Standard Error	*p*-Value	Sample Size[a]
Year 1 (all districts)								
Grade 3	-0.13	0.08	0.094	2,410	0.01	0.07	0.865	2,413
Grade 4	0.04	0.08	0.633	2,239	0.00	0.10	0.991	2,279
Grade 5	-0.05	0.03	0.132	1,829	-0.01	0.04	0.871	1,913
Grade 6	-0.02	0.08	0.818	3,803	-0.12	0.06	0.081	1,019
Grade 7	0.16*	0.01	0.001	577	0.05	0.03	0.152	4,994
Grade 8	-0.07	0.07	0.336	4,207	0.04	0.04	0.444	1,248
Year 2 (cohort 1 only)								
Grade 3	-0.02	0.10	0.816	2,224	-0.10	0.06	0.106	2,510
Grade 4	0.07	0.07	0.325	2,076	0.12*	0.03	0.000	2,367
Grade 5	0.02	0.05	0.678	1,940	0.01	0.04	0.770	2,147
Grade 6	-0.11*	0.01	0.008	486	-0.06	0.05	0.278	687
Grade 7	0.32*	0.02	0.001	455	-0.01	0.07	0.913	761
Grade 8	-0.16*	0.04	0.012	847	-0.11*	0.01	0.000	1,450

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]The number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in the nonfocal-teacher analyses because they can be linked to more than one teacher. Students in the nonfocal-teacher analyses are weighted proportionately to the probability that the teacher they are linked to is the nonfocal teacher. Due to the uncertainty involved in focal-teacher identification, some teachers and their students were included in both the focal and nonfocal samples.

B. Test-Score-Scaling Issues

A methodological challenge for this study was the fact that students in two of the study districts could choose between different courses and were subsequently assigned different end-of-year tests in the same grade. Because our study design relies on treatment-control comparisons within grade and within district, the tests within these groups had to be placed on a common scale so we would have comparable outcome measures. For example, if the treatment-group members took a more challenging exam than their control-group counterparts and the degree of difficulty was not captured in the test scaling, estimates of the impact of TTI would be biased downward.

Each of the two districts had a different problem: one was related to middle school math difficulty and the other to bilingual instruction. Middle school students in the first district took different math tests, depending on which math course they completed. These were end-of-course tests rather than end-of-grade tests. Students in the other district could take either an English or Spanish version of the state assessment, depending on their English language skills. Although the English and Spanish versions of the assessment were designed to be on the same scale, we adjusted these scores to account for the different populations of students taking each test.

1. Linking End-of-Course Math Tests

As noted above, middle school students in one study district completed different math assessments depending on which math course they took during the school year. The assessments included a 7th-grade math assessment for students enrolled in the grade-level math course, a general math test for students taking a pre-algebra course, and algebra I and geometry tests for students enrolled in algebra or geometry. The most common course progressions in the study district are (1) 7th-grade math followed by a pre-algebra course in 8th grade or (2) 7th-grade math followed by an algebra I course in 8th grade. A less common progression is taking algebra I in 7th grade followed by a geometry course in 8th grade.

The study sample for the district includes two grades—grade 6, in which all students take a grade-level math test, and grade 8. Because fewer than 2 percent of 8th-grade students in the study sample took algebra I in 7th grade and geometry in 8th grade, we excluded those students from the analysis. Of the remaining students, the 8th graders in study teams overwhelmingly took algebra I: 92 percent took the algebra I assessment and 8 percent took the general math assessment (Table F.5). However, there were statistically significant differences in test-taking across treatment and control groups: students on treatment teams were more likely to take the algebra I test than students on the control teams.

Table F.5. Percentage of 8th-Grade Students on Study Teams Taking General Math and Algebra I Post-Tests, by Treatment Status, Program Year 1

	Students on Treatment Teams	Students on Control Teams	All Study Students
General math post-test	2	22	8
Algebra I post-test	98	78	92

Notes: N = 3,482 students

Differences between treatment and control in the percentage of students taking the general math post-test are significant at the .05 level.

Although 8th-grade students in this district took different tests, which are on different scales, we needed a common measure of student achievement to estimate the impact of TTI. Ideally, we would have information equating scores on the algebra I test and the general math test, making it possible to translate scores from one test into scores on the other. However, technical reports on the state assessment that are publicly available offer no information about efforts to equate the two state tests (for example, by having the same students take both tests). The method used by the state to compare proficiency rates on the two different exams equates "proficient" scores on algebra I with "advanced" on pre-algebra, "basic" on algebra I with "proficient" on pre-algebra, and "below basic" with "basic" on pre-algebra. The linking function implied by this rule created an implausible and unusable mapping for the purposes of this study.

As a result, we used a two-step process to link scores on the pre-algebra test and algebra I test. First, we estimated the relationship between algebra I scores and student characteristics for the population of all students in the district who took the algebra I test to impute algebra I scores for all 8th-grade students in the district taking the pre-algebra test. The background characteristics we used to predict algebra I scores were prior math achievement, including 6th- and 7th-grade math test scores and 7th-grade math course letter grades, 7th-grade reading test

scores, race and ethnicity, FRL status, an indicator of ELL status, SPED status, and education level of the student's parent. We also included school fixed effects. We use this same approach to estimate the relationship between pre-algebra scores and student characteristics for the student population taking the pre-algebra test, to impute a pre-algebra score for all 8th-grade students taking the algebra I exam. The student background characteristics and school fixed effects explain 66 percent of the variation in algebra I scores and 65 percent of the variation in pre-algebra scores.

Second, we estimated the relationship between algebra I scores and pre-algebra scores by regressing the actual and imputed scores from one test on the other. The resulting prediction equation was then used to predict algebra I scores for students who took the pre-algebra test.

We used the linked scores for the impact analysis, but checked the sensitivity of the middle school results by excluding the 8th-grade students from this district. The middle school comparisons at the team, focal-teacher, and nonfocal-teacher levels were larger when excluding these students from the analysis, although they remain statistically insignificant.

Table F.6. Test-Score Impacts in Middle Schools, Omitting 8th-Grade Math Students from One District, Program Year 1

	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Math						
Team	0.03	0.026	14,198	537	128	0.00
Focal teacher	0.13*	0.042	6,089	190	128	0.10*
Nonfocal teacher	-0.04	0.028	11,877	379	119	-0.06*

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

2. English and Spanish Versions of State Assessment

In another study district, 3rd- and 4th-grade students could take either an English or Spanish version of the state assessment in reading or math. The Spanish version includes items translated directly from the English version as well as items that are created specifically for the Spanish version of the test. The two versions of the assessment test the same objectives and have the same number of questions. Statewide, 6 to 10 percent of 3rd and 4th-grade students took the Spanish version of the assessment. Students who took the Spanish version of the assessment differed from the full population of students. For example, 96 percent of 4th-grade students taking the Spanish version of the assessment statewide were economically disadvantaged, compared with 60 percent of students taking the English version.

Forty percent of 3rd- and 4th-grade students in the study district took the Spanish version of the assessment. There were significant differences in the percentage of students who took the Spanish version across treatment and control groups: 44 percent of treatment students and 37 percent of control students in the math sample took the Spanish version, and 37 percent of treatment students and 45 percent of control students in the reading sample did. That amounted to differences of 7 and 8 percentage points, respectively. Corresponding differences in percentages of ELL students in that district were 8 percentage points for math and 3 percentage points for reading, although the differences in percentages of ELL were not statistically significant.

Although student test scores for the English and Spanish versions of the assessment are reported to be on the same scale, we used a revised approach to standardizing the scores. The statewide test score means and standard deviations are reported separately for the English and Spanish versions of the assessment. The statewide means for the Spanish version of the assessment were lower than the English version, possibly because students taking the Spanish version of the assessment were more likely to be disadvantaged. In addition, the statewide standard deviations for the Spanish version of the assessment were lower than for the English version because students taking the Spanish version represent a less diverse (lower variance) subset of the full population of students. Standardizing student test scores using the lower means and standard deviations on the Spanish assessment could inflate the scores for students taking the Spanish version compared with students taking the English version. As a result, we used the means and standard deviations for the English version of the assessment to standardize the Spanish version of the assessment.

We tested the sensitivity of our results to this decision by adding a control to the benchmark model to account for whether students in this district took the English or Spanish version of the assessment (Tables F.7 and F.8). When including this additional control for the year 1 analysis, the team-level impacts in reading and math changed by less than 0.01 and remained nonsignificant, the focal teacher impacts changed by less than 0.01 of a standard deviation and remained significant. The negative nonfocal teacher impact in math was no longer significant. In year 2, the team-level impacts changed by less than 0.01, although the team-level reading impact became nonsignificant. The significance of the focal and nonfocal teacher impacts did not change in year 2 after adding the control for students taking the Spanish version of the test in this district.

