Working Paper

Group Work Is Not Cooperative Learning

An In-Depth Look at the 2014-2015 Academic Year

A Working Paper from the Investing in Innovation (i3) Evaluation

Shelley Rappaport
Jean Grossman
Ivonne Garcia
Pei Zhu
Osvaldo Avila
Kelly Granito

with

Deni Chen
Ashley Kennedy
Joseph Quinn

June 2017

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

Funding for this report came from the U.S. Department of Education under its Investing in Innovation (i3) initiative. The i3 grant called for the Success for All Foundation to scale up PowerTeaching, its middle school math initiative and for MDRC to conduct an independent evaluation of the implementation and impacts of that expansion.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, Charles and Lynn Schusterman Family Foundation, The Edna McConnell Clark Foundation, Ford Foundation, The George Gund Foundation, Daniel and Corinne Goldman, The Harry and Jeanette Weinberg Foundation, Inc., The JPB Foundation, The Joyce Foundation, The Kresge Foundation, Laura and John Arnold Foundation, Sandler Foundation, and The Starr Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

For information about MDRC and copies of our publications, see our website: www.mdrc.org.

Abstract

PowerTeaching is an evidence-based, structured cooperative learning program. It aims to prepare students to meet the stringent demands of new state math standards, both the knowledge standards and the 21st-century skills standards, such as around communication and collaboration. Awarded a scale-up grant in 2012 by the federal government, the program expanded to more than 100 high-need middle schools to train their math teachers how to implement cooperative learning in their classrooms. The evaluation of this expansion effort found that while program teachers learned to organize their students into long-standing mixed-ability groups, which are conducive of cooperative learning, teachers did not consistently implement the instructional supports necessary to transform group work into cooperative learning. The experimental impact study found that students' math performance did not differ dramatically between program and non-program schools. A likely cause for the weak implementation was that the ongoing professional development that is an integral part of PowerTeaching mostly did not occur or was more focused on how to teach the new content required by recently adopted education standards rather than how to implement cooperative learning. The impacts were not large and significant (as earlier studies had found), in part because students in the control schools were also working in groups in response to the communication and collaboration standards embedded in the new standards. Indeed, this study is one of the first to document this shift toward group work. Given that prior studies have shown that cooperative learning teams increase the academic performance of students more than group work does, the evaluation points to the potential for increasing math performance if teachers, who are now routinely using groups, can be trained to consistently provide the supports needed to transform groups into cooperative learning teams. This type of ongoing support, missing in this trial of PowerTeaching, is likely to be key to shifting teachers' instructional practices.

Contents

Abstract iii
List of Exhibits vii
Acknowledgments ix

Chapter

1 Introduction 1
The Value of Cooperative Learning 2
PowerTeaching 4
Overview of the PowerTeaching Evaluation 5
Evaluation Challenges 7

2 Study Design and Study Sample 9
Study Design and Research Questions 9
Recruitment and Sample Characteristics 10
Random Assignment, Analysis Sample, Baseline Equivalence, and Methods 12
Analytical Approaches 15
Exploratory Analyses 16

3 Implementation of the PTi3 Program 17
Were the Key Structures of PTi3 Implemented? 17
What Did Implementation of the Math Program Look Like in Study Schools? 18
Implementation of the Key Instructional Elements of PowerTeaching 20
Were Cooperative Learning Teams Created? 23
Summary of the Implementation of the Key Elements of Effective Cooperative Learning 27

4 Impact of the Success for All PowerTeaching Program 31
Estimation Method, Outcome Measures, and Description of Impact Tables 31
Impact Findings for the Confirmatory Analysis Sample 32
Impact Findings for Student Subgroups 33

5 Scale-Up 35
Recruitment Goals 35
Recruitment and Expansion Strategies 36
Costs: Economies of Scale 36
Implementation Fidelity 37

6 Conclusion 39

Appendix A 43

References 51

List of Exhibits

Table

2.1	Background Characteristics for Schools in the Study Sample, Schools in the PowerTeaching Scale-Up Sample, and Similar Schools in the National Population (2012-2013 Academic Year)	57
2.2	Confirmatory Impact Analysis Sample, by Treatment Status (Spring 2015)	59
2.3	Background Characteristics for Study Sample Schools, by Treatment Status	60
2.4	Background Characteristics for 6th-Graders Students in Confirmatory Analysis Sample, by Treatment Status	61
2.5	Teacher Background Characteristics (2014-2015 Academic Year)	63
3.1	School Achievement Snapshot Scores for Items Related to Schoolwide Structures and Instructional Practices, Study Schools (2014-2015 and 2015-2016 Academic Years)	64
3.2	Program-Control Group Comparisons Related to Support Received by Teachers (2014-2015 Academic Year)	65
3.3	Program-Control Group Comparisons Related to Student Teams (2014-2015 Academic Year)	67
3.4	Impact of PowerTeaching on Minutes Students Spent Doing Group Activities During the Math Block (Activities Are Not Mutually Exclusive)	69
4.1	Impact of PowerTeaching on Student Math Achievement for 6th-Graders in Analysis Sample (2014-2015 Academic Year), by Full Sample and Cohort	71
4.2	Impact of PowerTeaching on Student Math Achievement for 6th-Graders in Analysis Sample (2014-2015 Academic Group), by Student Subgroup	72
A.1	Math State Test Information, by District and Grade	73
A.2	Background Characteristics for 7th-Graders in Analysis Sample, by Treatment Status	74
A.3	Background Characteristics for 8th-Graders in Analysis Sample, by Treatment Status	76
A.4	Impact of PowerTeaching on Minutes Students Spent Doing Group Activities During the Math Block (Activities Are Not Mutually Exclusive), by Student Rank	77

A.5 Impact of PowerTeaching on Student Math Achievement in Grade 7 for Analysis Sample (2014-2015 Academic Year), by Full Sample and Cohort 79

A.6 Impact of PowerTeaching on Student Math Achievement in Grade 8 for Analysis Sample (2014-2015 Academic Year), by Full Sample and Cohort 80

A.7 Impact of PowerTeaching on Student Math Achievement in Grades 6-8 (2014-2015 Academic Year), Robustness Checks 81

A.8 Impact of PowerTeaching on Student Math Achievement in Grades 6-8 for Analysis Sample (2014-2015 Academic Year), Stable Samples 83

A.9 Impact of PowerTeaching on Student Math Achievement in Grade 7 for Analysis Sample (2014-2015 Academic Year), by Student Subgroup 85

A.10 Impact of PowerTeaching on Student Math Achievement in Grade 8 for Analysis Sample (2014-2015 Academic Year), by Student Subgroup 86

Figure

1.1 Logic Model for the Success for All Math PowerTeaching Program in Middle Schools 88

2.1 Timeline and Sample Configuration of the Evaluation 89

3.1 School Achievement Snapshot Scores, by Category and School Year, Study Schools Only 90

3.2 Average Minutes Students Spent During Math Block, by Collaboration Type 91

5.1 2015-2016 School Achievement Snapshot Scores, by Category, Scale-Up Schools Only 92

Box

1.1 Cooperative Learning in Action: Becoming a Super Team 93

Acknowledgments

The evaluation of the Investing in Innovation (i3) scale-up initiative of the Success for All Foundation (SFAF) middle school PowerTeaching program, and this resulting report, reflect the efforts of many people. Our first debt of gratitude is to the principals, SFAF facilitators, research liaisons, and teachers at the schools that participated in the study; to staff members in the central offices of the five school districts who provided us with critical student records data as well as staffing information; and to district coaches who took part in the interviews. The assistance and cooperation of these individuals were vital for enabling the study to go forward and for providing the rich and detailed information on which this report is based.

At Old Dominion University, John Nunnery and Pamela Arnold-Puchalski and at the Success for All Foundation, Nancy Madden and Paul Miller responded promptly and patiently to our many information requests. In particular, Paul Miller spent countless hours giving us a strong understanding of the PowerTeaching program. John Nunnery, Pamela Arnold-Puchalski, Nancy Madden, Robert Slavin, and Paul Miller provided useful critiques at each stage of the report drafting process.

Thomas J. Smith, Melissa Comerchero, Zach Pinto, Amanda Ferrandino, and Cammie Brown played an active role in the field research data collection effort. Melissa Comerchero, Zach Pinto, and Amanda Ferrandino also participated in the coding effort. Helping the team with the principal and teacher surveys as well as the teacher logs were Zach Pinto, Seth Muzzy, Bulent Can, Sandhya Bandhi, Usha Krishnan, and Matthew Au. Jillian Verillo helped with the preparation of this report.

Fred Doolittle ensured that the team received both material and moral support at every turn. He, along with Janet Quint, Marie-Andree Somers, and Robert Ivry, carefully reviewed earlier drafts of this report and made comments that improved the final product. Mario Flecha handled a variety of administrative and other tasks. Kelly Granito kept the project on task and on budget and both Gina Price and Alpesh Shah provided fiscal oversight. Rebecca Bender and Christopher Boland edited the paper and Carolyn Thomas prepared it for publication.

The Authors

Chapter 1

Introduction

Changes in information technology over the last several decades have enabled the creation of the modern world economy (Grossman and Rossi-Hansburg, 2008). Now that companies communicate instantaneously with individuals far away from their physical location, they can more easily outsource and offshore many tasks once done locally. Thus, jobs once done in the United States are moving to countries that have cheaper wages. While this means U.S. customers are reaping the benefits of less expensive goods (Fajgelbaum and Khandelwal, 2016), the set of jobs for which U.S. youth need to prepare has changed.

The jobs of the future will be jobs that cannot be performed abroad — such as health care — and jobs where U.S. firms have a competitive advantage. The fastest growing occupations in the next decade are projected to be in science, technology, engineering, and math (STEM) areas that require advanced mathematical and scientific knowledge (Hanushek, Peterson, and Woessmann, 2010). Unfortunately, given the low level of math achievement among many U.S. students today, especially those in low-income schools, many students will not be prepared to take these STEM jobs. If students in Title I schools are to be given a chance to compete in this new economic arena, math instruction needs to change.

Research has found that math performance also seems to function as a filter, even beyond general academic ability, for many career outcomes since a number of prestigious career pathways (even those that do not involve math) require that a student has completed high school or college math prerequisites (Sherman, 1982). Shapka, Domene, and Keating (2006) found that, in particular, math achievement at the start of high school has a significant effect on students' career aspirations and the courses they choose to take, even controlling for overall achievement. Thus, middle school is a critical time to shore up and strengthen math ability, so that students feel comfortable taking the higher-level high school math courses they will need to pursue many attractive careers, including in medicine, engineering, and computing technology.

To improve the math abilities of students in low-income schools, math instruction at all levels should improve, but middle school is an especially critical period. While math achievement among U.S. students has increased since 1999, a large drop-off remains between elementary and middle school. According to the 2015 Nation's Report Card, the percentage of students achieving proficiency in mathematics dropped from 40 percent to 33 percent between fourth and eighth grades.[1] Anderman, Maehr, and Midgley (1999) and many others have shown that

[1]The Nation's Report Card (2017).

student engagement and grades in general significantly decline in middle school. Much research has been done to understand why so many middle school students start disengaging from school in different subjects. The most commonly accepted explanation relates to the large change in instructional practices that occurs when students move from elementary to middle school and the mismatch between the students' developmental needs and instruction. Developmentally, middle school students (early adolescents) are changing rapidly — physically and psychologically. They increasingly desire more autonomy and control over their actions (for example, Steinberg, 1990), are focusing more on peers and social acceptance (Juvonen, 2006; Wigfield, Byrnes, and Eccles, 2006), and are growing cognitively. However, while in elementary schools there is an emphasis on flexible, small-group work, in middle school many teachers emphasize whole-group instruction accompanied with a focus on individual competition and sorting by perceived aptitude (Midgley, Anderman, and Hicks, 1995). This shift creates the expectation among students and teachers that only a select few have the ability to understand and apply mathematical concepts and reinforces students' disengagement with math.

The Value of Cooperative Learning

A number of studies have documented the positive effects on achievement of cooperative learning in K-12 settings (e.g., Johnson and Johnson, 1994; Slavin, 1996a, 1996b). Indeed, the impact appears to be strongest among students who are not the highest achievers.[2] Given that the role of peers — peer acceptance and peer norms — becomes relatively more important in middle school (Juvonen and Wentzel, 1996), cooperative learning practices may be particularly powerful instructional tools at this age. In a recent meta-analysis of middle school achievement, the quality of peer relationships was found to account for 33 percent to 40 percent of the variance in achievement of middle school students (Johnson, Johnson, and Roseth, 2010). Spurred by this research, over the last 25 to 30 years, cooperative learning practices (structured group and pair activities) have greatly spread throughout U.S. schools (National Mathematics Advisory Panel, 2008).

Many cooperative learning practices have been developed and formalized (Jigsaw, Student Team-Achievement Divisions, Kagan, Learning Together, think-pair-share, Team-Games-Tournaments, and so on). However, most teachers do not use cooperative learning as a core *strategy* (Pianta et al., 2007), nor do they implement the practices in such a way as to create an effective cooperative learning setting. Studies have found that while many teachers claim to be

[2]Slavin's analysis of cooperative learning research indicated that students at different performance levels benefit similarly from cooperative learning (Slavin, 1996b). However, other researchers (Hampton and Grudnitski, 1996; Kenneth, Young, and Berrill, 1999; Stockdale and Williams, 2004) have shown that high achievers are less likely to benefit from cooperative learning than are low achievers. Stockdale and Williams (2004) reported that high achievers actually declined slightly under cooperative learning.

using cooperative learning, most "cooperative learning activities" in mathematics generally consist of unstructured group work, with no group goals and little individual accountability (Hiebert and Wearne, 1993; Stein, Grover, and Henningsen, 1996). This tends to lead to simply sharing answers without explaining how a student derived the answers, which inhibits learning and effort (Antil et al., 1998; Emmer and Gerwels, 2002; Webb, 2008).

For cooperative learning activities to improve academic achievement, research shows that two key elements are needed: positive interdependence, that is, a group *learning* goal that is achieved only through group effort, and individual accountability, where a student's grade is based solely on his or her own performance (Slavin, 1995; Slavin, Hurley, and Chamberlain, 2003). When these two elements are both in place, cooperative learning produces achievement gains. Johnson and Johnson (2009) identify three other necessary features: group members support and encourage each other (promotive interaction), group members have good small-group skills (such as communication, decision-making, and conflict management skills), and the group discusses how well they are achieving their goals and maintaining group relationships.

When done correctly, cooperative learning strategies align well with the developmental needs of adolescents. Within the group, students have autonomy and control to accomplish the group learning task as they wish. This learning process should be more engaging than listening to a teacher or working individually because it involves peer interaction. Students who have grasped a topic more quickly can explain it to group mates who still do not understand it. The students can build on each other's knowledge of the material to jointly solve a problem or learn a skill. Finally, it should also create a peer norm of academic success, which can be a strong motivator for adolescents.

This study examines the impact of scaling up of a specific well-studied cooperative learning pedagogy, PowerTeaching (PT; formerly known as Student Team-Achievement Division). It is a way of teaching math that can be applied to any math curriculum. Nunnery, Chappell, and Arnold (2013) identified 14 evaluations of this strategy in either primary or secondary schools. The average impact on math test scores was a positive shift of 0.60 of a standard deviation for secondary school students and a 0.13 standard deviation shift for primary school students. The average impact of the studies that met What Works Clearinghouse (WWC) evidence standards was 0.42. Given the strength of this evidence, the U.S. Department of Education awarded an Investment in Innovation (i3) grant to Old Dominion University (ODU) and the Success for All Foundation (SFAF) to fund scaling up PT so that it would reach an additional 130,000 students over five years. In the next section, the PowerTeaching (PTi3) model is described.

PowerTeaching

The PowerTeaching model, developed by SFAF, has been refined based on research for over 25 years. Its structures are intended to provide the scaffolding to ensure that teachers learn how to incorporate cooperative learning techniques into their instructional practices such that the critical interlocking elements needed for effective cooperative learning are in place.

Figure 1.1 presents the logic model underlying the PTi3 model. SFAF recruiters find school districts that are interested in adopting PTi3. The leadership in middle schools that will receive PTi3 has to commit to supporting the program for three years and the school must provide a part-time school-based math facilitator for each school. For this scale-up project,[3] SFAF modified its training component to a train-the-trainer model. In the i3 version of the PT model, SFAF provides training to these school-level facilitators, who in turn train the math teachers in the PTi3 schools. Because it takes time to understand how to do cooperative learning effectively, facilitators and teachers are expected to meet biweekly to learn and improve upon cooperative learning practices, discuss classroom challenges, and review progress data. SFAF also provides the program's supporting material to all participating schools.

This training component of the PTi3 program was planned to differ from the earlier version of the model, where SFAF staff trained and provided ongoing support to the teachers on-site. It was hoped that a model in which much of the support was provided electronically by SFAF would reduce the amount of time SFAF coaches had to spend traveling to train and support the schools, as well as lower costs, while still allowing the schools to implement the program with fidelity.

If the training and ongoing support are adequately done, it is expected that the teachers will be able to incorporate PTi3's key practices into their math classes. That is, they will provide structured cooperative learning opportunities to students in long-standing heterogeneous skill groups that feature the three key elements[4] of effective cooperative learning.

- Team buy-in: Teams earn recognition daily when team members do well.

[3]There were three different levels of funding available through i3. Scale-up grants supported programs that (1) had already shown, through previous research, that they had positive impacts on students and (2) had already been implemented somewhat widely. Programs receiving scale-up funds had to be rigorously researched and use much of their funds to bring the program to scale; thus, scale-up grantees received the highest level of funding in order to bring on many new sites. For PTi3, there were pilot and scale-up schools in addition to the study schools.

[4]SFAF identifies the three central concepts as team recognition, equal opportunities for success, and individual accountability. The terms have been changed slightly in this report to be more straightforward for a reader unfamiliar with the program.

- Equal opportunities for all students to help the team: All team members, no matter their ability, can contribute to the team goals by improving on their past performance.
- Team interdependence: Team success depends on the individual learning of all team members, while an individual's grade depends only on his or her own performance.

Slavin (1995), reviewing the literature, concludes that group study alone has no effect on achievement. Student achievement is improved by cooperative learning opportunities only if students earn recognition and rewards by achieving group learning goals. Students care about how their teams perform, but individuals' grades depend on students' own academic performance. To operationalize these principles in PTi3, *team buy-in* is achieved by publicly awarding and posting team points when teams successfully achieve a learning or behavioral goal. These points enable teams to achieve "good," "great," or "super" team status. To ensure that all team members feel they can contribute — that is, providing *equal opportunities for all students to help the team* — the team goals must incorporate the learning of all group members, such as "all team members hand in their homework" or "all team members will score at least one point higher on the next quiz." But the lynchpin that makes PTi3 effective at increasing achievement is the combination of team buy-in and equal opportunities for all students to help the team with *team interdependence*. PTi3 creates interdependence by making the group's success depend on the answer of a randomly chosen individual in the group. This mechanism works when students do not want to "let the team down" and understand that they may be randomly selected to answer a question on behalf of their team. When these three essential elements of PTi3 are simultaneously in place, prior research cited above has shown that PTi3 increases academic performance. Box 1.1 provides an example of what the combination of the key elements could look like in the classroom.

Overview of the PowerTeaching Evaluation

In 2012, the PTi3 program staff began their four-year effort recruiting middle schools to "scale up" PTi3 and serve students; ultimately, 106 schools joined. A small pilot study was conducted in school year (SY) 2012-2013, with the study schools starting operations in SY 2013-2014. As part of the i3 grant, MDRC studied the effort, conducting (a) a school-level randomized controlled trial with the first 58 non-pilot recruited schools to determine PTi3's impact, (b) an implementation study to examine how PTi3 worked, and (c) a scale-up study to examine if and how SFAF and ODU were able to reach their scale-up goals.

Chapter 2 provides much more detail about the study and its samples. Briefly, the study schools entered the study in two cohorts — 24 in SY 2013-2014 and 34 in SY 2014-2015 — across five districts. Schools were randomized by district.

Key Research Question

The main confirmation impact question of the study is this:

- What are the impacts of PTi3 on sixth-grade students' math performance on the high-stakes math tests after one year of exposure to the program?

The confirmatory impacts for this report are based on sixth-graders who entered the program schools in SY 2014-2015 and experienced math taught according to the PTi3 model for one year. Unfortunately, data availability issues made it impossible to report on longer-term findings for the full sample.