Table F.7. Test-Score Impacts in Elementary Schools, Adjustment for Taking Test in Spanish in One District

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.01	0.030	17,052	583	134	0.00
Focal teacher	0.09*	0.043	6,578	196	134	0.10*
Nonfocal teacher	-0.06	0.031	15,065	418	125	-0.06*
Reading						
Team	0.03	0.020	15,909	622	135	0.03
Focal teacher	0.07*	0.033	7,065	210	135	0.07
Nonfocal teacher	0.00	0.023	13,866	451	129	0.02
Year 2 (cohort 1 only)						
Math						
Team	0.04	0.031	10,192	408	103	0.05
Focal teacher	0.11*	0.050	4,902	152	103	0.12*
Nonfocal teacher	0.00	0.035	8,028	301	96	-0.01
Reading						
Team	0.04	0.022	10,969	455	110	0.05*
Focal teacher	0.11*	0.031	5,291	158	110	0.13*
Nonfocal teacher	-0.03	0.025	9,922	336	108	-0.01

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.8. Test-Score Impacts in Middle Schools, Adjustment for Taking Test in Spanish in One District

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	-0.02	0.056	8,875	169	30	-0.02
Focal teacher	0.04	0.086	2,827	38	30	0.04
Nonfocal teacher	-0.05	0.053	8,549	136	30	-0.05
Reading						
Team	0.02	0.027	7,812	195	31	0.02
Focal teacher	0.01	0.052	3,261	49	31	0.01
Nonfocal teacher	0.03	0.028	7,224	154	31	0.03
Year 2 (cohort 1 only)						
Math						
Team	-0.02	0.057	2,627	42	13	-0.02
Focal teacher	0.03	0.060	1,575	19	13	0.03
Nonfocal teacher	-0.07	0.069	1,788	31	13	-0.07
Reading						
Team	-0.02	0.022	3,488	72	20	-0.02
Focal teacher	-0.06*	0.024	2,090	25	20	-0.06*
Nonfocal teacher	-0.05	0.028	2,898	50	20	-0.05

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

3. Multiple Study Teams Within a School

One may be concerned that TTI may have a downward-biased impact on schools with multiple study teams, where those teams had been assigned to both treatment and control status, than on schools with only one study team or a single treatment status. This concern about downward bias would be based on fears of contamination.

To explore these concerns, we divided our elementary and middle school analysis samples into blocks that contain only schools with a single treatment status (either treatment or control) and blocks that contain schools with one or two treatment statuses. We present the results of these subgroup analyses in Tables F.9–F.12. The results suggest that our full-sample estimates may indeed understate the true impacts. When we restrict our sample to blocks that contain only schools with a single treatment status, team impacts generally become more positive as compared with full sample results. On the other hand, impacts are more negative when we restrict our sample to blocks containing schools with one or two treatment statuses. We obtained

similar results when we compared blocks containing only schools with one study team to blocks containing schools with one or more study teams.

Table F.9. Test-Score Impacts in Elementary Schools, Benchmark Model, Blocks Containing Schools with Single-Treatment Status Only

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.11*	0.045	5,822	289	70	0.05
Focal teacher	0.28*	0.071	2,624	108	70	0.18*
Nonfocal teacher	-0.02	0.050	4,807	203	64	-0.07
Reading						
Team	0.08*	0.040	5,756	302	70	0.03
Focal teacher	0.19*	0.072	2,658	109	70	0.10*
Nonfocal teacher	0.03	0.054	5,046	216	67	-0.01
Year 2 (cohort 1 only)						
Math						
Team	0.10*	0.042	5,422	262	64	0.08*
Focal teacher	0.20*	0.078	2,539	97	64	0.22*
Nonfocal teacher	0.05	0.050	4,456	192	57	0.00
Reading						
Team	0.12*	0.033	5,352	277	64	0.07*
Focal teacher	0.24*	0.068	2,309	96	64	0.25*
Nonfocal teacher	0.06	0.035	5,117	205	62	0.00

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.10. Test-Score Impacts in Middle Schools, Benchmark Model, Blocks Containing Schools with Single-Treatment Status Only

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.07*	0.033	4,683	84	16	-0.02
Focal teacher	0.08	0.072	1,376	17	16	0.04
Nonfocal teacher	0.07*	0.024	4,301	66	16	-0.05
Reading						
Team	0.06	0.034	3,358	91	18	0.02
Focal teacher	0.05	0.054	1,908	33	18	0.01
Nonfocal teacher	0.05	0.048	2,792	66	18	0.03
Year 2 (cohort 1 only)						
Math						
Team	-0.02	0.067	1,831	28	9	-0.02
Focal teacher	0.00	0.063	1,099	14	9	0.03
Nonfocal teacher	-0.05	0.093	1,228	20	9	-0.07
Reading						
Team	0.01	0.018	2,764	56	16	-0.02
Focal teacher	-0.04	0.028	1,688	20	16	-0.06*
Nonfocal teacher	-0.02	0.031	2,424	39	16	-0.05

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.11. Test-Score Impacts in Elementary Schools, Benchmark Model, Blocks Containing Schools with Both Treatment and Control Teams

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.04	0.075	4,192	85	14	-0.02
Focal teacher	0.23	0.186	1,451	21	14	0.04
Nonfocal teacher	-0.02	0.044	4,248	70	14	-0.05
Reading						
Team	-0.01	0.035	4,454	104	13	0.02
Focal teacher	0.00	0.083	1,353	16	13	0.01
Nonfocal teacher	0.01	0.034	4,432	88	13	0.03
Year 2 (cohort 1 only)						
Math						
Team	-0.03	0.041	796	14	4	-0.02
Focal teacher	0.15	0.072	476	5	4	0.03
Nonfocal teacher	-0.07	0.046	560	11	4	-0.07
Reading						
Team	-0.14*	0.011	724	16	4	-0.02
Focal teacher	-0.17*	0.029	402	5	4	-0.06*
Nonfocal teacher	-0.13*	0.030	474	11	4	-0.05

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.12. Test-Score Impacts in Middle Schools, Benchmark Model, Blocks Containing Schools with Both Treatment and Control Teams

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	-0.10	0.053	2,355	125	34	0.05
Focal teacher	0.01	0.056	1,127	50	34	0.18*
Nonfocal teacher	-0.21*	0.050	1,709	81	31	-0.07
Reading						
Team	-0.08	0.048	2,341	125	34	0.03
Focal teacher	-0.03	0.041	1,146	52	34	0.10*
Nonfocal teacher	-0.12	0.063	1,596	82	31	-0.01
Year 2 (cohort 1 only)						
Math						
Team	0.05	0.080	2,143	104	26	0.08*
Focal teacher	0.29*	0.103	788	36	26	0.22*
Nonfocal teacher	-0.07	0.063	1,784	79	26	0.00
Reading						
Team	-0.04	0.055	2,129	106	26	0.07*
Focal teacher	0.26*	0.071	892	37	26	0.25*
Nonfocal teacher	-0.13*	0.049	1,907	82	26	0.00

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

C. Sensitivity Analysis Tables

In this section, we provide details about the sensitivity analyses we referred to in Chapters II and V. In Table F.13, we summarize the complete list of sensitivity analyses conducted that relate to specification of the regression model and alternative rules for defining the sample. Tables containing the sensitivity results follow. We only presented selected results, to illustrate cases where the magnitudes of impact estimates or their significance levels may suggest a different qualitative conclusion from the benchmark analysis.

Table F.13. Complete List of Sensitivity Analyses

Sensitivity Analysis	Program Year	Location of Results	Location of Detailed Sample-Size Information
Addressing Course-Taking Patterns			
Benchmark model omitting grade 8 math students from one district	1	Table F.6	N.P.
Benchmark model + indicator of taking test in Spanish	1 & 2	Tables F.7 and F.8	N.P.
Alternative Definition of Focal Teachers			
Benchmark model, selective method for identifying focal teachers	1 & 2	Tables F.14 and F.15	N.P.
Alternative Handling of Errors in Test-Score Data			
Benchmark model excluding students with imputed pre-test	1 & 2	Tables F.16 and F.17	N.P.
Benchmark model restricting z-scores to be between -3 and 3	1 & 2	N.P.	N.P.
Benchmark model, error-in-variables correction for pre-test	1 & 2	N.P.	N.P.
Alternative Pre-Test Specification			
Benchmark model + grade*district*pre-test interactions	1 & 2	N.P.	N.P.
Benchmark model + opposite subject pre-test	1 & 2	Tables F.18 and F.19	N.P.
Benchmark model + pre-test^2 + pre-test^3	1 & 2	N.P.	N.P.
Benchmark model, moving pre-test variable to left-hand side, meaning the dependent variable is gain between pre-test and post-test	1 & 2	N.P.	N.P.
Alternative Specification of Student-Background Covariates			
Benchmark model + student background covariates*district interactions, where student background covariates refer to race/ethnicity, gender, ELL, SPED, and FRL	1 & 2	Tables F.20 and F.21	N.P.
Benchmark model—pre-test—student-background covariates, where student-background covariates refer to race/ethnicity, gender, ELL, SPED, and FRL	1 & 2	Tables F.22 and F.23	N.P.
Alternative Specification of Treatment			
Benchmark model replacing treatment variable with treatment*percent of year enrolled in school interaction	1 & 2	Tables F.24 and F.25	N.P.
Alternative Sample-Inclusion Rules			
Benchmark model adding pilot teams to the analysis sample	1 & 2	Tables F.26 and 27	N.P.

N.P. = not presented in this report. Results are available from the authors.

1. Alternative Definition of Focal Teachers

As discussed in Appendix D, the focal-teacher comparisons are based on an inclusive definition for cases where the identity of the teacher who filled the study vacancy identified for random assignment was ambiguous. The inclusive definition approximates a broad notion of the counterfactual, where teachers who were not necessarily new to the team may be part of the comparison. This is preferable to a more selective definition in which survey nonrespondents in the control group, or their entire teams, are disproportionately likely to be omitted from the test-score analysis. In such a case, the treatment-control differences may reflect differences in the characteristics of teachers who respond to surveys as much as the impact of the intervention. Nevertheless, we show what the impacts would have been had we used the selective definition, as a way to further understand the data.