In addition to the main confirmatory question, this report addresses exploratory questions intended to deepen the understanding of the overall average impact of PTi3. For example, it examines how impacts vary by baseline math ability, both for those who experienced PTi3 in sixth and seventh grades and for the seventh and eighth-graders who started PTi3 after sixth grade. In particular, the report examines the following:

a. What is the two-year impact on students in the first cohort who could have received PTi3 math instruction in sixth and seventh grade? (These are presented as exploratory because of the smaller, geographically constrained nature of that sample.)

b. Does PTi3 produce impacts on math achievements for seventh- or eighth-graders in the sample schools in general?

c. How do the impacts of PTi3 differ by cohort (perhaps reflecting changes in the scaling process as SFAF gained more experience)?

d. How do the impacts of PTi3 differ for students with below-proficient math pretest scores versus those with at- or above-proficient scores? For students of low socioeconomic status? For limited English proficient students? For students of various ethnic backgrounds? For boys and girls?

e. Does PTi3 produce greater impacts for students who are exposed to a longer period of PTi3 — that is, a "stable sample" of sixth-graders who remained in the PTi3 schools for the entire 2014-2015 school year and have both pre- and post-test measures?

Some of these exploratory analyses are presented fully in the text, while others are presented in Appendix A. Each state's standardized assessment of math skills was used to gauge the impacts on math achievement.

In addition to the impact questions, the study uses quantitative and qualitative data from a wide variety of sources to investigate how PTi3 was implemented, what program students' experiences in math class were, and how these experiences differed from those of control students.

Data sources include teacher surveys, implementation summaries completed by the coaches and SFAF staff, teachers' instructional logs, and interviews and focus groups conducted in the course of site visits with school personnel. Using this information, the study addresses three main questions:

- To what extent were PTi3's structures implemented during the 2014-2015 school year, and what factors were associated with more complete implementation?
- What were classroom and instructional practices in the program schools, and how did they differ from those in control schools?
- What did SFAF do to recruit the scale-up districts and schools, and how did these schools and their implementation of PTi3 compare with the study program schools?

The answers to these implementation questions are useful for understanding the type of support teachers and schools may need to implement the program. They are also helpful for generalizing the types of effects seen in this evaluation to other schools — the other scale-up schools and districts that might adopt the program in the future.

Evaluation Challenges

Most multiyear evaluations face issues that impinge on the evaluators' original plans. This evaluation is no different. The biggest challenge the evaluators had to deal with was the impact the Common Core State Standards (CCSS) were having on the study districts. California, New York, and Florida adopted the standards with modifications, while Illinois adopted them verbatim. One major implication of the Common Core was to dramatically increase the amount of information that students were expected to have mastered at each grade level. In addition, students were required to be able to not just give the correct answer but to explain and support how they came to their conclusion. To achieve these new learning goals, district and school leaders expected teachers to try new instructional practices and cover new material. The

teachers in the study schools, as will be described later, were not just being asked to adopt new PTi3 instructional practices, but the content they were expected to cover was changing. Thus, the educational environment the schools and teachers had as they tried to adopt PTi3 during the study period may not be representative of what would happen in a less stressful time.

Indeed, the press the study districts felt from the Common Core was a major topic of conversation during site visits. To be responsive to the district, over the first two years of the study, SFAF developed and offered the study schools math lessons that were consistent with CCSS grade-level standards and could be used throughout the school year. Not every district, school, or teacher used SFAF's lesson resources as a curriculum replacement, but many did. Thus, the study is about the first year's impact not just of a new teaching method, but also, in many classrooms, of adopting the PTi3 "curriculum."

Another challenge the CCSS posed for the evaluation related to data. To be aligned with the changing academic expectations of schools, standardized tests were being totally revamped. New tests were created that would be compatible with the Common Core standards or states' new standards, and thus would test a different set of skills than the previous standardized tests did. These new tests were being piloted and used by districts for the first time as the evaluation was unfolding. For example, the state of California piloted the new Smarter Balanced test in SY 2013-2014. The districts would not release these data since they were intended only for piloting purposes, to allow the test developers to refine the questions. The evaluation design called for the impacts to be based on state standardized math tests, yet no data were available in California in the 2013-2014 school year. As all of the Cohort 1 schools were in California, no information was available for first-year impacts. Chapter 2 discusses these issues in more detail.

Thus, in sum, the educational setting in which the study took place was challenging both for the schools and for the evaluators. However, this study sheds light on key issues that show how schools and teachers begin to integrate cooperative learning practices into their classrooms.

Chapter 2

Study Design and Study Sample

This chapter first summarizes the basic study design and the recruitment and random assignment processes. It then describes the characteristics of the evaluation schools and their students at the beginning of the study in order to compare these study schools with the broader group of Investment in Innovation (i3) scale-up schools and with schools nationally that serve low-income children. It also establishes that program and control schools were, as intended, similar to each other. The chapter concludes with a brief description of the analytic approach and related issues for impact evaluation.

Study Design and Research Questions

This study uses an experimental design that randomly assigns schools within each random assignment block (usually a school district or a subset of schools within a district) to either the program condition or the control condition. Under the program condition, all math teachers and their students in Grades 6 through 8 were provided training and support for the i3 PowerTeaching (PTi3) program; they could also continue to participate in business as usual. Under the control condition, the Grades 6-8 math teachers and students received the trainings and supports that they would have received in the absence of the study — that is, business as usual. The difference in outcomes between the program schools and the control schools can be interpreted as the effect of the PowerTeaching program relative to "business as usual" in each participating district.

To support this design, the study team recruited 58 middle schools in five school districts around the country to participate in the study. These schools were recruited in two cohorts: 24 Cohort 1 schools that started the program in SY 2013-2014 and 34 Cohort 2 schools starting in SY 2014-2015. Figure 2.1 demonstrates the timeline of the study for each of the cohorts.

For the 24 Cohort 1 schools, SY 2012-2013 is the baseline year and SY 2013-2014 and SY 2014-2015 are the first and second implementation years, respectively. For the 34 schools in Cohort 2, on the other hand, SY 2013-2014 is the baseline year, and SY 2014-2015 is the first and only implementation year by the time of the report. It would be ideal to pool information from the first implementation year across the two cohorts to assess the program impact after one year of implementation for the largest sample possible. However, the outcome measure — state math test scores — is not available for Cohort 1 schools for SY 2013-2014 due to a transition of state tests during that year. Given the timeline of the study, the only viable option that allows assessment of the program impact for all sample schools (and therefore has adequate statistical

power to detect meaningful program effect) is to pool information from both cohorts for SY 2014-2015 and estimate program impact on the sample for that year.

One drawback of this approach is that the estimated impact reflects mixed program maturity for the two cohorts of schools, with two years of implementation for Cohort 1 and one year of implementation for Cohort 2. This is an important caveat to keep in mind when interpreting the full sample findings. Impact findings by cohort are presented to provide information on the effect of program maturity. However, these results are based on small samples and therefore should be viewed as exploratory only.

Another potential complication for the interpretation of the impact findings is the amount of program exposure students experience. For example, seventh and eighth-graders in Cohort 1 schools could have had a maximum of two years of program exposure by the spring of 2015, while the students in Cohort 2 schools would have only one year of program exposure. The only grade level that has a consistent amount of potential program exposure is Grade 6, where all students will have had just one year of possible exposure to the program. Therefore, to minimize the complication of the interpretation, the study focuses on the following confirmatory research question:

> What are the effects of PTi3 on an average treatment school with respect to the math achievement of its sixth-graders after one year of exposure to the program?

This confirmatory analysis is based on district-administered spring math achievement test results of all sixth-graders who were not in self-contained classes in the sample schools in the spring of SY 2014-2015. These sixth-graders theoretically have the same maximum amount of exposure to PTi3 (one year, in sixth grade), but the maturity of the program they were exposed to will differ slightly: Students in Cohort 1 were exposed to the PTi3 program in its second year of implementation, while students in Cohort 2 were exposed in its first year of implementation.

In addition to the key research question, the study also explores the effects of PTi3 on subgroups of schools and students to see if the program impact on students' math achievement varies across different dimensions.

Recruitment and Sample Characteristics

The recruitment for the evaluation was conducted by the Success for All Foundation (SFAF) and occurred as part of the general outreach to schools, districts, and states for the i3 scale-up grant. Each school had to meet the following eligibility criteria: It had to serve students in

Grades 6 through 8, it had to be eligible for Title I status,[1] it had to be willing to comply with the study's data request, and at least 75 percent of its teachers had to vote to adopt the PTi3 program.

Recruitment for the study was initially a challenge because districts that were interested in the program often wanted the benefits for their entire school population as soon as possible and were reluctant to identify schools to serve as control schools. Additionally, although including demographic diversity, and specifically rural schools, was a core feature of the project, it was a unique challenge. Rural school districts often do not have multiple middle schools to serve as program and control sites. The initial round of recruitment efforts produced a sample of schools (Cohort 1 schools) that was not adequate for the impact study. The recruitment timeline was extended and more districts and schools were recruited for the following fall through a second round of efforts (Cohort 2 schools).

At the end of the recruitment phase, five school districts in four states agreed to participate in the study. The number of study schools provided by each district ranged between 4 and 24, producing a total sample of 58 schools.

The 58 study schools were predominantly from urban settings such as cities, large towns, and urban fringe, as shown in the first column in Table 2.1. More than half of the study schools were from the western region of the United States. The average rate of students eligible for free or reduced-price lunch was 72 percent, and 91 percent of the schools were designated Title I schools in SY 2012-2013. On average, across the sample schools, 11 percent of all students were White non-Hispanic, 31 percent were Black non-Hispanic, and 51 percent were Hispanic. Most schools served only students in Grades 6 through 8, with a few exceptions. There were about 52 full-time equivalent teachers in the study schools on average.

To illustrate the extent to which the recruitment process affected the basic characteristics of the sample, Table 2.1 provides a comparison of the study sample with all 71 schools in the scale-up sample and with a national sample of middle schools that serve students in Grades 6 through 8 and are designated Title I schools.[2]

To be recruited, scale-up schools had to serve a large fraction (more than 50 percent) of minority students or students who were eligible for free or reduced-price lunch in Grades 6 through 8. Compared with the scale-up sample, the study schools had a similar proportion of students eligible for free and reduced-price lunch and were equally likely to be designated Title

[1]One of the recruited study schools was serving a high proportion of low-income students but was not designated as a Title I school in the baseline year in the Common Core of Data (CCD. It became a Title I school in the subsequent year.

[2]Schools randomly assigned to the program condition are considered part of the SFAF scale-up sample.

I schools. However, the study schools differed from the scale-up schools in their geographical location, with more than half of the study schools in the western region, and a majority of them in urban areas. The study schools were also larger than the scale-up schools on average, in terms of both total student enrollment and the number of full-time employees in the school. Finally, the study schools had more Hispanic and Black students and fewer White non-Hispanic students than the scale-up sample schools. This is reflective of the fact that most of the study schools were located in large urban areas.

Similar patterns exist when the study schools are compared with a national sample of middle schools that serve students in Grades 6 through 8 and are designated Title I schools. Overall, the study schools differed from the national sample in most dimensions measured in Table 2.1.

Specifically, the study schools were more likely to be located in the West and in urban areas and to have more low-income and more Hispanic students. They also tended to be larger than the national sample of schools.

Random Assignment, Analysis Sample, Baseline Equivalence, and Methods

In the spring of 2013, 24 Cohort 1 schools participating in the study were randomly assigned to the program or the control condition within each of two study districts. The following spring, the 34 Cohort 2 schools were randomly assigned to the program or the control condition within each of four study districts.[3] To ensure that schools with particular characteristics (such as academic achievement) were equally represented in the program and control conditions, schools in some districts were grouped into two or three blocks with similar characteristics within each district, and about half of the schools within each block were randomly assigned to the program group. The random assignment process produced 30 program schools and 28 control schools.

Once schools were randomly assigned, all sixth-graders in their regular classes (that is, not in classes for students with special education needs) in these schools in the spring of 2014-2015 became members of the confirmatory student sample. In other words, they constitute the analysis sample for the confirmatory impact analysis. The study school and student samples are broken down into program and control groups in Table 2.3.

The purpose of random assignment is to produce a program group and a control group that are statistically equivalent on all characteristics of these schools at the start of the study. If the two groups are indeed equivalent at the outset, and if any attrition from the sample over the

[3]Six more schools were randomly assigned to one of the Cohort 1 districts.

course of the study is balanced across groups, one can be reasonably confident that any differences in outcomes between the two groups found later are due to the intervention.

Table 2.4 shows that, as intended, random assignment produced groups of schools that were very similar on all observed characteristics at the beginning of the study. There were no statistically significant differences in any school-level baseline characteristics between the program and control groups. In addition to testing for differences in each variable, an F-test for all school-level variables was conducted to see if there were any overall differences in baseline school characteristics between the two groups of schools.[4] The results indicate no such differences in school characteristics.

Using the demographic data received from students' district records, as well as baseline test scores, Table 2.4 presents the characteristics for students in the confirmatory analysis sample.

Results in the table shows that random assignment was fairly successful in creating two equivalent research groups at baseline. Overall, the students in the program group and control group were very similar to each other across a wide range of student background characteristics. The only exceptions were that the program group seemed to have about 3 percent fewer Black, non-Hispanic students and about 3 percent fewer male students than the control group. It is worth noting that 9 out of the 12 estimated baseline differences between the program and control groups are less than 0.05 standard deviations in effect size, which are considered as satisfying baseline equivalence by the What Works Clearinghouse standard (Institute of Education Sciences, 2014). The other three estimated baseline differences were between 0.05 and 0.09 standard deviations in absolute effect size, also considered satisfying baseline equivalence as long as they are controlled for in the impact estimation (Institute of Education Sciences, 2014).

Similar checks on baseline equivalence were carried out for the seventh- and eighth-grade student sample. There were no statistically significant differences between the program and control groups for these samples. See Appendix A for details.

In addition to these baseline equivalence checks, Table 2.5 presents the average educational attainment, certification, and teaching experience for the math teachers in the program and control schools based on responses from the teacher surveys conducted by the study in the spring of 2015. The data were not collected at baseline; therefore, results reported in this table should be viewed as a description of the teacher characteristics in the study schools during the

[4]This test was based on a logistic regression, predicting program status with the measured school-level baseline characteristics. The p-value for the F-test is 0.98.

implementation year rather than a baseline equivalence test, even though the variables examined in the table are likely to be time-invariant.

Results in the table show there was no significant difference between teachers in the program schools and those in the control schools in terms of these characteristics. In particular, the table shows that, on average, the math teachers in the study schools had postsecondary degrees, they were certified to teach both math and middle school, and they had more than 10 years of teaching experience, with about eight years of experience in teaching middle school math.

In order to understand the similarities and differences in math instruction that program group and control group students received, math teachers in both groups of schools were asked to complete **instructional logs**. The log is a close-ended instrument that has been shown in research to differentiate effectively between math instruction programs. Over a two-week period in the spring of 2015, teachers were asked to fill out a log for one student each day for up to eight randomly selected middle school students (1,567 for the PowerTeaching schools and 1,374 for the control schools).

Teacher surveys were administered at the schools during the spring of 2015. The survey data yielded information about teachers' backgrounds, team work in classrooms, and teachers' experiences and perceptions, especially as these related to math instruction and the school environment. The teacher surveys, in conjunction with the School Achievement Snapshots, discussed below, were intended to provide valuable insights into the conditions under which effective and faithful implementation of the program model was most likely to occur.

The **School Achievement Snapshot** is a rubric used by SFAF point coaches and school facilitators to assess program implementation and to guide schools in a continuous improvement process. When PowerTeaching coaches visit the schools, they meet with school personnel, visit classrooms, and examine program documents; they then use this information to complete the Snapshot, once per quarter if possible but at least at the end of each school year. In filling out the form, the coach rates the extent to which each school manifests a wide array of program practice. PowerTeaching coaches' ratings of Snapshot items from the end of the school year were then used to calculate implementation scores.

Focus groups were conducted in all the study schools from February to April of 2015. Focus group participants were selected to represent math teachers in sixth to eighth grade who taught grade-level math or algebra I. There were between two and eight teachers per focus group, depending on the size of the school. The main goal of the focus groups was to learn as much as possible about the nature of group work in math classrooms — how groups were composed, what the students were expected to do in them, what teachers' perceptions were of the purpose of group work, and how group work was manifested. Interview protocols were

designed to capture the interconnected nature of the three elements of cooperative learning within PowerTeaching, but questions were phrased using broad, non-program-specific language so that they were also applicable to teachers in control schools.

Analytical Approaches

Given this design, the basic analytic strategy for assessing the impacts of the PTi3 program is to compare outcomes for schools that were randomly assigned to the program group and received the program with those for schools that were randomly assigned to the control group and remained "business as usual." The average outcome in the control group schools represents an estimate of the achievement level that would have been observed in the program group schools if they had not been assigned to the program group. The difference in outcomes between the program and control groups provides an unbiased estimate of the impact of the PTi3 program.[5]

Analytically, the primary impact estimation model is a two-level hierarchical model with students nested within schools. The model uses data from all five study districts in a single analysis, treating random assignment blocks as fixed effects in the model. Separate program impact estimates are obtained for each district and then averaged across the five districts, weighting each district's estimate in proportion to the number of program schools in the district. Findings in this report therefore represent the impact on student math performance in the average program school within the study sample. The results do not necessarily reflect what the program effect would be in the wider population of districts.

Strategy for Dealing with Multiple Comparisons

It is important to recognize the potential problems associated with conducting multiple hypothesis tests, since the more impact estimates are tested, the greater the likelihood of falsely rejecting a true null hypothesis simply by chance.

Consequently, this evaluation has tried to keep the number of outcome measures and subgroups in the study to a minimum. The confirmatory impact analyses focus on a single measure of math achievement (standardized state achievement test scores), and thus no adjustment is needed. Note that impact findings on the probability of students performing at or above the state proficiency level are presented as exploratory.

[5]All impact estimates are based on an intent-to-treat analysis that includes all students in the sample schools at the time of outcome data collection. Therefore, the impact estimates reported here reflect the impact of assignment to the program. See Appendix A for a detailed discussion of the impact estimation model.

Exploratory Analyses

The study also explores other analyses to help readers understand how the impacts may differ. Specifically, the evaluation assesses the program impacts for the following alternative samples.

Stable Sample

To examine the impact of program exposure on students, the study looks at the program impacts for the subset of students who had attended their study school since the beginning of sixth grade and have both pre- and post-test scores, that is, *the stable sample*. These are the students who have had the maximum exposure to PTi3.

Grade 7 and Grade 8 Sample

The report also includes results from Grades 7 and 8 to see if the impacts differ for higher grade levels. The interpretation of these findings is more complicated than that of the confirmatory sample because for these two grades, both program maturity and student exposure differ by cohort (as marked in Figure 2.1). For example, seventh-graders from Cohort 1 schools would have been in the program for a maximum of two years, and by the end of their seventh grade, the program would have been two years old. On the other hand, seventh-graders from Cohort 2 schools would have been exposed to the program for only one year, and that would be the first year of the program implementation.

School-Level Subgroups by Cohort

Given that the study sample consists of two cohorts of schools that started program implementation at different times, it is of interest to look at the program impacts by cohort separately for each of Grades 6, 7, and 8. As illustrated in Figure 2.1, for Cohort 1 schools, the estimated program impacts represent the effects of PTi3 after two years of implementation, while for Cohort 2 schools, they represent the effects of PTi3 after the first year of implementation. Looking at school subgroups by cohort would therefore provide useful information on whether program impact varies with program maturity. Note, however, that these cohort-level estimates have limited statistical power because of the small number of schools in each cohort. Therefore, this exercise is primarily for hypothesis-generating purposes.

Student-Level Subgroups by Baseline Student Characteristics

Finally, the report presents the impacts of PTi3 on subgroups of students within the *confirmatory sample* of sixth-graders who were hypothesized to potentially be differentially affected by PTi3: students whose math achievement differed at baseline, students of different socioeconomic status, students with limited English proficiency at baseline, boys and girls, and students of various ethnic backgrounds.

Chapter 3

Implementation of the PTi3 Program

Chapter 3 focuses on the quality of the Investment in Innovation PowerTeaching (PTi3) implementation in the program schools. The chapter is split into two primary sections. The first describes whether the key structures of PTi3 were in place. In other words, did the Success for All Foundation (SFAF) provide all the training and materials the schools would need to implement the program, and did the schools commit the time and resources required for strong implementation at the classroom level? The second section examines whether the provision of these structures — time, training, and resources — led to the creation of cooperative learning teams in math classrooms.

Were the Key Structures of PTi3 Implemented?

In this section, the following key questions are addressed:

- What are the key structures of PTi3?
- Were the key structures of PTi3 in place?
- Did the program meet the standards for implementation that SFAF established?

The logic model shows that there are three key types of structures needed to place PTi3 into schools:

1. Program developer structures — training and materials from SFAF,
2. School structures — school leaders' commitment to the program, and
3. Continuous improvement processes — ongoing support of teachers implementing the program.

A more detailed explanation of the items within each type of structure is described below:

Program developer structures refer to the training and materials provided by SFAF to the schools. Leaders and school staff receive training about the program from SFAF before the start of the school year. The principal and the school-based math facilitator attend a week-long conference in Baltimore and SFAF trainers conduct two-day trainings at the school sites for math teachers. The materials provided by SFAF include an iPad or the equivalent for teachers to

track data, a flash drive containing program lessons and record-keeping tools, posters, and an apron to carry around program supplies.