In Tables F.14 and F.15, we present the elementary and middle school results of estimating the benchmark model using the selective definition to identify focal and nonfocal teachers. Using the selective definition, elementary school program year 1 nonfocal teacher-level comparisons decreased and are statistically significant. Aside from reading impacts in program year 1, elementary school focal teacher-level comparisons remained positive and statistically significant. Middle school impacts remain statistically insignificant, except for reading impacts in program year 2. Because we have determined that impacts are heterogeneous and the composition of focal and nonfocal teachers included in the analysis differs between the analyses using selective rather than inclusive definition, it is not surprising that impacts change when the analysis sample is altered.

Table F.14. Test-Score Impacts in Elementary Schools, Benchmark Model, Selective Method for Identifying Focal Teachers

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.05	0.036	8,177	414	104	0.05
Focal teacher	0.17*	0.058	2,211	99	84	0.18*
Nonfocal teacher	-0.15*	0.044	4,976	223	77	-0.07
Reading						
Team	0.03	0.032	8,097	427	104	0.03
Focal teacher	0.05	0.061	2,084	98	85	0.10*
Nonfocal teacher	-0.10*	0.043	4,922	235	78	-0.01
Year 2 (cohort 1 only)						
Math						
Team	0.08*	0.039	7,565	366	90	0.08*
Focal teacher	0.21*	0.081	1,952	76	65	0.22*
Nonfocal teacher	-0.03	0.050	4,865	213	70	0.00
Reading						
Team	0.07*	0.031	7,481	383	90	0.07*
Focal teacher	0.13*	0.055	1,643	74	64	0.25*
Nonfocal teacher	-0.05	0.038	5,466	230	76	0.00

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.15. Test-Score Impacts in Middle Schools, Benchmark Model, Selective Method for Identifying Focal Teachers

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	-0.02	0.056	8,875	169	30	-0.02
Focal teacher	-0.06	0.097	2,071	27	27	0.04
Nonfocal teacher	-0.08	0.064	7,793	125	27	-0.05
Reading						
Team	0.02	0.027	7,812	195	31	0.02
Focal teacher	-0.01	0.057	2,600	32	28	0.01
Nonfocal teacher	0.04	0.030	6,563	137	29	0.03
Year 2 (cohort 1 only)						
Math						
Team	-0.02	0.057	2,627	42	13	-0.02
Focal teacher	-0.01	0.044	929	9	9	0.03
Nonfocal teacher	-0.06	0.184	1,142	21	10	-0.07
Reading						
Team	-0.02	0.022	3,488	72	20	-0.02
Focal teacher	-0.11*	0.023	1,389	16	14	-0.06*
Nonfocal teacher	-0.07*	0.028	2,197	41	17	-0.05

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

2. Alternative Handling of Errors in Test-Score Data

As mentioned in Chapter V, the benchmark model considered extreme outliers, test-score values that were greater than 4 or less than -4 standard deviations, to be data errors and set them to missing. We explored a few additional ways of handling possible errors in the test-score data. For example, we re-estimated the test-score impacts by discarding observations where z-scores did not lie between -3 and 3 standard deviations. We also implemented an errors-in-variables correction for the pre-test.[108] In both cases, the impact estimates were the same as our benchmark impacts so they are not shown. In Tables F.16 and F.17, we show elementary and middle school test-score comparisons from yet another model alternative, one that drops students with imputed pre-tests.

[108] The errors-in-variables procedure assumes that the reliability of the pre-test is known and uses that information to adjust.

Table F.16. Test-Score Impacts in Elementary Schools, Benchmark Model Excluding Students with Imputed Pre-Test

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.05	0.038	6,253	391	97	0.05
Focal teacher	0.18*	0.053	2,834	139	96	0.18*
Nonfocal teacher	-0.03	0.040	4,961	251	87	-0.07
Reading						
Team	0.04	0.035	6,200	404	97	0.03
Focal teacher	0.14*	0.043	2,864	141	96	0.10*
Nonfocal teacher	0.00	0.050	5,021	265	91	-0.01
Year 2 (cohort 1 only)						
Math						
Team	0.10*	0.040	6,139	366	90	0.08*
Focal teacher	0.21*	0.066	2,856	131	90	0.22*
Nonfocal teacher	0.07	0.041	4,957	253	82	0.00
Reading						
Team	0.10*	0.031	6,103	383	90	0.07*
Focal teacher	0.21*	0.045	2,742	129	88	0.25*
Nonfocal teacher	0.04	0.037	5,509	272	88	0.00

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.17. Test-Score Impacts in Middle Schools, Benchmark Model Excluding Students with Imputed Pre-Test

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	-0.02	0.059	8,038	169	30	-0.02
Focal teacher	0.05	0.080	2,533	38	30	0.04
Nonfocal teacher	-0.05	0.056	7,754	136	30	-0.05
Reading						
Team	0.04	0.030	7,063	195	31	0.02
Focal teacher	0.03	0.048	2,916	49	31	0.01
Nonfocal teacher	0.05	0.030	6,491	152	31	0.03
Year 2 (cohort 1 only)						
Math						
Team	-0.01	0.053	2,355	42	13	-0.02
Focal teacher	0.03	0.043	1,403	19	13	0.03
Nonfocal teacher	-0.04	0.067	1,621	31	13	-0.07
Reading						
Team	-0.02	0.023	3,128	72	20	-0.02
Focal teacher	-0.06	0.030	1,841	25	20	-0.06*
Nonfocal teacher	-0.04	0.023	2,619	50	20	-0.05

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

3. Alternative Pre-Test Specification

Research has shown that a student's pre-test score is an important predictor of future performance. Given the importance of including the pre-test as a covariate in our impact-estimation model, we investigated the robustness of our impact estimates when we used alternative pre-test specifications. For example, we allowed for a more flexible relationship between the pre-test and post-test in our estimation model by including pre-test squared and cubed as covariates, and we also added interactions to the estimation that allow the effect of pre-test to differ by district and grade. Furthermore, we tried changing the dependent variable from the post-test to the "gain" (difference) between the pre- and post-test. The gain model imposes a strict assumption about the relationship between pre- and post-test, but it also avoids risk of bias that results from a pre-test being measured with error. For each of these model variations, the sign and significance of the impact estimates were very similar to the benchmark model so the results are not shown here. In addition, because we found that the baseline mean math pre-test was different across students in treatment and control groups for the middle school reading

analysis sample, it is important to make sure that estimated TTI impacts were robust to the inclusion of the opposite subject pre-test. Therefore, we also conducted a sensitivity analysis where we added the opposite-subject pre-test to the estimation model (Tables F.18 and F.19).

Table F.18. Test-Score Impacts in Elementary Schools, Benchmark Model Plus Opposite-Subject Pre-Test

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.05	0.036	8,177	414	104	0.05
Focal teacher	0.18*	0.049	3,751	158	104	0.18*
Nonfocal teacher	-0.06	0.041	6,516	282	95	-0.07
Reading						
Team	0.02	0.032	8,097	427	104	0.03
Focal teacher	0.10	0.050	3,804	161	104	0.10*
Nonfocal teacher	-0.01	0.039	6,642	297	98	-0.01
Year 2 (cohort 1 only)						
Math						
Team	0.08*	0.038	7,565	366	90	0.08*
Focal teacher	0.22*	0.061	3,327	133	90	0.22*
Nonfocal teacher	0.00	0.041	6,240	270	83	0.00
Reading						
Team	0.08*	0.031	7,481	383	90	0.07*
Focal teacher	0.25*	0.050	3,201	133	90	0.25*
Nonfocal teacher	0.01	0.033	7,024	286	88	0.00

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.19. Test-Score Impacts in Middle Schools, Benchmark Model Plus Opposite-Subject Pre-Test

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	-0.02	0.054	8,875	169	30	-0.02
Focal teacher	0.05	0.079	2,827	38	30	0.04
Nonfocal teacher	-0.05	0.052	8,549	136	30	-0.05
Reading						
Team	0.00	0.028	7,812	195	31	0.02
Focal teacher	0.00	0.051	3,261	49	31	0.01
Nonfocal teacher	0.00	0.030	7,224	154	31	0.03
Year 2 (cohort 1 only)						
Math						
Team	-0.03	0.052	2,627	42	13	-0.02
Focal teacher	0.02	0.048	1,575	19	13	0.03
Nonfocal teacher	-0.09	0.063	1,788	31	13	-0.07
Reading						
Team	-0.02	0.021	3,488	72	20	-0.02
Focal teacher	-0.05*	0.024	2,090	25	20	-0.06*
Nonfocal teacher	-0.05	0.024	2,898	50	20	-0.05

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

4. Alternative Specification of Student-Background Covariates

To address the baseline differences in a number of the student-background covariates described in Chapter II for the elementary school and middle school math analysis samples, we augmented our benchmark model with interaction terms of student background covariates (race/ethnicity, gender, ELL, SPED, and FRL) and district dummies. The results are presented in Tables F.20 and F.21. We also estimated a model where we removed the pre-test variable and student-background covariates from our benchmark model (see Tables F.22 and F.23).