School structures refer to schools' commitment to the program, which includes a fully involved principal and a part-time math facilitator. Principals are expected to support the program by organizing the school calendar to include time for program activities; providing a part-time, school-based math facilitator; making resources available; and having regular communication with SFAF coaches and the math facilitator. The math facilitator role includes supporting teachers implementing the program, tracking student progress, and communicating with the principal and SFAF coaches about the program.

Emphasis on *continuous improvement processes* refers to the type of ongoing support that teachers receive in biweekly PTi3 professional development sessions, referred to as "component team meetings." These meetings, which are led by the math facilitator, focus on setting teacher instructional goals relative to the program, monitoring teachers' implementation of the program, and tracking student progress. Data that show what progress has been made are shared and discussed. The math facilitator then shares meeting concerns, questions, and solutions with the SFAF coach for feedback.

What Did Implementation of the Math Program Look Like in Study Schools?

The extent to which these structures were in place was measured quantitatively using an instrument known as the School Achievement Snapshot. (See Appendix A.) The Snapshot is a form created by SFAF to guide schools in a continuous improvement process.[1] To contextualize the implementation ratings, the teacher survey that was fielded in the spring of 2015 to all study school teachers is used here to show what implementation of the math program might have looked like absent the program. To complete the Snapshot, SFAF coaches work with schools and rate the extent to which each school has implemented specific parts of the PTi3 program[2]

[1]A total of eight items were considered across the three categories to yield a score. Two items each were used to rate the program developer and school inputs categories. The remaining four items were used to rate continuous improvement. Five of the items, including both items related to SFAF and both items related to schools' commitment, are considered "priority items" because they are essential to basic program implementation. When implemented with fidelity, each of these five items receives a score of 2. The remaining items receive a score of 1 when implemented. The maximum possible score for implementing the structure categories is 13.

[2]The Snapshot was developed for SFAF coaches to be able to work with individual schools to identify specific implementation goals over time and track schools' progress on these. It was never intended to be used to rate how well PTi3 schools were comparatively implementing the program. As a result, no efforts were

(continued)

on a quarterly basis each year.[3] For example, the Snapshot shows whether or not the principal was fully involved with the program and whether program data were used to set instructional targets. (See Table 3.1 for a list of all the items that together make up the three key implementation structures.) The Snapshot did not measure dosage or quantity at the student level. Instead, the summative measures derived from the Snapshot were interpreted as school-level measures of the extent of program implementation.

Looking first at whether the program developer, SFAF, provided the schools with the training and materials called for in the model, Figure 3.1 shows that SFAF's provision of essential training and program materials to the schools received a high rating. In the 2014-2015 school year, the majority of middle schools received the initial training and all schools received the program materials (Table 3.1), for an overall average rating[4] of 92 percent.

Next, examining whether the schools provided their own resources to the program, the research from both the Snapshot and the teacher survey found that, on average, schools adequately provided what was needed for program implementation, receiving a rating of 75 percent (Figure 3.1). Table 3.1 shows the breakdown by the items that make up each structure. To put some meaning to various scores, at the beginning of the study, SFAF determined that for the purpose of this study, if schools achieve a total score of 50 percent or more of the maximum possible score per structure, they should be deemed to have implemented the structure with "adequate," although not necessarily high, fidelity. The total score for each structure is the average of all the item scores within a structure. Scores are reported for the 2014-2015 school year (the second year of implementation for Cohort 1 and the first year of implementation for Cohort 2). Overall, as shown in Table 3.1, principals were adequately involved with the program (70 percent). Similarly, the results from the teacher survey (Table 3.2) indicate that over half of the PTi3 teachers agreed that their principal was involved. However, while there was no significant difference in the level of principals' involvement as perceived by control and PTi3 teachers, the results suggest that PTi3 teachers were less satisfied with their principals' involvement than control teachers were. While principal involvement in the PTi3 schools was adequate but low, Tables 3.1 and 3.2 indicate that the majority of the PTi3 schools not only had a part-time math facilitator,[5] but they felt they had significantly more support from a facilitator than did control school teachers.

---

made to achieve inter-rater reliability across SFAF coaches. In addition, some of the Snapshot data contradict what was seen by the study team during site visits. Therefore, the results should not be viewed as definitive.

[3]MDRC researchers collected Snapshot scores for all 30 schools in the program sample at the end of the 2014-2015 school year.

[4]The average rating reflects the average proportion of the maximum possible points achieved by the average school.

[5]Facilitators, sometimes referred to as coaches, could be found in both PTi3 school and in control schools.

The overall ratings in the continuous improvement category were very low. Table 3.1 shows that the schools held component team meetings about half of the expected time, on average. And, Table 3.2 shows that a significantly greater proportion of PTi3 school teachers than control teachers reported having coaching support. However, the core purpose of the component team meetings — setting goals, monitoring program implementation, and going over student data — was not the focus of these meetings. Unfortunately, very few schools collected and used student assessment and teacher implementation data meant to drive the meetings, and very few schools used the coaching method prescribed by SFAF in order to create cooperative learning in math classrooms.

Thus, in summary, while SFAF and the PTi3 schools provided the time, staff, and materials needed to support teachers in their implementation of the program, very few schools collected and used student assessment data to drive instruction, and most teachers did not receive the kind of training and support that in theory should lead to the creation of cooperative learning teams in their classrooms. Given this, the next section discusses whether teachers were nevertheless able to create cooperative learning teams in their classrooms.

Implementation of the Key Instructional Elements of PowerTeaching

The logic model posits that the PTi3 structures (training, resources, and school commitment) should lead to the implementation of cooperative learning teams in classrooms. This section provides a description of the teaming of students and the key cooperative learning elements in PTi3. Each subsection presents an overview of what a key element of the model should look like in PTi3 when implemented as planned, analysis of how well cooperative learning was implemented in schools, and the extent of any contrast between the implementation of cooperative learning in PTi3 schools and control schools.[6]

Teaming Students

As indicated in the first chapter, cooperative learning practices have greatly spread throughout U.S. schools (National Mathematics Advisory Panel, 2008). Cooperative learning requires that students work together in teams.[7] Teacher survey results confirm that the majority

[6]The bulk of the data for this chapter comes from teacher focus groups that were conducted in all the study schools from February to April of 2015. Data from teacher logs that were fielded within the same period as the focus groups and teacher surveys that were fielded in May and June of 2015 were also used in this chapter.

[7]For this report, all groups are referred to as "teams." In practice, both terms are used to describe groups of three or more students working together. Under PTi3, teams are unique in that they stick together throughout the course of a unit, the members work together to establish and maintain a unique "team identity," and the (continued)

of study teachers reported that they had students working in teams (Table 3.3), although PTi3 teachers, on average, were significantly more likely to group students in teams (98 percent) compared to control teachers (84 percent). Closer examination of Table 3.3 shows that there were also significant differences in the amount of time spent in teams and the makeup of teams between PTi3 and control classrooms. The teacher survey results show that, on average, students in PTi3 classes spent significantly more time working in teams (59 percent of time in class) than students in control classes (40 percent of time in class). Teams in PTi3 classes were also more likely to have four students rather than three in a team. The PTi3 teams were more likely to stay intact for longer periods of time, thus allowing students to bond and care for their team, and PTi3 teachers were less likely to separate individual students from their teams when academic or behavioral issues arose. In focus groups, teachers across the study schools reported grouping students by mixed ability. However, teachers in program schools reported grouping students in this way more consistently than teachers in control schools did, and teachers in control schools more frequently reported grouping students in other ways (for example, randomly or by giving the students a choice). In more than half of the program schools and a few control schools, teachers considered other factors (in addition to ability level) when grouping students, including behavior, personality, and English language learner status.

Focus group data suggest that teachers in program schools used this type of mixed-ability teaming fairly consistently. However, they also reveal that teachers across study schools agreed that one of the drawbacks to heterogeneously grouping students based on math facility was that it was hard to use mixed-ability teams effectively in classrooms with student populations that were largely homogenous in terms of math skill level. In classrooms like this, where most students struggled with math, teachers reported feeling like the teams did not function the way they were supposed to, with students helping each other learn the material. In schools where teachers did say they had heterogeneous classrooms based on skill level, some teachers reported that the higher-level students ended up doing the majority of the team's work.

A few program school teachers suggested that grouping students based on mixed abilities was difficult because they had high numbers of students who were English language learners. However, they also said that it helped to have a higher-level student in each team whenever possible because those students had the "opportunity to peer teach the rest so they weren't lost and off task." Similarly, one program school teacher said that grouping students based on mixed abilities worked well because "[the lower-level students] eventually…try to live up to the standards of the team because they learn not to let the team down… [and] they're accountable [to their team]." Two program school teachers suggested that the roles

---

teams compete against themselves to win points and become "good," "great," or "super" teams. In some cases where "teams" is used, it may be that the teachers actually said "groups."

were "difficult to keep up with," and teachers in a few program schools said that they did not assign roles at all. One said, "I don't assign roles [so the students] take charge where they feel comfortable."

These findings indicate that on the surface, there were some clear differences in the makeup of and time spent in teams in PTi3 classrooms compared to those in control classrooms. However, what really matters is what happens within those teams. In order to understand the similarities and differences in the dynamics of and activities within team work between program and control group students, teachers of math in both groups of schools were asked to complete instructional logs. Over a two-week period, they were to fill out a log for one student each day for up to eight randomly selected middle school math students.[8]

A central question that the logs help answer is how students worked together in their teams. For each randomly selected student, study teachers were asked to report the proportion of time during the lesson that the selected student spent doing things such as helping team members solve a mathematical problem, asking others for clarification, applying mathematical concepts to a "real world" problem, and engaging in something other than the assigned task. Teachers were also asked to report the time the selected student spent in a team and to rate the math ability of the selected student relative to the class (top third, middle third, or bottom third).

Figure 3.2 graphically depicts the significant difference in the amount of time PTi3 students spent working in teams, or individually compared to control students. The findings confirm those from the teacher survey: On average, PTi3 students spent significantly more time in teams relative to control students. The log results in Table 3.4 show that the average PTi3 student, vis-à-vis the average control student, spent team time engaging in specific behaviors and activities. The findings show that the PTi3 program had little impact on the activities that students engaged in while in teams but that it did have some impact on student behavior. In particular, PTi3 students spent significantly more time, on average, jointly solving math problems by using an algorithm. An average PTi3 student also spent a little more than 2 minutes (out of 5 minutes total) more (out of the 31 minutes spent in teams) engaging in negative behaviors during team work. This is small but statistically significant. However, given the significant difference in the time that PTi3 students spent in teams (10 more minutes), even with the time spent on negative behavior, PTi3 students still spent more time in teams working on math. Findings from the teacher focus groups indicate that the most likely reason that students in PTi3 classes spent more average time engaging in negative behaviors was because PTi3 teachers did not feel that, given the program, they could easily have students work individually

[8]The logs employed for this study were adapted from those used by Brian Rowan, Eric Camburn, and Richard Correnti for the Study of Instructional Improvement conducted by the University of Michigan in partnership with the Consortium for Policy Research in Education.

when team work was not going well. In contrast, control teachers disbanded the teams when students were confused about how to do the math or when they were misbehaving. In addition, fewer control teachers than PTi3 teachers even began team work if they sensed that the activity was best done individually.

What the averages reported above do not indicate is the dynamics of students in different math levels within teams. If cooperative learning, as opposed to simple group work, is taking place, one would expect that the students who are stronger in math, according to their teachers, should be spending more time teaching the less-skilled students. The instructional logs (Appendix Table A.4 show that the students in the top third of the PTi3 classes spent more of their team time helping other students and explaining how to do the problems than those in the top third in control classes. Students in the middle third of the PTi3 classes spent more time than did students in the control classes asking other students for help in solving math problems. The students in the bottom third did not spend more time asking other students for help. The only difference between program and control schools for these students was that the students in the lower third of the PTi3 classes spent more than twice as much time on negative behavior — making fun of others and engaging in off-topic discussions — as students in the lower third of control school classes. (The negative behaviors are seen only among students in the bottom third.) Thus, while there are hints that students were engaged in some behaviors consistent with team work, the logs do not provide strong evidence for the study to conclude that the PTi3 program led to significant changes in instruction.

Were Cooperative Learning Teams Created?

In this section, the following key questions are addressed:

- What are the key elements of cooperative learning?
- Were the key elements of cooperative learning in place?
- How did the PTi3 schools compare with the control schools in terms of implementing the cooperative learning elements?

PTi3 teachers organized their students into teams for math and engaged students in some behaviors and practices consistent with cooperative learning — some more often than control teachers — but this does not mean that these were cooperative learning teams. As indicated in Chapter 1, while many teachers claim to be using cooperative learning, when assessed by researchers, most "cooperative learning activities" in mathematics generally consist of unstructured group work, with no group goals and little individual accountability (Hiebert and Wearne, 1993; Stein, Grover, and Henningsen, 1996). This section of the chapter is about

whether the math students in the study schools were in what are considered to be cooperative learning teams.

As described in Chapters 1 and 2, it is theorized that the skill heterogeneity of PTi3 teams coupled with the use of specific cooperative learning practices creates a context in which students are held accountable for their behavior and work, both to their team members and to themselves. In order for students to feel accountable, three cooperative learning elements must be simultaneously in place. Students need to care about the status and performance of their team — that is, they must have team buy-in. Each of the team members needs to feel that he or she has something of value to contribute to the team in order to stay engaged in the teaming process. The key strategy to promoting team accountability comes from individuals being able to contribute to the team's status — that is, to earn points for their team and help their team to be celebrated — only through the quality of their own work, whether on exams, in homework, or in their explanations of solutions to math problems when called on during class. PTi3 achieves this by having teachers award team points based on the individual performances of *randomly* chosen team members' work. The randomness motivates students to help each other understand the math to ensure that *all* team members understand the math and can represent the team well when their work is selected.

This next section describes whether program teachers were able to simultaneously implement the key elements of cooperative learning — encouraging team buy-in, creating equal opportunities for all students to contribute to the team's success, and facilitating team interdependence.

Team Buy-In

The PTi3 model asks teachers to create team buy-in through the use of specific strategies, including creating *team identity* and *rewarding and celebrating* team success. To build team identity, teachers can use a variety of team-building activities whenever new teams are formed.[9] For example, teachers could elect to use the "Boxing Match" strategy, in which they provide each new team a shoebox, scissors, glue, and old magazines and teams must cover the outside of the box with cutout pictures and words that relate to the members of that team (they then use the box as storage space for team supplies). PTi3 teams are rewarded through the accumulation of points over time for good work and team collaboration in order to earn the title of "Super Team" at the end of a unit. Thus, in theory, PTi3 team members are invested in continuously working to help each other learn the math to reach a common goal.

[9]As seen in Table 3.1, many PTi3 teachers (58.5 percent) stated that they rotated their student teams once or twice a semester.

In many PTi3 schools, at least one teacher explicitly expressed making an effort to promote activities to build team unity in the classroom. However, many of these teachers described making an effort at the beginning of the year only, when students were meeting each other for the first time. Similarly, in control schools, where teams were used less often and teams remained intact for much shorter periods of time (Table 3.3), very little was done in the way of team-building activities after the beginning of the school year. While teachers in most study schools described using a reward system (for example, assigning points, giving out candy, or granting homework passes) to keep students motivated, teachers in many more program schools than control schools specifically described using a cumulative point system as a reward. Another important difference found between program and control teachers was related to the recipients of the rewards, with many more program teachers than control teachers rewarding team successes as opposed to individual successes.

Teachers in a couple of program schools shared that they did not feel the need to do team-building activities when students formed new teams since students in the class were already well accustomed to one another. In a few schools, teachers claimed that team identity activities were too logistically challenging to implement every time students switched teams. As one program teacher stated, "I haven't been good with [establishing a team identity]. I like the idea but…I feel like I'm pressed for time and that coming up with a new name and a new team identity and a logo takes a whole period…I feel like I don't have a whole extra day each time I switch teams."

Moreover, teachers in some program schools indicated that they did not use rewards or point systems within the PTi3 framework because their students were not motivated by score sheets, celebration points, or team rewards. Although teachers in many schools did describe using a point system, some of these teachers suggested that it was burdensome to consistently keep track of points and provide rewards.

Thus, in summary, team buy-in strategies were used by some PTi3 teachers, but their use was not consistent.

Equal Opportunities for All Students to Help the Team

If work done in heterogeneous student teams is to lead to improvement in all students' individual performance, less-skilled students must not disengage from the learning tasks and just rely on the more-skilled students to do the work.

PTi3 includes two strategies to be used within heterogeneous teams that allow less-skilled team members to have meaningful roles in the team and create opportunities within the team structure for them to contribute to the team's overall success. PTi3 students are given different team roles throughout the school year so that all students have the opportunity to try

out and master different ways of engaging within a team. Additionally, teams work on meeting specific team goals that require individual team members to do better academically or behaviorally than they had done previously. For example, they can bring in their homework more often, increase their level of team collaboration, or improve their academic performance. Celebration points are awarded to teams based on their progress toward their goals.

Teachers in program schools more often described assigning roles to team members than did teachers in control schools. However, they reported doing so inconsistently, and they also reported some flexibility in how they assigned those roles. For example, some teachers said they only assigned one of the roles (the "captain," "leader," or "recorder" role) or that they let the students switch roles at their discretion.

Regarding goals, teachers in some of the program schools explicitly described setting formal team goals, while in other program schools they said they set goals for individuals or for the entire class, but not for teams. In the program schools in which teachers described setting team goals, some teachers described establishing goals at the beginning of the year but not continuing the practice over the long term. Teachers in control schools more often described setting individual goals or setting class-wide goals.

Thus, it appears that in the program schools some of practices designed to keep lower-level math students engaged were implemented, though not consistently. However, given that the logs showed the misbehaving was higher in the program schools and concentrated in the students in the lower third of the class in math ability, more needs to be done to fully engage this group.

Team Interdependence

Team interdependence is the heart of cooperative learning. Because the PTi3 model includes practices to ensure that individual performances affect the overall success of the team, interdependence is created among team members. For example, teams receive a celebration point for each person who completes the homework and an extra point when the whole team brings in their homework. Additionally, on a random day, about once a week, the teacher collects and scores the students' homework. Scores are recorded for individual students, but also, depending on the quality and completeness of the homework turned in, celebration points are generated for the team. Thus, team members encourage each other to complete and turn in their homework regularly. Another PTi3 practice for ensuring that each team member cares about the other team members' math skills (that is, team interdependence) is known programmatically as "Random Reporter." With Random Reporter, teachers randomly select students to share the team's solutions to problems worked on within their teams. In PTi3 schools, solutions

are scored[10] and recorded for individual students, but they also serve the purpose of generating celebration points for the team. While students are tested individually and their grades are based on their own scores, teams earn celebration points based on how well the randomly chosen team members do on exams.

Teachers in most of the program schools reported giving teams an extra point if all members brought in their homework, although this was not done consistently. In contrast, teachers in few control schools described giving any rewards for homework completion, and when they did, they gave these rewards to individuals rather than to teams. However, few teachers in program schools (and even fewer in control schools) described using Random Reporter in a way that held students accountable. Many teachers in both program and control schools agreed that they sometimes[11] used a random selection process to call on students to answer math questions. However, many of them also noted that they allowed randomly selected students to pass questions to a teammate or to confer with their team before answering, or they described only randomly picking students who had not yet had a chance to share a response. As a result, team interdependence was not created because it was often possible for answers to come from the most advanced team member, with no repercussions in the way of not acquiring team points or rewards when lower-level students could not provide an answer.

Fewer than five teachers in the study schools reported randomly selecting students' exams for the purpose of rewarding teams. Teachers in study schools were more likely to report letting students work on a test together before collecting them.

Summary of the Implementation of the Key Elements of Effective Cooperative Learning

When team interdependence is implemented well, students are held accountable for their own learning *because* they are liable for their team's elevation to "Super Team" status and overall success. For this element to work, students must (a) be personally invested in the concept of "not letting the team down" (team buy-in); (b) have meaningful ways that they can generate points for the team; and (c) be consistently aware that they may be randomly selected to answer a question on behalf of their team (thus earning their team points or being responsible for the team's stagnation). All three levers must be pulled for team interdependence to be realized, and if they are not, then PTi3 will fail to create true cooperative learning teams.

[10]Solutions are scored using a rubric based on the Common Core State Standards (CCSS) math practices.

[11]Teachers also had other ways of calling on students, including calling on volunteers, calling on students who rarely volunteered or who were not paying attention, or calling on lower-performing students when it was evident that they had a strong response to share.