Table F.20. Test-Score Impacts in Elementary Schools, Benchmark Model with Interactions of Student-Background Covariates and District Dummies

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.02	0.038	8,151	414	104	0.05
Focal teacher	0.17*	0.053	3,730	158	104	0.18*
Nonfocal teacher	-0.11*	0.044	6,027	282	95	-0.07
Reading						
Team	0.02	0.033	8,072	427	104	0.03
Focal teacher	0.09	0.051	3,552	161	104	0.10*
Nonfocal teacher	-0.01	0.046	6,143	297	98	-0.01
Year 2 (cohort 1 only)						
Math						
Team	0.07	0.041	7,547	366	90	0.08*
Focal teacher	0.24*	0.071	3,291	133	90	0.22*
Nonfocal teacher	-0.01	0.044	5,592	270	83	0.00
Reading						
Team	0.07*	0.032	7,463	383	90	0.07*
Focal teacher	0.29*	0.056	3,062	133	90	0.25*
Nonfocal teacher	0.00	0.032	5,966	286	88	0.00

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.21. Test-Score Impacts in Middle Schools, Benchmark Model with Interactions of Student-Background Covariates and District Dummies

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.05	0.045	8,849	169	30	-0.02
Focal teacher	0.16	0.092	2,789	38	30	0.04
Nonfocal teacher	0.02	0.034	7,234	136	30	-0.05
Reading						
Team	0.02	0.030	7,785	195	31	0.02
Focal teacher	0.03	0.061	2,914	49	31	0.01
Nonfocal teacher	0.03	0.030	5,981	154	31	0.03
Year 2 (cohort 1 only)						
Math						
Team	-0.03	0.031	2,621	42	13	-0.02
Focal teacher	0.01	0.050	1,568	19	13	0.03
Nonfocal teacher	-0.08*	0.028	1,741	31	13	-0.07
Reading						
Team	-0.07*	0.017	3,484	72	20	-0.02
Focal teacher	-0.10*	0.022	1,941	25	20	-0.06*
Nonfocal teacher	-0.41*	0.037	2,418	50	20	-0.05

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.22. Test-Score Impacts in Elementary Schools, Benchmark Model Without Pre-Test Variable and Student-Background Covariates

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.10*	0.043	8,177	414	104	0.05
Focal teacher	0.25*	0.062	3,751	158	104	0.18*
Nonfocal teacher	-0.04	0.058	6,516	282	95	-0.07
Reading						
Team	0.04	0.037	8,097	427	104	0.03
Focal teacher	0.17*	0.060	3,804	161	104	0.10*
Nonfocal teacher	0.01	0.057	6,642	297	98	-0.01
Year 2 (cohort 1 only)						
Math						
Team	0.07	0.044	7,565	366	90	0.08*
Focal teacher	0.21*	0.065	3,327	133	90	0.22*
Nonfocal teacher	0.02	0.054	6,240	270	83	0.00
Reading						
Team	0.06	0.040	7,481	383	90	0.07*
Focal teacher	0.28*	0.051	3,201	133	90	0.25*
Nonfocal teacher	0.01	0.046	7,024	286	88	0.00

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.23. Test-Score Impacts in Middle Schools, Benchmark Model Without Pre-Test Variable and Student-Background Covariates

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.17*	0.041	8,875	169	30	-0.02
Focal teacher	0.11	0.131	2,827	38	30	0.04
Nonfocal teacher	0.15*	0.026	8,549	136	30	-0.05
Reading						
Team	0.13*	0.049	7,812	195	31	0.02
Focal teacher	0.06	0.107	3,261	49	31	0.01
Nonfocal teacher	-0.09	0.058	7,224	154	31	0.03
Year 2 (cohort 1 only)						
Math						
Team	0.13	0.116	2,627	42	13	-0.02
Focal teacher	0.21	0.169	1,575	19	13	0.03
Nonfocal teacher	-0.01	0.106	1,788	31	13	-0.07
Reading						
Team	0.01	0.062	3,488	72	20	-0.02
Focal teacher	-0.02	0.080	2,090	25	20	-0.06*
Nonfocal teacher	-0.09	0.090	2,898	50	20	-0.05

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

5. Alternative Specification of Treatment

It may be important to distinguish between students who were exposed to a full year of treatment and those who were in the study team for only part of the academic year. We replaced the treatment indicator with the treatment indicator multiplied by the percentage of the year the student was enrolled in the school. The results from this alternative specification of the treatment variable are presented in Tables F.24 and F.25.

Table F.24. Test-Score Impacts in Elementary Schools, Benchmark Model Replacing Treatment Variable with Treatment Multiplied by Percent of Year Enrolled in School

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.06	0.036	8,175	413	104	0.05
Focal teacher	0.20*	0.049	3,751	158	104	0.18*
Nonfocal teacher	-0.05	0.042	6,515	282	95	-0.07
Reading						
Team	0.04	0.033	8,095	426	104	0.03
Focal teacher	0.12*	0.050	3,803	161	104	0.10*
Nonfocal teacher	-0.01	0.044	6,642	297	98	-0.01
Year 2 (cohort 1 only)						
Math						
Team	0.09*	0.040	7,418	366	90	0.08*
Focal teacher	0.23*	0.066	3,261	133	90	0.22*
Nonfocal teacher	0.01	0.043	6,130	270	83	0.00
Reading						
Team	0.08*	0.032	7,337	383	90	0.07*
Focal teacher	0.26*	0.051	3,136	133	90	0.25*
Nonfocal teacher	0.01	0.033	6,917	286	88	0.00

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.25. Test-Score Impacts in Middle Schools, Benchmark Model Replacing Treatment Variable with Treatment Multiplied by Percent of Year Enrolled in School

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	-0.01	0.055	8,874	169	30	-0.02
Focal teacher	0.05	0.087	2,827	38	30	0.04
Nonfocal teacher	-0.04	0.053	8,549	136	30	-0.05
Reading						
Team	0.02	0.026	7,809	195	31	0.02
Focal teacher	0.01	0.052	3,261	49	31	0.01
Nonfocal teacher	-0.05	0.029	7,224	154	31	0.03
Year 2 (cohort 1 only)						
Math						
Team	-0.01	0.058	2,627	42	13	-0.02
Focal teacher	0.03	0.062	1,575	19	13	0.03
Nonfocal teacher	-0.05	0.070	1,788	31	13	-0.07
Reading						
Team	-0.02	0.025	3,399	72	20	-0.02
Focal teacher	-0.07*	0.025	2,008	25	20	-0.06*
Nonfocal teacher	-0.04	0.031	2,816	50	20	-0.05

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

6. Alternative Sample Inclusion Rules

Finally, we added study teams from our pilot study to the analysis sample to see if results were robust to this slight increase in the sample. Results are shown in Tables F.26 and F.27.

Table F.26. Test-Score Impacts in Elementary Schools, Benchmark Model Adding Pilot Teams to the Analysis Sample

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	0.06	0.035	8,626	436	108	0.05
Focal teacher	0.18*	0.049	3,751	158	104	0.18*
Nonfocal teacher	-0.07	0.042	6,516	282	95	-0.07
Reading						
Team	0.03	0.031	8,543	450	108	0.03
Focal teacher	0.10*	0.050	3,804	161	104	0.10*
Nonfocal teacher	-0.01	0.044	6,642	297	98	-0.01
Year 2 (cohort 1 only)						
Math						
Team	0.08*	0.036	8,010	383	94	0.08*
Focal teacher	0.22*	0.064	3,327	133	90	0.22*
Nonfocal teacher	0.00	0.042	6,240	270	83	0.00
Reading						
Team	0.06*	0.030	7,921	400	94	0.07*
Focal teacher	0.25*	0.050	3,201	133	90	0.25*
Nonfocal teacher	0.00	0.032	7,024	286	88	0.00

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.27. Test-Score Impacts in Middle Schools, Benchmark Model Adding Pilot Teams to the Analysis Sample

Program Year, Subject, and Comparison Type	Impact	Standard Error	Number of Students[a]	Number of Teachers[b]	Number of Teams	Benchmark Impact
Year 1 (all districts)						
Math						
Team	-0.01	0.051	9,338	175	32	-0.02
Focal teacher	0.04	0.086	2,827	38	30	0.04
Nonfocal teacher	-0.05	0.053	8,549	136	30	-0.05
Reading						
Team	0.02	0.027	7,812	195	31	0.02
Focal teacher	0.01	0.052	3,261	49	31	0.01
Nonfocal teacher	0.03	0.028	7,224	154	31	0.03
Year 2 (cohort 1 only)						
Math						
Team	0.01	0.046	3,123	48	15	-0.02
Focal teacher	0.03	0.060	1,575	19	13	0.03
Nonfocal teacher	-0.07	0.069	1,788	31	13	-0.07
Reading						
Team	-0.02	0.022	3,488	72	20	-0.02
Focal teacher	-0.06*	0.024	2,090	25	20	-0.06*
Nonfocal teacher	-0.05	0.028	2,898	50	20	-0.05

Source: District administrative data.

*Statistically significant at the 0.05 level, two-tailed test.

[a]For the comparisons of focal and nonfocal teachers, number of students refers to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in identifying focal teachers, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

The results presented in this appendix point to the robustness of the benchmark impacts. The impact estimates fluctuated more when the sample changed (for example, when we used the alternate focal teacher definition in Tables F.14 and F.15); otherwise, the results showed the same pattern of heterogeneous impacts presented in Chapter V.

D. Sample-Size Tables

1. Sample-Size Information for Chapter V Tables and Figures

Table F.28. Sample-Size Information for Table V.1. Test-Score Impacts in Elementary Schools

Program Year, Subject, and Comparison Type	Number of Treatment Students[a]	Number of Control Students[a]	Number of Treatment Teachers[b]	Number of Control Teachers[b]	Number of Treatment Teams	Number of Control Teams
Year 1 (all districts)						
Math						
Team	4,236	3,941	212	202	53	51
Focal teacher	1,724	2,027	71	87	53	51
Nonfocal teacher	3,034	3,482	142	140	47	48
Reading						
Team	4,215	3,882	223	204	53	51
Focal teacher	1,661	2,143	70	91	53	51
Nonfocal teacher	3,325	3,317	152	145	49	49
Year 2 (cohort 1 only)						
Math						
Team	3,903	3,662	181	185	46	44
Focal teacher	1,576	1,751	59	74	46	44
Nonfocal teacher	2,857	3,383	126	144	42	41
Reading						
Team	3,856	3,625	196	187	46	44
Focal teacher	1,489	1,712	60	73	46	44
Nonfocal teacher	3,716	3,308	136	150	45	43

Source: District administrative data.

[a]For the focal- and nonfocal-teacher comparisons, sample sizes provided refer to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in focal-teacher identification, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.29. Sample-Size Information for Table V.2. Test-Score Impacts in Middle Schools

Program Year, Subject, and Comparison Type	Number of Treatment Students[a]	Number of Control Students[a]	Number of Treatment Teachers[b]	Number of Control Teachers[b]	Number of Treatment Teams	Number of Control Teams
Year 1 (all districts)						
Math						
Team	5,433	3,442	102	67	16	14
Focal teacher	1,214	1,613	16	22	16	14
Nonfocal teacher	5,229	3,320	84	52	16	14
Reading						
Team	4,205	3,607	97	98	16	15
Focal teacher	1,388	1,873	18	31	16	15
Nonfocal teacher	3,732	3,492	75	79	16	15
Year 2 (cohort 1 only)						
Math						
Team	1,591	1,036	26	16	7	6
Focal teacher	810	765	9	10	7	6
Nonfocal teacher	1,095	693	19	12	7	6
Reading						
Team	1,951	1,537	40	32	10	10
Focal teacher	1,110	980	12	13	10	10
Nonfocal teacher	1,666	1,232	27	23	10	10

Source: District administrative data.

[a]For the focal- and nonfocal-teacher comparisons, sample sizes provided refer to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in focal-teacher identification, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

Table F.30. Sample-Size Information for Table V.3. Test-Score Impacts in Elementary and Middle Schools Combined

Program Year, Subject, and Comparison Type	Number of Treatment Students[a]	Number of Control Students[a]	Number of Treatment Teachers[b]	Number of Control Teachers[b]	Number of Treatment Teams	Number of Control Teams
Year 1 (all districts)						
Math						
Team	9,669	7,383	314	269	69	65
Focal teacher	2,938	3,640	87	109	69	65
Nonfocal teacher	8,263	6,802	227	191	63	62
Reading						
Team	8,420	7,489	320	302	69	66
Focal teacher	3,049	4,016	88	122	69	66
Nonfocal teacher	7,057	6,809	226	225	65	64
Year 2 (cohort 1 only)						
Math						
Team	5,494	4,698	207	201	53	50
Focal teacher	2,386	2,516	68	84	53	50
Nonfocal teacher	3,952	4,076	145	156	49	47
Reading						
Team	5,807	5,162	236	219	56	54
Focal teacher	2,599	2,692	72	86	56	54
Nonfocal teacher	5,382	4,540	163	173	55	53

Source: District administrative data.

[a]For the focal- and nonfocal-teacher comparisons, sample sizes provided refer to the number of unique student-teacher combinations included in the analysis samples. Students may appear more than once in these analyses because they can be linked to more than one teacher. Students in these analyses are weighted proportionately to the probability that the teacher they are linked to is the focal (or nonfocal) teacher. Due to the uncertainty involved in focal-teacher identification, some teachers and their students were included in both the focal and nonfocal samples.

[b]For the team-level analyses, sample sizes provided refer to the number of unique teacher-team combinations included in the analysis samples. A teacher will show up in more than one team if he or she teaches students in multiple study teams.

APPENDIX G

SUPPLEMENTAL MATERIALS FOR CHAPTER VI

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING

In this appendix, we present additional detail on the methods used to estimate the retention impacts presented in Chapter VI, the results of sensitivity tests on the benchmark results, and subgroup impacts. We also include a non-experimental analysis of retention among teachers who received retention stipends through the Talent Transfer Initiative (TTI) and tables relating to sample sizes that correspond to tables in Chapter VI and other tables in this appendix.

A. Benchmark-Model-Estimation Equation

The benchmark retention impacts presented in Chapter VI were estimated with a linear probability model. The estimation equation is:

$$Y_j = \theta + \varphi_j T_j + \varphi_* R_j + \varphi_* D_j + e_j$$

where Y_j is a binary variable indicating if teacher j stayed in the same school. To examine one-year retention in the report, we determined if teachers remained at the same school in the fall of years 1 and 2, using the equation above. We repeated the analysis to examine two-year retention by comparing teachers' school location in the fall of years 1 and 3. T_j is a binary variable indicting if the teacher was assigned to teach on a treatment or control team in year 1. R_j is a binary variable indicating if the teacher was a retention-stipend teacher. We included this as a covariate because teachers receiving retention stipends through TTI would be expected to stay in their positions at higher rates than those not receiving retention stipends. D_j is a vector of indicators for the random assignment block. The coefficients φ, φ and φ were estimated. φ was used to predict the regression-adjusted retention rates for teachers on treatment and control teams. The standard errors are estimated assuming a common variance at the team level to account for teachers clustered within grades and subjects within schools.

The team-level impacts were estimated by including the full sample of focal and nonfocal teachers, and the focal-teacher impacts were estimated on the subsample of focal teachers, based on the inclusive definition of focal teachers (see Appendix D for an explanation).

B. Sensitivity Tests

The benchmark model represents the most efficient and practical way to estimate the impact of TTI on teacher retention, but it encompasses several assumptions about the appropriate functional form of the estimation model and the equivalence of treatment and control groups on baseline student characteristics. We also made assumptions about how to identify focal teachers. To test the robustness of the benchmark results reported in Chapter VI, we estimated the impacts of TTI on school retention with several alternative models. We describe in this section the alternative models and compare the results of these models to the benchmark-retention impacts. These alternative models include:

- **Alternative functional form.** Logistic regression model with random block effects
- **Additional covariates.** Inclusion of covariates controlling for team-level student characteristics
- **Alternative samples.** Blocks without closed or reconstituted schools, alternative focal-teacher definition

1. Alternative Functional Form of Retention-Estimation Model

The retention impacts reported in Chapter VI are estimated using a linear probability model (LPM) with fixed block effects. The LPM can produce out-of-range predictions (such as retention rates greater than 100 percent), but it also allows us to estimate block fixed effects in the presence of perfectly predicted outcomes for some blocks. As a check on the robustness of these results, however, we also estimated a logistic regression model with random block effects. This model included the same covariates as the benchmark LPM and assumed a common variance at the team level to account for teachers clustered within grades within schools. As with the benchmark model, we predicted one- and two-year retention rates for treatment and control teachers in the full sample and the focal-teacher sample.

The results of this model are presented in Table G.1 along with those of the benchmark model. The impacts are presented as marginal effects evaluated at the sample means for the entire estimation sample (treatment and control). The one-year retention estimates in the two models are within 2 percentage points of each other. The statistical significance is the same across the two models at the focal-teacher level. However, the team-level one-year retention impact estimated with the logistic model is not statistically significant at the 5 percent level ($p = 0.064$).

Table G.1. Logit-Model Retention Results Versus Benchmark-Model Retention Results

	Benchmark Impact (LPM)	Logistic Model Impact	Sample Size (teachers)
One-Year Retention			
All teachers on study teams	0.07*	0.07	725
Focal teachers	0.23*	0.25*	230
Nonfocal teachers	0.02	0.02	559
Two-Year Retention			
All teachers on study teams	0.04	0.05	498
Focal teachers	0.09	0.13	176
Nonfocal teachers	0.00	0.02	376

Source: School rosters.

Note: One-year retention impacts include data from both cohorts 1 and 2. Data from cohort 1 districts is from 2010–11 and 2011–12; data from cohort 2 districts is from 2011–12. Two-year retention impacts include data from cohort 1 only.

See Table G.10 for teacher and team sample sizes by treatment status.

*Difference is statistically significant at the 0.05 level using a two-sided test.

2. Additional Covariates in Retention-Estimation Model

We also conducted a robustness check by including additional covariates in the estimation model. The benchmark model included covariates controlling for a teacher's receipt of a retention stipend through TTI as well as for fixed-block effects. We estimated an alternative model with additional covariates controlling for student characteristics. Although the student characteristics should be similar on treatment and control teams because of randomization, there are some chance differences in student race, free and reduced-price lunch (FRL), and English-language learner (ELL) status between treatment and control teams, as reported in Chapter II.

This sensitivity analysis controls for these student characteristics. The model includes covariates for the demographic characteristics of students on each team, including race (white, black, Hispanic, and other race), gender, FRL status, ELL status, and special education (SPED) status.[109] These variables are equal to the percentage of students on each team that fall into each category. Because we use team-level percentages rather than teacher-specific percentages, the values of these variables should not be affected by treatment or focal status.

In Table G.2, we show the results of this model compared with those of the benchmark model. The impacts estimated when controlling for student characteristics are between 1 and 4 percentage points smaller than the benchmark impacts. The one-year focal-teacher impact is statistically significant in both models, but the one-year team-level impact is not significant when controlling for student characteristics ($p = 0.204$).

Table G.2. Student-Characteristic-Model Retention Results Versus Benchmark-Model Retention Results

	Benchmark Impact	Student Covariate Model Impact	Sample Size (teachers)
One-Year Retention			
All teachers on study teams	0.07*	0.04	725
Focal teachers	0.23*	0.19*	230
Nonfocal teachers	0.02	-0.01	559
Two-Year Retention			
All teachers on study teams	0.04	0.00	498
Focal teachers	0.09	0.08	176
Nonfocal teachers	0.00	-0.04	374

Source: School rosters.

Note: One-year retention impacts include data from both cohorts 1 and 2. Data from cohort 1 districts is from 2010–11 and 2011–12; data from cohort 2 districts is from 2011–12. Two-year retention impacts include data from cohort 1 only.