Focus groups revealed that teachers in many program schools and in some control schools incorporated at least some of the instructional practices that function under the PTi3 framework to encourage team buy-in and create equal opportunities for all students to contribute to team success (for example, students working collaboratively for some period of time in mixed-ability teams, or each student in the team having a designated "role"). That said, the instructional practices that were missing in control classrooms, as well as in most program classrooms, were those that function to create team interdependence — namely, a *truly random* mechanism for choosing a student to give the team response. For instance, students may have been randomly selected to share solutions with the class in order to earn celebration points, but in many classrooms, conditions were such that students could pass their question on to someone else or consult with their team after being selected, so there was little incentive to make sure that all team members understood the math solutions. Thus, celebration points were often awarded for practices that did not hold students accountable.

The Snapshot scores for instructional practices (Table 3.1) confirm what the focus groups revealed. Many program teachers were having students work in teams — what teachers thought was the basic cooperative learning practice — but these generally consisted of unstructured team work, with few, if any, group goals and little team interdependence. Teacher focus groups suggested that program teachers implemented more of the PTi3 type instruction and team practices than did control teachers. Therefore, the PTi3 program seemed to lead to some changes in teacher instruction as well as to changes in how student teams were manifested. However, key instructional practices that lead to cooperative learning were not implemented simultaneously or correctly. Table 3.1 shows that overall, in each cohort, the Snapshot scores for instructional practices were below 50 percent in PTi3 schools. While Snapshot scores were over 50 percent for teachers' use of the basic PTi3 lesson structure and objectives (58 percent) and for providing time for partner and team talk (56 percent), scores were below 50 percent for the other instructional practice items. Teachers overall scored below standard when it came to facilitating partner and team discussions (38 percent), using Random Reporter (31 percent), and conducting whole-class discussions that included thoughtful questioning by the teacher and addressing student misconceptions (20 percent). In other words, teachers had some structures in place that are necessary for cooperative learning, but they did not use these structures to generate the practices that change the ways that students collaborate and learn.

There are several possible explanations for why cooperative learning did not occur in PTi3 classrooms. One reason may be that SFAF coaches were more focused on the PTi3 "curriculum"[12] than they were on the cooperative learning elements. With the advent of Com-

[12]SFAF called these "lesson resources," but they are widely known to teachers and some SFAF coaches as the "PTi3 curriculum." Particularly in Snapshot write-ups, coaches often referred to the lesson resources as "the curriculum."

mon Core, most teachers felt that their first priority was teaching according to the Common Core standards. SFAF responded to the needs of teachers and provided the "curriculum." Indeed, many focus group teachers equated PTi3 with the "PTi3 curriculum" and were unaware that cooperative learning was a goal of PTi3. In fact, teachers in some PTi3 schools reported that they had sought out Kagan training, which is another cooperative learning method, because they wanted to include cooperative learning in their teaching. However, even when teachers understood that PTi3 was a cooperative learning program, they did not have the time to consistently use the three elements, or they did not realize that all three elements had to be in place for cooperative learning to occur. Teachers included many of the basic lesson structures referred to in the "curriculum," such as *Get the Goof*, wherein student teams try to figure out a mistake in a math problem, or *Lightning Round*, in which students are called on randomly to share solutions to math problems. But, they did not adhere to the cooperative learning elements; for example, students were not often chosen to represent their team at random.

Chapter 4

Impact of the Success for All PowerTeaching Program

Chapter 3 shows that during the evaluation period, the program schools implemented a superficial version of the Investment in Innovation PowerTeaching (PTi3) model: Fewer than half implemented with adequate fidelity, and in general the instruction that surrounded the group work did not embody the practices that create team interdependence incentives within the group settings (for example, there was a lack of strict adherence to Random Reporter). This chapter examines whether the PTi3 program, as implemented, had an impact on students' general performance in mathematics. The results for the full study sample are reported first, followed by the results for the subgroups of districts defined by the implementation starting time. Lastly, the chapter provides impact assessments for subgroups defined by a set of student baseline characteristics.

Estimation Method, Outcome Measures, and Description of Impact Tables

As explained in Chapter 2, this evaluation study is based on an experimental design that randomly assigned schools to the program and control conditions, and all impact estimates are based on an intent-to-treat analysis that includes all students in the sample schools at the time of outcome data collection. Thus, the impact estimates reflect the impact of assignment to the program conditions.

The primary outcome measure for students' math achievement is their test scores on the high-stakes standardized state math tests at the end of the 2014-2015 school year. The tests used by each school district differ in their contents and metrics.[1] To be able to aggregate the test score results across districts, student test scores were standardized within each district by grade cell, using the control group scores as the basis for standardization.[2] In addition, whether or not a student scored at or above the state-defined proficiency level is also used as an alternative way to measure the overall student performance level in math.

The main impact table in this chapter reports the estimated program impacts on targeted outcome measures as well as the p-value and corresponding confidence interval for each impact

[1] Appendix Table A.1 shows the specific tests used in each study district.

[2] See Appendix A for details on the standardization of the outcome variable.

estimate. The impact estimate for the standardized test score is an effect size metric, which indicates the magnitude of the estimated effect, calculated as a proportion of the standard deviation of the outcome measure for the control group. In addition, the impact estimate on students' probability of scoring at or above proficiency level, measured in percentage point units, is shown in some of the impact tables to supplement the main findings.

The p-value and 95 percent confidence intervals are complementary ways of showing the chance of obtaining an impact as large as the estimated impact if in fact there were no true impact. If a result is considered statistically significant at the 5 percent level (in other words, the p-value of the estimate is less than or equal to 0.05, or the 95 percent confidence interval does not include zero), it means that there would be no more than a 5 percent chance of obtaining an impact if there were no true effect. Results that are not statistically significant may have occurred by chance and thus do not provide strong evidence about the impact of the program.

Impact Findings for the Confirmatory Analysis Sample

Full Sample

The top panel in Table 4.1 presents the impacts of the PTi3 program on the confirmatory impact analysis sample of sixth-graders, comprising students who were present in a study school at the time of outcome data collection and who had a valid state math test score for the 2014-2015 school year.

For this sample of students, the PTi3 program did not produce any statistically significant impact on their math performances as measured by their state test scores. The estimated impact on the standardized score is virtually zero (with a p-value = 0.990). On average, the probability of scoring at or above the state proficiency level on the math test is about 23 percent for control students and 22 percent for program students, and their difference is small and not statistically significant.

The estimated impact for this sample represents the program's effect on students with a maximum of one year of program exposure. However, it also represents sixth-graders experiencing programs of mixed program maturity. For Cohort 1 schools, the reported finding demonstrates the program impact at the end of the second implementation year, while for Cohort 2 schools, it shows the impact at the end of the first implementation year.

By Cohort

To assess potential impact variation by program maturity, the bottom panel of Table 4.1 reports the program impact separately for these two cohorts of schools. Note, however, that the

study is not powered for this kind of school-level subgroup analyses, and therefore all results presented here are considered exploratory and need to be interpreted as such.

Overall, results in the table show that there is no statistically significant impact for either cohort. The magnitudes of the estimates are suggestive that Cohort 1 schools seem to have experienced more negative impacts (effect size = -0.07) than their counterparts in Cohort 2 (effect size = 0.04), even though they were more mature with the program implementation than the Cohort 2 schools. However, the differences in impacts between the two cohorts are not statistically significant across grades (p-value = 0.17).

Other Findings

Exploratory findings for seventh- and eighth-graders show similar patterns as those reported in Table 4.1.[3] Robustness test results (reported in Appendix Table A.6) show that these main findings are not sensitive to model specifications for all three grades: The estimated impacts do not change in any substantial way when estimated with no baseline covariates in the model (Panel A) or with only the baseline test as a covariate in the model (Panel B).

In addition, to examine the maximum potential effect of the PTi3 program, a subsample of "stable students" is constructed to include all students who remained in the study schools for the implementation year(s). Similar to the results reported in Table 4.1, none of the impact estimates for the stable sample are statistically significant.[4]

Impact Findings for Student Subgroups

The study team also explored potentially heterogeneous impacts across different student subgroups defined by their baseline demographic and socioeconomic characteristics. These subgroups include those defined by students' baseline math performance levels, as well as students' characteristics such as their gender, race/ethnicity, English language learner status, poverty status, and special education status.[5] Table 4.2 presents results of this exploration for the sixth-grade analysis sample.

Overall, the findings indicate that the PTi3 program did not produce any statistically significant impacts across a range of student subgroups, nor do these findings seem to vary

[3]See Appendix Tables A.4 and A.5 for details.

[4]See Appendix Tables A.7 for details.

[5]All subgroups are defined based on students' baseline characteristics. Students in the main analysis sample whose subgroup cannot be determined because of missing information about their baseline characteristics are excluded from the subgroup analyses.

across such subgroups. Similar results can be found for seventh- and eighth-grade samples as well.[6]

Overall, there is no evidence that the PTi3 program as implemented has affected middle school students' math skills in any substantial way. This finding holds for students across the grade levels and by various subgroups.

[6]Findings for seventh- and eighth-graders' analysis samples can be found in Appendix Tables A.8 and A.9.

Chapter 5

Scale-Up

In addition to conducting an evaluation of Investment in Innovation PowerTeaching (PTi3), scale-up grantees were required to bring the program to scale. For PTi3, this meant serving additional students beyond those in the study schools. This chapter examines whether the Success for All Foundation (SFAF) was able to meet its scale-up goals. It also compares the implementation efforts of the scale-up schools and PTi3 study schools using the Snapshot.

Recruitment Goals

At the outset of the project, the goal was to recruit a specific number of schools. These were to be schools in urban and rural districts. However, it was very challenging to recruit rural districts for the study because the study design called for random assignment of schools. In urban districts, that was not as much of an issue (but it was still challenging) because any district that agreed to be in the study could be sure that there would be some district schools that would be randomly selected to receive the program.[1] However, many rural districts only have one middle school, so random assignment could not happen at the district level. Instead, the schools in these small districts had to enter the random assignment pool along with schools in other districts that were similar to them. What this meant was that a rural middle school district that agreed to be in the study had only a 50 percent chance of receiving the program in its one middle school and a 50 percent chance of being in the control group. Districts were not willing to take the risk of being in a study that offered little other than the aggravation of participating in data collection. As a result, for the study portion of the project, only urban middle schools were recruited. These schools had very large student populations and therefore a large number of math teachers — up to about 20 math teachers in a school. It is more difficult and costly to serve a large middle school, so a decision was made to change the goal from recruiting a specific number of schools to recruiting a specific number of students.

As of October 10, 2016, Old Dominion University (ODU) calculated that 132,166[2] (111,233 served without mobility) students in 106 high-need schools were served over the course of the project, which put them at 98 percent of their target of 135,000 students. Most of these scale-up schools were smaller than the study schools. As a result of a six-month no-cost extension that the project received, it planned to dedicate additional residual resources to fund

[1] Random selection occurred at the school level.

[2] This number includes the eight pilot schools that were written into the grant.

another wave of recruitment to implement a technology-enhanced redesign of the program to serve an additional 5,000 students by the end of the project.[3]

Recruitment and Expansion Strategies

Almost all recruitment of districts and schools was done by one person at SFAF. Since SFAF is mostly focused on reading, there were not many people who could effectively recruit schools into a math program. The recruiter began by reaching out via email to networks that SFAF had already worked with, but very few leads came from this effort. This surprised SFAF as this strategy had worked well in the past. The recruiter then looked at regions with affiliations to organizations SFAF was familiar with, such as the Association for Middle Level Education (AMLE) and the National Council of Teachers of Mathematics (NCTM). Additionally, the recruiter did things like looking at conference brochures to see who the key people were to reach out to. About 90 percent of recruited schools were in districts that SFAF had never worked with, and recruitment was mostly done through cold calling. Of those who were cold-called, very few sites (about 10 percent) did not join the project. Districts and schools were eager to join the project as scale-up sites because they were very interested in the funding provided by the project.

In the 2015-2016 school year, six different people were brought in (replacing the original recruiter) to recruit districts and schools; the original recruiter came back to recruit for 2016. SFAF also became part of a science, technology, engineering, and mathematics (STEM) initiative in Iowa, which brought in new schools. Additionally, SFAF developed the PTi3 Common Core "curriculum" during the project, and this helped recruit more schools since so many were at a loss of what to do for Common Core. To recruit big districts, SFAF asked to do awareness presentations with interested school principals and then met individually with the principals who remained interested.

Costs: Economies of Scale

Traditionally, SFAF coaches visit each school about once a month during the first year, with slightly fewer visits in the following years. These site visits are costly as coaches are constantly on the road. The original project plan was to do coaching via technology in order to serve more

[3]By the 2016-2017 school year, a total of 157,183 students had been served in 134 schools, which put the number of students being served over the original target of 135,000 students. The 28 additional schools were the study's control group schools, which received access to the program after the final follow-up data was collected.

schools without spending time or money on travel. However, teachers were not ready for this type of Internet coaching, and the schools did not have the technology. In the final year of the project, SFAF and ODU are piloting the use of technology for coaching on a small scale.

Somewhat lower costs were achieved by the 2015-2016 school year. In 2013-2014, the cost per student was $114.40; the next year, 2014-2015, the costs went up to $120.99 per student. In 2014-2015, when the costs were higher, there was an investment in the lesson resources (the PTi3 "curriculum") development that ultimately made the project more scalable and drove costs down for the next year. By the 2015-2016 school year, the costs were $106.64 per student.

A main reason large economy of scale savings did not occur is because more resources than expected were needed to support the schools. The plan had been to deliver content and training to teachers via the Internet. However, the majority of schools did not have Internet access, so SFAF and ODU had to develop a platform so that content could be put on a flash drive and teachers did not have to depend on the Internet. In addition, training was provided regularly by SFAF coaches (with the number of days of coaching based on the number of teachers). As mentioned above, SFAF unexpectedly had to develop Common Core content to go with the PTi3 framework because this was an immediate need of teachers, and the Common Core content served as a bridge to the program framework. This content, which began as "lesson resources," came to be known as a "curriculum" by districts, schools, and teachers, with the effect of painting the program as a curriculum rather than a framework.

Implementation Fidelity

As with the study schools, the Snapshot was used to measure implementation in scale-up schools.[4] Figure 5.1 shows how well the scale-up schools implemented the program in 2015-2016.[5] As can be seen, the scale-up schools performed slightly better than the study schools but not to the level that defined "adequate" PTi3 implementation. The scale-up implementation scores looked very similar to the study school scores, with the exception of the score in the continuous improvement category.

As with the study schools, SFAF and the scale-up schools were able to get the more tangible structures in place. SFAF was rated highly for its provision of training and materials to both study schools and the scale-up schools. Similarly, both the study cohorts and the scale-up

[4]There were a total of 43 scale-up schools by 2015-2016, but Snapshot scores were only available for 39 schools.

[5]The study school Snapshot scores were from the previous school year, 2014-2015. Also, the scale-up schools varied in the number of years they had been implementing the program.

schools met the threshold for the school inputs, which include having an involved principal and a part-time facilitator.

Where both the study schools and the scale-up schools ran into trouble was taking the tangible structures and using them effectively to support teachers in their creation of cooperative learning teams. The scale-up school implementation score for continuous improvement was far above the study schools' scores, and over the threshold score that defined "adequate" (50 percent). However, the scale-up schools' score is below threshold in the instructional practices category — but still slightly higher than the study schools' scores. These findings suggest, since the evidence indicates that cooperative learning was not implemented in the study schools, that had the scale-up schools been in an impact study, no impacts would have been found.

It may be that the scale-up schools did slightly better than the study schools because many of the scale-up schools were recruited later in the project and had the benefit of getting a more cohesive package, including the complete set of content resources (curriculum) and the final versions of other program materials that were developed early in the project with feedback from study school teachers. It could also be that the scale-up schools were more committed to the program because, unlike the study schools, which were fully funded by the project, the scale-up schools had to pay for a portion of the program. However, if what matters is whether teachers were able to create cooperative learning teams in their classrooms, the scores suggest that, as with the study schools, this was not achieved.

Chapter 6

Conclusion

The Department of Education provided funds to scale up Investment in Innovation Power-Teaching (PTi3) — an instructional process program that numerous studies have shown trains teachers to effectively use cooperative learning practices to improve their students' math achievement — with the goal of partially addressing the poor math achievement of many economically disadvantaged middle school students. According to the 2015 National Assessment of Educational Progress (NAEP), 67 percent of eighth-grade students nationally had below-proficient math skills, and for poor students (those qualifying for the national school lunch program), the rate was 82 percent.

The current evaluation examined whether the Success for All Foundation (SFAF) could meet its scale-up goal of serving an additional 135,000 students over five years and determined the impact on students' math achievement. Chapter 5 shows that the program was able to meet its scale-up goals, with five districts encompassing 106 schools signing up for the program. A total of 132,166 students experienced PTi3 over the four years.[1] The impact findings presented in this report examine the effect of one year of PTi3 on sixth-graders who were enrolled in Cohort 1 and Cohort 2 program schools at the end of school year 2014-2015 (the schools' first or second year of operation). By the end of sixth grade, no impact on math performance was found.

Given that PTi3 (called Student Teams Achievement Divisions (STAD) in the prior research) has been shown to be effective in the past, what might explain the "no impact" finding from this study? The evaluation points to several likely reasons. The most likely is that this study shows that while many of the structures of PTi3 were in place in the study schools (teacher training, ongoing coaching, heterogeneous teams, and lots of team work), the conditions that must be present for PTi3 to be effective — namely, individual accountability and positive interdependence (Slavin, Hurley, and Chamberlain, 2003) — were not. Thus, this evaluation does not measure the impact of PTi3 when it is fully implemented.

The information collected during the research site visits points to a factor that likely led to the imperfect implementation of the program — the adoption of new state standards. New and much more difficult standardized tests that were aligned to the Common Core State Standards were being introduced into the study districts to assess student, teacher, and school performance. Principals and teachers were struggling with how best to teach under this new

[1] 157,183 students in 134 schools were served over five years.

regime of standards. Indeed, SFAF responded to this strong need by creating Common Core-aligned math lesson resources during the first several years of the evaluation. Thus, during the evaluation, teachers had much on their plates.

While the PTi3 program was attempting to improve its instructional processes and practices, teachers were also struggling to teach new and different content to prepare their students for the new assessments. The research found that several key components of the PTi3 model were not in place, perhaps as a result of all these demands. In particular, teachers did not spend time to record the data they were supposed to, nor did they participate in the biweekly meetings in which they were supposed to review data and work on improving their cooperative learning practices. Thus, the ongoing support and continuous improvement part of the PTi3 structure, which is critical to help teachers master the synergistic components of effective cooperative learning practices, did not occur at the level the model specifies.

PTi3 teachers did receive more coaching than control teachers, but despite this, interviews with program teachers showed that they did not understand how the concepts of group learning goals, team buy-in, equal opportunities for students to help the team, and team interdependence worked together to support cooperative learning while still retaining individual accountability. Similarly, they did not understand how specific instructional practices, such as Random Reporter, worked, often altering the details of a practice in such a way as to undermine its purpose. This could have been because the coaches failed to stress the critical importance of making each individual within the group understand that his or her performance alone may have to represent the group's, or because the teachers could not absorb this key point.

This study, while not a good test of PTi3, confirms other research that finds that simply having students work together on problems in groups does not improve academic performance. For true cooperative learning to occur, interdependence must be created so students want to teach each other and the other students want to learn.

This study also points to a large but yet unrecognized need to help teachers across the country understand how to use team work in a way to create effective cooperative learning settings. Unlike in the 1980s and 1990s, when middle school teachers mostly relied on traditional formats of teacher demonstration followed by individual student practice (McKinney and Frazier, 2008), middle school teachers of today almost all incorporate cooperative learning practices into their instruction. This study showed that 96 percent of control teachers were using peer-learning practices, namely pair or team work, in their classes. However, the qualitative data also show that, like the program teachers, control teachers for the most part were not creating environments of positive interdependency paired with individual accountability. Thus, convincing math teachers that team learning activities are a useful instructional practice is not an issue, but there is still a crucial need to help teachers turn all this peer-group work into effective

cooperative learning settings. If this can occur, large gains could be reaped. Slavin, Lake, and Groff (2009) found that compared to other commonly used practices for improving math performance of secondary school students — namely, changing math curricula or supplementing teacher instruction with computer-assisted instruction — improving the instructional practices of math teachers has the largest marginal effect. In particular, the subset of seven evaluations of cooperative learning programs they examined demonstrated the largest of the impacts (ES = +0.46). Thus, strengthening the cooperative learning practices that teachers are already implementing and getting them to use these practices in a synergistic manner could have a large impact on math achievement.

Scaling up PTi3 to more schools could be part of the solution. But, this study shows that no matter what instructional program is adopted to improve cooperative learning, close attention needs to be paid to whether teachers are actually able to create teams where the students have bought into their team's achievement, every team member has opportunities to help the team, and students in the teams are truly interdependent, but where individual and team recognition comes from all students' individual performance, not the performance of a few team members.