See Table G.10 for teacher and team sample sizes by treatment status.

*Difference is statistically significant at the 0.05 level using a two-sided test.

3. Alternative Samples

Another way of checking the robustness of the benchmark results is by altering the sample of teachers included in the analysis. The benchmark model used an intent-to-treat approach, estimating impacts on the full sample of teachers by using all of the blocks that were originally randomized. However, the full benchmark sample includes blocks in which TTI teachers were not hired by the treatment school as well as schools that were affected by closures or external staffing initiatives that could have affected retention.

We estimated retention on several alternative samples that exclude noncompliers and other special cases to check the robustness of the benchmark results. We found that although the point estimates vary by 2 to 5 percentage points under different sample definitions, the magnitude and

[109] The pre-imputed values of demographic variables are used, and students with missing data are not included in these team-level calculations.

statistical significance of the impacts are robust. Below we describe two alternative samples and compare the impact estimates from these samples to those reported in Chapter VI.

4. Blocks Affected by External Closures or Reconstitutions

This study was conducted over several years in high-need schools, which are often subject to district initiatives related to staffing. Consequently, some schools in the sample were closed; others implemented reconstitution in which some or all teachers were forced to re-apply for their positions and some were not re-hired. Between program years 1 and 2, four study schools closed or went through reconstitution. Between program years 2 and 3, an additional three study schools closed. These schools included both treatment and control teams.

Although these closures and reconstitutions were most likely unrelated to participation in the study or to treatment status, they almost certainly affected teacher retention. To check whether our benchmark impacts were affected by these changes, we estimated retention for an alternative sample that excluded blocks in closed or reconstituted schools. For this sensitivity check, we dropped five blocks in three districts from the one-year retention analysis, and a total of eight blocks in four districts from the two-year retention analysis.

A comparison between the impacts for the full sample and the nonclosure sample are presented in Table G.3. The one-year retention impacts for the two samples are within 2 percentage points of each other. Although the one-year focal-teacher impact did not change, the team impact decreased by 2 percentage points and is not significant at the 5 percent level ($p = 0.090$). The two-year impacts vary by between 3 and 6 percentage points, with smaller impacts in the nonclosure sample. The smaller impacts observed in the nonclosure sample analysis are primarily driven by higher predicted retention rates on control teams rather than lower predicted retention rates on treatment teams.

Table G.3. Impacts on Retention in School, Comparison of Alternative Samples

	Benchmark Sample Impact	Nonclosure Sample Impact
One-Year Retention		
All teachers on study teams	0.07*	0.05
Focal teachers	0.23*	0.23*
Nonfocal teachers	0.02	-0.01
Team analysis sample size	725	673
Two-Year Retention		
All teachers on study teams	0.04	0.01
Focal teachers	0.09	0.03
Nonfocal teachers	0.00	-0.01
Team analysis sample size	498	454

Source: School rosters.

Note: One-year retention impacts include data from both cohorts 1 and 2. Data from cohort 1 districts is from 2010–11 and 2011–12; data from cohort 2 districts is from 2011–12. Two-year retention impacts include data from cohort 1 only.

See Table G.12 for teacher and team sample sizes by treatment status.

*Difference is statistically significant at the 0.05 level using a two-sided test.

In addition to the impacts, this table also includes the sample sizes for the team-level analysis under each sample definition. The nonclosure sample is smaller than the benchmark sample because some blocks are excluded from this analysis. See Table G.12 for the teacher and team sample sizes by treatment status for the team- and focal-level analyses.

5. Alternative Definition of Focal Teacher

As described in Appendix C, we use an inclusive sample of focal teachers throughout this report. However, we also applied a more selective definition of focal teachers that includes only those for whom we had information confirming that they filled the study vacancy. Teams for which we could not confirm the identity of the focal teacher are excluded from the selective focal-teacher sample. As a result, the selective focal-teacher analysis includes only 146 teams, compared with 165 teams in the benchmark analysis.

Although the more inclusive definition was used for our benchmark estimates of teacher retention, we also estimated retention for the more selective focal-teacher sample. A comparison of the retention impacts for the inclusive and selective focal-teacher samples is presented in Table G.4. The impact estimates are somewhat smaller for the selective sample, but they are within 4 percentage points of each other, and the statistical significance is consistent across the two samples.

Table G.4. Impacts on Retention in School, Inclusive and Selective Focal Teacher Samples

	Inclusive Focal-Teacher Sample Impact (Benchmark)	Selective Focal-Teacher Sample Impact
One-Year Retention	0.23*	0.20*
Sample size	230	166
Two-Year Retention	0.09	0.05
Sample size	176	122

Source: School rosters.

Note: One-year retention impacts include data from both cohorts 1 and 2. Data from cohort 1 districts is from 2010–11 and 2011–12; data from cohort 2 districts is from 2011–12. Two-year retention impacts include data from cohort 1 only.

See Table G.13 for teacher and team sample sizes by treatment status.

*Difference is statistically significant at the 0.05 level using a two-sided test.

C. Subgroup Impacts

We also analyzed teacher retention within subgroups of the full sample. The following analyses split the teacher sample by cohort, district, and grade span.

1. Cohort Impacts

In Figure VI.1 in Chapter VI, we present the regression-adjusted retention rates of cohort 1 teachers. In Table G.5, we present the full data for two-year retention among cohort 1 teachers (one-year retention estimates for these teachers are shown in Table VI.1).

Table G.5. Two-Year Impacts on Retention in School, Cohort 1 Districts Only

Outcome and Sample	Treatment	Control	Impact	*p*-Value	N
All teachers on study teams	0.60	0.57	0.04	0.297	498
Focal teachers	0.60	0.51	0.09	0.286	176
Nonfocal teachers	0.60	0.60	0.00	0.921	374

Source: School rosters.

Note: Data from cohort 1 districts are from 2010–11 and 2011–12.

See Table G.10 for teacher and team sample sizes by treatment status.

*Difference is statistically significant at the 0.05 level using a two-sided test.

2. District Impacts

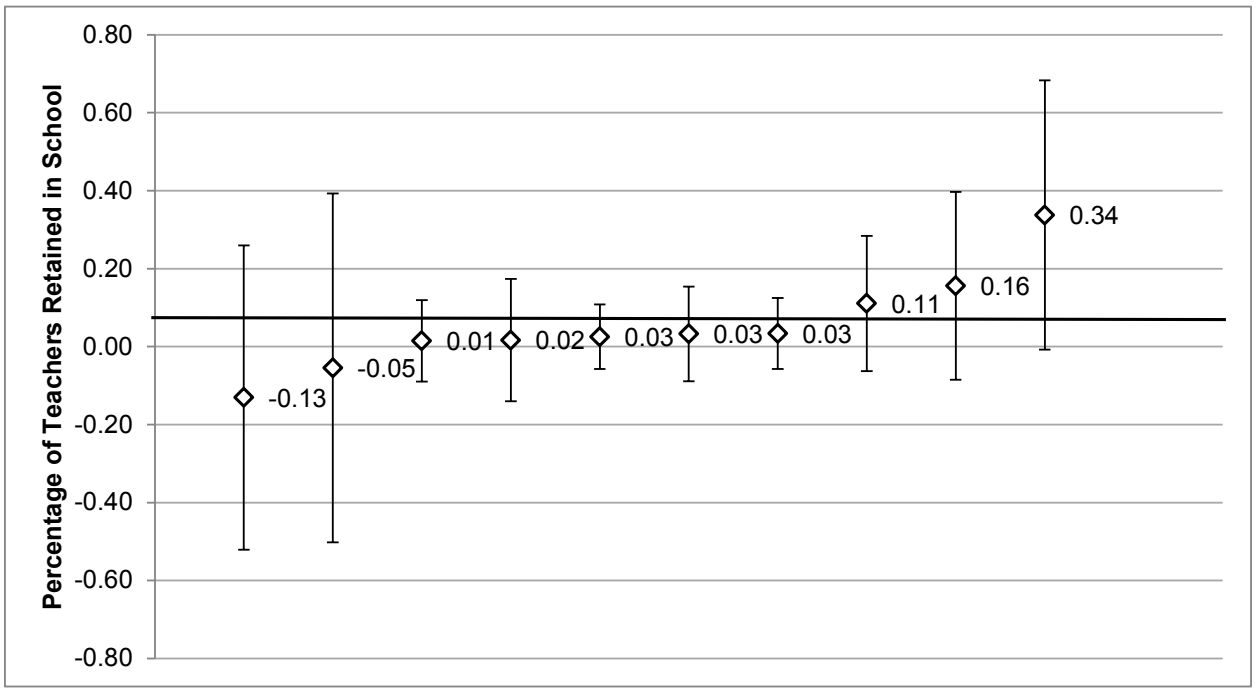
We estimated the benchmark model for individual districts to examine how the intervention affected teacher retention differently across the 10 districts. Sample sizes are small in the district-specific analyses, so there may be more noise in these estimates. However, the results are useful for understanding the extent of variation in impacts across the sample.[110]

The district-specific impacts on teacher retention are dispersed above and below the full-sample impact estimate. Figure G.1 shows the one-year team-level impacts for each of the 10 districts. The diamonds represent the impact estimates (hollow because none is statistically significant), and the bars represent 95 percent confidence intervals. The full-sample impact is included for comparison and is represented by the black line (solid because it is statistically significant). Seven of the district-level impacts are within 10 percentage points of the full-sample impact estimate, and all of the districts' 95 percent confidence intervals overlap the full-sample estimate. Eight of 10 districts have positive impacts. Although one district has an impact of 34 percentage points, which might appear to be inflating the overall impact, this district accounts for fewer than 5 percent of the teachers in the sample and has a 95 percent confidence interval of 70 percentage points. It does not appear that a single district is driving the overall impact.

We also ran a model on the full sample that included district-treatment interactions terms to test the equality of the district estimates. This test indicated that the district impact estimates are not significantly different ($p = 0.499$ for one-year retention team-level impacts; $p = 0.612$ for one-year retention focal-teacher impacts).

[110] District sample sizes ranged from 30 to 183 teachers. Eight of the 10 districts had sample sizes of fewer than 100.

Figure G.1. One-Year Team-Level Retention Impacts, by District



Source: School rosters.

Note: One-year retention impacts include data from both cohorts 1 and 2. Data from cohort 1 districts are from 2010–11; data from cohort 2 districts are from 2011–12.

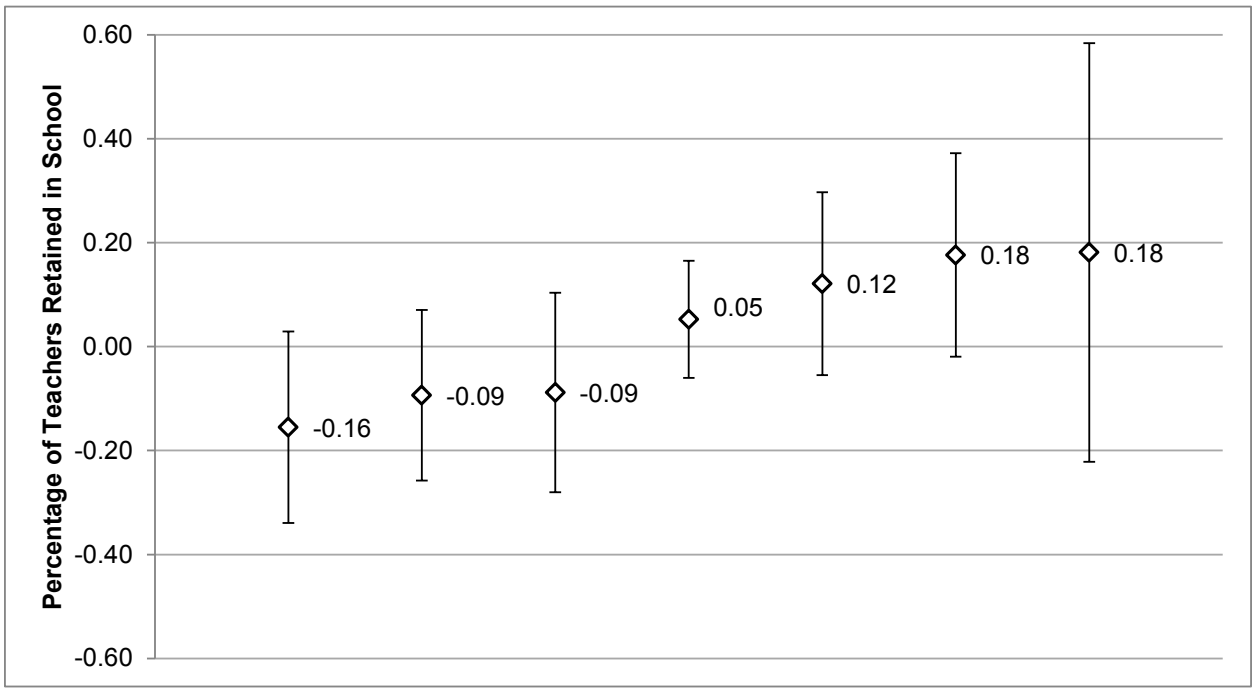
Bars represent 95 percent confidence intervals.

Sample sizes of teachers in each district range from 30 to 183. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

Figure G.2 presents the two-year retention impacts for the 7 districts in cohort 1. As with the one-year retention impacts, the estimates are dispersed above and below the full-sample impact estimate, ranging from -16 percentage points to 18 percentage points. Although 5 out of the 10 one-year district impacts were between 1 and 3 percentage points, the two-year estimates are more dispersed: only 2 district-level estimates fell within 10 percentage points of the full-sample impact estimate. Based on the district-interaction model, we found that the team-level estimates are statistically different from each other ($p = 0.014$). The focal-teacher two-year impacts are not statistically different from each other ($p = 0.376$).[111]

[111] The district-specific focal-teacher analysis has very small sample sizes and is not presented here.

Figure G.2. Two-Year Team-Level Retention Impacts, by District



Source: School rosters.

Note: One-year retention impacts include data from both cohorts 1 and 2. Data from cohort 1 districts is from 2010–11 and 2011–12; data from cohort 2 districts is from 2011–12.

Bars represent 95 percent confidence intervals.

Sample sizes of teachers in each district range from 36 to 183. We do not report sample sizes for specific data points in this figure to avoid linking results to specific districts.

3. Grade-Span Impacts

We also estimated the benchmark model for grade-span subgroups. In Table G.6, we present the one-year retention impacts for subgroups defined by grade span and subject at the team, focal, and nonfocal levels. As with the overall sample, we observe significant impacts in the focal-teacher sample. There are smaller, nonsignificant impacts at the team level, and the nonfocal impact estimates are close to zero.

In Table G.7, we present the two-year retention impacts by grade span and subject. None of the impacts is statistically significant. It is important to note the small sample sizes, particularly in the subject-specific middle school analyses, resulting in larger standard errors for the two-year estimates than for the one-year estimates.

Table G.6. One-Year Impacts on Retention in School, by Grade Span

	Treatment	Control	Impact	*p*-Value	N
Elementary					
All teachers on study teams	0.87	0.80	0.06	0.090	421
Focal teachers	0.92	0.70	0.22*	0.007	154
Nonfocal teachers	0.86	0.84	0.01	0.700	315
Middle School					
All teachers on study teams	0.72	0.65	0.07	0.172	304
focal teachers	0.83	0.59	0.25*	0.015	76
Nonfocal teachers	0.70	0.68	0.02	0.743	242
Middle School Math					
All teachers on study teams	0.76	0.64	0.08	0.135	161
Focal teachers	0.93	0.52	0.41*	0.001	42
Nonfocal teachers	0.74	0.70	0.03	0.752	133
Middle School English/language arts (ELA)					
All teachers on study teams	0.68	0.67	0.01	0.857	157
Focal teachers	0.75	0.64	0.11	0.493	34
Nonfocal teachers	0.66	0.68	-0.02	0.782	123

Source: School rosters.

See Table G.14 for teacher and team sample sizes by treatment status.

*Difference is statistically significant at the 0.05 level using a two-sided test.

Table G.7. Two-Year Impacts on Retention in School, by Grade Span

	Treatment	Control	Impact	*p*-Value	N
Elementary					
All teachers on study teams	0.64	0.60	0.04	0.403	391
Focal teachers	0.62	0.50	0.12	0.201	140
Nonfocal teachers	0.65	0.66	-0.01	0.786	299
Middle School ELA and Math					
All teachers on study teams	0.49	0.45	0.04	0.412	107
Focal teachers	0.52	0.53	-0.01	0.948	36
Nonfocal teachers	0.45	0.42	0.02	0.744	75
Middle School Math					
All teachers on study teams	0.38	0.26	0.13	0.106	45
Focal teachers	0.39	0.29	0.10	0.707	14
Nonfocal teachers	0.37	0.31	0.07	0.361	35
Middle School ELA					
All teachers on study teams	0.54	0.59	-0.04	0.602	69
Focal teachers	0.60	0.67	-0.07	0.766	22
Nonfocal teachers	0.51	0.53	-0.03	0.836	47

Source: School rosters.

Note: None of the impact estimates is statistically significant at the 0.05 level, two-tailed test.

See Table G.15 for teacher and team sample sizes by treatment status.

D. Non-Experimental Analysis of Retention-Stipend Teachers

Beyond the scope of the random assignment study, TTI also offered retention stipends to high-performing teachers who were already teaching at low-achieving schools in study districts. These teachers were identified through the value-added analysis that was used to identify transfer candidates. Because these teachers were not eligible to apply to transfer to a low-achieving school, they were each offered a $10,000 stipend paid in installments over two years to continue teaching at their current schools. These teachers were in treatment, control, and nonstudy potential receiving schools, and they were in study and nonstudy grades. Some, but not all, of the retention-stipend teachers were therefore included in the team-level and nonfocal benchmark estimates.

These stipends were not randomly assigned, so there is no control group to estimate the retention rates of these teachers in the absence of TTI. However, we can compare their retention rates to those of other teachers in their schools. This provides information on whether high-performing retention-stipend teachers stayed at their schools at higher rates than other teachers, but it does not provide evidence on the causal impact of retention bonuses for high-performing teachers on their school retention. Any differences in retention rates between groups of teachers cannot be attributed to TTI.

1. Data and Methods

As mentioned in Chapter VI, we collected teacher rosters from districts and schools in the fall of program years 1 and 2 for all 10 districts and in the fall of the year after the completion of the program for the 7 cohort 1 districts. We requested rosters from all schools that had a treatment team, a control team, or a teacher who was eligible for a retention stipend, so the sample for this analysis is larger than the experimental-analysis sample because it includes schools with retention-stipend teachers but no treatment or control teams.[112] Using the teacher rosters, we identified the TTI transfer teachers, TTI retention-stipend teachers, and all other academic teachers in the same schools. Academic teachers include those teaching self-contained classes, math, ELA, social studies, science, or foreign language.