Appendix A

Program Impact Estimation and Additional Findings

This appendix provides detailed description of the estimation model used for the impact estimation as well as for the baseline equivalence tests reported in Chapter 2. It then describes the outcome measure used for the impact analysis. Lastly, it presents results for additional baseline analyses and additional impact findings.

I. Estimation Model

The primary impact estimation model is a two-level hierarchical model with students nested within schools. The model uses data from all five study districts in a single analysis, treating districts as fixed effects in the model. Separate program impact estimates are obtained for each district and then averaged across the five districts, weighting each district's estimate in proportion to the number of treatment schools from each district in the sample. Findings in this report therefore represent the impact on student performance in the average treatment school within the study districts. The results do not necessarily reflect what the treatment effect would be in the wider population of districts from which districts participating in the study were selected.

Specifically, the following statistic model is used for all impact estimations reported in Chapter 4 of the report:

$$Y_{ik} = \sum_{m} \gamma_{0m} B_{mk} + \sum_{n} \gamma_{1n} T_k D_{nk} + \gamma_2 Y_{-1ik} + \sum_{l} \alpha_l X_{lik} + \mu_k + \varepsilon_{ik}$$

Where:

Y_{ik} = achievement measurement for student i from school k;

B_{mk} = one if school k is in random assignment block m (m = 1 to 10) and zero otherwise;

D_{nk} = one if school k is in district n (m = 1 to 5) and zero otherwise;

T_k = one if school k is assigned to receive the PowerTeaching program and zero otherwise;

Y_{-1ik} = pretest scores for student i from school k;

Y_{-1k} = average pretest scores for school k;

X_{lik} = student-level covariate l for student i from school k; and

μ_k , ε_{ijk} = school-level and student-level random error, respectively, assumed to be independently and identically distributed.

The error term structure reflects the "hierarchical" or "nested" structure of the data, which has students nested within schools. The model is estimated as a two-level hierarchical model with the MIXED procedure in SAS. The weighted average γ_1 (weighted by the number of treatment schools in each district) of the estimated γ_{1m} coefficients for the five districts is the estimated program effect on student achievement for the average treatment school in the study sample. A two-tailed t-test is used to assess whether γ_1 differs from zero. Impact results are reported both in terms of scaled scores and effect sizes.

Note that this is a fixed effect model instead of a random effect one. This model is chosen because this is a school-level randomized trial and schools in the evaluation sample are purposefully selected and are unlikely to be fully representative of a broader population of schools.

Also note that the impact estimates described above provide an intent to treat analysis of the impact of the program. In other words, the estimates reflect the program impact on all students in the targeted schools, with each student's treatment status determined by the status of the school in which he or she was enrolled at the time of the baseline tests.

A similar model is used to estimate the differences in student background characteristics between the program and control group schools. The only difference between the model for impact estimation and the model for baseline equivalence is that the baseline model does not include any student-level covariates.

II. Outcome Measure

The primary outcome measure of the study is the math test scores based on the state standardized math tests administered at the end of the 2014-2015 school year. Appendix Table A.1 reports the test names and their respective reliabilities.

Test scores from different state tests are naturally in different metrics. In order to pool these scores across districts and use them in one unified regression model, these scores were standardized at the district-by-grade level by using the control group sample means and the sample standard deviations for each district-by-grade combination.[1] This is because the control group means and standard deviations are not affected by the program. The following equation was used for the standardization of the outcome:

[1]No national or statewide norming sample means and standard deviations are available for all districts and grades in the study at the time of this report.

$$Y_{ikj} = \frac{(X_{ikj} - M_{kj})}{\sigma_{kj}}$$

Where:

Y_{ikj} = the standardized score for student i in grade k from district j;

X_{ikj} = the raw test score for student i in grade k from district j;

M_{kj} = the control group sample mean for grade k from district j; and

σ_{kj} = the control group sample standard deviation for grade k from district j.

The standardized score will be in effect size unit and therefore can be pooled across districts.

III. Additional Baseline Equivalence Test Results

Appendix Tables A.2 and A.3 present the baseline equivalence test results for the seventh- and eighth-graders in the impact analysis sample, respectively. The results show no substantively important difference between the program and control group students in these two grade levels.

IV. Analytical Approaches

Decision Rules for Inclusion or Exclusion of Covariates

The principles and rules for choosing covariates are the following:

- Choose covariates because they are related to outcomes. Do not choose covariates because there are big differences between program and control group members at baseline.
- In determining whether covariates are related to outcomes, consider theory, prior empirical evidence, and the data. In most cases, the best covariate is a baseline measure of the outcome.

Based on these principles, a consistent set of covariates are selected and included in all impact analysis. This list includes students' age, gender, race/ethnicity, free/reduced-price lunch (FRL) status, English language learner (ELL) status, and individualized educational program

(IEP) status at baseline. Their baseline math test scores are also included to improve the precision of the impact estimation.[2]

Treatment of Missing Data

The report will document the prevalence missing data and mobility for the confirmatory sample and conduct descriptive analyses to determine whether response rates, missing data, and in- and out-mobility differ by treatment condition or by any observed school or student characteristics (for testing, for example, whether attrition is related to students' pretest scores). These descriptive analyses will be conducted for the whole sample, by treatment group, and by district to detect differential patterns of missing data and mobility.

For purposes of the impact analyses, missing covariate values are imputed, but not missing outcome data. Imputation of missing data values is most relevant for baseline student achievement scores, which are used as precision-increasing covariates in the impact analysis.

Dummy variable adjustment will be used to impute or adjust for missing data values in the student analysis. For cases where the values of the covariates are missing, district mean values will be used to impute for the missing cases and a missing indicator variable interacted with the district indicator variable is included in the outcome model. The dummy variable adjustment approach includes the following steps:

- Replace missing values on a given covariate X with a constant (for example, the district mean);
- Create an indicator variable D for missing versus observed; and
- Include both the covariate and the indicator variable (interacted with the district indicator) in the impact model.

This approach will use the covariate values to improve precision for cases that have measured values, and it will incorporate an effect (different in each district) to reflect the degree to which cases with missing values perform above or below what would be expected if they scored at the district mean on the covariate.[3]

[2]For student subgroup analysis, the variable that defines a given subgroup will be dropped from the covariate list used for the impact estimation of that subgroup.

[3]There is little information on the relative advantages and disadvantages of different imputation methods for covariates in the context of randomized trials. See Puma, Olsen, Bell, and Price. (2009).

V. Additional Impact Findings

In addition to the impact findings for the confirmatory analysis sample of sixth-graders reported in Chapter 6, the study has conducted robustness checks and exploratory analyses on all relevant samples. This part of the appendix reports findings from these analyses.

Impact Findings for Grades 7 and 8

Appendix Tables A.5 and A.6 show the impact findings for the full sample and by cohort for seventh- and eighth-graders in the study sample, respectively. In general, there are no statistically significant findings for either one of these grades. The differences in the impact findings between the two cohorts are not statistically significant, either.

Robustness Checks for Full Sample Impact Findings

Appendix Table A.7 reports results from robustness checks for the full sample findings for all three grade levels. It shows that these main findings are not sensitive to model specifications for all three grades: The estimated impacts do not change in any substantial way when estimated with no baseline covariates in the model (Panel A) or with only baseline test as covariate in the model (Panel B).

Impact Findings for Stable Samples

The main findings of the Investment in Innovation PowerTeaching (PTi3) program presented earlier in the chapter are based on the full analysis sample of students who were present at the study schools at the time of outcome data collection. By construction, it includes students who transferred into the study schools during the school year ("in-movers") and therefore did not receive the "full dosage" of the program. To examine the maximum potential effect of the PTi3 program, a subsample of "stable students" is constructed to include all students who remained in the study schools for the implementation year(s).

Table A.8 presents impact estimations using such samples based on two different ways of defining such sample. One way of defining the "stable" sample is to include all students who stayed in the study schools for all implementation years. In other words, for Cohort 1, this sample would include students who were enrolled in the study schools for both the 2013-2014 and 2014-2015 school years.[4] For Cohort 2, this would include students who were enrolled in the study schools for the 2014-2015 school year. Results based on this definition are shown in

[4]However, this only applies to students in Grades 7 and 8. Sixth-graders in the 2014-2015 school year would have only been in the study schools for that one year at most.

the top panel of the table. Overall, the pattern of findings does not differ from that of the full sample findings, and none of the estimates are statistically significant.

The alternative way would be to define the "stable" sample as students who were enrolled in the study schools for the 2014-2015 school year, regardless of which cohort they were in. The bottom panel of the table shows results based on such a sample. Again, the findings do not differ from previous findings in any substantial way.

Subgroup Impact Findings for Seventh- and Eighth-Graders

Appendix Tables A.8 and A.9 report impact findings for various student subgroups defined by students' baseline characteristics for Grades 7 and 8, respectively. Overall, like what is reported in Table 4-2, there are not many significant findings across these subgroups in both grades. There are a handful of exceptional findings for Grade 7, however. Specifically, students who performed below proficiency level at baseline in math, girls, and students who were English language learners at baseline seem to have experienced negative and statistically significant program impacts. However, these impacts are not significantly different from the impact findings for their counterparts, which indicates that these significant findings might have happened by chance due to the large number of subgroups being examined.

References

Anderman, Eric M., Martin L. Maehr, and Carol Midgley. 1999. "Declining Motivation After the Transition to Middle School: Schools Can Make a Difference." *Journal of Research and Development in Education* 32, 3: 131-147.

Antil, Laurence R., Joseph R. Jenkins, Susan K. Wayne, and Patricia F. Vadasy. 1998. "Cooperative Learning: Prevalence, Conceptualizations, and the Relation Between Research and Practice." *American Educational Research Journal* 35, 3: 419-454.

Emmer, Edmund T., and Mary Claire Gerwels. 2002. "Cooperative Learning in Elementary Classrooms: Teaching Practices and Lesson Characteristics." *The Elementary School Journal* 103, 1: 75-91.

Fajgelbaum, Pablo D., and Amit K. Khandelwal. 2016. "Measuring the Unequal Gains from Trade." *The Quarterly Journal of Economics* 131, 3: 1113-1180.

Grossman, Gene M., and Esteban Rossi-Hansberg. 2008. "Trading Tasks: A Simple Theory of Offshoring." *The American Economic Review* 98, 5: 1978-1997.

Hampton, David R., and Gary Grudnitski. 1996. "Does Cooperative Learning Mean Equal Learning?" *Journal of Education for Business* 72, 1: 5-7.

Hanushek, Eric A., Paul E. Peterson, and Ludger Woessmann. 2010. "U.S. Math Performance in Global Perspective: How Well Does Each State Do at Producing High-Achieving Students?" PEPG Report No. 10-19. Cambridge, MA: Harvard Univeristy, Program on Education Policy and Governance.

Hiebert, James, and Diana Wearne. 1993. "Instructional Tasks, Classroom Discourse, and Students' Learning in Second-Grade Arithmetic." *American Educational Research Journal* 30, 2: 393-425.

Institute of Education Sciences. 2014. *What Works Clearinghouse Procedures and Standards Handbook: Version 3.0.* Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Johnson, David W., and Roger T. Johnson. 1994. *Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning.* Boston: Allyn and Bacon.

Johnson, David W., and Roger T. Johnson. 2009. "An Educational Psychology Success Story: Social Interdependence Theory and Cooperative Learning." *Educational Researcher* 38, 5: 365-379.

Johnson, David W., Roger T. Johnson, and Cary Roseth. 2010. "Cooperative Learning in Middle Schools: Interrelationship of Relationships and Achievement." *Middle Grades Research Journal* 5, 1: 1-19.

Juvonen, Jaana. 2006. "Sense of Belonging, Social Bonds, and School Functioning." Pages 655-674 in Patricia A. Alexander and Philip H. Winne (eds.), *Handbook of Educational Psychology*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Juvonen, Jaana, and Kathryn R. Wentzel. 1996. *Social Motivation: Understanding Children's School Adjustment*. New York: Cambridge University Press.

Kennett, Deborah J., Anna May Young, and D. P. Berrill. 1999. "Is Cooperative Learning Effective for High Achieving Entrance Students? Implications for Policy and Teaching Resources." *Journal of Research and Development in Education* 33, 1: 27-35.

McKinney, Sueanne, and Wendy Frazier. 2008. "Embracing the Principles and Standards for School Mathematics: An Inquiry into the Pedagogical and Instructional Practices of Mathematics Teachers in High-Poverty Middle Schools." *The Clearing House: A Journal of Educational Strategies, Issues and Ideas* 81, 5: 201-210.

Midgley, Carol, Eric Anderman, and Lynley Hicks. 1995. "Differences Between Elementary and Middle School Teachers and Students: A Goal Theory Approach." *The Journal of Early Adolescence* 15, 1: 90-113.

The Nation's Report Card. 2017. "2015 Mathematics and Reading Assessments." Website: https://www.nationsreportcard.gov/reading_math_2015/#?grade=8.

National Mathematics Advisory Panel. 2008. *Foundations for Success: The Final Report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education.

Nunnery, John A., Shanan Chappell, and Pamela Arnold. 2013. "A Meta-analysis of a Cooperative Learning Model's Effects on Student Achievement in Mathematics." *Cypriot Journal of Educational Sciences* 8, 1: 34-48.

Pianta, Robert C., Jay Belsky, Renate Houts, and Fred Morrison. 2007. "Opportunities to Learn in America's Elementary Classrooms." *Science* 315, 5820: 1795-1796.

Puma, Michael J., Robert B. Olsen, Stephen H. Bell, and Cristofer Price. 2009. *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.

Shapka, Jennifer D., José F. Domene, and Daniel P. Keating. 2006. "Trajectories of Career Aspirations Through Adolescence and Young Adulthood: Early Math Achievement as a Critical Filter." *Educational Research and Evaluation* 12, 4: 347-358.

Sherman, Julia A. 1982. "Mathematics the Critical Filter: A Look at Some Residues." *Psychology of Women Quarterly* 6, 4: 428-444.

Slavin, Robert E. 1995. *Cooperative Learning: Theory, Research, and Practice*. Boston: Allyn and Bacon.

Slavin, Robert E. 1996a. "Cooperative Learning in Middle and Secondary Schools." *The Clearing House* 69, 4: 200-204.

Slavin, Robert E. 1996b. "Research on Cooperative Learning and Achievement: What We Know, What We Need to Know." *Contemporary Educational Psychology* 21, 1: 43-69.

Slavin, Robert E., Eric A. Hurley, and Anne Chamberlain. 2003. "Cooperative Learning and Achievement: Theory and Research." Pages 177-198 in William M. Reynolds, Gloria E. Miller, and Irving B. Weiner (eds.), *Handbook of Psychology, Volume 7: Educational Psychology*. Hoboken, NJ: John Wiley and Sons.

Slavin, Robert E., Cynthia Lake, and Cynthia Groff. 2009. "Effective Programs in Middle and High School Mathematics: A Best-Evidence Synthesis." *Review of Educational Research* 79, 2: 839-911.

Stein, Mary Kay, Barbara W. Grover, and Marjorie Henningsen. 1996. "Building Student Capacity for Mathematical Thinking and Reasoning: An Analysis of Mathematical Tasks Used in Reform Classrooms." *American Educational Research Journal* 33, 2: 455-488.

Steinberg, Laurence. 1990. "Autonomy, Conflict, and Harmony in the Family Relationship." Pages 255-276 in S. Feldman and G. Elliot (eds.), *At the Threshold: the Developing Adolescent.* Cambridge, MA: Harvard University Press.

Stockdale, Susan L., and Robert L. Williams. 2004. "Cooperative Learning Groups at the College Level: Differential Effects on High, Average, and Low Exam Performers." *Journal of Behavioral Education* 13, 1: 37-50.

Webb, Noreen M. 2008. "Teacher Practices and Small-Group Dynamics in Cooperative Learning Classrooms." Pages 201-221 in Robyn Gillies, Adrian Ashman, and Jan Terwel (eds.), *The Teacher's Role in Implementing Cooperative Learning in the Classroom*. New York: Springer.

Wigfield, Allan, James P. Byrnes, and Jacquelynne S. Eccles. 2006. "Development During Early and Middle Adolescence." Pages 87-113 in Patricial A. Alexander Philip H. Winne (eds.), *Handbook of Educational Psychology*. Mahwah, NJ: Lawrence Erlbaum Associates.

Exhibits

Table 2.1

Background Characteristics for Schools in the Study Sample, Schools in the PowerTeaching Scale-Up Sample, and Similar Schools in the National Population (2012-2013 Academic Year)

Characteristics	Study Sample	Scale-up Sample	National[a] Population
Geographic region (% of schools)		†	†
Northeast	6.9	12.7	16.4 *
South	25.9	29.6	29.3
Midwest	15.5	22.5	27.4 *
West	51.7	35.2 *	26.7 *
Urbanicity (% of schools)		†	†
Large or mid-sized city	44.8	25.4 *	22.3 *
Urban fringe and large town	51.7	40.8	29.4 *
Small town and rural area	3.4	33.8 *	48.3 *
Title I status (% of schools)	91.4	91.5	100.0 *
Eligible for free or reduced-price lunch (average % of students)	72.0	68.7	61.1 *
Race/Ethnicity (average % of students)			
White non-Hispanic	11.5	38.0 *	51.5 *
Black non-Hispanic	31.3	23.3 *	17.2 *
Hispanic	50.5	28.4 *	23.0 *
Asian	4.9	3.6	2.8 *
Other	1.7	6.7 *	5.5 *
Male (average % of students)	48.8	50.7 *	52.1 *
Enrollment (average number of students in Grades 6-8)	961.3	686.8 *	251.7 *
Full-time teachers (average % of teachers of Grades 6-8)	51.6	43.2 *	29.7 *
Sample Size	58	71	31,102

(continued)

Table 2.1 (continued)

SOURCE: 2012-2013 Common Core of Data.

NOTES: Due to missing values for some variables, the number of schools included varies by characteristic.

"*" indicates a statistically significant difference (p-value < = 0.05) between the study sample and either the scale-up sample or the national population of schools for a given characteristics. A two-tailed t-test was applied to each comparison.

"†" indicates a statistically significant difference (p-value <= 0.05) between the study sample and either the scale-up sample or the national population of schools for categorical characteristics. A chi-square test was applied to each of such comparisons.

To examine whether there is any systematic difference between the study sample and the scale-up sample, an F-test was conducted in a regression model controlling all school characteristics reported in this table ($p = 0.892$). A similar test was conducted for systematic difference between the study sample and the national population ($p < 0.001$).

[a]The national sample includes Title 1 schools with Grades 6 through 8.

Table 2.2

Confirmatory Impact Analysis Sample, by Treatment Status (Spring 2015)

Treatment Status	Number of Schools	6th-Grade Students	Average per School
Program	30	8,534	284
Control	28	7,416	265
Total	58	15,950	275

SOURCES: Calculation based on school sample and student analysis sample.

Table 2.3

Background Characteristics for Study Sample Schools, by Treatment Status

Characteristics	Program Group	Control Group	Estimated Difference	P-value for Estimated Difference
Title I status (% of schools)	100.0	96.7	3.3	0.318
Eligible for free or reduced-price lunch (average % of students)	74.7	72.9	1.8	0.567
Race/Ethnicity (average % of students)				
White non-Hispanic	10.5	11.4	-0.9	0.687
Black non-Hispanic	31.9	33.8	-1.9	0.567
Hispanic	51.0	48.2	2.8	0.526
Asian	4.6	5.1	-0.5	0.764
Other	2.0	1.5	0.5	0.196
Male (average % of students)	48.8	48.5	0.3	0.562
Enrollment (average number of students in Grades 6-8)	987.4	905.2	82.2	0.437
Number of full-time teachers (all grades)	51.9	49.8	2.0	0.597
8th-grade math proficiency in state Accountability test (% of students)	36.9	41.2	-4.4	0.252
Sample Size	30	28		

SOURCES: Common Core of Data (CCD) for Years 2012-2013 (for Cohort 1 schools) and 2013-2014 (for Cohort 2 schools). Eigth-grade math performance measures are from school report cards in respective baseline years.

NOTES: CCD data are used from different years for Cohorts 1 and 2 to reflect their true baseline information. The estimated differences for school-level data are regression-adjusted using ordinary least squares regressions, controlling for indicators of random assignment blocks. The values in the column labeled "Program Group" are the weighted average of the observed district means for schools randomly assigned to the program group (using number of program group schools in each district as weight). The control group values in the next column are the regression-adjusted means using the observed distribution of the program group across blocks as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

To examine if there is any systematic difference between the program and control groups, a chi-square test was calculated for the full sample of 58 schools in a regression model controlling for the following variables: indicators of random assignment strata and all school characteristics reported in this table. The p-value of the test is 0.98.

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

Table 2.4

Background Characteristics for 6th-Graders in Confirmatory Analysis Sample, by Treatment Status

Characteristics	Program Group	Control Group	Estimated Difference	Standard Deviations in Effect Size of Estimated Difference	P-Value for Estimated Difference
Age (year)	9.7	9.7	0.0	0.018	0.311
Eligible for free or reduced-price lunch (average % of students)	83.0	82.0	1.0	0.025	0.594
Race/Ethnicity (average % of students)					
White non-Hispanic	11.1	10.5	0.6	0.017	0.680
Black non-Hispanic	30.1	33.5	-3.4	-0.093 *	0.034
Hispanic	52.0	49.5	2.5	0.051	0.245
Asian	5.6	5.8	-0.2	-0.009	0.841
Other	1.2	1.1	0.1	0.007	0.804
Male (average % of students)	46.9	49.4	-2.5	-0.050 *	0.047
English-language learner (average %	16.1	17.6	-1.5	-0.037	0.322
of students)	16.1	17.6	-1.5	-0.037	0.322
Special education status (average %	14.3	14.3	0.0	0.000	0.991
of students)	14.3	14.3	0.0	0.000	0.991
Baseline math achievement					
Standardized test score	-0.05	-0.06	0.01	0.006	0.888
At or above proficiency (average %	57.0	57.1	0.0	-0.001	0.986
of students)	57.0	57.1	0.0	-0.001	0.986

(continued)

Table 2.4 (continued)

SOURCES: District student records from the 2012-2013 school year for Cohort 1, and from the 2013-2014 school year for Cohort 2.

NOTES: Due to missing values, the number of students included varies by characteristics, ranging from 6,842 to 7,416 for the control group, and from 7,785 to 8,534 for the program group.

The estimated differences for student-level data are regression-adjusted using hierarchical linear models to account for the nested structure of the data (with students nested within classes and classes nested within schools). The models control for indicators of random assignment blocks.

The values in the column labeled "Program Group" are the weighted average of the observed district means for schools or students randomly assigned to the program group (using number of program group schools in each district as weight). The control group values in the next column are the regression-adjusted means using the observed distribution of the treatment group across blocks as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

To examine if there is any systematic difference between the treatment and control group students, an F-test was conducted for the full sample in a logistic regression model controlling for the following variables: indicators of random assignment strata, all student characteristics reported in this table, and corresponding missing indicators. The p-value of the test is less than 0.01.

Table 2.5

Teacher Background Characteristics (2014-2015 Academic Year)

Characteristics	Program Group	Control Group	Estimated Difference	P-value for Estimated Difference
Highest degree earned (% of teachers)				
Associate's degree	0.4	0.0	0.4	0.586
Bachelor's degree	38.9	46.4	-7.5	0.250
Master's degree	56.3	45.9	10.4	0.115
Educational specialist or professional diploma	3.1	6.3	-3.2	0.233
Doctorate or professional degree	1.3	1.4	0.0	0.990
Degree in STEM (%)	40.6	42.1	-1.5	0.818
Certified to teach math (%)	92.1	91.3	0.8	0.845
Certified to teach math in middle school (%)	88.9	88.7	0.2	0.973
Experience (years)				
Teaching total	11.7	11.0	0.7	0.575
Teaching at current school	6.5	5.8	0.7	0.500
Teaching middle school math	7.4	7.6	-0.2	0.836
Teaching elementary or high school math	3.9	2.9	1.0	0.166
Sample size (number of teachers)	229	203		

SOURCE: Spring 2015 Teacher Survey.

NOTES: STEM = Science, Technology, Math, and Engineering. The estimated differences for student level data are regression-adjusted using hierarchical linear models to account for the nested structure of the data (with students nested within schools). The models control for indicators of random assignment blocks.

The values in the column labeled "Program Group" are the weighted average of the observed district means for schools randomly assigned to the program group (using number of program group schools in each district as weight). The control group values in the next column are the regression-adjusted means using the observed distribution of the treatment group across blocks as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

Table 3.1

School Achievement Snapshot Scores for Items Related to Schoolwide Structures and Instructional Practices, Study Schools (2014-2015 and 2015-2016 Academic Years)

Item	2014-2015 Percent of maximum possible score	2015-2016 Percent of maximum possible score
Program developer		
All leaders and staff have received essential training	83.3	76.7
Materials for program implementation are complete	100.0	93.3
School inputs		
School-based math facilitator is a part-time position	80.0	56.7
The principal is fully involved with PowerTeaching	70.0	66.7
Continuous improvement processes		
Component teams meet at least twice a month	43.3	45.0
Each teacher submits a quarterly classroom assessment summary	18.3	5.0
Instructional component teams set targets, chart progress, and work to meet targets	25.0	31.7
The school-based math facilitator uses PowerTeaching coaching process	18.3	34.5
Instructional practices		
Teachers...		
Use basic lesson structure, objectives, and available media regularly and effectively	57.7	54.0
Use think-pair-share, whole group response, or random reporter frequently and effectively	45.3	46.7
Provide time for partner and team talk to allow mastery of learning objectives by all students	55.7	62.0
Facilitate partner and team discussion	38.3	46.3
Randomly select students to report for their teams during class discussion, use rubrics to evaluate responses, and awards teams with points	30.7	37.3
Effectively summarize and address misconceptions or inaccuracies during class discussion	20.0	14.7

SOURCES: 2014-2015 and 2015-2016 School Achievement Snapshots.

NOTES: The sample includes 12 schools in Cohort 1 and 18 schools in Cohort 2. For Cohort 1 schools, the 2014-2015 and 2015-2016 academic years were Years 2 and 3 of implementation. For Cohort 2 schools, the 2014-2015 and 2015-2016 academic years were Years 1 and 2 of implementation.

Table 3.2

Program-Control Group Comparisons Related to Support Received by Teachers (2014-2015 Academic Year)

	Program Group	Control Group	Estimated Difference	P-value for Difference
Coach support				
Percentage of teachers reporting being supported by a coach	78.4	37.5	40.9 *	0.000
Principal support				
Percentage of teachers who agree their principal:				
Communicates a clear vision for your school	65.1	70.0	-4.9	0.436
Communicates to staff what is expected of them	73.9	72.8	1.1	0.874
Makes clear to staff the expectations for meeting instructional goals	69.7	70.6	-0.9	0.880
Sets high standards for student learning	73.4	75.9	-2.5	0.664
Invites teachers to play a meaningful role in making decisions for this school	55.2	62.0	-6.8	0.345
Helps teachers at your school address student behavior issues	43.8	52.8	-9.0	0.205
Encourages teachers to implement what they have learned in professional development	69.9	77.2	-7.3	0.173
Knows what is going on in your classroom	59.1	60.6	-1.5	0.835
Participates in instructional planning with teams of teachers	39.9	50.5	-10.6	0.151
Gives you regular and helpful feedback about your teaching	43.0	50.8	-7.8	0.276
Builds time in school schedule for teachers to collaborate	77.0	73.2	3.8	0.537
Allocates sufficient resources to the math program	60.7	67.4	-6.7	0.270
Carefully monitors student academic progress	53.4	55.9	-2.4	0.730
Ensures that teachers use data to modify instruction	62.4	58.8	3.6	0.625
Encourages teachers to routinely use data to meet students' learning needs	69.6	68.5	1.1	0.852
Supports teachers in understanding and using data	62.7	62.4	0.2	0.971
Communicates a clear vision for school-wide data use	57.8	59.2	-1.4	0.826

(continued)

Table 3.2 (continued)

	Program Group	Control Group	Estimated Difference	P-value for Difference
Professional development				
Percentage of teachers who agree the professional development they received has:				
Prepared them to work with students achieving below grade level	34.8	41.9	-7.1	0.340
Prepared them to work with students achieving above grade level	45.3	52.4	-7.1	0.273
Prepared them to work with students with special needs	25.7	29.4	-3.6	0.555
Helped them learn new techniques for organizing and managing the classroom	62.2	60.2	2.0	0.777
Teacher satisfaction				
Percentage of teachers reporting being satisfied with overall quality of math program	68.6	78.9	-10.3	0.119

SOURCE: Spring 2015 teacher survey.

NOTES: Items on the teacher survey that asked about levels of agreement were on a four-point scale: 1 = Strongly disagree, 2 = Disagree, 3 = Agree, 4 = Strongly agree. The percentages of teachers who agree with an item were obtained by taking the number who responded 3 or 4.

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

Completed surveys were received from 229 teachers at program group schools and 203 teachers at control group schools.

Table 3.3

Program-Control Group Comparisons Related to Student Teams (2014-2015 Academic Year)

	Program Group	Control Group	Estimated Difference	P-Value for Estimated Difference
Percentage of teachers who arrange students in:				
Teams	97.7	84.4	13.3 *	0.003
Pairs	40.7	78.1	-37.4 *	0.000
Neither	0.4	3.7	-3.3	0.096
Average number of students in a team	4.1	3.5	0.5 *	0.000
Average percentage of class time students work in teams	58.6	39.5	19.2 *	0.000
Percentage of teachers who change teams:				
Daily	3.1	8.1	-5.0	0.060
Weekly	0.2	13.1	-12.9 *	0.000
Monthly	20.3	22.9	-2.5	0.643
Once or twice a semester	58.5	26.6	32.0 *	0.000
Percentage of teachers reporting changing teams because:				
Students not working well in their teams	39.4	49.3	-9.9	0.170
New cycle/unit of instruction begins	45.5	42.4	3.1	0.649
New task or assignment	14.6	58.0	-43.4 *	0.000
Students socialize too much in their teams	46.6	51.9	-5.3	0.472
Percentage of teachers reporting separating individual students from their team	70.7	84.9	-14.2 *	0.029
Percentage of teachers who agree that:				
Students are well-behaved during math class	72.7	77.1	-4.5	0.478
Students are intellectually engaged during math class	80.7	90.2	-9.5	0.102
Students are respectful towards one another	67.2	77.8	-10.6	0.082
Classroom management problems limit the types of math activities you can do with your students	47.7	40.3	7.4	0.344

(continued)

Table 3.3 (continued)

SOURCE: Spring 2015 teacher survey.

NOTES: Items on the teacher survey that asked about levels of agreement were on a 4-point scale: 1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree. The percentages of teachers who agree with an item were obtained by taking the number who responded 3 or 4.

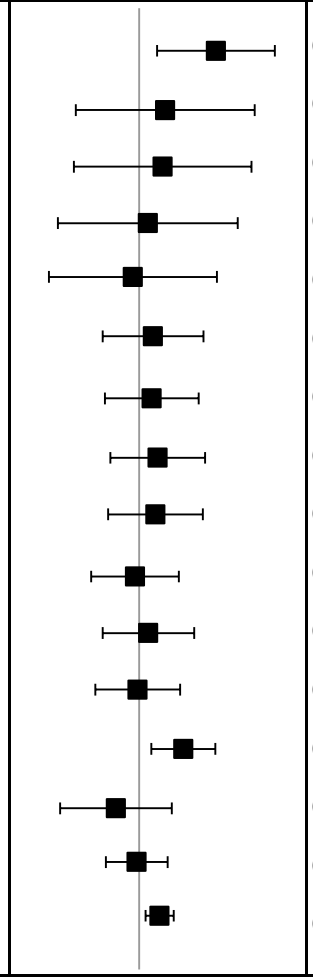
A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

Completed surveys were received from 229 teachers at SFA schools and 203 teachers at control group schools.

Table 3.4

Impact of PowerTeaching on Minutes Students Spent Doing Group Activities During the Math Block (Activities Are Not Mutually Exclusive)

Activity	Mean Program Minutes	Mean Control Minutes	Estimated Impact[a]	90% Confidence Interval
Solving mathematical problems by using an algorithm	8.42	5.22	3.19 *	(0.74 5.64)
Discussing and working on a problem with multiple solution methods	8.06	6.98	1.08	(-2.64 4.80)
Applying mathematical concepts to "real world" problems	7.82	6.84	0.98	(-2.71 4.67)
Representing and analyzing relationships using tables or models	6.71	6.35	0.36	(-3.38 4.10)
Analyzing data to make inferences or draw conclusions	5.54	5.80	-0.26	(-3.76 3.24)
Explaining a solution to a problem to other students	5.47	4.89	0.57	(-1.52 2.67)
Helping other students solve math problems	5.20	4.68	0.52	(-1.42 2.47)
Asking other students clarifying questions	5.06	4.29	0.77	(-1.20 2.73)
Asking other students for help in solving a math problem	4.85	4.17	0.67	(-1.29 2.64)
Exchanging work with other students for review and checking	4.82	4.99	-0.17	(-2.00 1.65)
Suggesting a strategy to a partner or group members	4.61	4.23	0.38	(-1.52 2.28)
Building on or challenging the ideas of other students	3.93	3.99	-0.06	(-1.82 1.70)
Engaging in discussion or activities not related to assigned activity	3.77	1.94	1.83 *	(0.51 3.16)
Jointly reading a textbook or supplementary materials	2.48	3.44	-0.97	(-3.28 1.35)
Discussing and jointly working on multiple choice exercises	2.23	2.33	-0.11	(-1.39 1.18)
Making fun of, belittling, or bothering a partner or group members	1.49	0.64	0.85 *	(0.27 1.43)



(continued)

Table 3.4 (continued)

SOURCE: Teacher logs administered in spring 2015.

NOTES: Sample consists of 2,941 logs (1,567 in the program group and 1,374 in the control group). There was a low response rate from teachers in one school district. As a result, all instructional logs (n = 84) from this district were dropped.

All estimations are based on a three-level hierarchical model with individual logs nested within teachers and teachers nested within schools. A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

In the instructional logs, teachers were asked how much time randomly selected students spent working in three different configurations during the mathematics period: in groups, in pairs, and individually. If teachers indicated that selected students worked in a group during the mathematics period, they were asked approximately what proportion of group time the students spent engaging in specific activities using a five-category scale: 0 percent, less than 10 percent, 10 to 25 percent, 26 to 50 percent, and more than 50 percent. In the analysis of the logs, the research team converted the proportion of group time into minutes by calculating the midpoint of the ranges in this scale and then multiplying the midpoint of the selected range by the total amount of time spent working in a group. If selected students did not spend time working in a group, the time spent on each activity was set to zero.

[a]The horizontal lines on each side of the impact estimate represent the "confidence interval" — that is, the range of estimated values of the impact, within which there is a 90 percent probability that the true value falls. The impact estimate is statistically significant when the range of the confidence interval (defined by the upper and lower bounds) crosses the vertical line.

Table 4.1

Impact of PowerTeaching on Student Math Achievement for 6th-Graders in Analysis Sample (2014-2015 Academic Year), by Full Sample and Cohort

Sample	Program Group	Control Group	Estimate Impact[a]	P-Value	95% Confidence Interval of the Impact[b] (in Standard Deviation Units) -0.25 0 0.25
Full sample (confirmatory)					
Grade 6					
Standardized state math test score	-0.05	-0.05	0.00	0.991	
% at or above proficiency level	21.9	22.9	-1.1	0.551	
Cohort (exploratory)					
Cohort 1					
Standardized state math test score	-0.06	0.01	-0.07	0.285	
% at or above proficiency level	22.2	24.1	-1.9	0.494	
Cohort 2					
Standardized state math test score	-0.05	-0.10	0.04	0.386	
% at or above proficiency level	35.8	35.9	-0.1	0.975	

SOURCES: District student records data from the 2014-2015 and 2012-2013 academic years (for Cohort 1), and the 2013-2014 academic year (for Cohort 2).

NOTES: The analysis sample consists of students from 58 schools (30 program group schools and 28 control group schools) and includes any student who had a valid spring test score in the spring of 2015.

The student sample size for Grade 6 is 15,950 students (8,534 in the program group schools and 7,416 in the control group schools). The student sample size for Cohort 1 is 8,326 students (4,743 in the program group schools and 3,583 in the control group schools). The student sample size for Cohort 2 is 7,624 students (3,791 in the program group schools and 3,833 in the control group schools).

The estimated impacts are based on a two-level model with students nested within schools, controlling for random assignment block and school- and student-level covariates. The program group and control group columns display regression-adjusted mean outcomes for each group, using the mean covariate values for students in the program group as the basis for the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

The difference between the impact estimates for Cohort 1 and Cohort 2 is not significant at the 5 percent level (p-value = 0.17)

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

[a]Impact estimates are measured in standard deviations in effect size or percentage point.

[b]The confidence intervals are for the impact estimates on the standardized test scores and proficiency classification.

Table 4.2

Impact of PowerTeaching on Student Math Achievement for 6th-Graders in Analysis Sample (2014-2015 Academic Year), by Student Subgroup

Subgroup	Number of Observations	Program Group	Control Group	Estimated Impact (in Standard Deviations in Effect Size)	P-Value	P-value for Difference Between Subgroups
By performance rank at baseline						1.000
Top third	4,910	0.48	0.50	-0.01	0.766	
Middle third	5,192	-0.02	-0.03	0.01	0.823	
Bottom third	4,525	-0.56	-0.56	0.00	0.998	
By proficiency level at baseline						0.412
At or above proficiency	9,615	0.29	0.32	-0.04	0.465	
Below proficiency	4,976	-0.41	-0.42	0.02	0.680	
By gender						0.743
Boys	8,139	-0.09	-0.07	-0.02	0.651	
Girls	7,781	-0.03	-0.03	0.00	0.988	
By race/ethnicity						1.000
Hispanic	9,714	-0.08	-0.07	-0.01	0.845	
White, Non-Hispanic	2,200	0.13	0.12	0.01	0.871	
Black, Non-Hispanic	2,622	-0.18	-0.21	0.03	0.549	
By family income						0.448
Eligible for free and reduced-price lunch	11,893	-0.10	-0.11	0.01	0.801	
Not eligible for free and reduced-price lunch	2,949	0.22	0.27	-0.05	0.464	
By English-language learner (ELL) status						0.332
ELL	2,988	-0.35	-0.29	-0.06	0.271	
Non-ELL	11,880	0.02	0.01	0.01	0.873	
By special education (SPED) status						0.278
SPED	2,124	-0.35	-0.29	-0.07	0.274	
Non-SPED	13,804	0.01	-0.01	0.02	0.721	

SOURCES: School district student records data from the 2014-2015 and 2012-2013 academic years (for Cohort 1), and the 2013-2014 academic year (for Cohort 2).

NOTES: The analysis sample consists of students from 58 schools (30 program group schools and 28 control group schools) and includes any student who had a valid spring test score in the spring of 2015.

The estimated impacts are based on a two-level model with students nested within schools, controlling for random assignment block and school- and student-level covariates. The program group and control group columns display regression-adjusted mean outcomes for each group, using the mean covariate values for students in the program group as the basis for the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each estimated impact. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

Appendix Table A.1

Math State Test Information, by District and Grade

Test Name	Grade	Reliability
Chicago, IL[a]		
Illinois Northwest Evaluation Association Measure of Academic Progress	6	0.88
	7	0.89
	8	0.87
Los Angeles, CA[b]		
California Assessment of Student Performance	6	0.93
and Progress (CAASPP) Smarter Balanced Test	7	0.90
	8	0.91
Marino Valley, CA[b]		
CAASPP Smarter Balanced Test	6	0.93
	7	0.90
	8	0.91
New York, NY[c]		
New York City Common Core Mathematics Test	6	0.95
	7	0.94
	8	0.93
Orlando, FL[d]		
Florida Comprehensive Achievement Test (FCAT)	6	0.92
	7	0.93
	8	0.87

NOTES: [a]Test reliability is based on the 2013-2014 Northwest Evaluation Association Measure of Academic Progress. Source: The 2010 Technical Manual for Measures of Academic Progress.

[b]Test reliability is based on the 2014-2015 CAASP Smarter Balanced Test. Source: Smarter Balanced Assessment Consortium 2014-2015 Technical Report.

[c]Test reliability is based on the 2014-2015 NYC Common Core Mathematics Test. Source: New York State Testing Program 2015: English Language Arts and Mathematics Grades 3-8 Technical Report.

[d]Test reliability is based on the 2014-2015 FCAT. Source: Florida Standards Assessments 2014-2015 Evidence of Reliability and Validity.

Appendix Table A.2

Background Characteristics for 7th-Graders in Analysis Sample, by Treatment Status

Characteristics	Program Group	Control Group	Estimated Difference	Standard Deviations in Effect Size of Estimated Difference	P-value for Estimated Difference
Age (years)	11	11	0.02	0.022	0.278
Eligible for free or reduced-price lunch (average % of students)	83.8	81.4	2.4	0.061	0.301
Race/Ethnicity (average % of students)					
White non-Hispanic	10.1	11.1	-1.0	-0.030	0.537
Black non-Hispanic	30.6	31.2	-0.6	-0.016	0.753
Hispanic	53.2	50.3	2.9	0.060	0.287
Asian	5.2	6.2	-1.0	-0.040	0.498
Other	0.9	1.1	-0.2	-0.021	0.449
Male (average % of students)	48.7	46.6	2.1	0.042	0.144
English-language learner (average % of students)	12.4	15.2	-2.8	-0.075	0.068
Special education status (average % of students)	15.5	14.0	1.5	0.044	0.205
Math achievement at baseline					
Standardized test score	-0.01	0.04	-0.05	-0.048	0.376
At or above proficiency (average % of students)	51.5	53.0	-1.5	-0.031	0.504

(continued)

Appendix Table A.2 (continued)

SOURCES: School district student records from the 2012-2013 academic year for Cohort 1, and from the 2013-2014 academic year for Cohort 2.

NOTES: Due to missing values, the number of students included varies by characteristics, ranging from 7,265 to 7,918 for the control group schools, and 8,231 to 8,985 for the program group schools.

The estimated differences for student-level data are regression-adjusted using hierarchical linear models to account for the nested structure of the data (with students nested within classes and classes nested within schools). The models control for indicators of random assignment blocks.

The values in the column labeled "Program Group" are the weighted average of the observed district means for schools or students randomly assigned to the program group (using number of program group schools in each district as weight). The control group values in the next column are the regression-adjusted means using the observed distribution of the program group across blocks as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

To examine if there is any systematic difference between students in the program and control group schools, an F-test was conducted for the full sample in a logistic regression model controlling for the following variables: indicators of random assignment strata, all student characteristics reported in this table and corresponding missing indicators. The p-value of the test is 0.51.

Appendix Table A.3

Background Characteristics for 8th-Graders in Analysis Sample, by Treatment Status

Characteristics	Program Group	Control Group	Estimated Difference	Standard Deviations in Effect Size of Estimated Difference	P-value for Estimated Difference
Age (years)	12	12	0.02	0.028	0.135
Eligible for free or reduced-price lunch (average % of students)	83.9	84.2	-0.2	-0.006	0.901
Race/Ethnicity (average % of students)					
White non-Hispanic	9.8	9.8	0.0	0.001	0.979
Black non-Hispanic	30.2	30.3	-0.1	-0.003	0.960
Hispanic	53.6	53.5	0.0	0.001	0.988
Asian	5.1	4.9	0.2	0.006	0.894
Other	1.3	1.3	-0.1	-0.006	0.836
Male (average % of students)	47.4	47.6	-0.1	-0.003	0.916
English-language learner (average % of students)	12.1	12.8	-0.7	-0.021	0.653
Special education status (average % of students)	14.6	12.0	2.7	0.081 *	0.009
Math achievement at baseline					
Standardized test score	-0.10	-0.04	-0.06	-0.057	0.267
At or above proficiency (average % of students)	40.8	40.5	0.2	0.005	0.915

SOURCES: School district student records from the 2012-2013 academic year for Cohort 1, and from the 2013-2014 academic year for Cohort 2.

NOTES: Due to missing values, the number of students included varies by characteristics, ranging from 7,136 to 7,847 for the control group schools, and 8,171 to 8,843 for the program group schools.

The estimated differences for student-level data are regression-adjusted using hierarchical linear models to account for the nested structure of the data (with students nested within classes and classes nested within schools). The models control for indicators of random assignment blocks.

The values in the column labeled "Program Group" are the weighted average of the observed district means for schools or students randomly assigned to the program group (using number of program group schools in each district as weight). The control group values in the next column are the regression-adjusted means using the observed distribution of the program group across blocks as the basis of the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

To examine if there is any systematic difference between students in the program and control group schools, an F-test was conducted for the full sample in a logistic regression model controlling for the following variables: indicators of random assignment strata, all student characteristics reported in this table, and corresponding missing indicators. The p-value of the test is less than 0.01.

Appendix Table A.4

Impact of PowerTeaching on Minutes Students Spent Doing Group Activities During the Math Block (Activities Are Not Mutually Exclusive), by Student Rank

	Top Third				Middle Third				Bottom Third			
Item	Program Mean Minutes	Control Mean Minutes	Impact Estimate	P-Value	Program Mean Minutes	Control Mean Minutes	Impact Estimate	P-Value	Program Mean Minutes	Control Mean Minutes	Impact Estimate	P-Value
Instructional activities												
Discussing and jointly working on multiple choice exercises	2.70	3.16	-0.46	0.69	2.47	2.26	0.21	0.75	1.52	1.67	-0.15	0.81
Discussing and jointly working on a problem with multiple solution methods	9.21	8.74	0.47	0.84	7.82	6.58	1.23	0.51	7.46	5.53	1.93	0.30
Jointly solving mathematical problems by using an algorithm	9.72	6.43	3.30	0.07	8.50	4.82	3.69	0.00 *	7.11	3.70	3.41	0.01 *
Representing and analyzing relation-ships using tables, charts, or models	8.22	7.67	0.55	0.83	6.73	6.08	0.65	0.73	5.24	5.06	0.18	0.92
Jointly reading a textbook or supplementary materials	3.33	4.07	-0.75	0.62	2.19	3.38	-1.18	0.30	2.00	2.67	-0.67	0.55
Jointly applying mathematical concepts to "real world" problems	9.74	7.97	1.77	0.49	7.78	6.64	1.14	0.53	5.99	5.58	0.41	0.82
Jointly analyzing data to make inferences or draw conclusions	6.85	7.16	-0.31	0.89	5.75	5.28	0.47	0.78	4.11	4.98	-0.87	0.61
Asking other students clarifying questions	5.73	4.85	0.88	0.49	5.62	4.09	1.53	0.18	3.90	3.90	0.00	1.00
Building on or challenging the ideas of other students	6.31	6.51	-0.20	0.91	3.86	3.67	0.19	0.84	1.79	1.80	0.00	0.99
Making fun of, belittling, or bothering a partner or group members	0.79	0.50	0.29	0.27	0.76	0.53	0.23	0.39	2.79	1.00	1.79	0.01 *
Suggesting a strategy to a partner or group members	7.84	7.19	0.65	0.75	4.45	3.72	0.74	0.46	1.85	1.60	0.25	0.55

(continued)

Appendix Table A.4 (continued)

Item	Top Third				Middle Third				Bottom Third			
	Program Mean Minutes	Control Mean Minutes	Impact Estimate	P-Value	Program Mean Minutes	Control Mean Minutes	Impact Estimate	P-Value	Program Mean Minutes	Control Mean Minutes	Impact Estimate	P-Value
Asking other students for help in solving a math problem	4.15	3.95	0.20	0.86	5.40	3.72	1.69	0.08	4.87	4.72	0.16	0.91
Helping other students solve math problems	9.01	8.03	0.98	0.64	5.11	4.23	0.88	0.39	1.78	1.70	0.08	0.86
Exchanging work with other students for review and checking	6.97	6.81	0.16	0.91	4.77	4.45	0.32	0.72	2.88	3.58	-0.70	0.43
Explaining a solution to a problem to other students	9.01	7.76	1.25	0.53	5.54	4.75	0.80	0.51	2.20	2.13	0.07	0.90
Engaging in discussion or activities not not related to assigned activity or task	2.09	1.30	0.80	0.07	2.60	1.62	0.98	0.16	6.44	2.96	3.48	0.01 *
Sample size				918				1018				957

SOURCE: Teacher logs administered in the spring of 2015.

NOTES: All estimations are based on a three-level hierarchical model with individual logs nested within teachers and teachers nested within schools. A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

In the instructional logs, teachers were asked how much time randomly selected students spent working in three different configurations during the mathematics period: in groups, in pairs, and individually.

If teachers indicated that selected students worked in a group during the mathematics period, they were asked approximately what proportion of group time the students spent engaging in specific activities using a five-category scale: 0 percent, less than 10 percent, 10 to 25 percent, 26 to 50 percent, and more than 50 percent. In the analysis of the logs, the research team converted the proportion of group time into minutes by calculating the midpoint of the ranges in this scale and then multiplying the midpoint of the selected range by the total amount of time spent working in a group. If selected students did not spend time working in a group, the time spent on each activity was set to zero.

Appendix Table A.5

Impact of PowerTeaching on Student Math Achievement in Grade 7 for Analysis Sample (2014-2015 Academic Year), by Full Sample and Cohort

Sample	Program Group	Control Group	Estimate Impact[a]	P-Value	95% Confidence Interval of the Impact[b] (in Standard Deviation Units) -0.25 0 0.25
Full sample (confirmatory)					
Grade 7					
Standardized state math test score	-0.10	-0.02	-0.08	0.112	
% at or above proficiency level	23.2	24.8	-1.6	0.338	
Cohort (exploratory)					
Cohort 1					
Standardized state math test score	-0.13	-0.01	-0.11	0.098	
% at or above proficiency level	21.3	25.1	-3.8	0.121	
Cohort 2					
Standardized state math test score	-0.14	-0.08	-0.06	0.368	
% at or above proficiency level	37.9	37.9	0.0	0.990	

SOURCES: School district student records data from the 2014-2015 and 2012-2013 academic years (for Cohort 1), and the 2013-2014 academic year (for Cohort 2).

NOTES: The analysis sample consists of students from 58 schools (30 program group schools and 28 control group schools) and includes any student who had a valid spring test score in the spring of 2015.

The student sample size for Grade 7 is 16,903 students (8,985 in the program group schools and 7,918 in the control group schools). The student sample size for Cohort 1 is 9,117 students (5,105 in the program group schools and 4,012 in the control group schools). The student sample size for Cohort 2 is 7,786 students (3,880 in the program group schools and 3,906 in the control group schools).

The estimated impacts are based on a two-level model with students nested within schools, controlling for random assignment block and school- and student-level covariates. The program group and control group columns display regression-adjusted mean outcomes for each group, using the mean covariate values for students in the program group as the basis for the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

The difference between the impact estimates for Cohort 1 and Cohort 2 is not significant at the 5 percent level (p-value = 0.59).

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

[a]Impact estimates are measured in standard deviations in effect size or percentage point.

[b]The confidence intervals are for the impact estimates on the standardized test scores and proficiency classification.

Appendix Table A.6

Impact of PowerTeaching on Student Math Achievement in Grade 8 for Analysis Sample (2014-2015 Academic Year), by Full Sample and Cohort

Sample	Program Group	Control Group	Estimate Impact[a]	P-Value	95% Confidence Interval of the Impact[b] (in Standard Deviation Units) -0.25	0	0.25
Full sample (confirmatory)							
Grade 8							
Standardized state math test score	-0.09	-0.01	-0.08	0.19			
% at or above proficiency level	22.3	24.8	-2.5	0.23			
Cohort (exploratory)							
Cohort 1							
Standardized state math test score	-0.08	0.00	-0.08	0.23			
% at or above proficiency level	20.5	23.7	-3.2	0.22			
Cohort 2							
Standardized state math test score	-0.14	-0.05	-0.09	0.34			
% at or above proficiency level	37.0	39.0	-2.0	0.53			

SOURCES: School district student records data from the 2014-2015 and 2012-2013 academic years (for Cohort 1), and the 2013-2014 academic year (for Cohort 2).

NOTES: The analysis sample consists of students from 58 schools (30 program group schools and 28 control group schools) and includes any student who had a valid spring test score in the spring of 2015.

The student sample size for Grade 8 is 16,690 students (8,843 in the program group schools and 7,847 in the control group schools). The student sample size for Cohort 1 is 9,529 students (5,221 in the program group schools and 4,308 in the control group schools). The student sample size for Cohort 2 is 7,161 students (3,622 in the program group schools and 3,539 in the control group schools).

The estimated impacts are based on a two-level model with students nested within schools, controlling for random assignment block and school- and student-level covariates. The program group and control group columns display regression-adjusted mean outcomes for each group, using the mean covariate values for students in the program group as the basis for the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

The difference between the impact estimates for Cohort 1 and Cohort 2 is not significant at the 5 percent level (p-value = 0.92).

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

[a]Impact estimates are measured in standard deviations in effect size or percentage point.

[b]The confidence intervals are for the impact estimates on the standardized test scores and proficiency classification.

Appendix Table A.7

Impact of PowerTeaching on Student Math Achievement in Grades 6-8 (2014-2015 Academic Year), Robustness Checks

A. No Covariates in Estimation Model

Grade	Program Group	Control Group	Estimated Impact (in Standard Deviations in Effect Size)	P-Value
Grade 6				
Standardized state math test score	-0.05	-0.03	-0.02	0.821
Grade 7				
Standardized state math test score	-0.10	0.02	-0.12	0.094
Grade 8				
Standardized state math test score	-0.09	0.03	-0.12	0.144

B. With Only Baseline Test Score as Covariates in Estimation Model

Grade	Program Group	Control Group	Estimated Impact (in Standard Deviations in Effect Size)	P-Value
Grade 6				
Standardized state math test score	-0.05	-0.05	0.00	0.935
Grade 7				
Standardized state math test score	-0.10	-0.02	-0.08	0.125
Grade 8				
Standardized state math test score	-0.09	-0.01	-0.08	0.170

(continued)

Appendix Table A.7 (conitinued)

SOURCES: Schools district student records data from the 2014-2015 and 2012-2013 academic years (for Cohort 1), and the 2013-2014 academic year (for Cohort 2).

NOTES: The analysis sample consists of students from 58 schools (30 program group schools and 28 control group schools) and includes any student who had a valid spring test score in the spring of 2015.

The student sample size for Grade 6 is 15,950 students (8,534 in the program group schools and 7,416 in the control group schools). The student sample size for the Grade 7 is 16,903 students (8,985 in the program group schools and 7,918 in the control group schools). The student sample size for Grade 8 is 16,690 students (8,843 in the program group schools and 7,847 in the control group schools).

The estimated impacts are based on a two-level model with students nested within schools, controlling for random assignment block and school- and student-level covariates. The program group and control group columns display regression-adjusted mean outcomes for each group, using the mean covariate values for students in the program group as the basis for the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

Appendix Table A.8

Impact of PowerTeaching on Student Math Achievement in Grades 6-8 for Analysis Sample (2014-2015 Academic Year), Stable Samples

Grade	Program Group	Control Group	Estimate Impact[a]	P-Value	95% Confidence Interval of the Impact[b] (in Standard Deviation Units) -0.25 0 0.25
Definition 1					
Grade 6					
Standardized state math test score	-0.03	-0.04	0.01	0.841	
% at or above proficiency level	22.3	23.3	-1.0	0.575	
Grade 7					
Standardized state math test score	-0.08	0.00	-0.08	0.130	
% at or above proficiency level	23.9	25.7	-1.8	0.323	
Grade 8					
Standardized state math test score	-0.08	0.00	-0.08	0.190	
% at or above proficiency level	22.7	25.4	-2.6	0.219	
Definition 2					
Grade 6					
Standardized state math test score	-0.03	-0.04	0.01	0.841	
% at or above proficiency level	22.3	23.3	-1.0	0.575	
Grade 7					
Standardized state math test score	-0.09	0.00	-0.08	0.105	
% at or above proficiency level	23.7	25.5	-1.8	0.290	
Grade 8					
Standardized state math test score	-0.08	0.00	-0.08	0.194	
% at or above proficiency level	22.7	25.2	-2.6	0.225	

(continued)

Table A.8 (continued)

SOURCES: District student records data from the 2014-2015 and 2012-2013 academic years (for Cohort 1), and the 2013-2014 academic year (for Cohort 2).

NOTES: The stable analysis sample consists of students from 58 schools (30 program group schools and 28 control group schools) and includes any student who had a valid test score in the spring of 2015 and stayed in the study schools for all implementation years (Definition 1), or any student who had a valid test score in the spring of 2015 and were enrolled in the study schools for the 2014-2015 academic year (Definition 2).

The student sample size for Grade 6 for both panel is 15,174 students (8,134 in the program group schools and 7,040 in the control group schools). The student sample size for the Grade 7 is 14,273 students (7,706 in the program group schools and 6,567 in the control group schools) for Definition 1, and 16,058 students (8,593 in the program group schools and 7,465 in the control group schools) for Definition 2. The student sample size for Grade 8 is 15,099 students (8,051 in the program group schools and 7,048 in the control group schools) for Definition 1, and 15,849 students (8,450 in the program group schools and 7,399 in the control group schools) for Definition 2.

The estimated impacts are based on a two-level model with students nested within schools, controlling for random assignment block and school- and student-level covariates. The program group and control group columns display regression-adjusted mean outcomes for each group, using the mean covariate values for students in the program group as the basis for the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

[a]Impact estimates are masured in standard deviations in effect size or percentage point.

[b]The confidence intervals are for the impact estimates on the standardized test scores and proficiency classification.

Appendix Table A.9

Impact of PowerTeaching on Student Math Achievement in Grade 7 for Analysis Sample (2014-2015 Academic Year), by Student Subgroup

Subgroup	Number of Observations	Program Group	Control Group	Estimated Impact (in Standard Deviations in Effect Size)	P-Value		P-Value for Difference Between Subgroups
By performance rank at baseline							1.000
Top third	5,164	0.39	0.48	-0.08	0.322		
Middle third	5,146	-0.04	0.02	-0.07	0.199		
Bottom third	5,186	-0.65	-0.56	-0.09	0.108		
By proficiency level at baseline							0.885
At or above proficiency	8,974	0.36	0.42	-0.07	0.349		
Below proficiency	6,522	-0.39	-0.31	-0.08	0.107		
By gender							0.391
Boys	8,540	-0.12	-0.08	-0.04	0.470		
Girls	8,363	-0.08	0.02	-0.10	0.039	*	
By race/ethnicity							1.000
Hispanic	10,626	-0.11	-0.02	-0.09	0.078		
White, Non-Hispanic	2,102	0.10	0.17	-0.08	0.392		
Black, Non-Hispanic	2,782	-0.23	-0.13	-0.10	0.082		
By family income							0.632
Eligible for free and reduced-price lunch	12,754	-0.14	-0.06	-0.08	0.095		
Not eligible for free and reduced-price lunch	3,005	0.16	0.28	-0.12	0.116		
By English-language learner (ELL) status							0.511
ELL	2,487	-0.36	-0.24	-0.12	0.034	*	
Non-ELL	13,266	-0.04	0.03	-0.07	0.180		
By special education (SPED) status							0.568
SPED	2,273	-0.51	-0.39	-0.12	0.111		
Non-SPED	14,596	0.00	0.06	-0.07	0.185		

SOURCES: District student records data from the 2014-2015 and 2012-2013 academic years (for Cohort 1), and the 2013-2014 academic year (for Cohort 2).

NOTES: The analysis sample consists of students from 58 schools (30 program group schools and 28 control group schools) and includes any student who had a valid spring test score in the spring of 2015.

The estimated impacts are based on a two-level model with students nested within schools, controlling for random assignment block and school- and student-level covariates. The program group and control group columns display regression-adjusted mean outcomes for each group, using the mean covariate values for students in the program group as the basis for the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

Appendix Table A.10

Impact of PowerTeaching on Student Math Achievement in Grade 8 for Analysis Sample (SY 2014-2015), by Student Subgroup

Subgroup	Number of Observations	Program Group	Control Group	Estimated Impact (in Standard Deviations in Effect Size)	P-Value	P-Value for Difference Between Subgroups
By performance rank at baseline						1.000
Top third	4,492.00	0.38	0.47	-0.09	0.181	
Middle third	5,354.00	-0.06	-0.01	-0.05	0.478	
Bottom third	5,461.00	-0.52	-0.46	-0.06	0.295	
By proficiency level at baseline						0.940
At or above proficiency	6,944.00	0.32	0.39	-0.07	0.341	
Below proficiency	8,365.00	-0.27	-0.21	-0.07	0.212	
By gender						0.505
Boys	8,360.00	-0.16	-0.05	-0.11	0.101	
Girls	8,330.00	-0.05	0.00	-0.05	0.399	
By race/ethnicity						1.000
Hispanic	10,813.00	-0.08	-0.04	-0.04	0.528	
White, Non-Hispanic	1,905.00	0.07	0.13	-0.06	0.561	
Black, Non-Hispanic	2,669.00	-0.18	-0.08	-0.11	0.117	
By family income						0.759
Eligible for free and reduced-price lunch	12,655.00	-0.10	-0.03	-0.07	0.207	
Not eligible for free and reduced-price lunch	2,862.00	0.04	0.14	-0.10	0.236	
By English language learner (ELL) status						0.598
ELL	2,269.00	-0.24	-0.19	-0.05	0.446	
Non-ELL	13,251.00	-0.05	0.04	-0.09	0.124	
By special education (SPED) status						0.329
SPED	2,088.00	-0.48	-0.49	0.01	0.921	
Non-SPED	14,580.00	0.00	0.09	-0.09	0.150	

(continued)

Table A.10 (continued)

SOURCES: District student records data from the 2014-2015 and 2012-2013 academic years (for Cohort 1) , and the 2013-2014 academic year (for Cohort 2).

NOTES: The analysis sample consists of students from 58 schools (30 program group schools and 28 control group schools) and includes any student who had a valid spring test score in the spring of 2015.

The estimated impacts are based on a two-level model with students nested within schools, controlling for random assignment block and school- and student-level covariates. The program group and control group columns display regression-adjusted mean outcomes for each group, using the mean covariate values for students in the program group schools as the basis for the adjustment. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

Figure 1.1

Logic Model for the Success for All Math PowerTeaching Program in Middle Schools

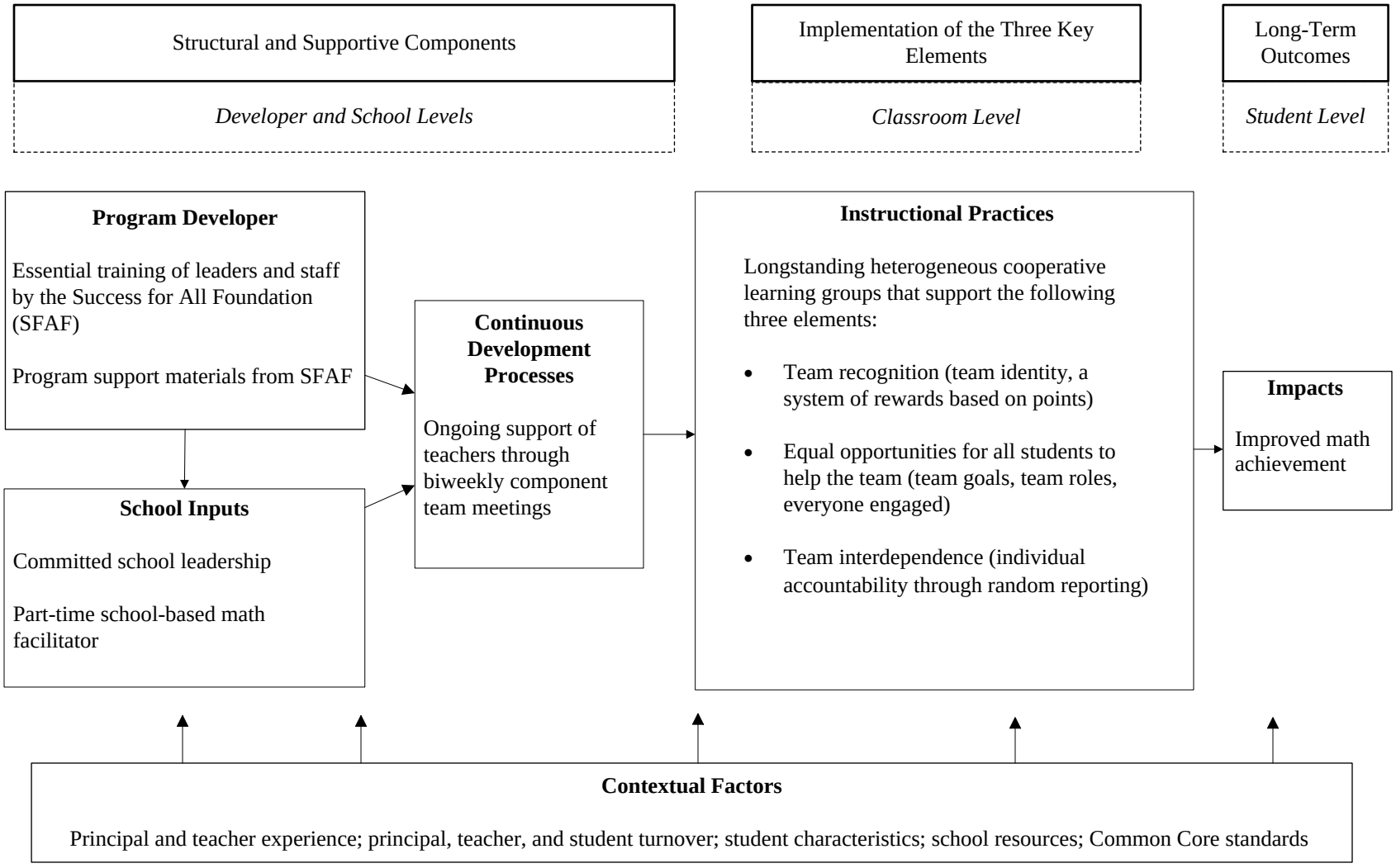


Figure 2.1

Timeline and Sample Configuration of the Evaluation

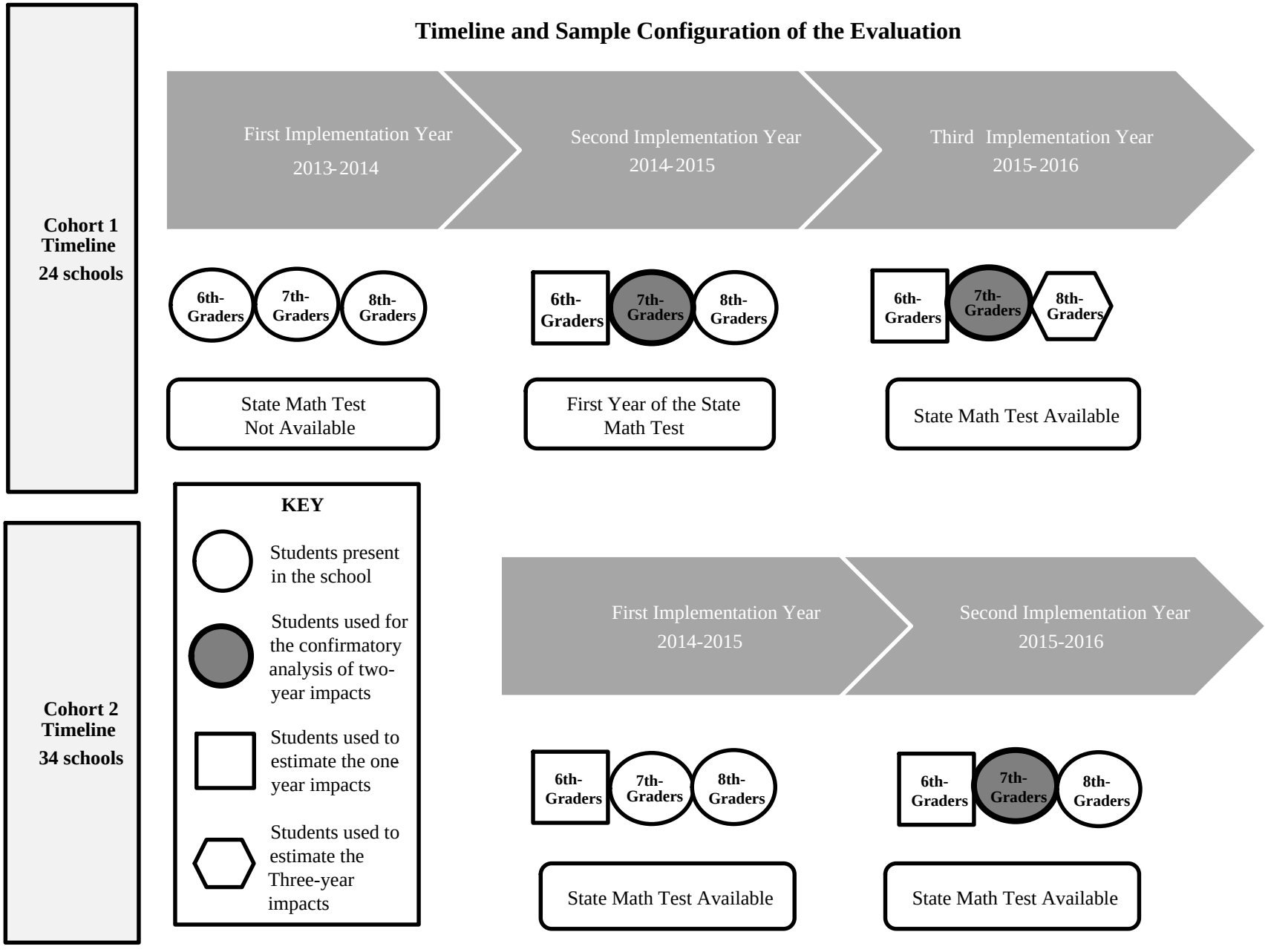
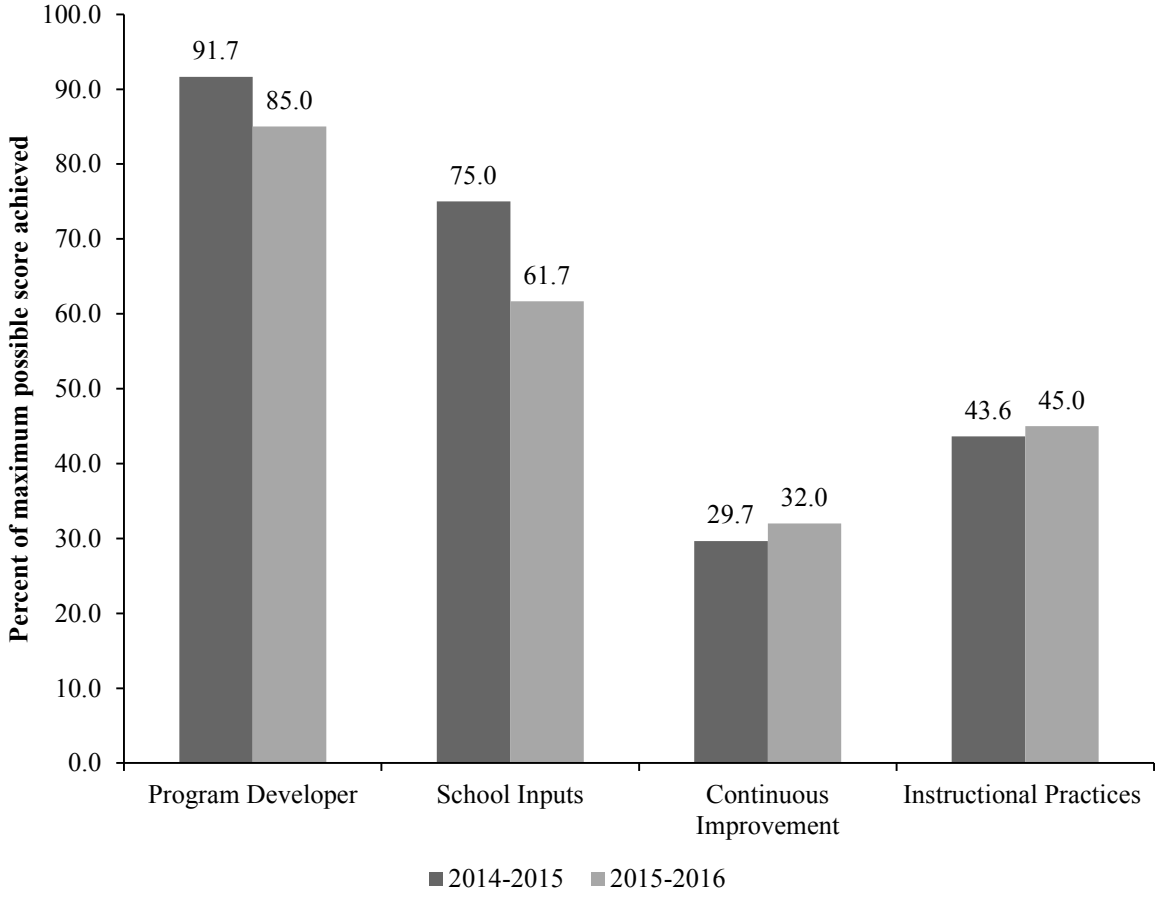


Figure 3.1

School Achievement Snapshot Scores, by Category and School Year, Study Schools Only

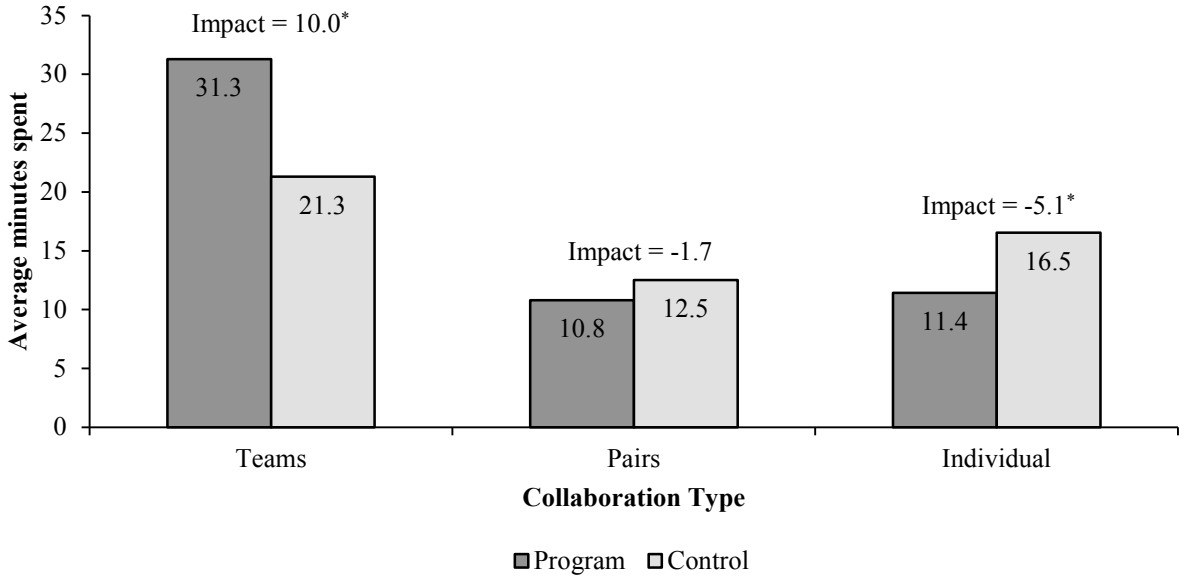


SOURCE: 2014-2015 and 2015-2016 School Achievement Snapshots.

NOTE: The sample includes 12 schools in Cohort 1 and 18 schools in Cohort 2. For Cohort 1 schools, the 2014-2015 and 2015-2016 academic years were Years 2 and 3 of implementation. For Cohort 2 schools, the 2014-2015 and 2015-2016 academic years were Years 1 and 2 of implementation.

Figure 3.2

Average Minutes Students Spent During Math Block, by Collaboration Type

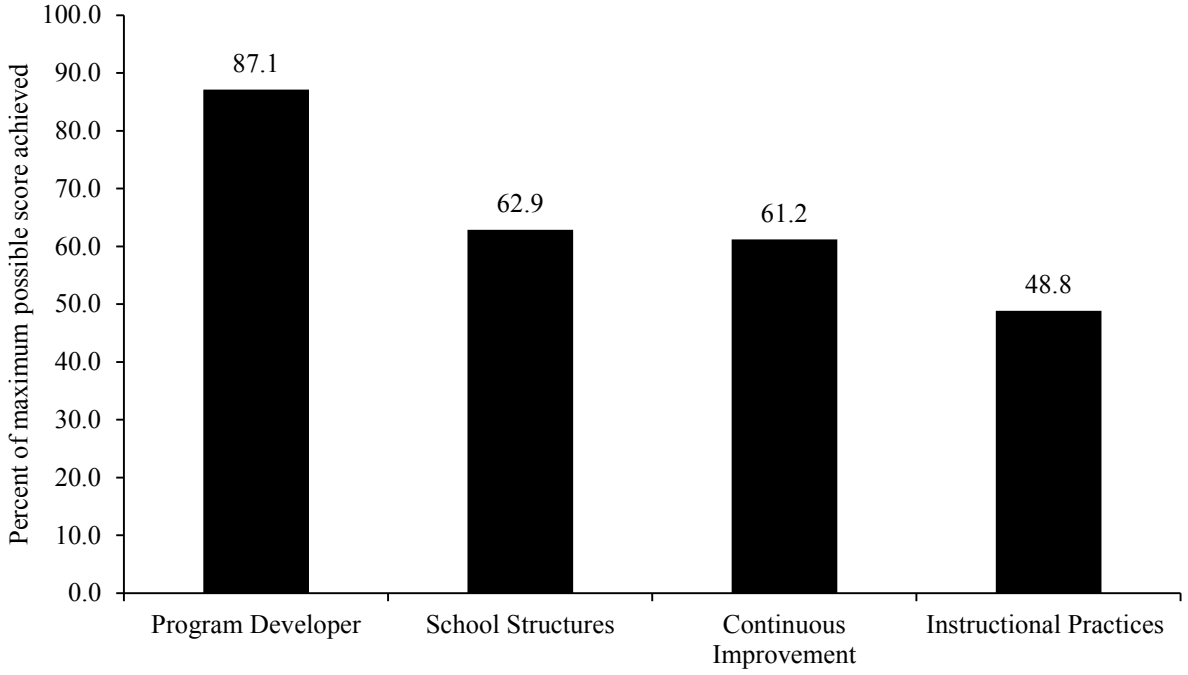


SOURCE: Teacher logs administered in spring 2015.

NOTES: All estimations are based on a three-level hierarchical model with individual logs nested within teachers and teachers nested within schools. A two-tailed t-test was applied to each estimated difference. Statistical significance is indicated by an asterisk (*) when the p-value is less than or equal to 5 percent.

Figure 5.1

2015-2016 School Achievement Snapshot Scores, by Category, Scale-Up Schools Only



SOURCE: 2015-2016 School Achievement Snapshots.

NOTE: In total, there were 41 non-study scale-up schools by 2015. Snapshot scores were available for 39 of these schools.

Box 1.1

Cooperative Learning in Action: Becoming a Super Team

Students are working in teams of four in Ms. Martin's seventh-grade math class. When it comes time to collect homework, it turns out that three members of the purple team have brought in their homework but one, Rudy, has not. The three team members that did their homework are disappointed; not only will they not receive a team "celebration point" for Rudy's homework, but they will not receive the extra point that teams earn when all members of the team bring in their homework. What is more, the team's goal for the week was to improve on completing homework, so now they are behind in their progress toward reaching their goal and accumulating enough points by the end of the math unit to be rated as a Super Team. Rudy promises to make a greater effort to bring in his homework. Ms. Martin then gives the teams a math problem to solve. Rudy and another team member, Malia, have some ideas about how to approach the problem, but the other two are stumped. Malia and Rudy share their ideas with the others to help them understand the problem and how to solve it, because they know that Ms. Martin might randomly call on any one of them to represent the team and explain its solution to the math problem. After giving the teams enough time to work on the problem, Ms. Martin randomly calls on a member of the purple team, Rosario, to share the solution. Rosario's explanation is clear and correct and he receives a high score that counts towards his grade. His team also receives celebration points. By the end of the unit, Rudy has gotten much better at bringing in his homework and the team continues to work collaboratively on math problems. As a result, the purple team receives enough celebration points to become a Super Team and the class celebrates their achievement.

More Typical Group Work That Is Not Cooperative Learning

Ms. Martin collects everyone's homework at the start of class. Most, but not all, have finished it. She then gives the class a math problem to solve and asks the students to form groups of three. Malia, Rosario, and Marie — who are good friends — get together. Malia has some ideas about how to solve the problem, but the other two are stumped. Malia tells the others her solution and assures them that the answer is right, so Rosario and Marie relax. Ms. Martin calls on Rosario to share the solution to the math problem. Rosario tries his best to repeat Malia's answer but knows he is getting it wrong. "Malia can explain it better," he says. "Ok, Malia, what's the answer?" asks Ms. Martin. Malia's answer is correct and very clearly explained. Ms. Martin is pleased because the class has gotten to hear the correct answer explained well.

About MDRC

MDRC is a nonprofit, nonpartisan social and education policy research organization dedicated to learning what works to improve the well-being of low-income people. Through its research and the active communication of its findings, MDRC seeks to enhance the effectiveness of social and education policies and programs.

Founded in 1974 and located in New York; Oakland, California; Washington, DC; and Los Angeles, MDRC is best known for mounting rigorous, large-scale, real-world tests of new and existing policies and programs. Its projects are a mix of demonstrations (field tests of promising new program approaches) and evaluations of ongoing government and community initiatives. MDRC's staff members bring an unusual combination of research and organizational experience to their work, providing expertise on the latest in qualitative and quantitative methods and on program design, development, implementation, and management. MDRC seeks to learn not just whether a program is effective but also how and why the program's effects occur. In addition, it tries to place each project's findings in the broader context of related research — in order to build knowledge about what works across the social and education policy fields. MDRC's findings, lessons, and best practices are proactively shared with a broad audience in the policy and practitioner community as well as with the general public and the media.

Over the years, MDRC has brought its unique approach to an ever-growing range of policy areas and target populations. Once known primarily for evaluations of state welfare-to-work programs, today MDRC is also studying public school reforms, employment programs for ex-offenders and people with disabilities, and programs to help low-income students succeed in college. MDRC's projects are organized into five areas:

- Promoting Family Well-Being and Children's Development
- Improving Public Education
- Raising Academic Achievement and Persistence in College
- Supporting Low-Wage Workers and Communities
- Overcoming Barriers to Employment

Working in almost every state, all of the nation's largest cities, and Canada and the United Kingdom, MDRC conducts its projects in partnership with national, state, and local governments, public school systems, community organizations, and numerous private philanthropies.