As with the impact analysis presented in Chapter VI, we measured one-year school retention by tracking whether study teachers taught in the same school in the fall of program years 1 and 2 and two-year school retention based on teaching in the same school in the fall of program years 1 and 3. Two-year retention was measured for the seven cohort 1 districts only.

[112] 154 schools are included in this analysis, compared with 114 in the experimental analysis. In Chapter III, we reported the retention patterns of the full sample of retention-stipend teachers over the two study years based on program records of incentive payments. The retention rates reported in Chapter III include a larger sample of teachers because some nonstudy schools did not provide teacher rosters and so could not be included in the roster analysis. Four schools refused to provide rosters for at least one year, and 11 schools were not included in the roster analysis because they closed, merged, or reconstituted during the study period.

Linear probability models were used to compare the retention rates of TTI retention-stipend teachers to those of other teachers in their schools, as well as the retention rates of TTI transfer teachers to those of other teachers in their schools. The equation to estimate retention rates for retention-stipend teachers is:

$$Y_j = \theta + \varphi_j R_j + \psi S_j + e_j$$

where Y_j is a binary variable indicating if the teacher stayed in the same school between two time points. We use the fall of program years 1 and 2 (one-year retention) or programs years 1 and 3 (two-year retention) as the time points in this report. R_j is a binary variable indicating if the teacher was a retention-stipend teacher. S_j is a vector of school dummies. The coefficients φ_j and ψ were estimated. φ_j was used to predict the regression-adjusted retention rates for retention-stipend teachers and other teachers. The standard errors are estimated assuming a common variance at the school grade level to account for teachers clustered within grades within schools. This model is similar to the benchmark model described earlier in this appendix, but it includes school dummies instead of randomization block dummies. This is because this analysis is not part of the random assignment study and not all teachers taught on study blocks.

A similar model was used to compare the retention rates of TTI transfer teachers to other teachers in their schools. Retention was estimated for the full sample and for elementary and middle school teachers separately.

2. Retention Rates of Stipend Teachers

Teachers receiving transfer or retention stipends through TTI had a higher school retention rate between program years 1 and 2 than teachers not receiving stipends (see Table G.8). The transfer teachers, who were eligible to receive $20,000, had the highest one-year retention rates (92 percent).[113] The retention-stipend teachers, who were eligible for $10,000, had significantly higher retention rates than teachers who were not eligible for any TTI incentives (83 percent compared with 76 percent).[114] The difference between retention-stipend teachers and other academic teachers is statistically significant for the full sample, and the *p*-value is 0.056 for the elementary subgroup estimate. It is not statistically significant for middle school teachers, but the sample of retention-stipend teachers in this analysis is only 63 teachers, and the standard error on this estimate is larger than the all-grades and elementary samples.

In Table G.9, we present the two-year retention rates for transfer-stipend teachers, retention-stipend teachers, and all other teachers. Two-year retention was measured in the year after the completion of the program, after both the transfer and retention stipends had ended. There are no significant differences in two-year retention rates between stipend and nonstipend teachers.

[113] This sample is slightly different than the focal sample. The transfer-stipend-teacher sample includes all TTI teachers in original or updated grades. The focal sample includes focal teachers on teams that did not hire TTI teachers, and it also does not include TTI teachers who changed grades before the start of the school year.

[114] The sample of "Other Academic Teachers" differs from the nonfocal sample because it includes all academic teachers in the same schools as retention-stipend teachers. The nonfocal sample includes only nonfocal teachers on study teams.

Table G.8. One-Year Retention Rates of Transfer and Retention-Stipend Teachers

	Retention Rate	*p*-Value[a]	N
All Grades			
Transfer-stipend teachers	0.92	0.000	78
Retention-stipend teachers	0.83	0.041	159
Other academic teachers	0.76	n.a.	4263
Elementary School			
Transfer-stipend teachers	0.94	0.003	52
Retention-stipend teachers	0.88	0.056	96
Other academic teachers	0.82	n.a.	2593
Middle School			
Transfer-stipend teachers	0.89	0.003	26
Retention-stipend teachers	0.74	0.330	63
Other academic teachers	0.69	n.a.	1670

Source: School rosters.

Note: One-year retention impacts include data from both cohorts 1 and 2. Data from cohort 1 districts is from 2010–11 and 2011–12; data from cohort 2 districts is from 2011–12.

[a]Statistical significance of difference between transfer stipend teachers and all others, and difference between retention stipend teachers and all others.

n.a. = not applicable

Table G.9. Two-Year Retention Rates of Transfer and Retention-Stipend Teachers

	Retention Rate	*p*-Value[a]	N
All Grades			
Transfer-stipend teachers	0.61	0.741	58
Retention-stipend teachers	0.62	0.865	108
Other academic teachers	0.63	n.a.	2,967
Elementary School			
Transfer-stipend teachers	0.61	0.492	43
Retention-stipend teachers	0.69	0.636	81
Other academic teachers	0.67	n.a.	2,374
Middle School			
Transfer-stipend teachers	0.55	0.658	15
Retention-stipend teachers	0.38	0.234	27
Other academic teachers	0.48	n.a.	593

Source: School rosters.

Note: Two-year retention impacts include data from cohort 1 only. Data from cohort 1 districts is from 2010–11 and 2011–12.

[a]Statistical significance of difference between transfer-stipend teachers and all others, and difference between retention stipend teachers and all others.

n.a. = not applicable

E. Sample-Size Tables

In this section, we show sample-size tables that correspond to tables presented in Chapter VI and throughout Appendix G.

Table G.10. Sample Sizes for Table VI.1, Table G.1, Table G.2, and Table G.5

	Number of Treatment Teachers	Number of Control Teachers	Number of Treatment Teams	Number of Control Teams
Cohort 1				
All teachers on study teams	260	238	64	60
Focal teachers	80	96	64	60
Nonfocal teachers	193	183	59	58
Cohorts 1 and 2				
All teachers on study teams	389	336	85	80
Focal teachers	102	128	85	80
Nonfocal teachers	300	259	78	76

Source: School rosters.

Note: There are fewer teams in the nonfocal-teacher analysis because there are 11 teams for which all teachers on the team in the study sample are classified as focal teachers under the selective definition.

Table G.11. Sample Sizes for Table VI.2

	Number of Treatment Teachers	Number of Control Teachers	Number of Treatment Teams	Number of Control Teams
Elementary	69	85	53	51
Middle School	33	43	32	29
Middle school math	16	26	16	14
Middle school ELA	17	17	16	15

Source: School rosters.

Table G.12. Sample Sizes for Table G.3

	Number of Treatment Teachers	Number of Control Teachers	Number of Treatment Teams	Number of Control Teams
One-Year Retention				
Benchmark Sample				
All teachers on study teams	389	336	85	80
Focal teachers	102	128	85	80
Nonfocal teachers	300	259	78	76
Nonclosure Sample				
All teachers on study teams	363	310	78	75
Focal teachers	95	118	78	75
Nonfocal teachers	281	240	72	71
Two-Year Retention				
Benchmark Sample				
All teachers on study teams	260	238	64	60
Focal teachers	80	96	64	60
Nonfocal teachers	193	183	59	58
Nonclosure Sample				
All teachers on study teams	237	217	57	54
Focal teachers	73	86	57	54
Nonfocal teachers	176	169	53	52

Source: School rosters.

Table G.13. Sample Sizes for Table G.4

	Number of Treatment Teachers	Number of Control Teachers	Number of Treatment Teams	Number of Control Teams
One-Year Retention				
Inclusive focal-teacher sample	102	128	85	80
Selective focal-teacher sample	89	77	82	64
Two-Year Retention				
Inclusive focal-teacher sample	80	96	64	60
Selective focal-teacher sample	67	55	61	45

Source: School rosters.

Table G.14. Sample Sizes for Table G.6

	Number of Treatment Teachers	Number of Control Teachers	Number of Treatment Teams	Number of Control Teams
Elementary				
All teachers on study teams	220	201	53	51
Focal teachers	69	85	53	51
Nonfocal teachers	163	153	47	47
Middle School ELA and Math				
All teachers on study teams	169	135	32	29
Focal teachers	43	32	32	29
Nonfocal teachers	137	106	31	29
Middle School Math				
All teachers on study teams	96	65	19	17
Focal teachers	16	26	19	17
Nonfocal teachers	80	53	16	14
Middle School ELA				
All teachers on study teams	81	76	19	16
Focal teachers	17	17	16	15
Nonfocal teachers	65	59	18	16

Source: School rosters.

Note: There are fewer teams in the middle school ELA focal analysis than the team analysis because several middle school math teams include at least one teacher who teaches both math and ELA. These teachers are not focal teachers, but are included in both the ELA and math-subgroup analyses.

Table G.15. Sample Sizes for Table G.7

	Number of Treatment Teachers	Number of Control Teachers	Number of Treatment Teams	Number of Control Teams
Elementary				
All teachers on study teams	204	187	47	44
Focal teachers	62	78	47	44
Nonfocal teachers	154	146	43	42
Middle School ELA and Math				
All teachers on study teams	56	51	17	16
Focal teachers	18	18	17	16
Nonfocal teachers	39	37	16	16
Middle School Math				
All teachers on study teams	27	18	8	7
Focal teachers	7	7	7	6
Nonfocal teachers	20	15	8	7
Middle School ELA				
All teachers on study teams	33	36	11	11
Focal teachers	11	11	10	10
Nonfocal teachers	23	25	12	11

Source: School rosters.

Note: There are fewer teams in the subject-specific middle school focal analyses than the team analyses because several middle school teams include at least one teacher who teaches both math and ELA. These teachers are not focal teachers, but are included in both the ELA and math-subgroup analyses.

PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA