SCALING UP THE SUCCESS FOR ALL MODEL OF SCHOOL REFORM

Final Report from the Investing in Innovation (i3) Evaluation

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

Janet Quint
Pei Zhu
Rekha Balu
Shelley Rappaport
Micah DeLaurentis

September 2015

Scaling Up the Success for All Model of School Reform

Final Report from the Investing in Innovation (i3) Evaluation

Janet Quint
Pei Zhu
Rekha Balu
Shelley Rappaport
Micah DeLaurentis

with

Emma Alterman
Colin Bottles
Emily Pramik

September 2015

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 its elementary school whole-school reform 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 JBP 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.

Overview

Success for All (SFA) is one of the best-known school reform initiatives. Combining a challenging reading program, whole-school reform elements, and an emphasis on continuous improvement, it seeks to ensure that every child learns to read well in the elementary grades. In 2010, the Success for All Foundation (SFAF) received a scale-up grant under the U.S. Department of Education's Investing in Innovation (i3) program. This third and final report from the independent evaluation of the i3 scale-up examines the program's implementation and impacts over three years, its incremental cost, and the scale-up process itself. Thirty-seven evaluation schools in five school districts were randomly assigned either to a program group of 19 schools that received SFA or to a control group of 18 schools that used alternative reading programs. This design supports causal impact findings for the average school assigned to SFA. Overall, the evaluation led to several key findings:

- Although SFA was implemented with adequate fidelity at the great majority of schools that adopted it, resource constraints prevented some schools from putting in place some of its key features, including a full-time facilitator and SFA's computerized tutoring program.

- Program group and control group schools were different in some respects (for example, SFA schools were unique in placing students in cross-grade ability groups for reading, and SFA teachers made greater use of cooperative learning) but similar in others.

- SFA is an effective vehicle for teaching phonics. In the average SFA school, the program registered a notable, statistically significant impact on a measure of phonics skills for second-graders who had been in SFA for all three years, compared with their control group counterparts. Students in the average SFA school did not outperform their counterparts in the average control group school on tests of reading fluency or comprehension.

- For a subgroup of special concern to policymakers and practitioners — students entering school with low preliteracy skills — SFA appears to be especially effective. Second-graders in the average SFA school who had started kindergarten in the bottom half of the sample in terms of their knowledge of the alphabet and their ability to sound out words registered significantly higher scores on measures of phonics skills, word recognition, and reading fluency than similar students in control group schools. The impact on comprehension for this group was also positive but not statistically significant. The program did not significantly affect outcomes for the subgroup of students who started kindergarten in the top half of the sample in terms of phonetic skills.

- In a case study district, the direct expenditures for additional reading facilitator time, after-school tutoring, materials, and professional development were estimated to cost $119 more per student per year in SFA schools than in control group schools. Including the extra time that SFA principals devoted to the program and that coaches and teachers spent in training, the extra cost of space for storing SFA materials, and other factors, program group schools spent about $227 worth of resources per student per year more than control group schools.

- Through the fourth year of the i3 grant, SFA was put in place in 447 new schools and reached an estimated 276,000 students. These numbers fell below SFAF's ambitious goals but represent a notable achievement in a period of staff layoffs and other cutbacks in many schools and districts.

Contents

Chapter

Appendix

List of Exhibits

Table

Table

Table

Table

Figure

Figure

Figure

Box

Preface

Low reading skills remain a pressing problem in the United States. With the emergence of more demanding academic expectations — as represented by the Common Core State Standards — students will need stronger reading skills to master more complex content. Despite more than a decade of concentrated effort to improve reading instruction, elementary student reading achievement scores on the National Assessment of Educational Progress ("The Nation's Report Card") have increased only modestly, signaling the need for continued improvement. The Success for All Foundation (SFAF) has been active in reading instruction for some three decades, building evidence of the effectiveness of its approach in improving the reading skills of elementary school students. Thus, it was not a surprise that the Success for All reading program was one of the initial recipients of a federal Investing in Innovation (i3) scale-up grant designed to provide support for the expansion of evidence-based programs. This report, the last in a series of three, completes MDRC's evaluation of the SFA elementary reading program under i3.

SFAF works with schools to put in place a comprehensive approach to reading instruction. It involves a highly structured curriculum that covers the five key elements identified by the National Reading Panel (phonemic awareness, phonics, fluency, vocabulary, and comprehension); a strong emphasis on cooperative learning; cross-grade grouping of students by reading skills; extensive use of assessment data to adjust groupings and instruction; and tutoring for students in need of additional help. SFA also includes schoolwide structures that address attendance, behavior, and parental involvement to support student learning. Full implementation of this comprehensive approach has required substantial effort on the part of school staff already facing the challenges of funding cuts resulting from the recession as well as the new demands of the Common Core standards. Thus, the scale-up of SFA under i3 was especially ambitious, but also timely in light of current efforts to encourage evidence-based funding decisions and the expansion of proven programs.

The report provides findings on all aspects of the SFA reading scale-up initiative: Was the program implemented with fidelity as it was scaled up? Did SFA produce a difference in reading instruction compared with instruction in business-as-usual schools in the study? Did schools implementing SFA have better student achievement than alternative reading programs, on average and for key subgroups of students? What were the extra costs of implementing SFA? And how did SFA fare in its efforts to increase the number of schools implementing the program in a time of severe economic constraints? The answers to these questions are highly relevant as policymakers strive to increase the use of evidence in decision-making.

Gordon L. Berlin
President, MDRC

Acknowledgments

This third and final report of the Investing in Innovation (i3) scale-up evaluation of Success for All reflects the efforts of a great many people. As with the two previous reports, our first debt of gratitude is to the principals, SFA coordinators, research liaisons, and teachers at the schools that participated in this study; to staff members in the central offices of the five school districts who provided us with critical student records data as well as fiscal and staffing information; and to district coaches who took part in 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 the Success for All Foundation, Nancy Madden, Sharon Fox, and Jill Hanson responded promptly and patiently to our many information requests. Nancy Madden and Robert Slavin provided useful critiques at each stage of the report drafting process.

Pamela Wells and her capable team at Decision Information Resources, Inc., produced the student assessment data that are at the heart of the impact study. MDRC staff members Jo Anna Hunter and Nicole Morris worked with DIR to ensure that these data were collected on schedule and with minimal disruption to school operations. Thomas J. Smith played an active role in field research, data collection, and coding.

Within MDRC, Seth Muzzy helped prepare the principal and teacher surveys and the instructional logs that constitute important data sources for the report. Shirley James, Zuleka Abakoyas, Donna George, and Carmen Troche keyed surveys and logs. Emma Alterman, Colin Bottles, Cammie Brown, Zachary Pinto, Emily Pramik, and Gary Reynolds played important roles in data collection, analysis, and the preparation of this report.

Fred Doolittle ensured that the team received both material and moral support at every juncture. He, along with Gordon Berlin, Howard Bloom, Charles Michalopoulos, Jean Grossman, Robert Ivry, and Leigh Parise, carefully reviewed earlier drafts of the report and made comments that improved the final product. Kelly Granito kept the project on task and on budget, and Gina Price provided general fiscal oversight. Mario Flecha was his usual unflappable self in handling a variety of administrative and other tasks. Christopher Boland provided advice on preparing exhibits, and Stephanie Cowell prepared the manuscript for publication.

Jennie Kaufman edited the report with great sensitivity, and John Hutchins proffered welcome assistance as report production drew to a close.

The Authors

Executive Summary

In 2013, almost one-third of fourth-grade students in the United States scored below the "basic" level in reading, according to the National Assessment of Educational Progress (NAEP), also known as "The Nation's Report Card."[1] They could not locate relevant information in a text, make simple inferences based on what they read, or identify details to support an interpretation or conclusion. Students performing at this level lack the skills needed to demonstrate solid academic performance or to master challenging subject matter.

Success for All (SFA), one of the best-known school reform models, aims to improve the reading skills of all children but is especially directed at schools that serve large numbers of students from low-income families. First implemented in 1987, SFA combines a challenging reading program, whole-school reform elements, and an emphasis on continuous improvement, with the goal of ensuring that every child learns to read well in the elementary grades. SFA includes several specific features:

- A kindergarten through grade 6 reading program that emphasizes phonics for beginning readers and comprehension for all students
- Instruction that is characterized by "scripted," briskly paced lesson plans that make extensive use of cooperative learning in pairs and small groups
- Cross-grade ability grouping for reading, with many students leaving their homeroom to receive reading instruction from another teacher, and quarterly regrouping
- Frequent assessments of student learning
- Computerized small-group tutoring and individual tutoring for students who need additional assistance
- Staff committees ("Solutions Teams") that address academic, behavior, and attendance issues and that promote parent and community involvement
- Schoolwide and classroom programs to develop social and conflict resolution skills

[1] See National Assessment of Educational Progress, "Nation's Report Card: 2013 Mathematics and Reading" (2013), http://nationsreportcard.gov/reading_math_2013. The percentage of students performing below the basic level was 32 percent. NAEP is a congressionally authorized project of the National Center for Education Statistics within the Institute of Education Sciences in the U.S. Department of Education. NAEP tests have been conducted periodically in a number of subject areas since 1969.

- Initial and ongoing professional development for teachers and use of data to monitor progress and set goals

This is the third and final report from an independent evaluation of the scale-up demonstration of the SFA elementary school reading program. Both the demonstration and the evaluation have been funded under the U.S. Department of Education's Investing in Innovation (i3) competition. Conducted by MDRC — a nonprofit, nonpartisan education and social policy research organization — the evaluation examines SFA's implementation and impacts in five school districts over a three-year period (the 2011-2012 school year through the 2013-2014 school year). It also includes an analysis of program costs. Finally, it considers the scale-up process itself — the methods employed and the extent to which the Success for All Foundation (SFAF), the organization that developed and provides technical assistance to schools operating the program, achieved its scale-up goals.

Previous evaluations, both experimental and quasi-experimental, showed that students in SFA schools performed better on standardized tests than students receiving other reading programs. The most salient of these evaluations was a three-year randomized experiment involving 35 schools serving low-income families. In that study, schools were randomly assigned to use Success for All either in kindergarten through grade 2 (K-2) or grades 3 through 5, with the schools that received the program in the later grades serving as a control group for the K-2 schools.[2] Children in the K-2 schools scored significantly higher than their counterparts in the 3-5 schools on the main outcomes measured — three tests that assessed children's phonetic skills and comprehension. In other large-scale studies, results for students in SFA schools outstripped those for students in matched comparison schools.[3] The strength of this evidence was critical to the selection of the Success for All Foundation as one of only four recipients of five-year scale-up grants awarded in 2010 in the initial i3 funding competition.

The Evaluation Design

The i3 evaluation of SFA employs an experimental design, in which 37 schools in five school districts that participated in the scale-up effort were assigned at random to a program group or to a control group. The 19 program group schools received SFA in all grades. The 18 control group schools did not get the intervention and, instead, either continued with the same reading

[2] See Geoffrey D. Borman, Robert E. Slavin, Alan C. K. Cheung, Anne M. Chamberlain, Nancy A. Madden, and Bette Chambers, "Final Reading Outcomes of the National Randomized Field Trial of Success for All," *American Educational Research Journal* 44, no. 3 (2007): 701-731.

[3] See, for example, Brian Rowan, Richard Correnti, Robert J. Miller, and Eric M. Camburn, *School Improvement by Design: Lessons from a Study of Comprehensive School Reform Programs* (Philadelphia: Consortium for Policy Research in Education, 2009).

program that they had used previously or, in the case of some schools, adopted a new one. This design supports causal impact findings for the average school assigned to SFA.

Context for the Evaluation

It is useful to consider the economic and instructional contexts in which the SFA scale-up demonstration has unfolded. These contexts provide a framework through which to view the participating schools' ability to implement the full program model and SFAF's ability to meet its ambitious expansion goals. They also help to define the "counterfactual" — what happens in the absence of the program. Only to the extent that SFA differs from the counterfactual is the program likely to produce impacts. Two trends are worth noting:

- **The effects of the Great Recession and its aftermath.** At the point that SFAF was recruiting schools for the i3 scale-up, many schools and districts were trying to restore positions and services that had been cut as a result of the recession. Furthermore, principals felt that they had less discretion in spending their schools' allocations than had been the case in the past. These circumstances added a new dimension to the challenges already associated with selecting and implementing a new and demanding reading program in high-poverty schools.

- **Heightened focus on reading instruction.** Over the period since SFA was first developed in 1987, reading instruction in the United States has changed markedly. For example, the influence of the National Reading Panel report of 2000, the passage of No Child Left Behind in 2001 (and, as a result, the creation of the Reading First program and the advent of high-stakes testing for grades 3 through 8), the rise of Response to Intervention reading support strategies, and the introduction of the Common Core standards have all contributed to an increased emphasis on phonics and additional interventions for struggling readers. These developments have had the effect of narrowing the differences between schools adopting SFA and schools using other reading programs and have made it harder than it used to be for SFA to "beat the competition."

Findings

Implementation

While Success for All was implemented with adequate fidelity at the large majority of schools that adopted it, resource constraints prevented some schools from putting in place some of the program's key features, including a full-time program facilitator and SFA's own computerized tutoring for students in need of instruction beyond the classroom. Reading instruction in program group schools (or "SFA schools") and control group schools was markedly different in some respects: Placing students by ability level in reading groups that crossed grade levels was unique to the program group schools, and program group teachers used cooperative learning as an instructional method more frequently than their counterparts in the control group schools. In other respects, the two groups of schools did not differ greatly. Despite some implementation challenges, 93 percent of the principals and 70 percent of the teachers in SFA schools who responded to surveys agreed that the SFA program benefited their schools.

Effects on Phonics

The evidence indicates that SFA is an effective vehicle for teaching phonics. In the average SFA school, the program registered a positive and statistically significant third-year impact on a strong measure of phonetic abilities (one that asks students to read phonetically regular nonsense words) for second-graders who had been in SFA classrooms for all three years, compared with their counterparts in control group schools — the groups that make up the "confirmatory sample" for the i3 study. This effect was also found in the previous two years of the demonstration, when these students were kindergartners and first-graders. In the second year, SFA also produced a positive and statistically significant effect on another measure of phonetic skills; in Year 3, this effect remained positive but was no longer statistically significant.

Effects on Comprehension

In Year 3 as in previous years, in the average SFA school, students did not outperform their counterparts in the average control group school on measures of reading fluency or comprehension. The comprehension finding is true for second-grade students in the confirmatory sample as well as for students in the upper elementary grades, for whom the analysis is considered exploratory. (The comprehension finding contrasts with that of the previously cited experimental study of SFA, which found a positive and statistically significant effect on comprehension for second-graders, as well as with several well-regarded quasi-experimental studies that found positive although not statistically significant effects.)

Effects on Students with Low Preliteracy Skills

Students who start school with low preliteracy skills are of special concern to policymakers and practitioners. An exploratory analysis indicates that the program had notable third-year impacts on a subgroup of second-graders who, at the start of kindergarten, scored in the bottom half of the sample in terms of their knowledge of the alphabet and their ability to sound out words. In the average SFA school, the program produced positive and statistically significant impacts on measures of phonics skills, word recognition, and reading fluency for these students. The impact on comprehension was also positive, although it fell shy of meeting conventional standards of statistical significance. The program did not significantly affect these outcomes for the subgroup of students who started kindergarten in the top half of the sample in terms of phonetic skills.

Effects on Special Education and Grade Retention Rates

The program did not affect the rates at which students were held back to repeat a grade or at which they were identified for or declassified from special education.

Cost Analysis

The cost analysis makes use of the random assignment design within one case study district to assess the extent to which the district's SFA schools required additional resources to implement the program, relative to those used for alternative reading programs in the control group schools. In the study district, the direct expenditures for school-based reading facilitator time, after-school tutoring time, materials, and professional development were estimated to cost $119 more per student per year in SFA schools than in control group schools. Adding to this the additional time that SFA principals devoted to the program, the additional time that coaches and teachers spent in training, the extra cost of devoting space to storing SFA curriculum materials, and other factors, program group schools spent about $227 worth of resources per student per year more than control group schools to implement their respective reading programs.

Scale-Up

During the first four years of scale-up, SFA was put in place in 447 new schools with a total enrollment of some 218,000 students and, taking into account student turnover, is estimated to have reached some 276,000 students. While these numbers fell below the Success for All Foundation's ambitious initial goal of recruiting 760 schools within the first four years, they represent a notable achievement, especially in a period when many schools and districts were laying off staff and cutting back on programs.

Conclusion

The i3 scale-up has heightened the prominence of Success for All on the educational landscape. In an economic climate characterized by budgetary cutbacks that forced many school districts to cut staff and restrict program offerings, the Success for All Foundation projects that it will have reached almost 400,000 students in 540 schools by the end of the i3 grant. The scale-up findings show that, for a modest investment, SFA reliably improves the decoding skills of students in kindergarten through second grade, and that it is especially beneficial for students who begin in the lower half in these skills.

Continuous improvement is a key element of the Success for All program. With a greater focus on improving comprehension and broader implementation of its tutoring component, Success for All might make an even bigger difference, and for more students, than it already does.

Chapter 1

Introduction

In 2013, almost one-third of fourth-grade students in the United States scored below the "basic" level in reading, according to the National Assessment of Educational Progress (NAEP), also known as "The Nation's Report Card."[1] They could not locate relevant information in a text, make simple inferences based on what they read, or identify details to support an interpretation or conclusion. Students performing at this level lack the skills needed to demonstrate solid academic performance or to master challenging subject matter. Although NAEP scores have improved modestly over the last two decades for all students, stubborn gaps remain between white and Asian students and their black and Hispanic counterparts, and between students from low-income families and their more affluent peers of the same age. Thus, 48 percent of children eligible for free lunch and 32 percent of children eligible for reduced-price lunch read below the basic level in 2013, compared with only 17 percent of children ineligible for lunch subsidies. Because reading ability is critical to subsequent academic success, and because education is widely viewed as an essential stepping-stone to upward mobility, improving the reading skills of poor children may be an important vehicle for promoting their long-term economic well-being.

Success for All (SFA), one of the best-known school reform models, aims to improve the reading skills of all children but is especially directed at schools that serve large numbers of students from low-income families. First implemented in 1987, SFA combines a challenging reading program, whole-school reform elements, and an emphasis on continuous improvement with the goal of ensuring that every child learns to read well in the elementary grades. Specific features include:

- A kindergarten through grade 6 (K-6) reading program that emphasizes phonics for beginning readers and comprehension for all students
- Instruction that is characterized by "scripted," briskly paced lesson plans that make extensive use of cooperative learning in pairs and small groups
- Cross-grade ability grouping for reading, with many students leaving their homeroom to receive reading instruction from another teacher, and quarterly regrouping
- Frequent assessments of student learning

[1] See National Assessment of Educational Progress (2013). The percentage of students performing below the basic level was 32 percent. NAEP is a congressionally authorized project of the National Center for Education Statistics within the Institute of Education Sciences in the U.S. Department of Education. NAEP tests have been conducted periodically in a number of subject areas since 1969.

- Computerized small-group tutoring and individual tutoring for students who need additional assistance

- Staff committees ("Solutions Teams") that address academic, behavior, and attendance issues and that promote parent and community involvement

- Schoolwide and classroom programs to develop social and conflict resolution skills

- Initial and ongoing professional development for teachers and use of data to monitor progress and set goals

This is the third and final report from an independent evaluation of the scale-up demonstration of the SFA elementary school reading program. Both the demonstration and the evaluation have been funded under the U.S. Department of Education's Investing in Innovation (i3) competition. Conducted by MDRC — a nonprofit, nonpartisan education and social policy research organization — the evaluation examines SFA's implementation and impacts in five school districts over a three-year period (the 2011-2012 school year through the 2013-2014 school year). It also includes an analysis of program costs. Finally, it considers the scale-up process itself — the methods employed and the extent to which the Success for All Foundation (SFAF), the organization that developed and provides technical assistance to schools operating the program, achieved its scale-up goals.

Previous evaluations, both experimental and quasi-experimental, showed that students in schools operating the SFA program ("SFA schools") performed better on standardized tests than students receiving other kinds of reading instruction. The most salient of these evaluations was a three-year randomized experiment that analyzed data from 35 schools receiving funding under Title I, the federal funding stream designated for schools serving low-income students. In that study, schools were randomly assigned to use Success for All either in kindergarten through grade 2 (K-2) or in grades 3 through 5, with the schools that received the program in the later grades serving as a control group for the K-2 schools.[2] Children in the K-2 schools scored significantly higher than their counterparts in the 3-5 schools on the main outcomes measured — scales from the Woodcock Reading Mastery Test that assessed children's phonetic skills and comprehension. In other large-scale studies, results for students in SFA schools outstripped those for students in matched comparison schools.[3] The strength of this evidence was critical to the selection of SFAF as one of only four recipients of five-year scale-up grants awarded in 2010 in the initial i3 funding competition.

[2] See Borman et al. (2007).

[3] See, for example, Rowan, Correnti, Miller, and Camburn (2009).

Ongoing evaluation of the program is called for, however, because, as discussed below, both the SFA program and many other school reading programs have changed since the earlier SFA studies were conducted. Moreover, the Great Recession imposed economic constraints on schools and districts that were less evident at the time of the earlier evaluations. Whether SFA continues to have an impact on early reading achievement in a changed environment is therefore a critical open question.

This report finds, in brief, that while SFA was implemented with adequate fidelity at the large majority of schools that adopted it, resource constraints prevented some schools from putting in place some of the program's key features. Program group and control group schools were different in some respects — SFA schools were unique in placing students in cross-grade ability groups for reading, and SFA teachers made greater use of cooperative learning — but otherwise, the two groups of schools did not differ greatly. The program registered a positive impact on a strong measure of phonetic abilities in the average SFA school, as it did in the previous two years of the demonstration. Students in the average SFA school did not outperform their counterparts in the average control group school on measures of reading fluency or comprehension. For students entering school with low preliteracy skills, however, SFA appears to be especially effective at teaching phonics skills, word recognition, and reading fluency.

The next section of this chapter describes the program more fully and presents the logic model on which the report is based. The third section sets the context for the present study, briefly describing the economic climate in which the SFA i3 demonstration has unfolded and reviewing shifts in literacy instruction that have occurred over the last 25 years. The fourth section summarizes the findings of the two earlier MDRC reports on SFA's i3 scale-up, and the fifth section lays out the questions addressed by this report and the report's contents.

The Success for All Program and Logic Model

Success for All's cornerstone initiative is a reading program for students in kindergarten through grade 6.[4] The program includes three levels: KinderCorner (for kindergartners), Reading Roots (for beginning readers, usually first-graders), and Reading Wings (for more advanced readers, usually second-graders and up); Roots and Wings are further divided into multiple levels. At the lower levels, there is a strong emphasis on phonics instruction, and all levels emphasize vocabulary and comprehension.

[4] SFAF's work has evolved to include reforms to math instruction. MDRC is also conducting an evaluation of the i3-funded scale-up of the SFA math program for middle schools.

Figure 1.1 presents a logic model that describes the implementation process, the program's key components, and the expected outcomes.[5] The leftmost column of the figure shows that implementation of the program comes as the result of a schoolwide vote. To ensure that there is adequate support for the program from the outset, SFAF proceeds with the program only if, after hearing a presentation about the program, at least 75 percent of a school's teachers elect to adopt it.[6]

Both SFAF and the participating schools supply important inputs. At the outset, SFAF provides all school personnel with the essential training they need to launch the initiative. It also supplies the necessary materials: the curriculum, a data system (called Member Center) for monitoring students' progress, and a computerized tutoring system (known as Team Alphie). SFAF used its i3 grant in part to subsidize the cost of these materials and of coaching assistance; schools participating in the evaluation received materials and coaching at no cost for all three years of the demonstration, while other schools recruited in the scale-up effort could receive a subsidy of up to 50 percent of first-year costs. For its part, each SFA school is expected to provide a full-time program facilitator. While the school principal's support for SFA cannot be ensured, SFAF leaders see it as critical to successful implementation.

As Figure 1.1 indicates, the program's elements fall into three broad categories. The reading program itself combines instructional features (for example, a strong focus on cooperative learning) with structural ones (for example, cross-grade ability-grouped reading classrooms). A computerized tutoring program, along with additional interventions, is an important dimension of SFA's approach to serving struggling students. Achievement of progress — whether at the level of the student, the school, or points in between — is celebrated.

A second category consists of program elements that respond to the fact that students' ability to learn is often impeded by other issues they face — attendance-related, behavioral, and the like. To address these issues, as well as to enlist the support of parents and of the broader community, SFA schools establish Solutions Teams composed of teachers, administrators, and other staff members. The program also includes strategies for helping students recognize and express their feelings appropriately, as well as for resolving conflicts, that are intended for use in all classrooms and throughout the building.

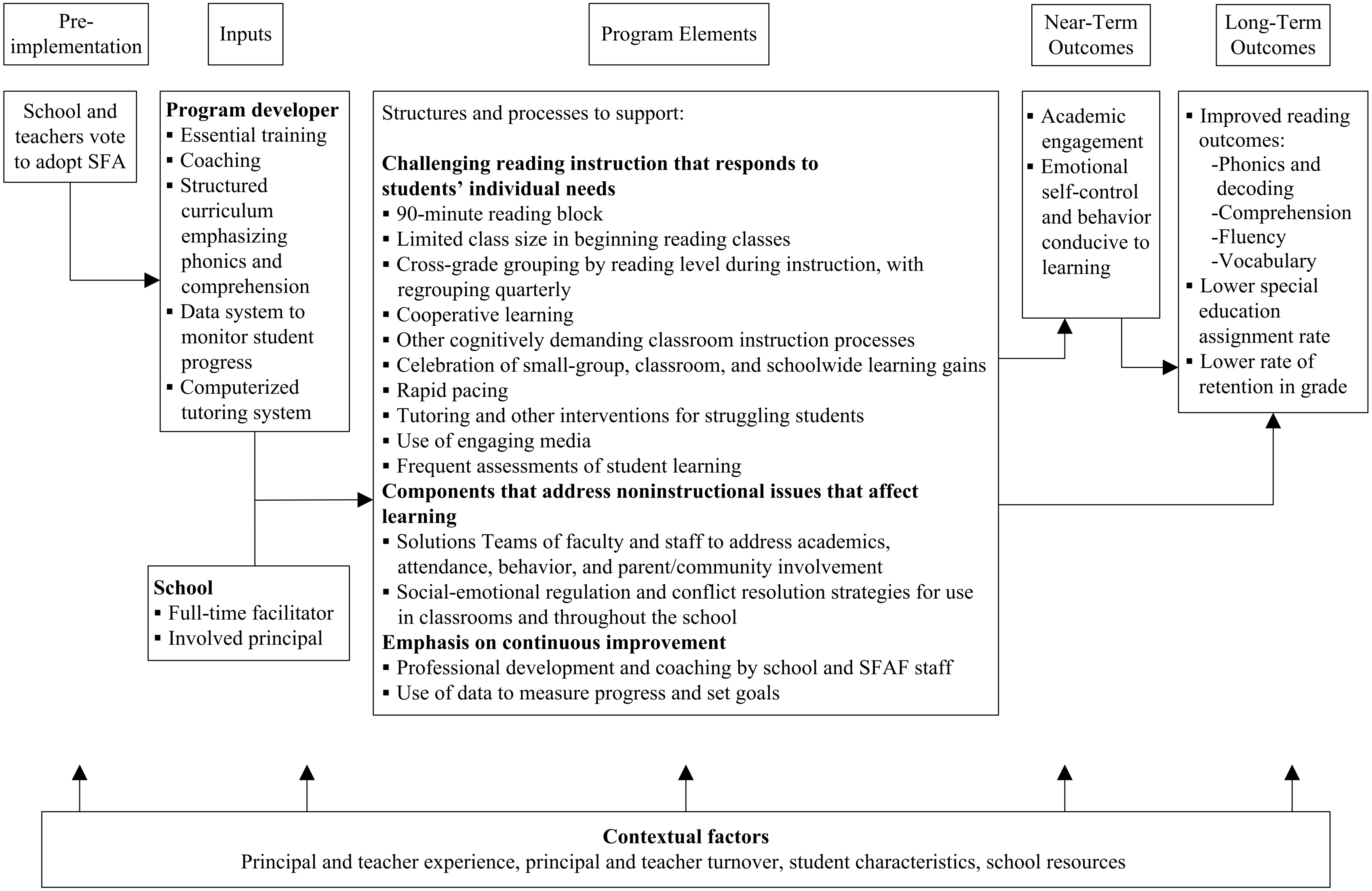
Processes related to the program's emphasis on continuous improvement constitute the third category of program elements. The SFAF "point coach" (also referred to as the "SFAF

[5] The logic model shown here differs from the one presented in earlier reports and reflects the researchers' greater understanding of the program and its essential elements.

[6] At the evaluation schools, the adoption vote preceded random assignment: Staff members at the schools knew that their school had a 50-50 chance of being chosen to operate the program and were willing to do so if selected.

Figure 1.1

Logic Model for the Success for All Reading Program in Elementary Schools



Pre-implementation

School and teachers vote to adopt SFA

Inputs

Program developer
- Essential training
- Coaching
- Structured curriculum emphasizing phonics and comprehension
- Data system to monitor student progress
- Computerized tutoring system

School
- Full-time facilitator
- Involved principal

Program Elements

Structures and processes to support:

Challenging reading instruction that responds to students' individual needs
- 90-minute reading block
- Limited class size in beginning reading classes
- Cross-grade grouping by reading level during instruction, with regrouping quarterly
- Cooperative learning
- Other cognitively demanding classroom instruction processes
- Celebration of small-group, classroom, and schoolwide learning gains
- Rapid pacing
- Tutoring and other interventions for struggling students
- Use of engaging media
- Frequent assessments of student learning

Components that address noninstructional issues that affect learning
- Solutions Teams of faculty and staff to address academics, attendance, behavior, and parent/community involvement
- Social-emotional regulation and conflict resolution strategies for use in classrooms and throughout the school

Emphasis on continuous improvement
- Professional development and coaching by school and SFAF staff
- Use of data to measure progress and set goals

Near-Term Outcomes

- Academic engagement
- Emotional self-control and behavior conducive to learning

Long-Term Outcomes

- Improved reading outcomes:
  -Phonics and decoding
  -Comprehension
  -Fluency
  -Vocabulary
- Lower special education assignment rate
- Lower rate of retention in grade

Contextual factors

Principal and teacher experience, principal and teacher turnover, student characteristics, school resources coach") assigned to each school and the school's SFA facilitator provide professional development to teachers to help them hone their classroom practice. Administrators and teachers regularly examine data on reading achievement and other outcomes to measure progress, identify problems, and establish new and higher goals.

The logic model posits that in the short term, SFA will increase students' academic engagement and enhance their ability to regulate their own behavior so that they can focus on learning. These intermediate outcomes, in turn, are expected to result in improved reading skills, lower rates of assignment to or retention in special education, and lower rates of retention in grade.

Many elements of the SFA model, including cooperative learning, grouping, frequent assessments, and tutoring, have remained constant since the program's inception. At the same time, the program has continued to evolve, with greater emphasis placed on the use of engaging technology in the classroom and, as part of the i3 scale-up, on the deployment of school district personnel trained by SFAF to provide professional development and technical assistance to program schools along with SFAF coaches.

At the bottom of the logic model are contextual factors that may affect implementation and outcomes. These include staff turnover, student characteristics, and schools' access to resources.

The Economic and Instructional Context of the Demonstration

It is useful to consider the economic and instructional contexts in which the SFA scale-up demonstration has unfolded. These contexts provide a framework through which to view the participating schools' ability to implement the full program model and SFAF's ability to meet its ambitious expansion goals. They also help to define the "counterfactual" — what happens in the absence of the program. Only to the extent that SFA differs from the counterfactual is the program likely to produce impacts.

An Economy in Recession

While the first years of the twenty-first century were marked by relative economic stability, the onset of the financial crisis in 2007 precipitated large spending cuts in K-12 education, the effects of which would last for many years.[7]

Between 2007 and 2009 (the official years of the "Great Recession," according to the National Bureau of Economic Research), the number of unemployed persons aged 16 and older nearly doubled, and federal and state revenues shrank.[8] Education was hit hard: Between 2008

[7] Oliff, Mai, and Leachman (2012).

[8] National Bureau of Economic Research (2010), Gordon (2012).

and 2011, 300,000 educator jobs were lost, and the student-teacher ratio in public schools increased by 4 percent.[9] Districts and schools also cut their counseling staffs, their after-school programs, and their instructional offerings.[10] All four states with school districts participating in the present i3 impact evaluation spent less per pupil in fiscal year 2013 than they spent in fiscal year 2008.[11]

To counteract potential reductions to the Title I funding stream, short-term federal stimulus funds were added to it through the American Recovery and Reinvestment Act.[12] These grants were intended to limit the effects of the recession between 2009 and 2011, but they expired thereafter.

Thus, at the point that SFAF was recruiting schools for the i3 scale-up, many schools and districts were trying to restore positions and services that had been cut or to cope with their losses. Furthermore, principals felt that they had less discretion in spending their schools' allocations than had been the case in the past. These circumstances added a new dimension to the challenges already associated with selecting and implementing a new and demanding reading program in high-poverty schools.

A Heightened Emphasis on Reading Instruction

Over the period since SFA was developed, reading instruction in the United States has changed markedly.[13] One impetus was the academic standards movement: During the early 1990s, 48 states and the District of Columbia promulgated academic standards for the various content areas, including reading and language arts, and the majority of states also developed assessments aligned with the standards. Professional development and new materials were seen as the vehicles for ensuring that these standards were incorporated into teachers' classroom practice. In addition, phonics-based instruction gained currency over other approaches to early reading instruction. By the late 1990s, 36 states had bills passed or pending that allocated funding for the purchase of materials and for professional development that emphasized phonics and phonemic awareness. At the federal level, the Reading Excellence Act, passed in 1998, allowed for competitive grants to states to provide professional development to staff in low-performing and high-poverty school districts. The express purpose of the grants was to improve instructional practice and to boost students' reading skills through the use of methods grounded in scientifically based reading research.

[9] U.S. Executive Office of the President (2011); National Center for Education Statistics (2013).

[10] The largest district participating in the i3 evaluation lost 272 staff positions between 2009 and 2011.

[11] Oliff, Mai, and Leachman (2012).

[12] U.S. Department of Education (2009).

[13] Please see Coburn, Pearson, and Woulfin (2011) for a useful summary of the changing policy context.

The focus on reading intensified during the first decade of the twenty-first century due to several factors. First, in 2000, the congressionally mandated National Reading Panel issued an influential report that focused attention on five aspects of reading and reading instruction — phonemic awareness, phonics, fluency, vocabulary, and comprehension — for which there was reasonably rigorous research evidence.[14]

Second, the enactment by Congress of the No Child Left Behind (NCLB) law in 2001 influenced reading policy and practice in several respects. The Reading First grant program created as part of that law provided competitive funding to states to assist Title I schools in increasing students' reading skills. States seeking this funding were required to adopt specific scientifically based reading programs and to put in place related professional development. By April 2007, nearly 6,000 schools had received Reading First grants. Reading First also called attention to the need for early identification of and provision of assistance to struggling readers.

NCLB put in place another mechanism — high-stakes accountability — for ensuring attention to reading. A further condition of receiving Title I funding was that states institute annual tests in reading (along with math and, later, science) for students in grades 3 through 8, beginning in the 2005-2006 school year. Schools that persistently failed to make sufficient progress on these tests were subject to serious sanctions, including state takeover or even school closure. Teachers' performance ratings were increasingly linked in part to their students' performance on these tests, further raising the stakes attached to the tests, as well as teachers' anxiety about them.

Third, the emphasis on struggling readers was reinforced when, under the 2004 reauthorization of the Individuals with Disabilities Act, states and school districts were permitted to use a portion of federal special education funding to provide early intervention services to students at risk of reading or other academic or behavioral problems. One of the key approaches to early intervention, called Response to Intervention (RtI), involves increasingly intensive tiers of support — first, small pull-out groups, followed by one-to-one tutoring — for students who are not making adequate progress in the classroom. RtI has achieved wide diffusion: In the 2008-2009 school year, 70 percent of districts with elementary schools reported using RtI in reading/language arts classes.[15]

Finally, the Common Core State Standards, enunciated in 2009 and since adopted by 43 states and the District of Columbia, call out the importance of instruction in reading comprehension beginning in the early grades.[16]

[14] National Reading Panel (2000).

[15] Bradley et al. (2011).

[16] The Common Core standards — established by the Common Core State Standards Initiative led by the National Governors Association Center for Best Practices and the Council of Chief State School Officers —

Several national studies indicate that Reading First and NCLB had little or no effect on student reading achievement.[17] But even if these specific policy and programmatic initiatives did not have the desired impacts, it nonetheless seems plausible that heightened attention to reading instruction over the past two decades may help to account for the statistically significant, although relatively modest, improvements in students' NAEP reading scores over time.[18] In terms of this evaluation, too, it seems plausible that an increased emphasis on phonics and additional interventions for struggling readers have narrowed the differences between schools adopting SFA and schools using other reading programs, making it harder than it used to be for SFA to "beat the competition."

The i3 Evaluation and Findings from Its Earlier Reports

The i3 evaluation of SFA employs an experimental design, in which 37 schools in five school districts that participated in the scale-up effort were assigned at random to a program group or to a control group. The 19 program group schools received SFA in all grades. The 18 control group schools did not get the intervention and, instead, either continued with the same reading program that they had used previously or, in the case of some schools, adopted a new one. As expected, the random assignment design produced two groups of schools that, at the outset of the demonstration, had similar characteristics (although, as discussed in Chapter 2, the 37 evaluation schools were not fully representative of all the schools participating in SFA's i3 scale-up).

The evaluation uses qualitative and quantitative data from a variety of sources. Implementation data collected from both program and control group schools include teacher and principal surveys, logs completed by teachers describing the instruction that they provided to individual students, classroom observations, and interviews with principals conducted in the course of research visits to all the study sites. Implementation information collected only from program group schools includes implementation summaries completed by SFAF coaches, interviews conducted with SFA facilitators, and information from teacher focus groups. SFAF program manuals were consulted as well. Data informing the impact analysis include demographic and other information contained in school district databases and individual and group assessments of students' reading skills. For this third report, data concerning program costs and the scale-up initiative were also collected from SFAF administrative records, additional principal interviews,

specify what students at each grade level from kindergarten through grade 12 should know and be able to do in the foundational areas of English and mathematics. See Common Core State Standards Initiative (2015).

[17] See Dee and Jacob (2010). Also see Herlihy et al. (2009), who suggest that the Reading First evaluation had no overall effects on student achievement because the instructional practices that were called for were put in place both in schools that received Reading First funds and those that did not receive this funding and served as comparison schools for the study.

[18] The percentage of fourth-grade students reading at the below-basic level declined from 38 percent in 1990 to 32 percent in 2013, while the percentage of students reading at the proficient or advanced level increased from 28 percent to 35 percent.

public data on staffing and expenditures, and interviews with program coaches and with representatives from both adopting and nonadopting schools. Appendix A provides further detail on the various sources of data.

As might be expected given the program's complexity and its highly structured curriculum, during the first year, the majority of teachers acknowledged struggling to implement it. They felt that they had received inadequate preparation for teaching in an SFA classroom; they worried about whether classes were moving too quickly for struggling students; and they found SFA's Member Center data system complicated and demanding. As the year drew to a close, however, many teachers reported feeling more comfortable with the program. By the end of the year, all but one school were deemed to have met the minimum first-year implementation standard that SFAF established, although there was plenty of room for growth and improvement. During the second year, that improvement did take place. Program schools put in place more sophisticated practices that they had not previously implemented, and more classrooms within the year evinced the desired practices. By the end of the year, 16 of the 19 schools were judged to have met SFAF's more demanding standards for adequate implementation fidelity. Teachers reported feeling much more at ease with the initiative, although they continued to express concerns about the program's pacing and grouping practices and about whether the program was adequately serving special education students.[19]

In both years, SFA reading classes were distinguished from reading classes in the control group schools by greater use of cooperative learning, more extensive grouping of students by reading level (with cross-grade grouping far more prevalent in SFA schools), and closer adherence to the curriculum. No differences were found between SFA schools and control group schools with respect to other elements of instructional practice that SFA's developers consider to be important: an extended class period for reading instruction, use of data, and tutoring for students who are not keeping up with their peers. SFA and control group school principals were equally likely to report that their schools had personnel and processes addressing a variety of issues not narrowly tied to reading instruction, such as attendance, parental involvement, and behavior problems.

The impact analysis centers on a group of children who entered kindergarten in the 37 study schools in fall 2011 and whose reading ability was assessed in the spring of each following year. During the first two years, SFA produced effects on these children's reading ability that were consistent with those of prior studies. Impacts on measures of students' phonetic abilities were positive and statistically significant. At this relatively early point (students were in first grade at the end of the study's second year), the program did not increase scores on measures of fluency and comprehension, which are more advanced reading skills. It also did not affect the

[19] For earlier findings from the present study, see Quint et al. (2013, 2014).

skills of students in grades 3 through 5, who had not started learning to read "the SFA way" and were instead first exposed to the program in second, third, or fourth grades.

Key Questions and Contents of This Report

This report addresses five key questions:

1. What was the experience of schools in the evaluation in implementing SFA, and did they do so with adequate fidelity?
2. To what extent did SFA schools differ from control group schools in their districts that served similar students, and in what ways were they similar?
3. Did SFA have an impact on students' reading skills?
4. How do the costs of implementing SFA compare with those of an alternative reading program in an evaluation district?
5. How did SFAF seek to scale up the initiative, to what extent did it succeed, and what factors contributed to or impeded success?

The 2013-2014 school year was the third and final year in which program group schools operated SFA under the i3 grant. By this time, most students who entered the research sample in kindergarten in both the program and control group schools had reached second grade.

This report builds on the methods used in the earlier reports to update the previous implementation and impact findings. These updated findings are the substance of the report's next several chapters.

Chapter 2 describes the characteristics of the program and control group schools and the students they served at baseline. The analysis demonstrates that random assignment, as expected, produced two groups of schools that were substantially similar at the outset and thereby provides reassurance that subsequent differences in outcomes were the result of the SFA intervention.

Chapter 3 examines program implementation in the third year. It discusses the extent to which program schools implemented SFA at a level of fidelity to the model that its developers consider to be adequate. It also considers in depth two program elements — cooperative learning and tutoring — that are especially important to the program model.

A "treatment contrast" is essential for producing program impacts; if program and control group schools provide similar environments for students, instructionally and otherwise, then impacts would not be expected. Chapter 4 investigates ways in which instructional and other practices differed between SFA schools and schools in the control group. It also compares teachers' opinions about their reading programs (either SFA or the programs used in the control group schools) and the ability of these programs to reach struggling students. In addition, it examines student engagement.

Chapter 5 presents the impact findings. It looks first at whether SFA increased the reading skills of students in program schools. It also examines impacts on two other important indicators of student progress: the extent to which students were judged to need special education services and the degree to which they were held back rather than promoted to the next grade. The discussion also compares the results of the i3 evaluation with those found in previous studies of SFA.

Chapters 6 and 7 go beyond the earlier reports to pose additional questions of interest to the research and policy communities. Chapter 6 looks at the cost of the initiative, focusing on the cost ingredients associated with SFA implementation in one of the five districts participating in the evaluation. Chapter 7 considers the i3 scale-up process itself: the strategies that SFAF used and the extent to which it realized the number of schools targeted in its initial proposal. It also compares implementation in the 19 SFA schools that participated in the evaluation with implementation in the i3 scale-up schools that were not evaluation sites.

Chapter 8 concludes the report with reflections on the findings and their implications.

Chapter 2

Schools and Students in the Success for All Impact Evaluation

The 37 schools participating in the Success for All (SFA) impact evaluation are a subset of all schools that were recruited for the scale-up of SFA funded under U.S. Department of Education's Investing in Innovation (i3) competition. This chapter first summarizes the recruitment and random assignment processes. It then describes the characteristics of the impact evaluation schools and their students at the beginning of the study in order to compare these study schools with the broader group of i3 scale-up schools and with schools nationally that serve children from low-income families, and to establish that SFA and control group schools were, as intended, similar to each other.

Key Findings

- The 37 study schools are mostly located in or on the outskirts of large or midsize cities in the Northeast, South, and West, serve 550 students on average, and enroll students who are two-thirds Hispanic and primarily from low-income families.
- Compared with other SFA schools in the i3 scale-up and with a national sample of schools serving students from low-income families, the study schools are more geographically concentrated in the South, are more likely to be located in large or midsize cities, and serve more Hispanic students.
- Random assignment produced two groups of schools that were very similar on school-level characteristics at baseline.
- No statistically significant differences on any baseline characteristics were found between students at program group schools and students at control group schools in the primary analysis sample.

Recruitment and Random Assignment of Schools in the Impact Evaluation

As noted in Chapter 1, the study uses an experimental design with random assignment of a roughly equal number of schools either to a program group, which put in place the SFA program, or a control group, which implemented the reading programs in regular use by their schools. The difference in outcomes between the program group schools and the control group schools can be interpreted as the average effect of the SFA program relative to "business as usual" across all participating districts.

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. The i3 grant presented the opportunity to offer considerable financial benefits to schools interested in taking on the SFA model: Essentially, SFAF was able to offer the intervention at half the usual first-year cost, and schools willing to participate in the evaluation received program materials, training, and technical assistance gratis.[1] Evaluation schools had to meet the same criteria as all SFA schools: At least 40 percent of the students had to be eligible for the free or reduced-price lunch program, 75 percent of the teachers had to vote to adopt SFA, and the school had to be willing to hire and fund the position of an SFA facilitator. In addition, each evaluation school had to serve students from kindergarten through fifth grade, and the school had to be willing to participate in a random assignment experiment.

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 17, producing a total sample of 37 schools. In spring 2011, the schools within each district were randomly assigned to program or control conditions. Random assignment produced 19 program group schools and 18 control group schools. Table 2.1 presents the number of participating schools from each district and the number assigned to each research group.

Characteristics of the Study Schools

Table 2.2 shows the average characteristics of the 37 schools in the study sample, and how the study schools compared with other scale-up schools and a national sample of elementary schools serving students in kindergarten (K) through grade 5, at a minimum, and in which at least 40 percent of enrolled students were eligible for free or reduced-price lunch. Table 2.2 uses data from the fall of 2010, one year before the current study began; at this point the schools' characteristics could not have been affected by the study itself and therefore represent true baseline values.

The study schools are located in the West, South, and Northeast regions of the country. The majority of these schools are in large or midsize cities or on their outskirts. The average school enrolls about 550 students. The majority of students in the study schools are Hispanic and eligible for free or reduced-price lunch.

[1] Nonetheless, recruitment of schools proved difficult. As described in Chapter 7, many districts and schools faced straitened economic conditions and were unwilling to take on new initiatives, even at a greatly reduced cost.

Table 2.1

Distribution of the Study Schools Across Districts

District	Number of Study Schools	Number of Program Group Schools	Number of Control Group Schools
A	4	2	2
B	17	9	8
C	4	2	2
D	6	3	3
E	6	3	3
Number of schools	37	19	18

SOURCE: Success for All evaluation data.

NOTE: Letters are used in place of district names so that the identities of districts in the study are not revealed.

The study schools differ in their geographical location compared with the scale-up sample: More than half are located in the South, and none are located in the Midwest. About 62 percent of study schools are located in cities, compared with 30 percent of scale-up schools, which are evenly spread across cities, towns, and rural areas. Study schools have fewer students eligible for free or reduced-price lunch and about three times as many Hispanic students as schools in the scale-up sample.

Overall, the study schools differ from schools in the national sample on most dimensions measured in Table 2.2. Specifically, the study schools are more likely to be located in the South and in urban areas; they have a higher percentage of Hispanic students and a lower percentage of students eligible for free or reduced-price lunch; and they tend to be larger than schools in the national sample.

The Research Samples Used in the Analysis

Once schools were randomly assigned, all kindergarten students who were in regular classes (that is, not in separate classes for students with special education needs) in the fall of the 2011-2012 school year and who could be tested in English (all except three students) were included in

Table 2.2

Selected Characteristics of Study Schools, Other Schools in the SFA Scale-Up, and the National Population of Schools Serving Students from Low-Income Families (2010-2011)

Selected Characteristics	Study Sample	Other Schools in Scale-Up Sample	National Population of Schools Serving Students from Low-Income Families[a]
Geographic region (% of schools)			
Northeast	16.2	11.9	13.1
South	67.6	44.8 ***	24.6 ***
Midwest	0.0	23.3 ***	24.8 ***
West	16.2	20.0	37.5 ***
Urbanicity (% of schools)			
Large or midsize city	62.2	30.4 ***	29.1 ***
Urban fringe or large town	21.6	32.9	40.0 **
Small town or rural area	16.2	36.7 **	30.9 *
Title I status (% of schools)	100.0	95.2	94.7
Free or reduced-price lunch (school average % of students)	56.8	71.9 ***	68.3 ***
Race/ethnicity (school average % of students)			
White	13.8	31.8 ***	41.8 ***
Black	22.6	39.8 ***	18.4
Hispanic	61.8	19.1 ***	31.4 ***
Asian	0.1	0.2 **	4.1 **
Other	0.2	0.2	4.2 ***
Male (school average % of students)	51.3	51.7	51.4
Total school enrollment	546.5	483.2	456.1 ***
Ratio of students to full-time-equivalent teachers (all grades)	16.9	15.6 **	16.6 **
Number of schools	37	428	6,047

SOURCE: 2010-2011 Common Core of Data.

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

A two-tailed t-test was applied to differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The national population includes schools that serve grades K-5 and in which at least 40 percent of students are eligible for free or reduced-price lunch.

the "full baseline sample."[2] This sample includes 2,956 students across the 37 schools. Students in the full baseline sample were tested in the fall of 2011 using the Peabody Picture Vocabulary Test (PPVT) and Woodcock-Johnson Letter-Word Identification test (WJLWI). About 96 percent of the SFA students and control group students in the full baseline sample — 2,831 students in all — had valid scores on both these tests and form the "baseline analysis sample."[3] This sample is particularly important because the "primary analysis sample," described below, is a subset of the baseline analysis sample.

The analysis of SFA's impacts focuses on the primary analysis sample. This sample consists of 1,635 students who were in the baseline analysis sample and who had at least one valid test score from each follow-up test period, with testing taking place annually in the spring. Students in the primary analysis sample were likely to have remained continuously enrolled in their schools from kindergarten on; SFA students in this sample, therefore, had the best chance of receiving the full amount of the SFA program.[4]

The "spring sample" consists of all students with at least one valid test score from the spring of 2014. It includes students in the primary analysis sample, students who were enrolled in the schools at baseline but lacked valid scores on one or both baseline tests and/or a valid follow-up score each year, and students who enrolled in the evaluation schools at some point after the baseline assessments were conducted — whether as kindergartners, first-graders, or second-graders. Results for the spring sample show how the implementation of SFA affected the average performance level of all students in the schools at the time of testing. While the spring sample is not the primary sample for the impact analysis, it is nonetheless of interest because it takes into account students who transferred into a study school at some point during the study. Student mobility is a phenomenon over which school administrators exercise little or no control and which is especially widespread in schools serving students from low-income families.

[2] For a comparison of baseline characteristics between program and control group students in the full baseline sample, see Appendix Table B.1.

[3] Out of 1,542 program group students eligible for baseline testing, 1,480 (96.0 percent) had valid WJLWI scores and 1,468 (95.2 percent) had valid PPVT scores. Out of 1,414 control group students eligible for baseline testing, 1,369 (96.8 percent) had valid WJLWI scores, and 1,367 (96.7 percent) had valid PPVT scores. For a comparison of baseline characteristics between program and control group students in the baseline analysis sample, see Appendix Table B.2.

[4] The analysis could not track whether students left their schools during the school year but returned in time for the spring assessments. For a detailed examination of how the primary analysis sample was formed over each year of the study, starting with the full baseline sample, see Appendix Figure B.1.

This study also explores the program effects on various subgroups defined by student baseline characteristics such as race/ethnicity, gender, poverty status, special education status, and English language learner (ELL) status.[5]

Equivalence of Baseline Characteristics Between Program and Control Groups

The purpose of random assignment is to produce a program group and a control group that are equivalent on all characteristics 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 course of the study is balanced across groups, one can be confident that any differences in outcomes between the two groups found later are due to the intervention.

Table 2.3 shows that, as intended, random assignment produced groups of schools that were very similar on all observed school-level characteristics at the beginning of the study. There were no statistically significant differences in any school-level baseline characteristics between 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 whether there were any overall differences in baseline school characteristics between the two groups of schools, and none were found.[6]

Using the demographic data received from students' district records, as well as baseline test scores, Table 2.4 summarizes the baseline characteristics for program and control group students in the primary analysis sample. On average, these students were five and a half years old as of the fall of 2011, a majority of the students were Hispanic, and a majority were eligible for free or reduced-price lunch. About 26 percent of the SFA students in this sample were ELLs, compared with about 21 percent of control group students. About 6 percent of SFA students and 6.4 percent of control group students were classified as having special education status. None of these differences is statistically significant. In fact, no statistically significant differences were found between SFA and control group students on any individual characteristic. However, an overall F-test did indicate a statistically significant difference when all baseline characteristics were analyzed simultaneously.[7]

[5] A small proportion of students in the primary analysis sample were mainly instructed in Spanish and were tested in Spanish as well as English on both the baseline and follow-up tests. This sample forms the "Spanish analysis sample," which is analyzed separately as a subgroup of the primary analysis sample.

[6] 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 0.999.

[7] This was obtained by using a logistic regression model to predict program membership using the student-level baseline characteristics presented in Table 2.4. The p-value of the F-test is 0.048.

Table 2.3

Selected Characteristics of the Study Schools, by Program or Control Group Status (2010-2011)

	Program Group	Control Group	Estimated Difference	P-Value for Estimated Difference
Title I status (% of schools)	100.0	100.0	0.0	
Students eligible for free or reduced-price lunch (school average % of students)	56.1	56.3	-0.2	0.928
Race/ethnicity (school average % of students)				
White	13.1	13.9	-0.7	0.496
Black	23.0	21.3	1.8	0.671
Hispanic	62.1	63.1	-1.0	0.823
Asian	0.6	0.8	-0.2	0.542
Other	1.2	1.0	0.2	0.436
Male (school average % of students)	51.6	51.0	0.6	0.407
Total school enrollment	558.4	533.8	24.6	0.548
Number of full-time teachers	32.8	31.7	1.1	0.598
Percentage of students at or above reading proficiency level (deviation from state mean, %)	-9.8	-11.3	1.4	0.595
Number of schools: 37	19	18		

SOURCES: 2010-2011 Common Core of Data; district-provided state reading test data, 2010-2011; state reading test records, 2010-2011; and demographic data collected from the five districts in the study sample.

NOTES: The estimated differences for school-level data are regression adjusted using ordinary least squares regression, controlling for indicators of random assignment blocks.

The values for the 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 differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

To examine whether there are any systematic differences between the program and control groups, an F-test was calculated for the full sample of 37 schools in a regression model controlling for indicators of random assignment strata and all school characteristics reported in this table. The p-value of the test is 0.999.

Due to differences in the estimation models used to create this table and Table 2.2, the means are not directly comparable across variables. For example, the overall mean of the 19 SFA and 18 control group schools for free or reduced-price lunch was estimated to be 56.8. In this table the SFA and control group means are both smaller than 56.8. This disparity occurs because in the model used to create Table 2.2, the means for SFA and control group schools were estimated together and without weights, whereas the model used for this table estimates SFA and control group means separately and applies weights.

Table 2.4

Selected Characteristics of Students in the Primary Analysis Sample at Baseline (Fall 2011)

	Program Group	Control Group	Estimated Difference	P-Value for Estimated Difference
Age (years)	5.5	5.5	0.0	0.743
Students in poverty (%)	87.5	88.5	-1.0	0.629
Race/ethnicity (%)				
White	12.4	12.6	-0.2	0.868
Black	18.9	17.8	1.2	0.807
Hispanic	65.8	66.9	-1.2	0.792
Asian	1.3	0.9	0.4	0.700
Other	1.6	1.4	0.3	0.672
Male (%)	49.0	48.9	0.0	0.990
English language learners (%)	26.4	20.6	5.8	0.170
Special education status (%)	5.9	6.4	-0.5	0.685
Peabody Picture Vocabulary Test				
Scaled score	92.3	92.7	-0.3	0.805
Percentile equivalent	30	32		
Woodcock-Johnson Letter-Word Identification Test, raw score	10.7	11.3	-0.6	0.212
Number of students	854	781		

SOURCES: MDRC calculations based on baseline test scores on the Peabody Picture Vocabulary Test (PPVT) and the Woodcock-Johnson Letter-Word Identification (WJLWI) test, administered to the baseline student sample in fall of the 2011-2012 school year; demographic data collected from the five districts in the study sample.

NOTES: The “primary analysis sample” consists of students from 37 schools (19 program group schools and 18 control group schools) and includes any student who had at least one valid spring test score in each of the three implementation years, and who had valid scores on the fall baseline 2011 PPVT and fall baseline 2011 WJLWI test.

A two-tailed t-test was applied to differences between program and control groups. Although there was no significant difference on any individual baseline characteristic, there was a statistically significant difference in the joint distribution of these baseline characteristics. This finding is based on a logistic regression predicting program group status from student-level baseline characteristics. The p-value for the F-test is 0.048.

The values for the 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.

Table 2.5 shows that the rate of student attrition was statistically equivalent for program and control group students. Attrition, in this context, occurred when students in the full baseline sample became ineligible for membership in the primary analysis sample; these students are referred to as "sample out-movers."[8] In order to be in the primary analysis sample, a student had to be in the baseline analysis sample (that is, have valid scores on both baseline tests) and be in every yearly analysis sample (that is, have at least one valid test score each spring). After three years, about 44 percent of program group students in the full baseline sample were no longer part of the primary analysis sample, and the percentage is nearly identical for control group students.[9]

[8] Most of the attrition that occurred (about 80 percent) was due to students transferring from a study school into a nonstudy school, not due to enrolled students missing baseline or follow-up tests. When attention is restricted to the reason for attrition (transferring or not having the required tests), the rate of attrition is still statistically equivalent among program and control group students.

[9] For more information about attrition, see Appendix Tables B.3 and B.4. For information about an auxiliary sample of students in grades 3 to 5, see Appendix Table B.5.

Table 2.5

Percentage of Sample Out-Movers, by Program Status

	Program	Control	Estimated	P-Value for
	Group	Group	Difference	Estimated
	(%)	(%)	(%)	Difference
Percentage of students in the full baseline sample *not* in the:				
Baseline analysis sample[a]	5.1	3.6	1.5	0.142
Year 1 analysis sample	15.0	13.7	1.3	0.562
Year 2 analysis sample	32.3	32.3	-0.1	0.983
Year 3 analysis sample (primary analysis sample)	44.1	44.4	-0.3	0.929

SOURCES: 2011-2012, 2012-2013, 2013-2014 student test and demographic data collected from the five districts in the study sample.

NOTES: The full baseline sample includes students who were eligible for testing in the fall of 2011. Only three students were ineligible for testing at baseline, either because they could not be tested in English or because of learning disabilities that prevented proper test administration. There were 1,414 control group students and 1,542 program group students in the full baseline sample.

Out-movers are defined as students in the full baseline sample who became ineligible for membership in the primary analysis sample. To be in the primary analysis sample, a student needs to have valid scores on both baseline tests administered in the fall of 2011. In addition, a student is required to have at least one valid test score from each follow-up test administration period, occurring annually in the spring. As soon as one of these conditions is not met, the student is no longer part of the primary analysis sample and is classified as an out-mover. Therefore, the percentages of out-movers are cumulative over the years. So, for example, by the end of Year 3, 44.1 percent of program group students in the full baseline sample were not part of the primary analysis sample, including all out-movers from prior years.

The values for the 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. Because weighted averages are used, the actual percentages of out-movers, by program group or control group status, are slightly different from the estimates.

A two-tailed t-test was applied to differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a] Students in the baseline analysis sample are those who have valid test scores on both baseline tests administered in the fall of 2011.

Chapter 3

Implementing Success for All

Chapter 3 describes the Success for All (SFA) program in operation — the inputs and program elements and contextual factors from the program's logic model depicted in Chapter 1 (Figure 1.1). This chapter addresses key research questions about the program's implementation:

1. How did implementation change over the course of the demonstration?
2. Did the initiative meet the standards for implementation that the Success for All Foundation (SFAF) established?
3. What were the schools' experiences in implementing various program elements?
4. What were the attitudes of school personnel toward the SFA program?

Both quantitative and qualitative data inform the answers to these questions. The quantitative data come from principal and teacher surveys and from the School Achievement Snapshot, a form used by SFAF point coaches. As described in Box 3.1, coaches' ratings of 67 Snapshot items were used to calculate implementation scores. Qualitative data include interviews and focus groups used to examine the perspectives of school personnel regarding the challenges of implementing SFA program elements and the program-specific and contextual factors that made SFA implementation easier or more difficult.

Key Findings

- Over the course of the demonstration, schools were able to implement not only more practices but also more sophisticated ones, although they still had room to improve the breadth and depth of their implementation.
- All but 2 of the 19 program group schools were judged to have achieved an adequate level of implementation fidelity, although there was significant variation.
- Over time, there was an increase in teachers' use of cooperative learning and celebration of learning gains, as well as improvement in teachers' ability to pace their reading lessons as suggested by SFAF.
- While most schools offered tutoring to struggling students, fewer than half the schools used SFA's computerized tutoring program, Team Alphie.

Box 3.1

Using the School Achievement Snapshot to Measure Program Implementation

The School Achievement Snapshot is a rubric used by the SFAF coaches and school facilitators to assess program implementation and to guide schools in a continuous improvement process.* When SFA 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 practices. The rating that the coach assigns to each *school-level* practice takes into account whether the practice has been implemented; the rating for each *classroom-level* practice reflects the proportion of the school's reading classrooms demonstrating use of the practice. MDRC worked closely with SFAF to convert the item ratings into numerical scores. Item scores are then summed to yield, for each school, an overall implementation score. The maximum score any school can achieve is 97; in the analysis that follows, a school's overall score is presented as a percentage of this maximum score.

Each item rated on the Snapshot represents one of three levels of practice. (SFAF refers to these levels of practice as "mechanical," "routine," and "refined." This report simply calls them Levels 1, 2, and 3.) Level 1 items represent basic practices (for example, scheduling a 90-minute reading block, regrouping students by ability across grades, and providing daily tutoring to students who need it) that schools are expected to put in place immediately. Level 2 items involve a greater degree of familiarity with the reading program and include some of SFA's whole-school features (for example, the use of SFA's own tutoring program and the establishment of Solutions Teams and the schoolwide behavior plan). Level 3 practices require even more experience and skill to implement (for example, classroom discussions that require students to think deeply and support their positions). It is possible to calculate a school's total score for items at each level of practice.

All Snapshot items used in the implementation analysis relate to one of the program developer inputs, school inputs, or program elements shown in the logic model (Figure 1.1 of this report). The program elements can be subsumed under three categories: challenging reading instruction, components to address noninstructional issues, and continuous improvement. Category scores have been calculated from the sum of scores on the items within each category. Although the Snapshot form includes 26 items about student engagement, these items are excluded from the analyses presented here, since the student engagement items may be better seen as reflecting the *results*, or *outputs*, of implementation.

*A copy of the School Achievement Snapshot can be found in Quint et al. (2013).

- Despite the challenges, 93 percent of the principals and 70 percent of the teachers responding to surveys agreed that the SFA program benefited their schools.[1]

Tracking Implementation over Time

In general, program implementation improved over time. Figure 3.1 indicates the percentage of the maximum possible implementation score of 97 that schools achieved, on average, in each implementation year. Because SFAF did not expect all program practices to be put in place in the first year, the SFAF coaches were instructed not to rate some of the Snapshot items in Year 1. In calculating the Year 1 scores, unrated items were counted as zeros. Because some of these unrated items may in fact have been put in place that year in some schools, the figure may understate Year 1 implementation. It is nonetheless apparent that the major growth in implementation came between Years 1 and 2; implementation continued to improve but much more slowly between Years 2 and 3.[2]

SFAF hopes that schools will maintain the practices they have put in place while moving on to tackle more ambitious ones. As Figure 3.2 makes clear, the increase in implementation scores is associated with the increased implementation of more sophisticated (Levels 2 and 3) school- and classroom-level practices over each successive year of the demonstration. In contrast, the percentage of the maximum possible score for the basic (Level 1) practices increased from the first year to the second year and then leveled off between the second and third years.

Finally, Figure 3.3 shows that improved implementation in Year 2 occurred along all dimensions but was especially marked with respect to components addressing noninstructional issues. This is largely accounted for by the fact that, as discussed below, many schools waited until the second year to put Solutions Teams in place.

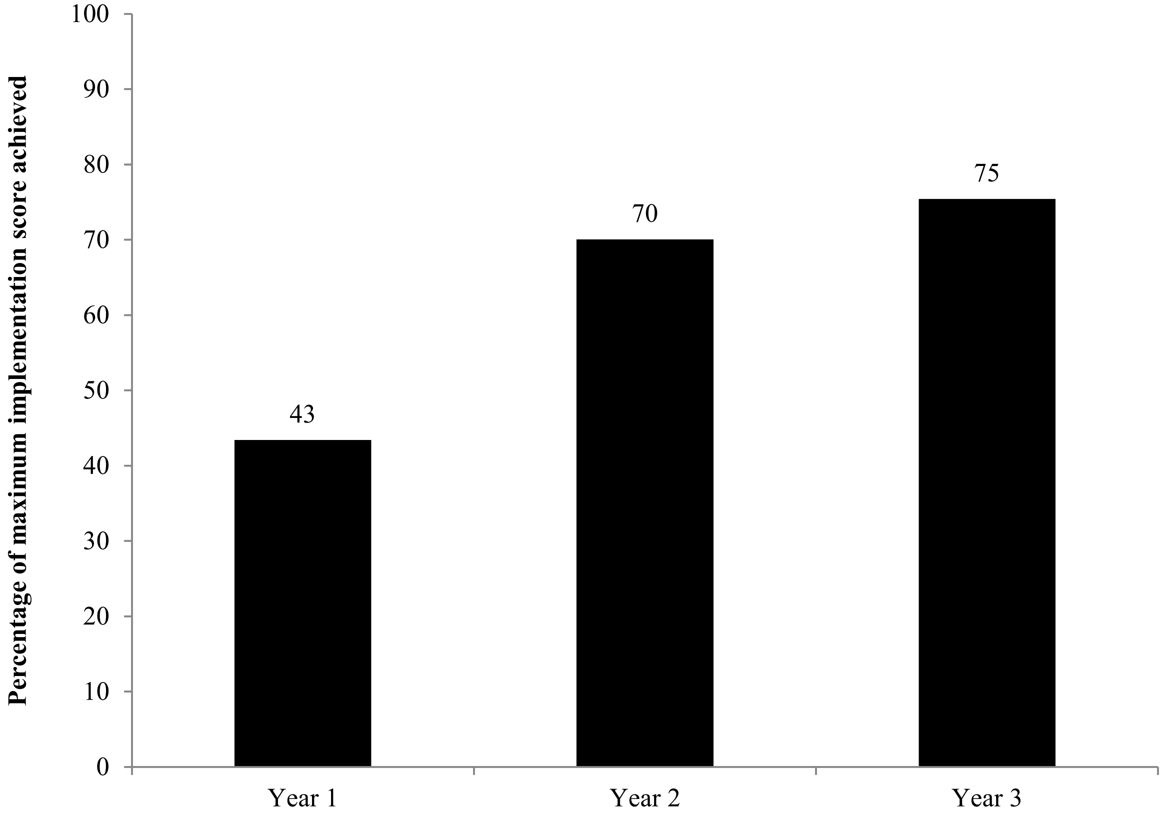
While the overall pattern was one of substantial improvement over time, at individual schools there were both improvements and dips. Schools' progress in implementing key elements is explored in more detail below.

[1] The survey response statistics presented in this chapter are not weighted. That is, they represent the percentage of all principals or teachers across all the program group schools who answered the survey in a certain way, regardless of the number of teachers in a program group school or the number of program group schools in a district.

[2] The average implementation score for Year 2 was significantly higher than that for Year 1, and the average score for Year 3 was significantly higher than the average Year 2 score.

Figure 3.1

Percentage of Maximum Implementation Score Achieved, by SFA Implementation Year



SOURCE: 2013-2014 School Achievement Snapshot.

NOTE: In Year 1, many items on the Snapshot were not rated because schools were not yet expected to have put those items into place; instead, attention was given to items that concerned the most fundamental aspects of implementation. For the purposes of calculating scores, however, unscored items were treated as if they were not in place on any level (that is, given a score of 0). Therefore, 43 percent is a lower bound on the schools' actual average score in Year 1. In Years 2 and 3 all Snapshot items were scored, so this issue does not arise.

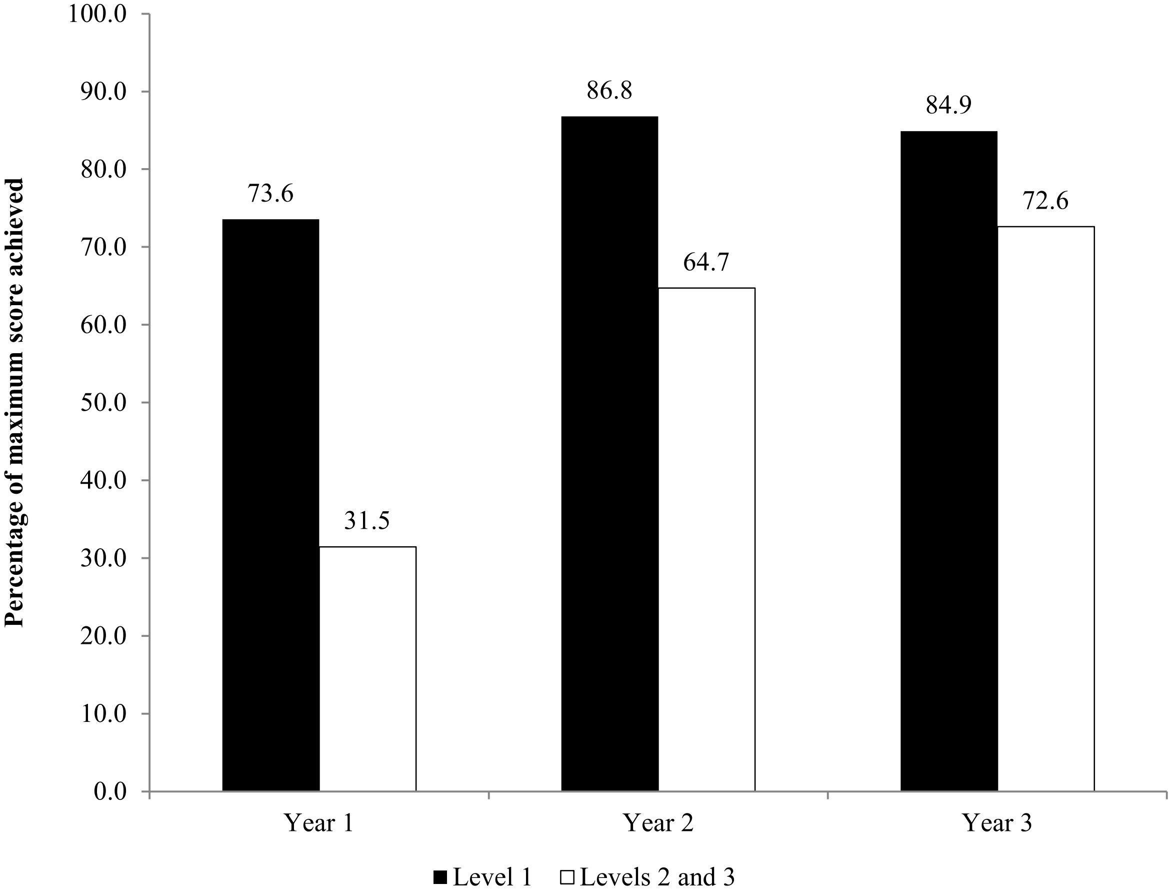
Implementation Fidelity

The evaluation makes use of the School Achievement Snapshot to measure implementation fidelity. Schools are considered to have implemented SFA with adequate, although not necessarily high, fidelity if their total score is 50 percent or more of the maximum possible score of 97.[3]

[3] As previously noted, the Snapshot is intended to pinpoint areas where schools need to improve implementation, and it is difficult to achieve a very high score on the instrument. A school in which all Level 1 items and 75 percent of items in Levels 2 and 3 were in place in 75 percent of the classrooms could fail to meet the fidelity threshold (depending on which specific elements were implemented).

Figure 3.2

Average Percentage of Maximum Snapshot Score, by Level of Practice Sophistication and SFA Implementation Year



SOURCES: 2011-2012, 2012-2013, and 2013-2014 School Achievement Snapshots.

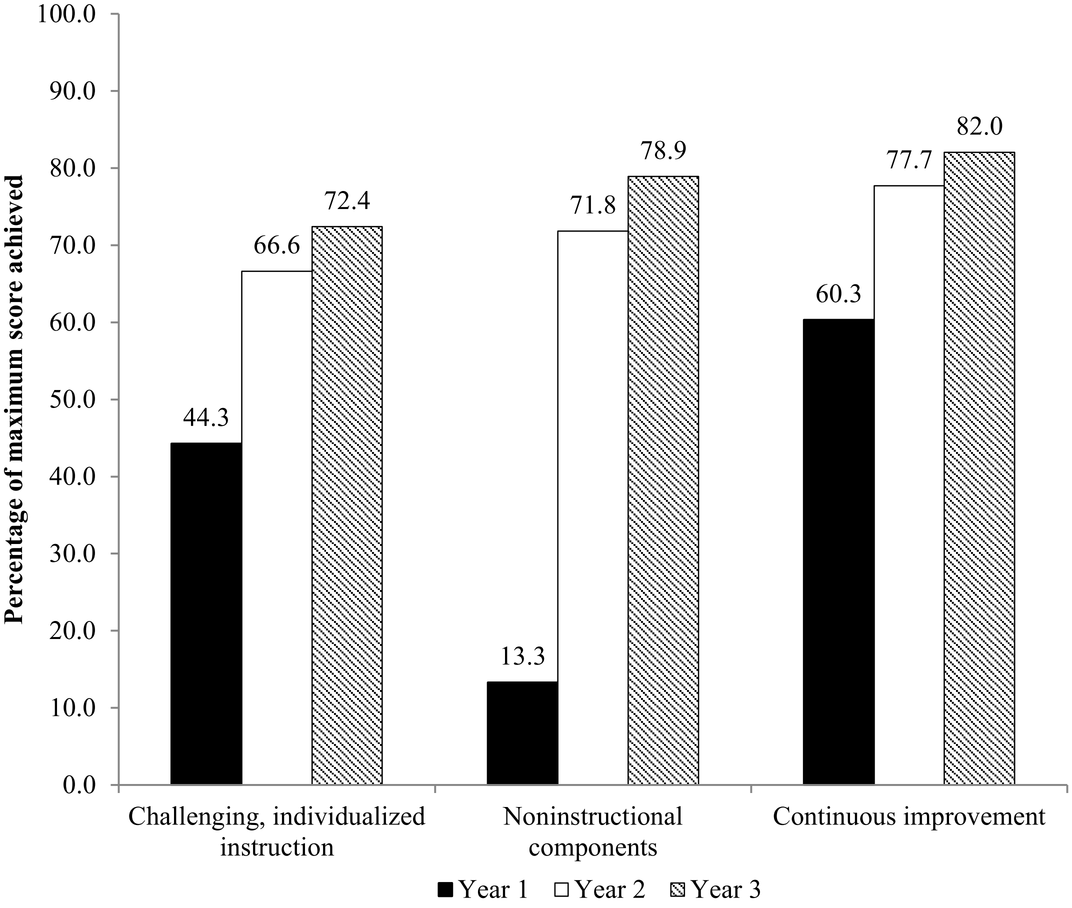
NOTES: Level 1 items are Snapshot items the SFA program considers to be critical to successful functioning implementation, while Level 2 or 3 items reflect more complex and sophisticated implementation tasks.

In Year 1, many items on the Snapshot were not rated because schools were not yet expected to have put those items into place; instead, attention was given to items that concerned the most fundamental aspects of implementation. For the purposes of calculating scores, however, unscored items were treated as if they were not in place on any level (that is, given a score of 0). The percentages pertaining to Year 1 therefore represent lower bounds.

As Figure 3.4 shows, 17 schools met that threshold during the third implementation year. (The same number met the threshold in Year 2, but the average implementation scores were higher in Year 3.) At the same time, there was considerable variation among the schools counted as implementing SFA with fidelity. Two schools achieved better than 95 percent of the total maximum score, but the majority still showed room for improvement.

Figure 3.3

Average Percentage of Maximum Snapshot Score, by Type of Practice and SFA Implementation Year



SOURCES: 2011-2012, 2012-2013, and 2013-2014 School Achievement Snapshots.

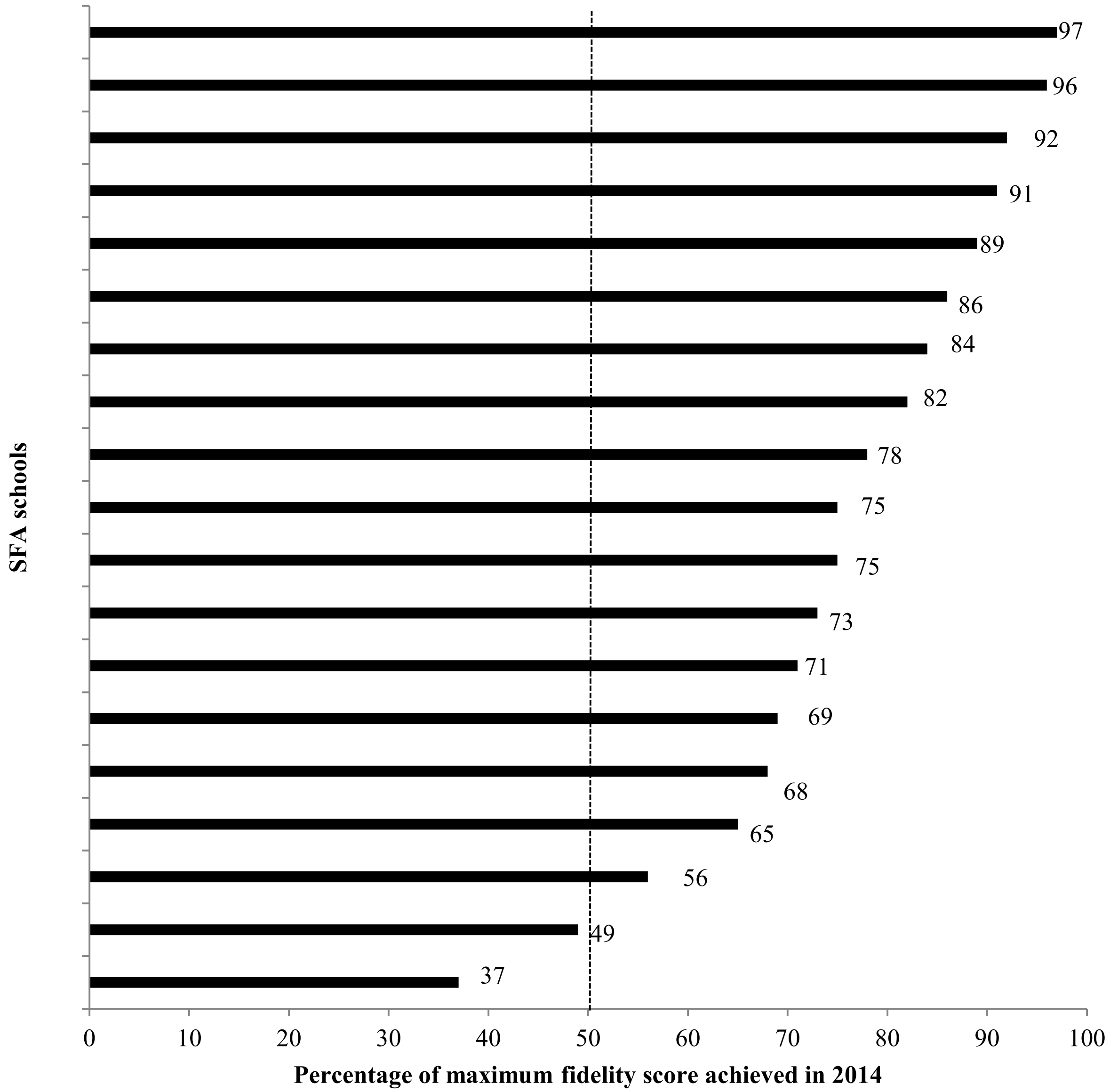
NOTE: In Year 1, many items on the Snapshot were not rated because schools were not yet expected to have put those items into place; instead, attention was given to items that concerned the most fundamental aspects of implementation. For the purposes of calculating scores, however, unscored items were treated as if they were not in place on any level (that is, given a score of 0). The percentages pertaining to Year 1 therefore represent lower bounds.

Schools' Experiences in Implementing SFA

In the following sections, the Snapshot is paired with survey and interview data to examine the implementation of a number of the program developer inputs, school inputs, and program elements shown in the logic model (Figure 1.1). The discussion assesses the extent of implementa-

Figure 3.4

Percentage of Maximum Possible Snapshot Score Attained in 2013-2014, by School



SOURCE: 2013-2014 School Achievement Snapshot.

NOTES: For program implementation to be considered adequate across all schools in Year 3, 80 percent of all schools had to achieve a score of at least 50 percent of the maximum possible Snapshot score. The mean percentage of the maximum score achieved was 75.4, and the standard deviation was 16.0.

tion of specific program features, with a focus on the final year, and also presents the perspectives of school personnel regarding the promise and problems associated with implementing these features.

Developer Inputs

School leaders and teachers depend on SFAF for two things: the materials needed for program implementation (classroom printed and media materials, specific guides, online resources and disks), and what SFAF deems to be "essential training for all staff." Teachers must receive training appropriate to the reading level they teach, and school leaders must receive training in virtually all aspects of the program.

Materials

In all three years of the study, Snapshot ratings show that staff in all the schools operating the SFA program ("SFA schools") received the materials necessary for program implementation. According to the teacher survey, 76 percent of teachers expressed satisfaction with the overall quality of the reading materials, including technology, that they used. During focus groups, some teachers commented that they liked the program videos. They also appreciated having enough books so that each student could look at his or her own book.

Teachers also had a number of complaints about the materials. The most common of these concerned the quality and availability of materials for English language learners (ELLs). Some teachers found the few books available for ELLs to have grammatical and other errors; they also commented that the Spanish used in the SFA books sometimes included words unfamiliar to the students in their region. Bilingual teachers in four SFA schools said that they used the district text or created their own materials instead of depending on Spanish materials provided by SFA. More broadly, some teachers found the volume of materials confusing and/or overwhelming; some asserted that the books were not always age-appropriate for the students reading them; and some reported that because there were not enough titles pitched at a particular level, students who were unable to advance to the next level or those already at the highest level had to reread books that they had already completed.

Essential Training

School leaders and teaching staff at 19 schools received essential training from SFA at some point during the demonstration, but they varied in when they received it, who received it, and how much they got. Fifteen of the 19 schools were rated on the Snapshot as having received this training during the first year of the demonstration, while the remaining 4 first received it at a later point. According to the Snapshot, the staff at 3 schools received only one year of training, 5 schools received two years, and the remaining 11 received training in all three years. Training after the first year often focused on new teachers rather than returning ones, but it also varied by district: In one district, new teachers received two to three days of training and returning teachers had the option of taking a daylong refresher course, while in another district, teachers (both new and returning) received no training at all.

By Year 3, teachers generally found the SFA training that they received at the start of the year, whether directly from SFAF or not, to be only somewhat useful. That said, there is no real evidence that SFA's provision of essential training was associated with a school's Snapshot score in any given year. Some higher-scoring schools received training and others did not, and the same is true of lower-scoring schools.[4]

School Inputs

The two school-level inputs measured on the Snapshot include the presence of a principal who is committed to the program and the existence of a full-time SFA facilitator at each school — both considered by SFAF to be critical to the program's success.

Committed Principal

As the school leader, the principal has prescribed SFA responsibilities. These include setting and supporting the expectation that the SFA program will be fully implemented; scheduling and leading SFA meetings; visiting classrooms; and meeting with the facilitator, the SFAF coaches, and the Solutions Teams coordinator to plan and troubleshoot.

There was no Snapshot rating for principals' commitment in the first year of program implementation. In the second year of implementation, 18 of the 19 schools were rated as having a principal who was fully involved; in the third year, that number dipped to 17 schools.

The extent of SFA program implementation seems to be associated with the person at a school's helm. In each instance in which a school was rated as not having a fully involved principal, the overall Snapshot score was low.[5] As discussed in Box 3.2, the school without a committed principal in Year 2 had a change in leadership in Year 3, and implementation improved. In Year 3, one of the two schools that was rated as not having a fully involved principal had its

[4] According to the Snapshot, one school that received no training in Year 3 had a low overall implementation score in that year, but that school was also rated by the SFAF coach as lacking a principal who was committed to the program.

[5] One of these schools was below the fidelity threshold with a score of 37 percent of the maximum, and another, with a score of 56 percent of the maximum, was only slightly above the threshold.

Box 3.2

Leadership Makes a Difference

As underscored by the staff of Roosevelt Elementary School [a pseudonym], a strong, supportive principal is crucial to fostering high implementation fidelity. For two years, the school's staff implemented SFA under an absent, unhelpful principal. According to the school's SFA facilitator, "It was very difficult to get teachers to listen to you, because they realize I'm not administration, I have no authority. And so without administrative support, this role was very difficult."

By the third year of implementation, however, the district replaced this principal with a more involved and accommodating leader. Citing positive changes that resulted that year, the facilitator commented, "With administrative support, it's very different. The overall morale and feel of the school is very different. … I think making sure that you have a strong administrator who's willing to support it is the most important." In line with the facilitator's observations, Roosevelt's implementation score increased more than that of any other school in the sample. Although in Year 2, the school had achieved only 42 percent of the maximum possible score (and was judged not to be implementing SFA with fidelity), in Year 3, it achieved 69 percent of the maximum score, well surpassing the 50 percent needed to be considered a faithful implementer of the program model.

principal replaced at midyear and was in the process of closing down entirely at the end of the school year.[6]

Facilitator

A precondition of a school adopting SFA is that the principal agree to fund the position of a full-time facilitator. The facilitator is key to the implementation process. The responsibilities of a facilitator are many and varied and include coaching and supporting teachers, monitoring student achievement, placing students in appropriate reading groups, serving on the leadership team, and managing the program materials.

While program guidelines specify that the position of facilitator should be full-time, economic pressures on schools have meant that facilitators have had to take on more and more responsibilities. Schools, whether in the evaluation or not, are now considered to have a full-

[6] The other school rated as not having a fully involved principal was nonresponsive to all attempts by the study team to collect data.

time facilitator as long as a full-time staff member without teaching responsibilities is designated as the facilitator.

Because the definition has shifted and is more ambiguous than it was in previous years, it is hard to compare the number of schools with a full-time facilitator over time. Notably, in the third year, 17 of the 19 SFA schools were rated as having a full-time facilitator in place, but in only 2 of these schools did data from the Snapshot, the principal survey, and facilitator interviews all indicate that the facilitator was full-time (that is, at the same school five days a week) with no responsibilities other than SFA. In most of the other schools, facilitators had other responsibilities, including lunch and bus duty, test coordination and preparation, supporting teachers in subject areas other than reading, and even teaching reading classes. At 2 schools, facilitators were only on campus three days a week at the most. Notably, by Year 3, schools without a facilitator received an average implementation score of 59 percent of the maximum possible score while schools that did have a facilitator had an average implementation score of 89 percent of the maximum possible score.

According to surveys, teachers generally thought that facilitators were knowledgeable about SFA, and 83 percent of teachers and 93 percent of principals responded that the facilitators provided teachers with useful feedback.

Program Elements

Attention now turns from inputs to the program elements that schools are charged with putting in place. As shown in the logic model (Figure 1.1), the program elements are subsumed under three categories: challenging reading instruction, components to address noninstructional issues, and continuous improvement. The program elements discussed below were chosen either because of their uniqueness to SFA or because they present interesting implementation choices and challenges.

CHALLENGING READING INSTRUCTION

Elements found under the "challenging reading instruction" category are directly related to what goes on in classrooms (for example, class size, how students are grouped, what cooperative learning looks like, lesson pace).

Cross-Grade Grouping

From first grade on, students in SFA schools are tested quarterly and, based on their performance on those tests as well as informal assessments and other measures, are placed in groups by level of reading ability. Placement is not limited by the grade level of the students.

Rather, the goal is to place students at the highest possible level that is appropriate, and, where necessary, to provide extra supports such as lesson modifications and tutoring.

Cross-grade grouping was a new strategy for all the SFA schools, but they implemented it from the first year forward (although in the third year, two principals told interviewers that their schools were grouping students within grade only). All principals and 89 percent of teachers who responded to the Year 3 surveys agreed that grouping helped students in their reading class become better readers.

Cross-grade grouping was not without its challenges, however. During interviews, staff members at half the schools noted that sometimes either too few or too many students tested into certain levels to create classes of equal size. Facilitators in some schools said that they looked at age and other social factors to avoid placing students in classes where they would be much older or much younger than most of the other students, in order to facilitate classroom management and to preserve students' self-esteem. A number of facilitators also commented that they tried to assign students who were having an especially hard time advancing to teachers they considered to be more able. In general, cross-grade grouping appeared to be as much an art as a science; perhaps inevitably, teachers participating in focus groups at seven schools opined that some students were placed in the wrong group.

Cooperative Learning

In cooperative learning, students, working together in either pairs or groups of four or five, hold each other accountable for their learning. Teachers in the SFA program are taught strategies and given tools to help their students work cooperatively, including modeling for students the kinds of behaviors and questions that lead to effective teamwork and helping students to set goals so that all will participate. To provide students with incentives, the teacher awards points to teams for meeting their goals, and they later celebrate the points they have earned. Sometimes the teacher calls on a student at random to represent his or her team. Therefore, to maximize their team points, students must first discuss the work together to make sure that all their team members understand it.

Snapshot scores indicate an improvement in the implementation of cooperative learning over time, although there remained room for improvement in teachers' use of this instructional method. Nearly all of the SFA teachers surveyed responded that their students worked in pairs or small groups daily, and the large majority (88 percent) agreed that cooperative learning helped students in their reading class become better readers. In focus groups, teachers also expressed positive views about the strategy, noting that it keeps students engaged, makes them responsible for each other, and gives them a chance to lead discussions; some even reported using it in teaching other subject areas. Teachers at a number of schools did note that it was difficult to get some students to participate in their teams and that behavior issues increased as a result, especially when the students were performing at a low level and did not have a clear understanding of what was expected of them or how to approach the work they were assigned.

Rapid Pacing

Success for All requires that teachers adhere to a quick pacing schedule. The 90-minute reading block is divided into short activities, and the lessons are grouped into six-day units. The pace is intended to allow teachers to complete daily lessons on time and to keep students engaged. The program is designed to review previously taught skills in subsequent units, so teachers are encouraged to move on during the lesson even when students have not demonstrated full mastery of a skill. Because of the strict time allotments and because many teachers had been taught to teach for mastery, teachers found pacing to be one of the most difficult parts of the program to implement in the first year.

Snapshot scores indicate an improvement in pacing over time. Facilitators or principals from half the schools reported that they monitored pacing or made it a focus of SFA implementation in the third year, and participants in focus groups held that year generally reported that pacing was easier than in the past, largely because the teachers were more familiar with the materials and the flow of the lessons.

Nonetheless, teachers at a number of schools, while noting that they followed the pacing guidelines, continued to express frustration or concern with the program's quick pace. They reported that getting through lessons or units remained a challenge, and they still expressed discomfort with moving on with the lessons before teaching skills to mastery. In response to teacher survey questions, only 47 percent of the teachers agreed that the reading program allowed most students in the class to learn critical concepts, and only 45 percent agreed that the pacing of the reading program allowed them to get through almost all the materials they needed to cover in class. Focus group participants at a number of schools acknowledged adapting the program's pacing guidelines to their own classrooms, sometimes having been advised to do so by SFAF staff.

Tutoring

SFA guidelines call for the use of an SFAF-developed computerized tutoring program, called Team Alphie, in tutoring students in first through third grades who struggle with reading. The Snapshot ratings indicate that there was fluctuation in the implementation of Team Alphie tutoring across the program years; by the third year, Team Alphie was used for tutoring at only 9 of the 19 program schools. Moreover, in Year 3, fewer than half the schools had the capacity to tutor 30 percent of first-grade students, 20 percent of second-grade students, and 10 percent of third-grade students, as specified by SFA guidelines.

Resource issues partly account for schools' difficulties in putting Team Alphie in place. Of the principals surveyed, 60 percent reported not being able to implement tutoring because of insufficient funding and/or staffing. Staff members at several schools noted that technical issues, such as malfunctioning computer equipment, made it difficult to provide Team Alphie.

It would be wrong to conclude that program schools did not provide tutoring, however. Rather, as an alternative (or, in some cases, supplement) to Team Alphie, many schools provided tutoring using programs provided or required by their districts, and to students in all grades rather than mainly those in the early grades. The large majority of principals responded that their school used a system of increasingly intensive reading interventions (for example, Response to Intervention, or RtI) for students who struggled with reading.

Components That Address Noninstructional Issues That Affect Learning

The elements in the "components to address noninstructional issues" category are strategies that can be used both in the classroom and schoolwide to address noninstructional factors that affect student academic achievement, such as student behavior and attendance.

Solutions Teams

In order to address noninstructional issues that affect student learning, SFA schools are expected to form five different committees, known as Solutions Teams, to address student absenteeism, "cooperative culture" (student behavior), family members' involvement in their children's academic experiences, community involvement, and interventions for students with reading difficulties. All teams are further expected to have an identified coordinator and specific meeting times, and all staff are encouraged to be members of at least one Solutions Team.

In the first year of implementation, only 10 schools were able to establish their teams and get them to meet regularly. By the third year, Solutions Teams were in place in the large majority of the schools.[7]

It is not clear, however, how much of a difference the SFA program made in how schools addressed noninstructional issues. As noted by school staff during first-year interviews, a number of the Solutions Teams predated their schools' adoption of SFA, albeit operating under different names. (For example, some schools' intervention teams were what had previously been their RtI teams, and the cooperative-culture teams had previously been their "positive behavioral intervention support" [PBIS] teams.) Schools with attendance issues were already ad-

[7] Snapshot scores show that 17 schools had an attendance team, 17 had a community involvement team, 16 had a parental involvement team, 15 had a behavior team, and 13 had an intervention team. Interviews with school staff yielded similar results.

dressing these before they launched SFA. And 12 schools in two districts were provided "family and community liaisons" by their districts.

Emphasis on Continuous Improvement

The "continuous improvement" elements include strategies that inform the implementation of the elements in the other two categories to increase student achievement. Through the use of data, teachers know which students to tutor, how quickly to pace lessons, how to group students, how to address behavior issues, and the like, while professional development teaches them how to put these elements in place.

Professional Development

In addition to the initial training that SFAF provides to school staff at the beginning of each school year, professional development is provided by SFAF throughout the school year. For both principals and facilitators, the bulk of the training is provided by SFAF point coaches, who regularly visit SFA schools.[8] A key activity during the coaches' visits involves classroom observations in which the principal and facilitator participate in order to get a sense of how the curriculum and SFA structures, such as cooperative learning, are being implemented. The point coaches then meet with the principal and the facilitator to go over what they saw, review goals using the Snapshot, and discuss strategies for improving implementation. The point coaches also support the principal and facilitator in their use of data, coaching and leadership practices, the implementation of Solutions Teams, and other SFA structures.

Both principals and facilitators were generally pleased with their point coaches. The principal and/or the facilitator at all 17 schools visited by the researchers during the third year of operations reported that the feedback from point coaches was useful.[9] Point coaches were described as supportive and available. Point coaches also gave feedback to teachers, either during in-person meetings or, more frequently, in notes in their school mailboxes. Teachers did not regard the point coaches as highly as their principals and facilitators did; the majority of the teachers surveyed reported that the feedback they had gotten from point coaches that year was only somewhat adequate or not at all adequate.

The professional development provided by the point coaches is mostly meant to be shared with the teachers via the facilitator during meetings of all staff teaching KinderCorner, Reading Roots, or Reading Wings (the three levels of SFA). Snapshot items in the second and third years of implementation show that component meetings were held and that facilitators

[8] The number of visits per year decreases each year since schools are expected to need less support as they gain experience in implementing SFA.

[9] The principal survey shows that 86 percent of principals found the feedback from both facilitators *and* coaches useful.

used specific SFA coaching strategies in which they had been trained. Survey results show that teachers thought that their facilitators were quite knowledgeable about SFA, and 83 percent of them agreed that the feedback they provided was useful.

Use of Data

Data-driven instruction is a foundation of many current reading programs and district reforms. In response to survey questions, SFA teachers reported that their schools most frequently used data to identify and monitor struggling students, to examine schoolwide instructional issues related to reading, and to communicate with and inform parents about student reading performance. Slightly more than half the teachers said that their school used data to identify teachers who needed instructional improvement. The Snapshot rating shows that six SFA schools were not using Member Center (SFA's data system) or equivalent data collection and reporting tools consistently by the end of Year 3. The interviews suggest that some teachers may not have used Member Center because they found the program difficult to navigate, the amount of data they were required to input burdensome, the technology slow, and the training on how to use it inadequate.

Overall, the analysis shows that implementation increased over time, but the biggest improvement was seen between Years 1 and 2. Some teachers continued to hold reservations about the program; nonetheless, by Year 3, the majority of staff members (93 percent of the principals and 70 percent of the teachers) responding to surveys agreed that the SFA program had benefited their schools. Principals were more enthusiastic about SFA than teachers, however, with 57 percent of the principals but only 17 percent of the teachers *strongly* agreeing that the program had benefited their schools.

Factors Promoting Program Implementation

The lower section of the logic model in Figure 1.1 presents a number of contextual factors that may affect implementation regardless of the quality or level of effort of program developers or school personnel. The data were examined to see if there were any meaningful correlations between each school's overall Snapshot score and principal and teacher characteristics, including years of experience and years at their current school. No interpretable patterns were found.

Chapter 4

SFA Schools and Control Group Schools Compared

Chapter 3 assesses the extent to which central features of the Success for All (SFA) program were implemented during the third year of operations at the 19 program group schools participating in the SFA evaluation ("SFA schools"). Implementation that is faithful to the program model is hypothesized to be essential to SFA's ability to produce positive impacts on student achievement and other outcomes.

Impacts, however, are driven not only by what happens in program group schools but also by what happens in control group schools. Unless SFA schools are distinct from control group schools in at least some features, there is no reason to expect differences between the two groups of schools in student outcomes.

This chapter covers much of the same territory as the previous one, and it is similarly grounded in the program's logic model. It addresses a different set of questions, however, focusing on the extent to which the program schools and control group schools resembled or differed from each other, especially with regard to features that are central to the SFA intervention. To address this issue, the chapter makes use of various data sources. Quantitative information comes from teacher logs and teacher and principal surveys. Interviews and focus groups supply qualitative data.[1]

The chapter begins by comparing program and control group schools in terms of selected characteristics identified as "contextual factors" in the logic model. The discussion then turns to how the two groups of schools compare with respect to structures and practices that fall in the "Inputs" and "Program Elements" columns of the logic model, with special attention to aspects of reading that teachers emphasize in day-to-day instruction.[2] The attitudes of SFA and control group teachers toward various aspects of their reading programs, as well as their beliefs about the extent to which these programs engage their students, are examined. The chapter concludes by summarizing the key areas of difference and similarity in the program's third year of implementation.

[1] Four SFA schools and two control group schools did not return surveys. There is no way to know how the findings would have been affected had they supplied this information. However, using survey data from 2012-2013, the research team checked whether teachers in schools that are missing data for 2013-2014 were more dissatisfied with the reading program at their schools than teachers from schools that returned surveys. In fact, teachers in schools that did not return surveys were *more* satisfied, on average, than teachers from schools that did return surveys.

[2] The research team does not have reliable measures for teacher pacing practices, the frequency of student assessment, or use of socio-emotional regulation and conflict resolution strategies. SFA and control group schools could not be compared on these elements, so this chapter does not discuss them.

Key Findings

- SFA teachers were less likely to focus on grammar, writing, and spelling during reading instruction, and at least for second-grade students, more likely to focus on reading comprehension.
- SFA schools made more extensive use of ability-grouping for reading than control group schools. In particular, SFA schools used cross-grade grouping, whereby students from different grades were placed in the same reading class. No control group schools used cross-grade grouping.
- Students in SFA classrooms were more likely to work together in small groups.
- Program and control group schools looked quite similar on several program elements that SFA deems critical: professional development, data use, tutoring, and the presence of individuals or groups to address noninstructional issues that are thought to affect learning.
- SFA teachers were just as satisfied with the overall quality of their reading program as control group teachers, and both were generally satisfied.
- Nonetheless, SFA teachers were more skeptical than control group teachers about the adequacy of their reading program with respect to serving students with particular learning challenges, including special education students, English language learners (ELLs), and students with behavior problems.

With a few exceptions noted below, these third-year findings are consistent with those of the study's first two years.

Contextual Factors: Selected Characteristics of Teachers and Principals

As Table 4.1 shows, principals in the SFA schools and in the control group schools had similar amounts of experience as principals (7.5 years for principals in program group schools and 6 years for principals in control group schools). However, SFA principals had headed their current school for about 2.8 years more, on average, than control group principals. This difference is statistically significant at the 5 percent level.

For teachers, the story was quite different: Teachers in SFA schools had fewer years of teaching experience, on average, than teachers in control group schools (6.7 years and 9.0 years,

Table 4.1

SFA-Control Group Comparisons on Selected Characteristics of Teachers and Principals (Implementation Year 2013-2014)

	Program Group	Control Group	Estimated Difference	P-Value	
Principal Characteristics					
Average number of years of experience as a principal	7.5	5.9	1.5	0.392	
Average number of years as a principal at current school	5.6	2.9	2.8	0.049	**
Teacher Characteristics					
Average number of years of experience as a teacher	6.7	9.0	-2.3	0.050	**
Percentage of teachers with 1 year of experience or less teaching at any elementary school	13.8	6.3	7.5	0.038	**
Percentage of teachers who have been at current school for 1 year or less	23.8	11.8	12.0	0.005	***

SOURCES: Spring 2014 teacher and principal surveys.

NOTES: A two-tailed t-test was applied to differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

Completed surveys were received from 15 out of 19 SFA schools and 16 out of 18 control group schools. Completed surveys were received from 297 teachers at SFA schools and 233 teachers at control group schools.

respectively). Teacher surveys indicate that, at the end of the first implementation year, a significantly higher proportion of teachers at SFA schools than at control group schools (12 percent and 6 percent, respectively) reported that they were first-year teachers. (Because these measures were taken at the end of the first year rather than at the point of random assignment, it is not possible to determine whether this difference predated SFA's introduction into the schools or came about during the first year). This gap continued into the third year: About 13 percent of SFA teachers were new to teaching as of the 2013-2014 school year, compared with 6 percent of control group teachers. Conversely, 43 percent of control group teachers in the first year had been teaching for 10 or more years, compared with 40 percent of SFA teachers, and this difference increased substantially over time: By 2014, 49 percent of control group teachers compared with 36 percent of SFA teachers had 10 years of experience in the classroom.

Although available measures of turnover are inconclusive, these differences may reflect higher rates of turnover in SFA schools. One notable Year 3 finding is that whereas three-quarters of the control group teachers agreed that teacher morale had been high at their school since the beginning of the year, fewer than half of the SFA teachers agreed with the same statement. Lower morale among SFA teachers might have caused increased turnover, creating the observed disparity in the proportions of new teachers in the two groups of schools.

Attention now focuses on the logic model and the extent to which structures and practices are similar or different in program and control group schools.

Developer Inputs

The logic model specifies two inputs related to SFA implementation that the Success for All Foundation is responsible for providing: program materials and essential training. Their analogues in control group schools are discussed below.

Materials

SFA provided program group schools with materials that included books of graduated difficulty, worksheets, and instructional videos. Interviews with control group school principals supplied information about their schools' reading curricula. The majority of control group schools used common basal reading programs available from leading educational publishers (including Macmillan/McGraw-Hill, Houghton Mifflin Harcourt, and Scott Foresman). The reading programs, like SFA, generally include a teacher's guide to help organize lesson plans, stories that are intended to engage children in reading, assessments, and suggested materials and strategies for struggling readers. The programs are also similar to SFA in covering the five reading components (phonics, phonemic awareness, vocabulary, fluency, and reading comprehension) identified by an influential National Reading Panel report as essential to reading instruction, in striking a balance between decoding and comprehension skills, and in integrating the Common Core standards.[3] The discussion that follows in this chapter suggests that it is not the curriculum per se but the specific ways in which it is enacted that differentiate SFA schools from control group schools.

Essential Training

As noted in Chapter 3, "essential training" refers to the initial training that school leaders and staff receive about the SFA reading program. The research team did not collect data about training at control group schools that would correspond to the essential training delivered

[3] National Reading Panel (2000). The Common Core standards — established by the Common Core State Standards Initiative led by the National Governors Association Center for Best Practices and the Council of Chief State School Officers — specify what students at each grade level from kindergarten through grade 12 should know and be able to do in the foundational areas of English and mathematics. Intended to prepare students more adequately for college and the workplace, the standards have been adopted by 43 states and the District of Columbia. See Common Core State Standards Initiative (2015).

at the program group schools. (A later section in this chapter uses teacher survey data to measure receipt of professional development more generally at the two groups of schools.)

School Inputs

Two school inputs into the SFA reading program include a full-time SFA facilitator and a committed principal. Interviews with principals at both sets of schools were the chief source of information about reading program facilitators, and teacher surveys supplied data about principal leadership in general and leadership of the reading program in particular.

Reading Program Facilitators

Systematic information on the presence of reading coaches in control group schools was not collected. Interviews with control group school principals conducted for the cost study indicate that many of the schools had reading coaches, although, in contrast to the SFA schools, the coaches' time was usually split across more than one school.

Principal Leadership and Involvement with the Reading Program

Teachers in program and control group schools held similar views about their principals' leadership capacities. A number of survey items asked teachers the extent to which they agreed that their principal was involved with the school's reading program and had played a general leadership role.[4] Seventeen of these items were clustered together to form a scale with a maximum value of 4. The mean score for SFA school principals was 3.04, while for control group school principals it was 3.10. This difference is not statistically significant.

Program Elements

Challenging Reading Instruction

Table 4.2 compares SFA schools and control group schools on a number of instructional elements that are fundamental to SFA reading instruction. These elements are intended to provide students with instruction that is both challenging and adapted to individual needs.

SFA guidelines call for a 90-minute reading block and small reading classes, especially in the early grades. The average length of the reading block in SFA schools (89 minutes) was

[4] For example, one item asks teachers the extent to which they agree that their principal provides them with sufficient time for planning reading instruction. Another asks the extent to which teachers agree that their principal makes expectations for meeting student learning goals clear to them. The scale has a Cronbach's alpha reliability coefficient of 0.95.

Table 4.2

SFA-Control Group Comparisons
Related to Challenging, Individualized Instruction (Implementation Year 2013-2014)

	Program Group	Control Group	Estimated Difference	P-Value	
Length of the reading block					
Average length of the reading block (in minutes) on a typical day (excluding grammar and writing), according to teacher reports	88.6	98.5	-9.9	0.001	***
Small class size					
Average number of students in reading class, according to teacher report	18.7	22.1	-3.3	0.002	***
Grouping					
Percentage of principals reporting that:					
Students in the same reading class are divided into smaller groups	61.5	50.0	11.5	0.564	
Students in the same grade are grouped into different reading classes by ability level	100.0	38.5	61.5	0.000	***
Students who are in different grades, but at the same ability level, are sometimes grouped together in the same reading class	100.0	0.0	100.0	0.000	***
Percentage of teachers reporting that students are periodically regrouped for reading by ability level	97.8	74.4	23.4	0.006	***
Cooperative learning					
Percentage of teachers who agree that students work in pairs or small groups daily or almost daily	99.0	70.0	29.0	0.000	***
Percentage of classrooms in which students were observed working in small groups	84.9	61.7	23.2	0.055	*
Percentage of classrooms where teams were rewarded	23.0	0.0	23.0	0.013	**

(continued)

Table 4.2 (continued)

	Program	Control	Estimated	P-	
	Group	Group	Difference	Value	
Tutoring					
Percentage of principals reporting that their school has a designated time for tutoring in addition to the regularly scheduled reading block	92.9	92.9	0.0	1.000	
Percentage of second-graders receiving tutoring	16.0	21.1	-5.1	0.199	
Percentage of principals reporting that students are tutored one on one	35.7	38.4	-2.7	0.888	
Percentage of principals reporting that students receive tutoring in pull-out groups	85.7	76.9	8.8	0.574	
Percentage of principals reporting that tutoring is scheduled every day for all students assigned to tutoring	57.1	23.0	34.1	0.077	*
Average length of a tutoring session (in minutes)	25.7	32.9	-7.1	0.107	
Percentage of principals reporting that their school uses a system of increasingly intensive interventions for students who are struggling with reading	85.7	85.7	0.0	1.000	
Use of educational media/technology					
Percentage of teachers who agree that they use educational media/technology as part of the reading program at their school	90.6	81.8	8.8	0.109	
Average amount of time teachers report using educational media/technology in their most recent reading class	44.5	18.0	26.5	0.000	***

(continued)

Table 4.2 (continued)

SOURCES: Spring 2014 teacher and principal surveys and classroom observations.

NOTES: Items on the teacher and principal surveys 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 or principals who agree with an item were obtained by taking the number who responded 3 or 4 and dividing by the total number of respondents to that item.

The means reported for teacher survey items are means of school means. First, means are taken within each school at the teacher level, then the mean across school means is taken. This was done to prevent overweighting schools with more teachers.

A two-tailed t-test was applied to differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

Completed surveys were received from 14 out of 19 principals at SFA schools and 14 out of 18 principals at control group schools. Completed teacher surveys were received from 15 out of 19 SFA schools and 16 out of 18 control group schools. Completed surveys were received from 297 teachers at SFA schools and 233 teachers at control group schools.

Response rates for all teacher survey items presented were above 98 percent for both SFA and control group teachers. The response rate for both SFA and control group principals was at least 11 principals out of 14 for all items presented in this table.

significantly shorter than the length of the average reading block in control group schools (about 99 minutes).[5] SFA classrooms were smaller, with an average of about 19 students, about 3 fewer than the average control group classroom. This difference is highly statistically significant.

A central aspect of the SFA reading program is the use of ability grouping for reading, which all SFA and control group schools reported in some fashion.[6] However, SFA schools distinguished themselves from control group schools by making use of a cross-grade grouping strategy whereby students at the same ability level, but from different grades, may be placed in the same reading class. Out of the 14 SFA principals surveyed, all 13 who answered the relevant item indicated that their school used cross-grade grouping.[7] No control group principals indicated that their school used cross-grade grouping.

Almost 98 percent of SFA teachers reported regrouping students throughout the school year based on the students' progress in reading. Although a substantial percentage of control group teachers also said they regrouped students (about 74 percent), the difference is highly statistically significant.

As discussed in Chapter 3, cooperative learning is a key SFA instructional method whereby students work together in small groups and take responsibility for making sure that everyone in their group is learning the relevant material. To foster this sense of shared responsibility, teachers are supposed to call on students randomly; if the student answers correctly, his or her group is rewarded with points or other forms of positive recognition. While the evaluation lacks a direct measure of this conception of cooperative learning, classroom observations and the teacher survey data supply measures of several of its key components. Almost all SFA teachers reported that students worked together in small groups daily or almost daily, compared with about 70 percent of control group teachers. MDRC researchers observed students working together in small groups in about 85 percent of the SFA classrooms they visited, compared with 62 percent of control group classrooms. Observations also indicate that teachers in SFA classrooms were significantly more likely than their control group counterparts to reward teams for their performance (23 percent compared with 0 percent), and to call on students randomly (a practice that occurred in 40 percent of SFA classrooms compared with 7 percent of control group classrooms that were observed).[8] (The latter result is not statistically significant but is

[5] This finding is based on a teacher survey item that asks specifically about time devoted to *reading* and not to writing or grammar instruction.

[6] Although some control group principals indicated on the survey that they did not use ability grouping, teachers from every school reported that students in their reading classes were either grouped by ability level or periodically regrouped for reading by ability level.

[7] In contrast to the survey results, interview data show that 2 of these 13 schools did not actually implement cross-grade grouping during the 2013-2014 school year.

[8] The result pertaining to teachers calling on students randomly is not presented in this chapter's tables.

nevertheless of interest.) All these results, taken together, suggest that the form of cooperative learning envisioned by SFA is taking place more frequently in program group schools.

Although tutoring is a critical element of the SFA program, it is perhaps not surprising that many tutoring practices in the two sets of schools were quite similar, given the implementation difficulties noted in the last chapter.[9] Thus, identical proportions (93 percent) of SFA and control group principals noted that their school had a designated time for tutoring that was distinct from the regular reading block, and similar percentages reported that students were tutored one on one (36 percent and 38 percent, respectively) and in pull-out groups (86 percent and 77 percent, a difference that is not statistically significant). Moreover, 86 percent of principals in both groups indicated that their schools used a system of increasingly intensive interventions (known as Response to Intervention, or RtI) to provide instruction to students who were struggling with reading.[10] Perhaps most salient is that, despite the importance that SFA attaches to tutoring, the proportion of second-grade students who received tutoring in the SFA schools was not significantly different from the comparable proportion in control group schools (16 percent and 21 percent, respectively).[11]

The SFA program emphasizes the importance of using educational media to help engage students in reading tasks. About 91 percent of SFA teachers reported that educational media were used in their reading program, compared with 82 percent of control group school teachers — a difference that just misses statistical significance at the 10 percent level. However, SFA teachers reported using educational media in their most recent reading class for 27 minutes more, on average, than control group teachers, and this result is highly statistically significant.

A More Fine-Grained Look at Reading Instruction in SFA Schools and Control Group Schools

In order to understand the similarities and differences in literacy instruction that program group and control group students received, teachers of literacy in both groups of schools

[9] The financial constraints that inhibited wider implementation of tutoring were not unique to SFA schools. A virtually identical proportion of program and control group school principals (some 60 percent) reported that their school could not implement tutoring because of insufficient staff or funding. (While the question asked principals about program elements that they could not implement “at all,” the pattern of responses suggests that some of them answered in terms of program elements that were implemented, but not at the level intended.)

[10] In two respects, tutoring practices at the two groups of schools differed: Principals at SFA schools were more likely to report that tutoring was scheduled daily for those students assigned to receive it (57 percent compared with 23 percent), a difference that is statistically significant at the 10 percent level, while tutoring sessions were slightly longer on average at control group schools than at SFA schools (33 minutes and 26 minutes, respectively).

[11] Similarly, the second-year report found that a virtually identical proportion of first-graders in the two groups of schools — 22 percent — received tutoring during the 2012-2013 academic year.

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 first- and/or second-grade students.[12]

A central question that the logs help answer is whether SFA teachers and control group teachers focused on different topics during their literacy instruction. Figure 4.1 graphically depicts the odds ratio that the average SFA teacher, vis-à-vis the average control group teacher, focused on each of seven topic areas during the literacy block: comprehension, word analysis, writing, reading fluency, vocabulary, grammar, and spelling.[13] (See Box 4.1 for an explanation of how to read the figures that illustrate odds ratios.) The figure shows that the instruction offered by SFA teachers differed significantly from that offered by control group teachers in three of the seven areas that were studied. SFA teachers were less likely to focus on spelling than their control group counterparts; the odds ratio for this area is 0.15. They were also less likely to focus on writing (odds ratio = 0.41) and grammar (odds ratio = 0.27).

Teachers were also asked more specifically about instruction that the selected students received in comprehension and on word analysis (the structure of words or the sounds and letters that make up words) when these were the focus of lessons. The first panel of Figure 4.2 shows the particular strategies for teaching comprehension that teachers used. SFA teachers were much more likely than control group teachers to elicit brief answers demonstrating students' understanding of text. Moreover, in line with SFA's use of cooperative learning, in the lessons in which comprehension was a focus, SFA teachers were more likely to have students discuss text with each other than were control group teachers (odds ratio = 1.80).

Some areas of comprehension instruction are more cognitively demanding than others. For instance, predicting what a text will be about based on the illustrations is far easier than analyzing characters' motivations. Figure 4.3 shows that SFA teachers were less likely than control group teachers to have students sequence information or events in text, which is somewhat cognitively demanding (odds ratio = 0.31). SFA and control group teachers did not differ in their use of any other cognitively demanding comprehension-related instructional strategies.

The second panel of Figure 4.2 indicates the particular teaching strategies that were used when word analysis was a focus of instruction. SFA teachers were somewhat less likely than control group teachers to focus on sight words (odds ratio = 0.60), which are words that students learn to recognize and read by sight, and on structural analysis (odds ratio = 0.42), which entails examining word families, prefixes, suffixes, contractions, and so on.

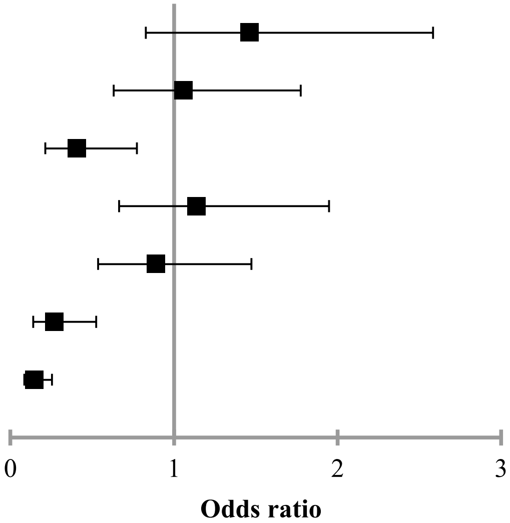
[12] The logs employed for this study were adapted from those used by Brian Rowan, Eric Camburn, and Richard Correnti for the Study of Instructional Improvement conducted by the University of Michigan in partnership with the Consortium for Policy Research in Education.

[13] It is worth noting that while writing, vocabulary, and grammar are treated as separate domains in this analysis, all of them can be taught in a way that enhances comprehension.

Figure 4.1

Instructional Differences Between SFA Classrooms and Control Group Classrooms in the Language Arts Topic Focus (2013-2014)

Area of Focus	Odds Ratio	Confidence Interval
Comprehension	1.46	(0.83 2.58)
Word analysis	1.06	(0.63 1.77)
Writing	0.41 **	(0.21 0.77)
Reading fluency	1.14	(0.66 1.95)
Vocabulary	0.89	(0.54 1.47)
Grammar	0.27 ***	(0.14 0.52)
Spelling	0.15 ***	(0.08 0.25)



Sample size: 1,771 logs (981 in program group and 790 in control group) from 224 teachers (124 in program group and 100 in control group).

SOURCE: Teacher logs administered in spring 2014.

NOTES: The analysis sample consists of 1,771 teacher logs (981 from program group schools and 790 from control group schools) collected from 224 grade 1 and grade 2 reading teachers (124 in the program group and 100 in the control group) in 29 schools (14 program group schools and 15 control group schools).

An odds ratio (OR) compares the odds of a certain practice being used in the average SFA school with the odds that it was used in the average control group school in the sample. Note that an OR of 1 for any outcome indicates that teachers in the SFA and control group schools were equally likely to have focused on that outcome across all logs in the study. An OR greater than 1 indicates that teachers in SFA schools were more likely to focus on that outcome, and an OR less than 1 indicates that teachers in SFA schools were less likely than teachers in control group schools to focus on that outcome.

In addition, the rightmost column presents the 90 percent confidence interval for these ORs. Instruction in SFA schools can be said to be statistically different from instruction in control group schools when the line representing the 90 percent confidence interval for the estimate does not cross the line representing an OR of 1.

All estimations are based on a three-level hierarchical linear model logistic regression with individual logs nested within teachers and teachers nested within schools. Results are presented only in cases in which teachers' logs indicated that the given area of focus was a "major or minor focus" of instruction.

A two-tailed t-test was applied to determine whether the estimated OR is statistically different from 1. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Box 4.1

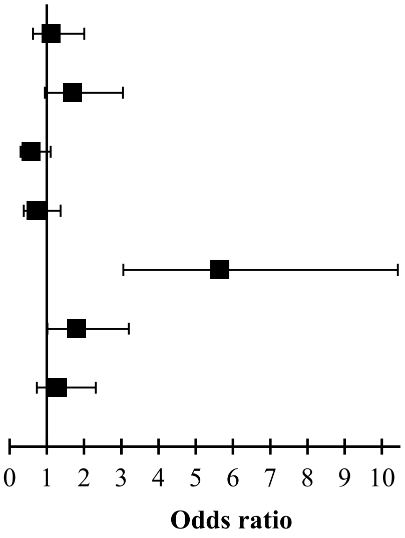
How to Read the Figures Depicting Odds Ratios

Logistic regression is the preferred statistical method when an outcome is binary: Either the event occurred or it did not. An odds ratio of 1 (as illustrated by the vertical line in Figure 4.1) indicates that the average SFA teacher and the average control group teacher were equally likely to have focused on a specific topic. An odds ratio greater than 1 indicates that the average SFA teacher was more likely to focus on the topic, and an odds ratio less than 1 indicates that the average control group teacher was more likely to focus on the topic. The horizontal lines on each side of the odds ratio represent the "confidence interval" — that is, the range of estimated values of the odds ratio, within which there is a 90 percent probability that the true odds ratio falls. The odds ratio is statistically significant when neither the upper bound nor the lower bound of the confidence interval surrounding it crosses the vertical line that represents the odds ratio of 1. Asterisks indicate whether an odds ratio is significant at the level of 10 percent, 5 percent, or 1 percent.

Figure 4.2

Impact of SFA on Teachers' Instruction, by Language Arts Construct

Comprehension Construct	Odds Ratio		Confidence Interval
Activate knowledge	1.12		(0.63 2.00)
Literal comprehension	1.70		(0.95 3.05)
Story structure	0.57		(0.30 1.10)
Analyze/synthesize	0.72		(0.38 1.37)
Brief answers	5.65	***	(3.06 10.43)
Students discuss text	1.80	*	(1.01 3.20)
Teacher-directed instruction	1.30		(0.73 2.32)

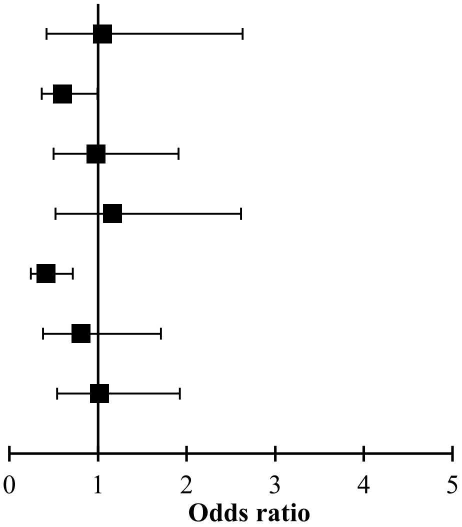


Sample size: 1,242 logs (715 in program group and 527 in control group) from 213 teachers (117 in program group and 96 in control group).

(continued)

Figure 4.2 (continued)

Word Analysis Construct	Odds Ratio	Confidence Interval
Letter-sound relationships	1.05	(0.42 2.63)
Sight words	0.60 *	(0.37 0.99)
Use picture/context cues	0.98	(0.50 1.91)
Use phonics cues	1.17	(0.52 2.61)
Structural analysis	0.42 **	(0.24 0.72)
Assess student ability	0.81	(0.38 1.71)
Teacher-directed instruction	1.02	(0.54 1.92)



Sample size: 781 logs (440 in program group and 341 in control group) from 175 teachers (92 in program group and 83 in control group).

SOURCE: Teacher logs administered in spring 2014.

NOTES: The constructs are taken from Correnti and Rowan (2007).

See notes to Figure 4.1 for an explanation of the odds ratio (OR) and confidence interval.

All estimations are based on a three-level hierarchical linear model logistic regression with individual logs nested within teachers and teachers nested within schools. The figure presents results based on a teacher log analysis sample that was restricted to include only logs indicating comprehension or word analysis, respectively, as a "major or minor focus" of instruction. The analysis sample for comprehension constructs is 1,242 logs (715 logs from program group schools and 527 from control group schools), completed by 213 teachers (117 from program group schools and 96 from control group schools). The analysis sample for word analysis constructs is 781 logs (440 from program group schools and 341 from control group schools), completed by 175 teachers (92 from program group schools and 83 from control group schools).

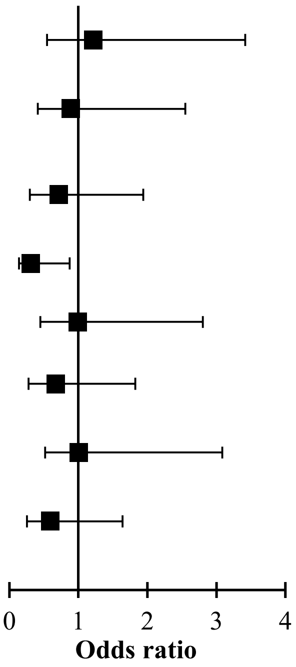
A two-tailed t-test was applied to determine whether the estimated OR is statistically different from 1. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Table 4.3 examines separately for first- and second-grade students the topical focus of reading instruction and the strategies used in teaching comprehension and word analysis skills. Most notable is SFA's greater emphasis on comprehension in second-grade classrooms. The average second-grade SFA teacher was 2.10 times as likely to focus on comprehension as the average second-grade control group teacher. In teaching comprehension, SFA teachers in both

Figure 4.3

Impact of SFA on Instruction of Cognitively Demanding Items

Item	Odds Ratio	Confidence Interval
Activate knowledge		
Activating prior knowledge	1.22	(0.67 2.20)
Previewing, predicting, surveying text	0.89	(0.47 1.66)
Story structure		
Summarizing important details in text	0.72	(0.42 1.22)
Sequencing information/events in text	0.31 ***	(0.18 0.56)
Using concept maps/frames	1.00	(0.55 1.81)
Identifying story structure	0.67	(0.39 1.15)
Analyze/synthesize		
Analyzing/evaluating text	1.01	(0.49 2.08)
Comparing/contrasting information in text	0.59	(0.34 1.05)



Sample size: 1,242 logs (715 in program group and 527 in control group) from 213 teachers (117 in program group and 527 in control group).

SOURCE: Teacher logs administered in spring 2014.

NOTES: The items are subcategories of the construct "comprehension," as discussed in Correnti and Rowan (2007).

See notes to Figure 4.1 for an explanation of the odds ratio (OR) and confidence interval.

All estimations are based on a three-level hierarchical linear model logistic regression with individual logs nested within teachers and teachers nested within schools. The figure presents results based on a teacher log analysis sample that was restricted to include only logs indicating comprehension as a "major or minor focus" of instruction.

A two-tailed t-test was applied to determine whether the estimated OR is statistically different from 1. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Table 4.3

Instructional Differences Between SFA Schools and Control Group Schools (Implementation Year 2013-2014)

	All Grades	Grade 1	Grade 2
Construct	Odds Ratio	Odds Ratio	Odds Ratio
Language arts focus[a]			
Comprehension	1.46	1.05	2.10 *
Word analysis	1.06	1.37	0.88
Writing	0.41 **	0.45	0.40 **
Reading fluency	1.14	2.36	0.60
Vocabulary	0.89	1.19	0.72
Grammar	0.27 ***	0.34 **	0.30 ***
Spelling	0.15 ***	0.28 **	0.08 ***
Comprehension[b]			
Activate knowledge	1.12	2.00	0.64
Literal comprehension	1.70	1.89	1.56
Story structure	0.57	0.45 *	0.60
Analyze/synthesize	0.72	0.76	0.80
Brief answers	5.65 ***	11.01 ***	2.28 *
Students discuss text	1.80 *	1.76	1.82
Teacher-directed instruction	1.30	1.87	0.76
Word analysis[c]			
Letter-sound relationships	1.05	0.90	2.38
Sight words	0.60 *	0.66	0.48
Use picture/context cues	0.98	1.25	0.88
Use phonics cues	1.17	1.47	0.74
Structural analysis	0.42 **	0.32 ***	0.67
Assess student ability	0.81	1.55	0.51
Teacher-directed instruction	1.02	2.02	0.41 *
Cognitively demanding items[d]			
Activate knowledge			
Activating prior knowledge	1.22	2.22	0.72
Previewing, predicting, surveying text	0.89	1.18	0.73
Number of schools	29	28	28

(continued)

Table 4.3 (continued)

Construct	All Grades Odds Ratio	Grade 1 Odds Ratio	Grade 2 Odds Ratio
Story structure			
Summarizing important details in text	0.72	0.74	0.61
Sequencing information or events in text	0.31 ***	0.27 ***	0.42 *
Using concept maps/frames	1.00	1.09	0.78
Identifying story structure	0.67	0.73	0.61
Analyze/synthesize			
Analyzing/evaluating text	1.01	0.88	1.22
Comparing/contrasting information	0.59	0.92	0.45
Number of schools	29	28	28

SOURCE: Teacher logs administered in spring 2014.

NOTES: Constructs are taken from Rowan, Camburn, and Correnti (2004).

An odds ratio (OR) compares the odds of a certain practice being used in the average SFA school with the odds that it was used in the average control group school in the sample. Note that an OR of 1 for any outcome indicates that teachers in the SFA and control group schools were equally likely to have focused on that outcome across all logs in the study. An OR greater than 1 indicates that teachers in SFA schools were more likely to focus on that outcome, and an OR less than 1 indicates that teachers in SFA schools were less likely than teachers in control group schools to focus on that outcome.

All estimations are based on a three-level hierarchical linear model logistic regression with individual logs nested within teachers and teachers nested within schools.

A two-tailed t-test was applied to determine whether the estimated OR is statistically different from 1. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]The analysis sample for language arts focus items consists of 1,771 teacher logs (981 from program group schools and 790 from control group schools) collected from 224 grade 1 and grade 2 reading teachers (124 in the program group and 100 in the control group) in 29 schools (14 program group schools and 15 control group schools). The grade 1 subset consists of 939 teacher logs (529 from program group schools and 410 from control group schools) collected from 133 teachers (81 in the program group and 52 in the control group) in 28 schools (14 program group schools and 14 control group schools). The grade 2 subset consists of 832 teacher logs (452 from program group schools and 380 from control group schools) collected from 131 teachers (82 in the program group and 49 in the control group) in 28 schools (14 program group schools and 14 control group schools).

[b]The analysis sample for comprehension constructs was restricted to include only those logs where teachers indicated comprehension as “a focus of instruction.” The sample consists of 1,242 teacher logs (715 from program group schools and 527 from control group schools) collected from 213 grade 1 and grade 2 reading teachers (117 in the program group and 96 in the control group) in 29 schools (14 program group schools and 15 control group schools). The grade 1 subset consists of 653 teacher logs (364 from program group schools and 289 from control group schools) collected from 125 teachers (75 in the program group and 50 in the control group) in 28 schools (14 program group schools and 14 control group schools). The grade 2 subset consists of 589 teacher logs (351 from program group schools and 238 from control group schools) collected from 121 teachers (74 in the program group and 47 in the control group) in 28 schools (14 program group schools and 14 control group schools).

(continued)

Table 4.3 (continued)

[c]The analysis sample for word analysis constructs was restricted to include only those logs where teachers indicated word analysis as "a focus of instruction." The sample consists of 781 teacher logs (440 from the program group schools and 341 from the control group schools) collected from 175 grade 1 and grade 2 reading teachers (92 in the program group and 83 in the control group) in 28 schools (14 program group schools and 14 control group schools). The grade 1 subset consists of 525 teacher logs (310 from the program group schools and 215 from the control group schools) collected from 112 teachers (68 in the program group and 44 in the control group) in 27 schools (14 program group schools and 13 control group schools). The grade 2 subset consists of 256 teacher logs (130 from program group schools and 126 from control group schools) collected from 87 teachers (47 in the program group and 40 in the control group) in 27 schools (14 program group schools and 13 control group schools).

[d]The analysis sample for cognitively demanding items was restricted to include only those logs where teachers indicated comprehension as "a focus of instruction." The items are subcategories of the construct "comprehension," as discussed in Correnti and Rowan (2004). The sample consists of 1,242 teacher logs (715 from program group schools and 527 from control group schools) collected from 213 grade 1 and grade 2 reading teachers (117 in the program group and 96 in the control group) in 29 schools (14 program group schools and 15 control group schools). The grade 1 subset consists of 653 teacher logs (364 from program group schools and 289 from control group schools) collected from 125 teachers (75 in the program group and 50 in the control group) in 28 schools (14 program group schools and 14 control group schools). The grade 2 subset consists of 589 teacher logs (351 from program group schools and 238 from control group schools) collected from 121 teachers (74 in the program group and 47 in the control group) in 28 schools (14 program group schools and 14 control group schools).

grades were more likely to ask students to supply brief answers than were their counterparts in the control group schools.

Components That Address Noninstructional Issues

SFA is a whole-school reform effort, not merely a reading program. The SFA program involves groups of teachers and other staff members (called "Solutions Teams") who are assigned to address noninstructional issues — poor attendance, disruptive behavior, lack of family support, and the like — that affect student learning.

As Table 4.4 shows, the SFA schools were not unique in this regard. The majority of control group principals also reported that their school had an individual or group of individuals charged with finding solutions to these problems. This is not altogether surprising, since, as noted in Chapter 3, teams addressing these issues predated Success for All in the program schools and presumably also existed in other schools in their districts. SFA principals were significantly more likely than their control group counterparts to report having a group or individual responsible for helping students with particular learning challenges and for helping teachers improve their reading instruction. Along other dimensions, the two groups of schools were similar, or the differences favored SFA schools but were not statistically significant.[14]

[14]These groups and individuals may have functioned differently in program and control group schools in ways that this evaluation could not assess. For example, a group of teachers charged with addressing community outreach could develop a strong sense of collegiality and heightened feelings of loyalty to their school that would not be expected if this responsibility were vested in a single individual.

Table 4.4

SFA-Control Group Comparisons
Related to Noninstructional Components That Affect Learning (Implementation Year 2013-2014)

	Program Group	Control Group	Estimated Difference	P-Value	
Percentage of principals who report that a group or individual at their school is responsible for:					
Developing schoolwide solutions for students with behavior problems	92.9	85.8	7.1	0.558	
Developing schoolwide solutions for students with learning problems	100.0	64.3	35.7	0.012	**
Helping teachers to improve their reading instruction of students	100.0	71.4	28.6	0.031	**
Implementing, monitoring, and improving a schoolwide program around social skills development and conflict resolution for all students	85.7	85.7	0.0	1.000	
Developing schoolwide solutions to improve student attendance	92.3	78.6	13.7	0.334	
Fostering closer relationships between the school and students' families	85.7	85.7	0.0	1.000	
Building relationships with local businesses and institutions to increase community involvement	71.4	57.1	14.3	0.449	

SOURCE: Spring 2014 principal survey.

NOTES: Items on the principal 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 principals who agree with an item were obtained by taking the number who responded 3 or 4 and dividing by the total number of respondents to that item.

A two-tailed t-test was applied to differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

Completed surveys were received from 14 out of 19 principals at SFA schools and 14 out of 18 principals at control group schools.

The response rate for both SFA and control group principals was at least 12 principals out of 14 for all items presented in this table.

Continuous Improvement

SFA's commitment to continuous improvement emerges in its provision of ongoing professional development to teachers and its implementation in each school of a data system to record individual and collective progress in reading achievement. These elements, and the extent to which they were found in control group schools, are explored in Table 4.5.

During the first year of the demonstration, teachers in the SFA schools were more likely to report having received professional development in reading instruction, to have received it on a greater number of reading-related topics, and to rate it as more helpful than their counterparts in the control group schools. By the third year, however, SFA and control group teachers were indistinguishable with respect to both the reported quantity and perceived quality of professional development. As Table 4.5 shows, similar proportions of SFA and control group teachers reported receiving professional development on each of several topics (for example, how to teach students at various levels of reading proficiency or how to use instructional time more effectively). The total number of topics on which they received professional development was also similar for both groups (6.4, on average, for SFA teachers and 6.0 for control group school teachers, out of 8 areas about which they were queried).[15] Finally, on a 4-point scale measuring the helpfulness of the professional development that they received, SFA and control group teachers registered similar average scores of about 2.75, an indication that they found the professional development somewhat helpful.[16]

Despite having similar overall ratings of the professional development received, teachers at SFA and control group schools differed in their ratings of professional development on particular topics. SFA teachers were more likely than control group teachers to rate as helpful the professional development they received on cooperative learning techniques and on how to implement their reading program properly. Control group teachers were more likely to see as helpful the professional development they got on how to teach students at different reading levels or in different reading groups — perhaps because the reading classes they taught were generally far more heterogeneous with regard to reading ability than those taught by their SFA counterparts. For further details on teachers' ratings of the helpfulness of professional development on specific topics, see Appendix Table C.1.

In promoting continuous improvement, the SFA program emphasizes the importance of collecting and analyzing student reading data to monitor progress. The "Use of Data" section in Table 4.5 suggests that control group schools gave similar attention to this practice. More than

[15] The surveys provide information about whether professional development about a particular topic was received, but not about the time that teachers spent in professional development.

[16] A score of 2.5 on this scale represents a neutral midpoint.

Table 4.5

SFA-Control Group Comparisons
Related to Continuous Improvement (Implementation Year 2013-2014)

	Program	Control	Estimated	P-
	Group	Group	Difference	Value
Professional Development				
Average number of domains in which professional development was received (out of 8)	6.4	6.0	0.3	0.373
Average rating of the helpfulness of professional development received (out of 4)	2.8	2.7	0.0	0.696
Percentage of teachers who, since the start of the 2013-2014 school year, received professional development in:				
How to implement the school’s reading program properly	80.6	74.7	5.9	0.221
New techniques for reading instruction	82.1	78.2	3.9	0.392
How to teach students at different reading levels or in different reading groups	77.9	77.0	0.9	0.858
How to develop strategies to better meet the needs of struggling students	80.2	78.1	2.1	0.650
How to use classroom materials, including technology, to improve reading instruction	77.2	73.0	4.2	0.400
How to better use the time allocated to reading instruction	79.7	72.2	7.5	0.128
How to implement cooperative learning techniques	79.5	73.5	6.0	0.242
How to use reading assessment data to guide instruction	80.5	79.4	1.1	0.809
Use of Data				
Percentage of teachers who agree that their school uses data to measure the reading progress of students over time	96.6	94.0	2.6	0.334
Percentage of teachers who agree that their school uses data to identify students struggling with reading	92.7	96.4	-3.7	0.305
Percentage of principals who report that student test scores are examined both individually and by grade level	78.6	78.6	0.0	1.000

(continued)

Table 4.5 (continued)

SOURCES: Spring 2014 teacher and principal surveys and classroom observations.

NOTES: Items on the teacher and principal surveys 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 or principals who agree with an item were obtained by taking the number who responded 3 or 4 and dividing by the total number of respondents to that item.

The means reported for teacher survey items are means of school means. First, means are taken within each school at the teacher level, then the mean across school means is taken. This was done to prevent overweighting schools with more teachers.

A two-tailed t-test was applied to differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

Completed surveys were received from 14 out of 19 principals at SFA schools and 14 out of 18 principals at control group schools. Completed teacher surveys were received from 15 out of 19 SFA schools and 16 out of 18 control group schools. Completed surveys were received from 297 teachers at SFA schools and 233 teachers at control group schools.

Response rates for all but one teacher survey items presented were above 98 percent for both SFA and control group teachers. The response rate on the item about the helpfulness of professional development was 74 percent for control group teachers and 80 percent for program group teachers. The response rate is lower on this item because only teachers who had received professional development in at least one area were counted.

90 percent of teachers in both groups of schools reported that their school used data to monitor the reading progress of students over time as well as to identify students who are struggling.

Teachers' Perceptions of and Attitudes Toward Their Reading Programs

Table 4.6 presents data from teacher surveys about teachers' perceptions of their reading programs. In comparison with the views of teachers in the control group schools, SFA teachers' views were quite mixed and not fully consistent.

Similar proportions of teachers in both groups of schools (63 percent of SFA teachers and 68 percent of control group school teachers) expressed general satisfaction with the overall quality of their reading programs. SFA teachers, however, were far more likely than control group school teachers to report that the program was too rigid and scripted (46 percent compared with 6 percent), and less likely to report that they changed parts of it that, in their opinion, did not work for students (54 percent compared with 98 percent).[17]

Fewer than half the teachers in both groups (45 percent of SFA teachers and 36 percent of control group school teachers, a difference that is not statistically significant) believed that their school's reading program adequately served the most struggling reading students. But SFA teachers were far less likely to believe that the program adequately served English language learners, special education students, and students with behavioral challenges. They were also significantly less likely to agree that tutoring at their school helped their students become better readers (perhaps because they were aware that SFA's tutoring component was not fully implemented).

Teachers' Perceptions of Student Engagement

In the logic model that guides the evaluation, student engagement is seen as a construct that mediates between receipt of the program and reading achievement and other outcomes. The evaluation did not seek to measure student engagement directly. Instead, the teacher survey included two items that asked teachers to assess the extent to which their students were engaged during their reading classes.

Table 4.7 shows the results. Although a substantial majority of teachers in both groups agreed that their students were engaged, the difference — 78 percent of teachers in the SFA

[17] The percentage of SFA teachers who reported that they changed parts of the program was higher in the third year than in either of the previous years (26 percent in Year 1 and 45 percent in Year 2).

Table 4.6

SFA-Control Group Comparisons on Teacher Perceptions of the Reading Program and Tutoring (Implementation Year 2013-2014)

	Program Group	Control Group	Estimated Difference	P-Value	
General perception of the reading program					
Percentage of teachers who are satisfied with the overall quality of the reading program at their school	63.1	67.9	-4.8	0.658	
Percentage of teachers who agree that their reading program is too rigid or scripted	46.4	5.6	40.8	0.000	***
Percentage of teachers who think their reading program is too time consuming or work intensive	34.3	28.8	5.5	0.456	
Percentage of teachers who agree that they change the parts of the reading program that don't work for their students	54.0	97.5	-43.5	0.000	***
Percentage of teachers who agree that teacher morale has been high since the start of the school year	47.2	73.7	-26.5	0.054	*
Tutoring and other services for students with special challenges					
Percentage of teachers who agree that tutoring practices at their school help students in their reading class become better readers	64.9	89.1	-24.2	0.001	***
Percentage of teachers who agree that their reading program adequately serves English language learners	50.1	80.4	-30.3	0.001	***
Percentage of teachers who agree that their reading program adequately serves special education students	55.3	73.3	-18.0	0.039	**
Percentage of teachers who agree that their reading program adequately serves students with behavioral challenges	44.3	57.3	-13.0	0.067	*
Percentage of teachers who agree that their reading program adequately serves students who struggle the most with reading	45.0	35.7	9.3	0.341	

(continued)

Table 4.6 (continued)

SOURCE: Spring 2014 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 and dividing by the total number of respondents to that item.

The means reported for teacher survey items are means of school means. First, means are taken within each school at the teacher level, then the mean across school means is taken. This was done to prevent overweighting schools with more teachers.

A two-tailed t-test was applied to differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

Completed surveys were received from 15 out of 19 SFA schools and 16 out of 18 control group schools. Completed surveys were received from 297 teachers at SFA schools and 233 teachers at control group schools.

Response rates for all teacher survey items presented were above 97 percent for both SFA and control group teachers, with one exception: 84 percent of control group teachers and 77 percent of SFA teachers responded to the item asking the extent to which they agreed that tutoring practices helped students in their reading class become better readers.

Table 4.7

SFA-Control Group Comparisons on Teacher Perceptions of Engagement and Behavior (Implementation Year 2013-2014)

	Program Group	Control Group	Estimated Difference	P-Value
Percentage of teachers who agree that their students are engaged during their reading class	77.6	95.7	-18.1	0.006 ***
Percentage of teachers who agree that the reading program gets students excited about reading or learning how to read	46.1	70.0	-23.9	0.011 **
Percentage of teachers who agree that their students are well-behaved during their reading class	77.3	84.9	-7.6	0.180

SOURCE: Spring 2014 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 and dividing by the total number of respondents to that item.

The means reported for teacher survey items are means of school means. First, means are taken within each school at the teacher level, then the mean across school means is taken. This was done to prevent overweighting schools with more teachers.

A two-tailed t-test was applied to differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

Completed surveys were received from 15 out of 19 SFA schools and 16 out of 18 control group schools. Completed surveys were received from 297 teachers at SFA schools and 233 teachers at control group schools.

Response rates for all teacher survey items presented were above 97 percent for both SFA and control group teachers.

group compared with 96 percent in the control group — is statistically significant, as is the difference in the percentage of teachers reporting that the reading program gets students excited about reading or learning to read (46 percent of SFA teachers compared with 70 percent of control group teachers). The differences in teachers' perceptions of student engagement may reflect the fact that SFA teachers were more negative in general about their reading program than were control group teachers, or it may be that SFA teachers were more negative about the program in part *because* their students really were less engaged.

Any differences in engagement did not result in serious behavior problems. The large majority of teachers in both groups said that their students were well-behaved during reading classes.

* * *

Figure 4.4 reproduces the "Program Elements" column of the SFA logic model. Elements with an asterisk are ones for which significant differences were found between SFA and control group schools, whereas elements without an asterisk indicate that either no or few significant differences were found. Elements in italics could not be evaluated due to insufficient measures.

As the figure shows, SFA incorporates a number of program elements that make it more than a reading program. Yet the figure suggests that it is the instructional core that most distinguishes SFA from control group schools. Both SFA and control group schools had individuals or groups that dealt with a wide range of noninstructional issues, such as improving attendance, building parent and community relations, and resolving problems with student behavior. Both sets of schools were focused on continuous improvement. Both collected extensive data on student reading and used it to guide instruction. By the third year of the study, teachers in control group schools were, like their SFA counterparts, receiving professional development on many areas of reading instruction and finding it generally helpful.

Differences in the teaching of reading had largely to do with instructional strategies. SFA schools made extensive use of cross-grade ability grouping; control group schools did not. Students in SFA classrooms were more likely to work together in small groups in ways consonant with SFA's emphasis on cooperative learning. SFA teachers used educational media more than teachers in control group schools. With respect to the content of what was taught, data from instructional logs suggest that SFA teachers, in upper-level early reading classes, were more likely to focus on comprehension. Teachers in control group schools were more likely to focus on grammar, spelling, and writing during reading instruction.

One important element of SFA reading instruction — SFA's own tutoring program for struggling readers — was not fully implemented, whereas both program and control group

Figure 4.4

Program Elements of the Logic Model, with Contrasts Between SFA and Control Group Schools

Structures and processes to support:

Challenging reading instruction that responds to students' individual needs

- 90-minute reading block*
- Limited class size in beginning reading classes*
- Cross-grade grouping by reading level during instruction, with regrouping quarterly*
- Cooperative learning*
- Other cognitively demanding classroom instruction processes
- Celebration of small-group, classroom, and school-wide learning gains*
- *Rapid pacing*
- Tutoring and other interventions for struggling students
- Use of engaging media*
- *Frequent assessments of student learning*

Components that address noninstructional issues that affect learning

- Solutions Teams of faculty and staff to address academics, attendance, behavior, and parent/community involvement
- *Social-emotional regulation and conflict resolution strategies for use in classrooms and throughout the school*

Emphasis on continuous improvement

- Professional development and coaching by school and SFAF staff
- Use of data to measure progress and set goals

NOTES: Elements with an asterisk are ones for which significant differences were found between SFA and control group schools. Either no or few significant differences were found in elements without an asterisk. Elements in italicized text could not be evaluated due to insufficient measures.

schools offered small-group and individual assistance in a Response to Intervention framework that calls for increasingly intensive supports for struggling students. The result is that tutoring looked quite similar in the two sets of schools. In particular, SFA and control group schools provided tutoring for similar proportions of students who were in second grade, the grade level that is the focus of the impact analysis in the next chapter.

Chapter 5

Impacts of the Success for All Program

Previous chapters in the report discuss the fidelity with which the Success for All (SFA) program was implemented and the resulting differences in reading instruction and in whole-school learning environments and supporting structures between the program group schools ("SFA schools") and the control group schools in the study. According to the logic model of the SFA program (Chapter 1, Figure 1.1), these observed differences are expected to lead to improved student performances in reading, as well as other related academic and behavioral outcomes. This chapter examines these hypotheses using outcome information collected at the end of the third implementation year.

The impact analyses focus on a *primary analysis sample* of students who started their kindergarten in these schools and have been in these study schools for the past three years. The program group in this sample of students had the maximum amount of exposure to the SFA program. All subgroups explored in this chapter are defined based on this sample. The chapter also explores the program impact on outcomes for all students who were present in the study schools in the spring of the third implementation year, regardless of how long they had been there (hereafter referred to as the *spring analysis sample*). Impacts for an *auxiliary analysis sample*, which consists of students in grades 3 through 5 in the study schools in the spring of 2014, are examined to assess the effects of the program on students in upper grades in elementary schools. In addition, the chapter examines program impacts on special education identification and grade retention rates at the school level.

Key Findings

- The program produced positive impacts for the average SFA school for one of the four reading outcomes, the Woodcock-Johnson Word Attack score, which measures students' phonics and decoding skills. This impact was registered for the sample of students who were with the study from kindergarten through second grade and therefore had the maximum possible amount of exposure to the SFA program. The estimated program impact is 0.15 standard deviation in effect size.

- The estimated program impacts on the average performances in word identification, reading fluency, and reading comprehension are not statistically significant.

- There is fairly consistent evidence suggesting that, for a subgroup of students who started kindergarten with letter-word identification skills below the median level of the primary analysis sample, the SFA program significantly improved performance in three of the four reading outcomes above and beyond the performance of their counterparts in the control group schools.

- The estimated program impacts for a range of demographic and socioeconomic subgroups are consistent with the whole-sample findings: The estimated impacts on Word Attack score are positive and statistically significant for some subgroups, but only sporadic significant impacts were detected for the other three outcomes.

- For the spring analysis sample, which includes all second-grade students present in the study schools with at least one valid spring test score from the 2013-2014 school year, there was a positive estimated impact on Word Attack but not on any of the other three outcomes.

- For schools' average reading performances in third, fourth, and fifth grades (the auxiliary sample) in the spring of 2014, the program did not produce any statistically significant impacts on tests of vocabulary and reading comprehension administered specifically for the study. There were also no statistically significant effects on state reading tests for third and fifth grades in the sample. There was a negative and significant impact on the state reading test for the average fourth grade in the average SFA school in the sample, which is no longer significant when adjustment is made to account for multiple hypothesis testing.

- By and large, there were no consistent patterns or significant findings of program impacts on special education identification rates and retention rates at the school level.

This chapter begins with a brief overview of the analytic approach used for the impact analysis and outcome measures collected for this final report of the study. It then presents the estimated cumulative program effects as well as effects for the subgroups and samples introduced above. It concludes by placing the primary analysis sample findings in the context of findings from previous years and the existing literature on the SFA reading program.

Analytic Approach and Related Issues

As briefly discussed in Chapter 1, this evaluation is a blocked school-level random assignment study involving 37 elementary schools. Within each of the five participating districts, or "blocks," schools in the evaluation were randomly assigned either to implement the SFA program or to carry on with "business as usual." Given this design, the impacts of the SFA program are the differences in outcomes between schools that were randomly assigned to the program group and those assigned to the control group.[1] Box 5.1 explains how to read the impact tables.

Box 5.1

How to Read the Impact Tables

The impact tables in this chapter provide the estimated effect size and p-value for each impact estimate. The *effect size* indicates the magnitude of the estimated effect, calculated as a proportion of the standard deviation of the outcome measure for the control group. The *p-value* is a measure of statistical significance. It indicates the likelihood that the estimated impact was obtained by chance, in the absence of a true effect. For example, if the p-value is 0.050, it indicates that there is a 5 percent chance that there was no true effect. This report uses asterisks to indicate findings that are statistically significant at the 10 percent (*), 5 percent (**), or 1 percent (***) level. Results that are not statistically significant do not provide strong evidence about the impact of the program.

In addition, to provide context for interpreting the estimated impacts, the impact tables show regression-adjusted mean outcomes (such as mean test scores) for the program and control groups. The regression adjustment uses a model to account for the design of the study and uses student baseline characteristics in the model to account for variability in the outcomes.*

*Black et al. (2008); Garet et al. (2010).

Two types of analytic models are used for analyzing outcomes measured at different levels. For outcomes measured at the student level, such as various reading test scores, the main impact estimation model is a two-level hierarchical linear model that accounts for the clustering of students within schools. For outcomes measured at the school level, such as schools' special education identification rate and retention rate, the estimation model is an ordinary least squares (OLS) regression model with the school as the unit of observation. Both types of models use 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 are averaged

[1] Note that all impact estimates reported in this chapter are based on an *intent-to-treat* analysis that includes all students who have valid outcome measures and who were in the sample schools at baseline. In other words, the impact estimates reflect the impact of assignment to the SFA program. Appendix D discusses the impact estimation model in detail.

across the five districts, with each district's estimate weighted in proportion to the number of SFA schools it has. Therefore, findings in this report represent the impact on student performance in the average SFA school within the five study districts. The results may not necessarily reflect what the program effect would be in the wider population of districts beyond the study sample.

Various developmentally appropriate outcome measures were used to assess students' different reading skills, ranging from alphabetics to reading comprehension. Table 5.1 summarizes the student-level outcomes used for the impact analysis for both the primary student sample and the auxiliary student sample.

In addition, school-level (by grade) special education identification rates and grade retention rates are used as outcomes to measure the impact of SFA on how schools address students' special needs and how students progress in school. Appendix D provides detailed descriptions of the data sources, construction procedure, and analytic issues related to these measures.

Impact Findings for the Primary Analysis Sample

As described earlier in the chapter, the primary analysis sample consists of students who started their kindergarten year in the study schools in the 2011-2012 school year and were with the schools through all three implementation years. In the program group, these are the students with the maximum possible exposure to the SFA program.

For the full sample of second graders in the average SFA school who had entered SFA schools as kindergartners and had been in SFA classrooms for three years, SFA produced a positive and statistically significant impact on one of two phonetic skill measures.

Table 5.2 shows the estimated program impacts on these measures for the full primary analysis sample. The four outcomes measure three distinct reading skills: phonics and decoding (Woodcock-Johnson Letter-Word Identification and Word Attack), reading fluency (TOWRE2), and reading comprehension (Woodcock-Johnson Passage Comprehension). Among the two phonics measures, the estimated impact on Woodcock-Johnson Word Attack score is 1.03 raw score points, or 0.15 standard deviation in effect size, with a p-value of 0.022.[2]

[2] Given that statistical tests were conducted on each of the two phonics skill measures (Word Attack and Letter-Word Identification), there is an increased probability that an impact that is not significant in reality will be identified as significant by chance. To address this, following the What Works Clearinghouse guidelines for multiple hypothesis testing (Version 3.0, What Works Clearinghouse, 2014), the Benjamini-Hochberg procedure was applied to the individual outcome findings reported in Table 5.2 by outcome domains. With this adjustment, the impact on the Word Attack score remains statistically significant at the 5 percent level.

Table 5.1

Description of Student Reading Achievement Outcome Measures

Outcome	Reading Skills Measured	Description	Reliability	Version
Primary analysis sample[a]				
Woodcock-Johnson Letter-Word Identification	Word identification skills	The student is asked to identify letters that appear in large type, and is then asked to pronounce words correctly. Items become increasingly difficult as the selected words appear less and less frequently in written English.	0.97 to 0.99 for ages 5-7	Woodcock-Johnson III Normative Update, Form B
Woodcock-Johnson Word Attack	Applying phonic and structural analysis skills to the pronunciation of unfamiliar words	The student is asked to produce the sounds for individual letters, then read aloud letter combinations that are regular patterns in English but are nonwords or low-frequency words.	0.92 to 0.94 for ages 5-7	Woodcock-Johnson III Normative Update, Form B
Test of Word Reading Efficiency	Efficiency of sight word recognition and phonemic decoding	Assessment is based on the number of real words the student can identify within 45 seconds, as well as the number of pronounceable nonwords the student can accurately decode within 45 seconds.	Average above 0.90	TOWRE-2
Woodcock-Johnson Passage Comprehension	Symbolic learning, comprehension	The student is asked to match pictographic representations of words with actual pictures of the object, choose pictures represented by a phrase, and read several short passages and identify missing key words.	0.96 for ages 5-7	Woodcock-Johnson III Normative Update, Form B

(continued)

Table 5.1 (continued)

Outcome	Reading Skills Measured	Description	Reliability	Version
Auxiliary analysis sample				
Gates-MacGinitie total score	General reading achievement	The exam consists of a vocabulary and a comprehension subtest, described below.	0.96 for grades 3-4[b]	Form S
Gates-MacGinitie Vocabulary test	Reading vocabulary	Each test word is presented in a brief context intended to suggest part of speech but not to provide clues to meaning. Students are expected to select the word or phrase that means most nearly the same as the test word.	Subtest reliability not available	Form S
Gates-MacGinitie Comprehension test	Ability to read and understand different types of prose	Content is selected from published materials and reflects the type of materials that students are required to read for their schoolwork and choose to read for recreation. Students are required to construct an understanding based on a literal understanding of a passage, or to make inferences or draw conclusions. The comprehension tests also measure the ability to determine the meaning of words in an authentic text context.	Subtest reliability not available	Form S
State reading tests	General reading achievement	Standardized state reading exams used for federal reporting of student reading proficiency. State test proficiency rates were calculated from student records provided by the study districts.	0.84 to 0.92 for grades 3-5	State exams administered in the spring of each year

(continued)

Table 5.1 (continued)

SOURCES: Mather and Woodcock (2001); McGrew, Schrank, and Woodcock (2007); Riverside Publishing (2011); PRO-ED (2012); Center on Response to Intervention (2015); and the technical reports for 2012-2013 state reading exams.

NOTES: All tests administered to students in the primary analysis sample were administered during one-on-one sessions with students. All tests administered to students in the auxiliary analysis sample were administered during group sessions with students.

[a] In addition to the English version of the tests in this table, a subgroup of Spanish-speaking students in the primary analysis sample were given the Spanish version of the Woodcock-Johnson Letter-Word Identification, Word Attack, and Passage Comprehension tests.

[b] Gates-MacGinitie reliability estimates were obtained from a technical summary of the exam available at the American Institutes for Research Center on Response to Intervention website.

Table 5.2

Impact of SFA on Average Second-Grade Reading Achievement for the Primary Analysis Sample (Implementation Year 2013-2014)

Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value
Woodcock-Johnson Letter-Word Identification	41.19	40.56	0.63	0.07	0.243
Woodcock-Johnson Word Attack	16.41	15.39	1.03	0.15	0.022 **
Test of Word Reading Efficiency	49.45	48.34	1.11	0.07	0.268
Woodcock-Johnson Passage Comprehension	21.58	21.44	0.15	0.03	0.603
Number of schools: 37	19	18			

SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2014), Woodcock-Johnson Word Attack test (Spring 2014), Woodcock-Johnson Passage Comprehension test (Spring 2014), Test of Word Reading Efficiency (Spring 2014), and student records data collected from the five districts in the study sample.

NOTES: The "primary analysis sample" consists of students from 37 schools (19 program group schools and 18 control group schools) and includes any student who had at least one valid spring test score in each of the three implementation years and who had valid scores on the fall baseline 2011 Peabody Picture Vocabulary Test and fall baseline 2011 Woodcock-Johnson Letter-Word Identification test.

The student sample size for the Woodcock-Johnson Letter-Word Identification test is 1,631 students (851 in the program group and 780 in the control group). The student sample size for the Woodcock-Johnson Word Attack test is 1,635 students (854 in the program group and 781 in the control group). The student sample size for the Test of Word Reading Efficiency is 1,625 students (847 in the program group and 778 in the control group). The student sample size for the Woodcock-Johnson Passage Comprehension test is 1,625 students (848 in the program group and 777 in the control group).

Students were tested using both Form A and Form B of the Test of Word Reading Efficiency. The scores reported above represent the average.

The impact analyses for reading achievement were conducted using raw scores. 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.

Effect sizes were computed using the full control group's standard deviations for the respective measures. The control group standard deviations are as follows: 8.81 for the Woodcock-Johnson Letter-Word Identification Test, 6.81 for the Woodcock-Johnson Word Attack test, 15.82 for the Test of Word Reading Efficiency, and 4.83 for the Woodcock-Johnson Passage Comprehension test.

A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

To put this finding into context, calculations based on national norming samples for seven major standardized tests show that, during the second-grade year, an average student's reading achievement test score grows 0.97 standard deviation in effect size.[3] This indicates that the impact on Word Attack score experienced by the students at an average SFA school in the study represents about 16 percent of the annual growth for an average second-grade student, or about one and a half months of learning.

More broadly, this finding is comparable to the impacts of other similar school reform programs. For example, Borman and colleagues used a meta-analysis to show that the overall effect across 29 of the most widely deployed comprehensive school reforms ranged between 0.09 standard deviation and 0.15 standard deviation in effect size.[4] Similarly, a synthesis of observational studies on the effectiveness of the federal Title I program put its effect at around 0.11 standard deviation.[5] Furthermore, the Tennessee Student-Teacher Ratio (STAR) study found that reducing early-grade classes from their standard size of 22 to 26 students to a size of 13 to 17 students significantly increased average reading performance in elementary schools by 0.11 to 0.22 standard deviation in effect size.[6]

- **The estimated impacts on the other phonetic skill measure and on two measures of more advanced reading skills — fluency and comprehension — were not statistically significant.**

The estimated impact on the other phonics measure — Letter-Word Identification — is positive (effect size = 0.07 standard deviation) but not statistically significant.

The improved performance on Word Attack by students in the SFA schools does not seem to translate into significantly better performances on reading fluency and comprehension than that of students in the control group schools. The estimated effects for these two outcomes are 0.07 standard deviation and 0.03 standard deviation, respectively, with p-values well above the 10 percent level (0.268 and 0.603, respectively).

However, the average SFA impacts on the full primary sample could be masking important heterogeneous effects on different types of students. Exploratory analyses on various student subgroups yield some interesting findings.

- **For a subgroup of students who started kindergarten with letter-word identification skills below the median level of the primary sample, the**

[3] Hill, Bloom, Black, and Lipsey (2007). Note that the annual gain for reading is calculated from seven nationally normed comprehensive reading tests: CAT5, SAT9, TerraNova-CTBS, MAT8, TerraNova-CAT, SAT10, and Gates-MacGinitie. These tests focus on multiple rather than single reading skills.

[4] Borman, Hewes, Overman, and Brown (2003).

[5] Borman and D'Agostino (1996).

[6] Nye, Hedges, and Konstantopoulos (1999).

SFA program had significantly improved their performances in three of the four reading outcomes above and beyond the performance of their control group counterparts by the end of the third year.

It is possible that the impacts of the SFA program vary by students' initial level of achievement. For example, students at different skill levels might have had different needs, which were differentially emphasized in the SFA program. To assess this possibility, students in the primary sample are divided into two subgroups: Those whose baseline test scores are below the sample median are in the "lower performing" group, and those with baseline test scores at or above the sample median are in the "higher performing" group.[7] Because two tests for different reading skills — Woodcock-Johnson Letter-Word Identification (WJLWI) for phonetics and decoding skills and Peabody Picture Vocabulary Test (PPVT) for vocabulary — were used to measure students' baseline performance levels, two sets of high and low performance subgroups based on each of these tests were defined and examined.

The WJLWI test measures a student's word identification and reading decoding skills, or put differently, it tests a student's cognitive ability to recognize letters and visual word forms, as well as to associate pronunciation with the words.[8] The PPVT, on the other hand, assesses a student's receptive vocabulary for the English language.[9] The correlation coefficient between the WJLWI and PPVT scores in the primary analysis sample is 0.457, suggesting that the skills assessed by these two tests, and thus a student's scores on them, do differ. As a result, the compositions of the two sets of subgroups differ.

Figure 5.1 uses the sample composition and overlap between the lower-performing subgroup as defined by WJLWI and the lower-performing subgroup as defined by PPVT as an example to illustrate this point. This figure shows that there are 759 students identified as lower performing by their baseline WJLWI test scores and 780 students identified as lower performing by the baseline PPVT scores. Among these students, 513 students (in Area A) are classified as lower performing no matter which baseline test is used for the classification. These students constitute about two-thirds of both lower-performing subgroups. The remaining one-third of the students in each of these two lower-performing subgroups (in Areas B and C) are different from

[7] The sample median was chosen to be the cut point because it provides a split of the full sample into two subgroups of almost equal size and therefore ensures the maximum possible statistical power for the impact estimation for both subgroups. Sensitivity checks were conducted to see whether the findings changed when grouping students with the median scores with the low-performing group. The results are robust to that change.

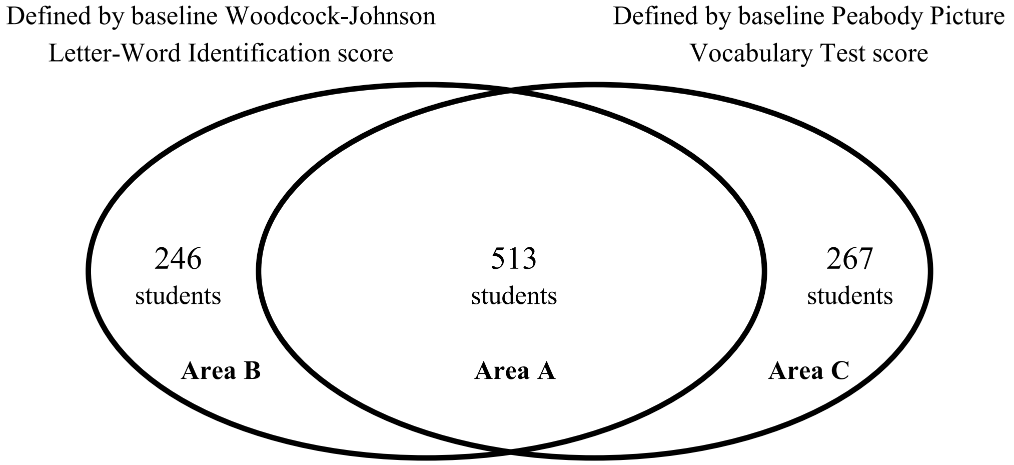
[8] Wendling, Schrank, and Schmitt (2007).

[9] Dunn and Dunn (2007).

Figure 5.1

Composition and Relationship Between Lower-Performing Subgroups Defined by Baseline WJLWI and PPVT Scores (Implementation Year 2013-2014)

Lower-Performing Subgroup Student Counts


Defined by baseline Woodcock-Johnson Letter-Word Identification score
Defined by baseline Peabody Picture Vocabulary Test score
246 students
Area B
513 students
Area A
267 students
Area C

SOURCES: Baseline Peabody Picture Vocabulary Test (PPVT) and Woodcock-Johnson Letter-Word Identification (WJLWI) test, administered to the baseline student sample in the fall of 2011, as well as the Woodcock-Johnson Letter-Word Identification test (Spring 2014), the Woodcock-Johnson Word Attack test (Spring 2014), the Test of Word Reading Efficiency (Spring 2014), the Woodcock-Johnson Passage Comprehension test (Spring 2014), and student records data collected from the five districts in the study sample.

NOTES: All student counts are based on the primary analysis sample, which consists of students from 37 schools (19 program group schools and 18 control group schools) and includes any student who had at least one valid spring test score in each of the three implementation years and who had valid scores on the fall baseline 2011 PPVT and fall baseline 2011 WJLWI test.

each other, and the difference between them determines the difference in the findings observed for the lower-performing subgroups defined by the WJLWI test and the PPVT.[10]

Table 5.3 reports the impact estimates by outcome for these two sets of subgroups separately. For each outcome, the table presents separate impact estimates for the lower- and higher-

[10] Students in Area B, who have higher baseline PPVT scores and lower baseline WJLWI scores, are less likely to be black or Hispanic and less likely to have poverty, special education, or English language learner status than students in Area C.

Table 5.3

Impact of SFA on Average Second-Grade Reading Achievement for the Primary Analysis Sample (Implementation Year 2013-2014), by Subgroup Defined by Baseline Reading Level

Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value	P-Value for Difference in Estimated Impact
Subgroup defined by baseline WJLWI scores						
Woodcock-Johnson Letter-Word Identification						0.032 **
Lower-performing students	38.40	36.86	1.54	0.17	0.074 *	
Higher-performing students	44.05	44.20	-0.16	-0.02	0.793	
Woodcock-Johnson Word Attack						0.090 *
Lower-performing students	14.65	13.06	1.58	0.23	0.014 **	
Higher-performing students	18.23	17.74	0.48	0.07	0.348	
Test of Word Reading Efficiency						0.030 **
Lower-performing students	44.65	41.69	2.96	0.19	0.099 *	
Higher-performing students	54.16	54.63	-0.46	-0.03	0.600	
Woodcock-Johnson Passage Comprehension						0.015 **
Lower-performing students	19.92	19.25	0.67	0.14	0.147	
Higher-performing students	23.24	23.56	-0.31	-0.06	0.254	

(continued)

Table 5.3 (continued)

Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value	P-Value for Difference in Estimated Impact
<u>Subgroup defined by baseline PPVT scores</u>						
Woodcock-Johnson Letter-Word Identification						0.929
Lower-performing students	38.21	37.73	0.48	0.05	0.579	
Higher-performing students	43.96	43.59	0.37	0.04	0.515	
Woodcock-Johnson Word Attack						0.986
Lower-performing students	14.49	13.74	0.76	0.11	0.228	
Higher-performing students	18.13	17.42	0.70	0.10	0.156	
Test of Word Reading Efficiency						0.983
Lower-performing students	45.41	44.77	0.63	0.04	0.698	
Higher-performing students	53.10	52.63	0.46	0.03	0.613	
Woodcock-Johnson Passage Comprehension						0.776
Lower-performing students	19.67	19.63	0.04	0.01	0.934	
Higher-performing students	23.40	23.26	0.14	0.03	0.654	

(continued)

Table 5.3 (continued)

SOURCES: Baseline Peabody Picture Vocabulary Test (PPVT) and the Woodcock-Johnson Letter-Word Identification (WJLWI) test, administered to the baseline student sample in fall of the 2011-2012 school year, as well as the Woodcock-Johnson Letter-Word Identification test (Spring 2014), Woodcock-Johnson Word Attack test (Spring 2014), the Test of Word Reading Efficiency (Spring 2014), the Woodcock-Johnson Passage Comprehension test (Spring 2014), and student records data collected from the five districts in the study sample.

NOTES: The baseline WJLWI subgroup sample size ranges from 853 to 855 students in the higher-performing group and from 770 to 780 students in the lower-performing group.

The baseline PPVT subgroup sample size ranges from 873 to 876 students in the higher-performing group and from 749 to 759 students in the lower-performing group.

The impact analyses for reading achievement were conducted using raw scores. 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 differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

performing subgroups, and then tests whether the difference in estimated impacts between these two subgroups is statistically significant. The first panel in Table 5.3 presents findings for the lower- and higher-performing subgroups defined by students' baseline WJLWI scores. Specific findings include the following:

- The impact estimates for the lower-performing subgroup are positive, ranging from 0.14 to 0.23 standard deviation in effect size, and are statistically significant at the 10 percent level for all but one of the four outcomes. The one exception comes from the Passage Comprehension test, which registered an estimated impact of 0.14 standard deviation in effect size with a p-value of 0.147.

- Consistently across all four outcomes, the difference in estimated impacts between these two subgroups is statistically significant, with the estimates for the lower-performing subgroup always higher than those for the higher-performing subgroup. This strongly suggests that students whose baseline performance on the WJLWI was below the median of the sample benefited more from the SFA program than did their counterparts who performed better on the test at baseline.

In addition, analysis using regression models that assume a linear relationship between students' baseline WJLWI scores and outcome scores detects consistent significant negative associations between students' baseline reading level and the program effect for all four outcomes. In other words, the lower the baseline WJLWI score, the larger the effect, confirming the subgroup findings reported above. Appendix D provides detailed results and graphical illustrations of this additional analysis.

However, this pattern is not observed when the subgroups are defined by students' baseline PPVT scores. The second panel in Table 5.3 reports results for this set of subgroups. In this case, there is no evidence of a significant difference in estimated impacts between the two subgroups on any of the four outcomes. In fact, none of the estimated impacts for either subgroup is significantly different from zero. Regression analysis similar to that described above confirms these findings. (See Appendix D for details.)

In addition to the subgroups defined by students' baseline reading performances, other student subgroups defined by various student characteristics are also examined.

- **The estimated program impacts for a range of demographic and socioeconomic subgroups are consistent with the whole sample findings.**

Table 5.4 summarizes the findings for different subgroups of students within the primary sample on the four reading outcomes. In this table, a plus sign indicates statistically signifi-

Table 5.4

Summary of the Impact of SFA on Average Second-Grade Reading Achievement for Subgroups of the Primary Analysis Sample (Implementation Year 2013-2014)

Subgroup	Woodcock-Johnson Letter-Word Identification	Woodcock-Johnson Word Attack	Test of Word Reading Efficiency	Woodcock-Johnson Passage Comprehension
Black	0	0	0	0
White	0	0	0	0
Hispanic	0	0	0	0
Female	0	+	0	0
Male	0	+	0	0
Special education	0	0	0	0
Non-special education	0	+	0	0
English language learner	0	0	0	0
Non-English language learner	+	+	0	0
Poverty	0	+	0	0
Non-poverty	0	0	0	0

SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2014), Woodcock-Johnson Word Attack test (Spring 2014), Woodcock-Johnson Passage Comprehension test (Spring 2014), Test of Word Reading Efficiency (Spring 2014), and student records data collected from the five districts in the study sample.

NOTES: In the table above, the plus sign ("+") indicates that positive and statistically significant estimated impacts were found for the program students within the subgroup. The minus sign ("-") would indicate that negative and statistically significant estimated impacts were found for the program students within the subgroup. A value of 0 indicates that no statistically significant impacts were found on the given measure for program students in the subgroup.

Program and control group sample sizes for each of the above subgroups, as well as more detailed information about subgroup effects, can be found in Appendix Table D.6. Due to small sample sizes, estimates could not be computed for race/ethnicity groups other than white, black, and Hispanic.

The estimated impacts and associated significance levels are based on a two-level model with students nested within schools, controlling for random assignment block and school- and student-level covariates.

cant and positive impact estimates; a zero indicates findings that are not statistically different from zero; and a minus sign would indicate statistically significant negative impact estimates. Positive and statistically significant impact estimates on the Woodcock-Johnson Word Attack test were observed for all large subgroups such as boys, girls, students who do not qualify for special education, students who are not English language learners, and students with poverty status.[11] Positive, statistically significant impacts were also found for non-English language learners on the Woodcock-Johnson Letter-Word Identification test. None of these subgroups registered significant impacts on the fluency and comprehension measures, which is consistent with the findings for the full primary analysis sample.[12]

Impact Findings for the Spring Analysis Sample

Throughout the three years of implementation, some students who were in the study schools at the beginning (fall 2011) left their schools, while others who originally were not in the study schools transferred into one of them. The first group of students (the "out-movers") was not tracked for outcome data collection and, therefore, is not in any of the impact analyses reported in this chapter; the second group (the "in-movers") was tested at the follow-up points. Thus, students who were tested in the spring of 2014 received varying amounts of the SFA intervention, ranging from less than one year to three years. Specifically, about 63 percent of the students in this sample were in the study sample for all three implementation years, about 18 percent were in the sample for two of the three implementation years, and the remaining 19 percent had one year of exposure to the program at most.[13] Therefore, the impact results for this sample of students who were present in the study schools at the time of spring 2014 data collection reflect the effects of SFA when taking student mobility into account.

Table 5.5 presents the impact estimates for this sample. The results are parallel to the impact findings for the primary analysis sample: The impact estimate for the Word Attack test is positive and statistically significant (effect size = 0.17), while the impacts for the other three outcomes are not significantly different from zero.

Impact Findings for Upper-Grade Students

Students in third, fourth, and fifth grades in the study sample schools are considered the "auxiliary sample." These students were in grades 1 through 3, respectively, when the study began. Therefore, they did not first learn to read "the SFA way." It would be interesting to see whether they benefited from the SFA instruction.

[11] Note that a small subset of English language learners in one district in the study sample were tested in both English and Spanish in the spring of 2014. Appendix Table D.7 reports the impact findings for them.

[12] Appendix D presents more detailed statistical information about these subgroup impact findings.

[13] Chapter 2 discusses the composition and comparison between the spring analysis sample and the primary analysis sample in more detail.

Table 5.5

Impact of SFA on Average Second-Grade Reading Achievement for the Spring Analysis Sample (Implementation Year 2013-2014)

Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value	
Woodcock-Johnson Letter-Word Identification	39.99	39.18	0.82	0.09	0.147	
Woodcock-Johnson Word Attack	15.53	14.37	1.15	0.17	0.009	**
Test of Word Reading Efficiency	46.96	46.15	0.81	0.05	0.392	
Woodcock-Johnson Passage Comprehension	21.03	20.88	0.15	0.03	0.558	
Number of schools: 37	19	18				

SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2014), Woodcock-Johnson Word Attack test (Spring 2014), the Test of Word Reading Efficiency (Spring 2014), the Woodcock-Johnson Passage Comprehension test (Spring 2014), and student records data collected from the five districts in the study sample were also used.

NOTES: The "spring analysis sample" is defined as the sample of students who had at least one valid score in the spring of 2014.

The student sample size for the Woodcock-Johnson Letter-Word Identification test is 2,902 students (1,553 in the program group and 1,349 in the control group).

The student sample size for the Woodcock-Johnson Word Attack test is 2,907 students (1,557 in the program group and 1,350 in the control group).

The student sample size for the Test of Word Reading Efficiency is 2,873 students (1,537 in the program group and 1,336 in the control group).

The student sample size for the Woodcock-Johnson Passage Comprehension test is 2,894 students (1,549 in the program group and 1,345 in the control group).

Students were tested using both Form A and Form B of the Test of Word Reading Efficiency. The scores reported above represent the average.

The impact analyses for reading achievement were conducted using raw scores. 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.

Effect sizes were computed using the full control group's standard deviations for the respective measures. The control group standard deviations are as follows: 8.81 for the Woodcock-Johnson Letter-Word Identification Test, 6.81 for the Woodcock-Johnson Word Attack test, 15.82 for the Test of Word Reading Efficiency, and 4.83 for the Woodcock-Johnson Passage Comprehension test.

A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

These students were administered grade-specific Gates-MacGinitie reading tests in vocabulary and reading comprehension. They also took the high-stakes state reading tests that are used to measure their comprehensive reading skills and establish school accountability. Table 5.6 reports the impact estimates on these outcomes by grade for students in the auxiliary sample.

For the average third- and fifth-graders in the sample, on none of these measures did SFA schools fare either better or worse than their control group counterparts. For students in grade 4, the impact estimate for the average SFA school on the state reading test is negative (effect size = -0.13) and is statistically significant at the 10 percent level. However, this is one of two comprehensive reading tests examined for fourth grade. If adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure, this estimate is no longer significant at the 10 percent level. Therefore, this finding should be interpreted with caution.

Impact Findings on School-Level Outcomes

The SFA logic model (Chapter 1, Figure 1.1) hypothesizes that the SFA program could potentially delay or reduce the incidence of special education (SPED) identification, especially identification for specific learning disabilities (SLD), the category most relevant to students' academic skills. To test this hypothesis, the study team collected information on the number of newly identified special education students for each school in the study sample and constructed new-identification rates for each school, one for all SPED categories and one for the SLD category only. By representing the proportion of students identified during a given school year, this measure is the one most directly affected by the SFA program. Additional analyses looking at schools' overall special education identification rates as well as declassification rates are reported in Appendix D.

Table 5.7 shows that the SFA program did not produce any statistically significant impact on schools' new SPED identification rates in any grade in these schools, regardless of whether the rate reflects new identifications across all SPED categories or the SLD category alone. These results should be interpreted with the following caveats in mind. First, new-identification rates are fairly low for all grade levels examined in the control group schools, hovering around 2 percent to 3 percent overall and below 2 percent for new identifications in the SLD category. The low incidence of occurrence could cause a "floor" on how much SFA would be able to reduce the rate of new identification. Second, the school-level new-identification counts include both students who were identified for these services during the given year and students who were identified for these services in another year somewhere else but who moved into the sample school in the given year. One would expect the SFA program to affect the identification of only the former type of students, but school districts could not provide school- or student-level information that would allow the study team to distinguish be-

Table 5.6

Impact of SFA on Average Reading Achievement in Grades 3-5 for the Auxiliary Analysis Sample (Implementation Year 2013-2014)

Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value
Grade 3					
Gates-MacGinitie Comprehension					
Scale score	447.10	446.01	1.09	0.03	0.699
Percentile rank	26	26			
Gates-MacGinitie Vocabulary					
Scale score	447.82	447.29	0.53	0.01	0.868
Percentile rank	29	29			
Gates-MacGinitie total					
Scale score	447.21	446.49	0.71	0.02	0.797
Percentile rank	27	27			
State reading test Z-score[a]	-0.02	-0.01	-0.02	-0.02	0.836
Grade 4					
Gates-MacGinitie Comprehension					
Scale score	468.75	470.53	-1.78	-0.05	0.466
Percentile rank	27	28			
Gates-MacGinitie Vocabulary					
Scale score	465.93	467.58	-1.65	-0.05	0.457
Percentile rank	28	29			
Gates-MacGinitie total					
Scale score	467.80	469.50	-1.70	-0.05	0.418
Percentile rank	27	28			
State reading test Z-score[a]	-0.12	0.02	-0.13	-0.13	0.053 *
Grade 5					
Gates-MacGinitie Comprehension					
Scale score	486.18	486.32	-0.14	-0.00	0.962
Percentile rank	29	29			
Gates-MacGinitie Vocabulary					
Scale score	486.89	485.12	1.78	0.05	0.426
Percentile rank	30	28			
Gates-MacGinitie total					
Scale score	486.37	485.79	0.58	0.02	0.811
Percentile rank	30	28			
State reading test Z-score[a]	-0.04	-0.02	-0.02	-0.02	0.608

(continued)

Table 5.6 (continued)

SOURCES: Gates-MacGinitie Reading Comprehension and Vocabulary subtests (Spring 2014) and student state testing records collected from the five districts in the study sample.

NOTES: The "auxiliary analysis sample" is defined as the set of students who were present in grades 3, 4, or 5 in the sample schools in the 2013-2014 school year and who have state testing scores or vocabulary or reading comprehension subtest scores from the Gates-MacGinitie Reading Test.

The sample of third-grade students ranges from 2,572 students (1,311 in the program group and 1,261 in the control group) to 2,841 students (1,450 in the program group and 1,391 in the control group).

The sample of fourth-grade students ranges from 2,604 students (1,329 in the program group and 1,275 in the control group) to 2,656 students (1,361 in the program group and 1,295 in the control group).

The sample of fifth-grade students ranges from 2,752 students (1,459 in the program group and 1,293 in the control group) to 2,789 students (1,478 in the program group and 1,311 in the control group).

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.

Effect sizes were computed using the full control group's standard deviations for the respective measures by grade level. For the Gates-MacGinitie reading comprehension subtest, the control group standard deviations are 39.27 for grade 3 students, 38.11 for grade 4 students, and 34.40 for grade 5 students. For the Gates-MacGinitie vocabulary subtest, the control group standard deviations are 41.87 for grade 3 students, 36.51 for grade 4 students, and 33.78 for grade 5 students.

A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

[a]Z-scores were computed based on control group means and standard deviations. The overall mean by grade was not exactly zero because weighted averages were used.

tween the two types of students. Therefore, the findings reported here reflect the program impacts on the new-identification rate both from truly new identifications *and* from "in-moving" students arriving at the sample schools with SPED status.[14]

It is also hypothesized that, by improving students' academic engagement and performance, the SFA program could potentially reduce the rate of retention in grade. Table 5.8 shows the results of the test for this hypothesis. The average retention rates in the sample

[14] There are other issues related to the data used for constructing this outcome measure. Appendix D provides a full account of those issues.

Table 5.7

Impact of SFA on Special Education and Specific Learning Disability New Identification Rates, by Grade (Implementation Year 2013-2014)

	All Special Education Categories				Specific Learning Disability (SLD) Category			
Outcome	Program Group (%)	Control Group (%)	Estimated Impact (%)	P-Value	Program Group (%)	Control Group (%)	Estimated Impact (%)	P-Value
New identification rate								
Kindergarten	4.94	2.55	2.39	0.146	0.18	0.40	-0.22	0.237
Grade 1	2.69	2.70	-0.01	0.990	0.80	0.32	0.47	0.139
Grade 2	3.23	2.57	0.66	0.459	1.30	1.51	-0.21	0.655
Grade 3	2.88	3.12	-0.24	0.781	1.45	1.77	-0.32	0.682
Grade 4	2.22	2.85	-0.63	0.478	1.41	1.52	-0.11	0.834
Grade 5	1.81	1.71	0.11	0.865	1.31	0.78	0.53	0.356
Number of schools: 37	19	18			19	18		

SOURCE: MDRC calculations based on special education records collected from the five districts in the study sample.

NOTES: The estimated impacts are based on an ordinary least squares (OLS) model with school-level data, controlling for random assignment block and school-level preprogram outcome measures. The program group and control group columns display regression-adjusted mean outcomes for each group. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Table 5.8

Impact of SFA on Student Retention Rate, by Grade (Implementation Year 2013-2014)

Outcome	Program Group (%)	Control Group (%)	Estimated Impact (%)	P-Value
Retention rate				
Kindergarten	1.02	1.73	-0.71	0.256
Grade 1	3.96	4.12	-0.16	0.878
Grade 2	2.42	3.23	-0.81	0.452
Grade 3	2.12	2.67	-0.56	0.541
Grade 4	0.59	0.68	-0.09	0.818
Grade 5	2.25	1.42	0.83	0.350
Number of schools: 37	19	18		

SOURCE: MDRC calculations based on student enrollment and retention records collected from the five districts in the study sample.

NOTES: The estimated impacts are based on an ordinary least squares (OLS) model with school-level data, controlling for random assignment block and school-level preprogram outcome measures. The program group and control group columns display regression-adjusted mean outcomes for each group. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

schools vary by grades, ranging from less than 1 percent to just above 4 percent. The implementation of the SFA program did not cause any discernible reduction in the retention rates across all grade levels.[15]

The Impact Findings in Context

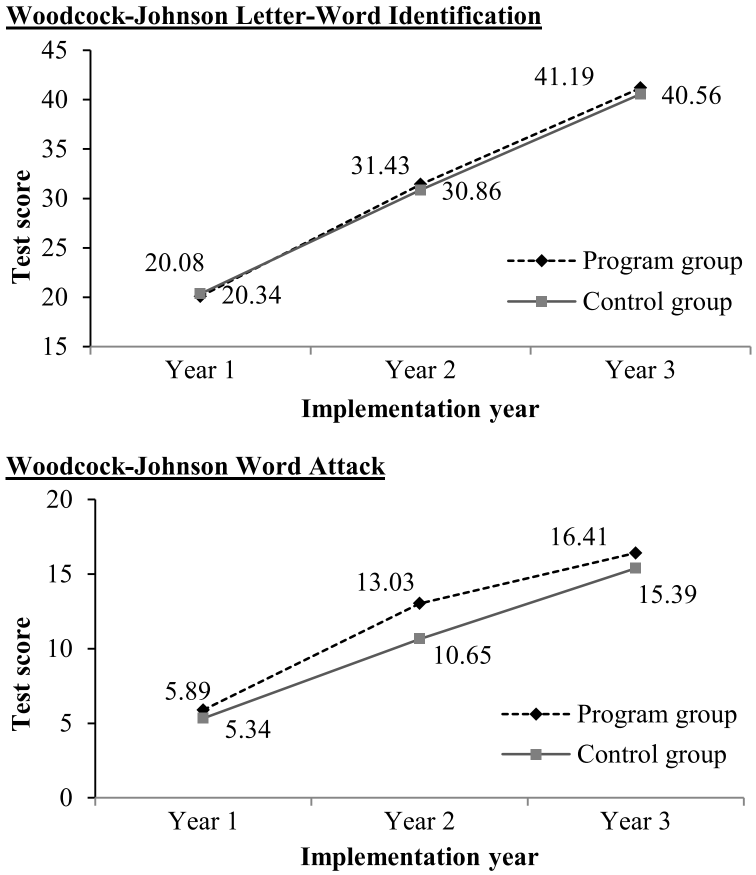
This section discusses the impact findings in the context of findings from prior years in this study as well as findings from existing literature. The discussion focuses on the findings for the primary analysis sample because the primary sample provides results that are most comparable to other studies.

Figure 5.2 presents the mean test scores for the program and control group students in the primary analysis sample over the three implementation years. The figure focuses on the

[15] Appendix D provides more information on data sources and specific issues related to this outcome.

Figure 5.2

Mean Test Scores for the Primary Analysis Sample over Time, by Program or Control Group Status



SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2012-2014), Woodcock-Johnson Word Attack test (Spring 2012-2014), and student records data collected from the five districts in the study sample.

Woodcock-Johnson Letter-Word Identification and Word Attack scores because these are the only two tests used in all implementation years. The top panel in this figure shows that the Letter-Word Identification scores for both the program and control group schools increased over time, with the control group scores closely tracking the program group scores over the years. The bottom panel of Figure 5.2 shows that the Word Attack scores for the two groups were at roughly the same point at the end of the first implementation year. They started to diverge in Year 2 with a higher growth rate for the program group than for the control group. The growth seems to have tapered off for the program group in Year 3, while the control group kept the pace in growth.

The observed patterns in the test score trends over time match the patterns of the impact findings. The top panel of Figure 5.3 presents the full primary sample impact findings on all reading outcomes for the three implementation years. The vertical bars in the figure are grouped by outcomes. The black, white, and diagonal bars within each outcome represent the estimated impacts of SFA at the end of the first, second, and third implementation years, respectively. All estimates are based on the primary analysis sample of students who were with the study schools for three years.[16] Two of the tests — Woodcock-Johnson Letter-Word Identification and Word Attack — were consistently used for all three years, while the TOWRE test and the Woodcock-Johnson Passage Comprehension test were used for the last two implementation years. This figure illustrates the following:

- For all outcomes other than Word Attack, the estimated impacts seem to increase over the years; however, none of these estimates is statistically significant. Therefore, it is not certain whether this observed pattern of growth in impacts reflects true growth or just random noise in the estimation.

- The impacts on Word Attack increased from Year 1 to Year 2 but then declined in Year 3 to about the same level as Year 1; this decline is statistically significant (p-value = 0.009). While the estimated impact is still statistically significant, this reversing pattern could suggest that the control group was gradually catching up with the program group on improving students' decoding skills by the end of the third year.

Is this pattern of findings comparable to findings from similar studies on SFA or is it unique to this evaluation? The bottom panel of Figure 5.3 shows the findings from the national randomized trial of Success for All conducted by Borman and colleagues with the program implementation in school years 2002-2003, 2003-2004, and 2004-2005.[17]

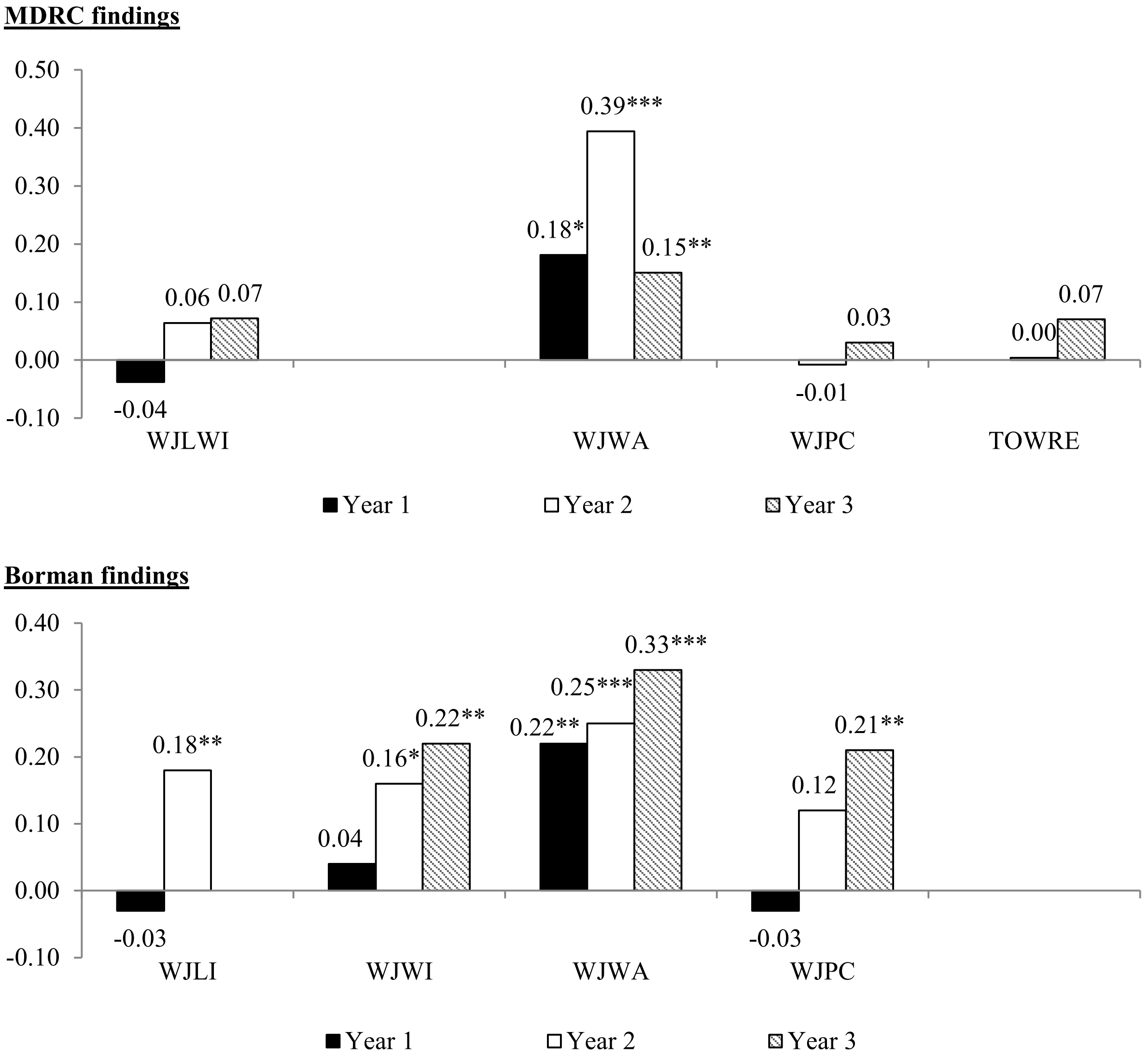
The Borman study was chosen as the focus of the comparison because it is the only other existing study of the SFA reading program that is based on an experimental design. Specifically, like the current study, it used a school-level randomization design that randomly assigned 18 of 35 elementary schools to receive the SFA program in kindergarten through grade 2. Also similar to this one, it used combinations of the Woodcock-Johnson Letter Identification,

[16] Appendix Figures D.3 and D.4 replicate Figure 5.2 and the top panel of Figure 5.3, respectively, for a consistent sample of students who were in the study for all three years. Both appendix figures demonstrate the same pattern of results for the consistent sample as were obtained for the primary analysis sample.

[17] Borman et al. (2006).

Figure 5.3

Comparison of Primary Impact Findings from the MDRC SFA Evaluation and the Borman SFA Evaluation



MDRC findings

Borman findings

(continued)

Figure 5.3 (continued)

SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2012-2014), Woodcock-Johnson Word Attack test (Spring 2012-2014), Woodcock-Johnson Passage Comprehension test (Spring 2012-2014), Test of Word Reading Efficiency (Spring 2012-2014), and student records data collected from the five districts in the study sample. Borman findings are taken from Borman et al. (2005a, 2005b, 2007).

NOTES: The student sample size for the Woodcock-Johnson Letter-Word Identification test ranges from 2,522 students (1,307 in the program group and 1,215 in the control group) in Year 1 to 1,631 students (851 in the program group and 780 in the control group) in Year 3.

The student sample size for the Woodcock-Johnson Word Attack test ranges from 2,522 students (1,310 in the program group and 1,212 in the control group) in Year 1 to 1,635 students (854 in the program group and 781 in the control group) in Year 3.

The student sample size for the Test of Word Reading Efficiency ranges from 1,905 students (993 in the program group and 912 in the control group) in Year 2 to 1,625 students (847 in the program group and 778 in the control group) in Year 3.

The student sample size for the Woodcock-Johnson Passage Comprehension test ranges from 1,980 students (1,035 in the program group and 945 in the control group) in Year 2 to 1,625 students (848 in the program group and 777 in the control group) in Year 3.

The MDRC impact analyses for reading achievement were conducted using raw scores. 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.

Effect sizes for the MDRC evaluation were computed using the full control group's standard deviations for the respective measures. A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Word Identification, Word Attack, and Passage Comprehension as outcome measures across the years, making the results directly comparable with this study.

The bottom panel of Figure 5.3 shows the Borman study findings by outcome, after the intervention students completed one, two, and three years of the program. Particular findings include the following:

- A positive and significant impact on Word Attack for kindergarten students, and null findings for all other outcomes after one year of SFA implementation[18]
- Positive and significant impacts on Word Attack, Word Identification, and Letter Identification, but not on Passage Comprehension, for first-graders after two years of SFA implementation[19]

[18] Borman et al. (2005a).

[19] Borman et al. (2005b).

- Positive and significant impacts on Word Identification, Word Attack, and Passage Comprehension for second-graders who had been with the study for three years[20]

While the current study started out on a similar trajectory, its findings diverge from the Borman study in Year 3: The current study finds smaller impacts on Word Attack and no impacts on comprehension in the third year of SFA implementation, while the Borman study finds continued growth in the impact on Word Identification and Word Attack, and a positive impact on comprehension.[21] The last chapter in this report dives into the study sample and the program implementation to explore possible explanations for this observed difference in impact findings in Year 3.

The current study and the Borman study are the only ones to date that use experimental designs to evaluate the impact of the SFA program on various reading skills for early-grade students. Combining the findings from the MDRC and Borman experimental studies through a fixed effects meta-analysis approach,[22] these findings show that, after three years of implementation, the SFA program registered a positive and statistically significant impact on second-graders' alphabetic skill as measured by the Woodcock-Johnson Word Attack test (effect size = 0.19, p-value = 0.0003). The combined impact on second-graders' reading comprehension skill is more muted (effect size = 0.08, p-value = 0.090).

In addition to the Borman study, there are six quasi-experimental studies of the effectiveness of SFA that meet the What Works Clearinghouse evidence standards, with reservation.[23] All of these studies focus on students from prekindergarten to grade 1, making them comparable in age to students in this study. On average, these quasi-experimental studies registered SFA impacts of 0.35 standard deviation, 0.17 standard deviation, and 0.27 standard deviation in effect size on students' alphabetics, reading comprehension, and general reading skill, respectively.[24] None of the findings from these studies are statistically significant at the 5 percent level, however.

[20] Borman et al. (2007).

[21] Two-tailed t-tests show that the differences in the third-year impact findings for Word Attack and Passage Comprehension between these two studies are borderline significant, with p-values at 0.095 and 0.157, respectively.

[22] This approach combines results by taking a weighted average of the estimates from different studies, using the inverse of the estimated variance from each estimate as the weight. For details about this approach, see Konstantopoulos and Hedges (2004).

[23] Dianda and Flaherty (1995); Madden et al. (1993); Ross, Alberg, and McNelis (1997); Ross and Casey (1998); Ross, McNelis, Lewis, and Loomis (1998); Smith et al. (1993).

[24] Ideally, one would conduct a meta-analysis using the impact findings and their corresponding standard errors to synthesize the findings from these studies. However, the standard errors of the impact estimates from these studies are not available to the study team. As an alternative, the study team used the sample size of each

- **Alphabetics.** All six studies showed positive impacts, with various magnitudes, of the SFA program on alphabetic outcomes (mostly measured by Woodcock-Johnson Word Attack or Letter-Word Identification tests): Two of the studies estimated the SFA impacts on alphabetics to be 0.13 to 0.14 standard deviation in effect size; two others put the estimates at around 0.30 standard deviation in effect size; and the last two studies estimated the impact to be around 0.56 to 0.58 standard deviation in effect size.
- **Reading comprehension.** The studies provide mixed findings for the impact of the SFA program on students' reading comprehension. Of the five studies that examined comprehension as an outcome, two studies estimate the impact to be below 0.10 in effect size (0.01 standard deviation and 0.08 standard deviation); one finds it to be 0.18 standard deviation in effect size; and the other two put the impact at above 0.20 in effect size (0.28 standard deviation and 0.44 standard deviation).
- **General reading.** The reported positive impacts on students' general reading achievement also vary in magnitude. One study found the impact to be 0.04 standard deviation in effect size; two studies put it around 0.15 standard deviation; and the remaining three studies reported findings ranging from 0.28 to 0.51 standard deviation in effect size.

In summary, SFA continues to show a positive effect on a measure of alphabetic and decoding skills in the third year of implementation.

study, which is closely related to the estimation standard error, as a proxy weight to calculate the average impacts reported here.

Chapter 6

A Cost Analysis of Success for All

Just as Chapter 4 and Chapter 5 showed the extent to which Success for All (SFA) program elements and student outcomes were distinct from those associated with reading programs in the control group schools, this chapter's analysis reflects the extent to which schools in the program group ("SFA schools") required different or additional resources in order to implement the program, relative to an alternative reading program. It presents the cost of implementing SFA as well as the extra cost or savings of using SFA rather than another reading program. (In this chapter, "reading program" refers to a school's core curriculum as well as reading interventions and other supports.)

The research questions address two different perspectives on how to think about costs.

- What is the difference in the annual per-student direct cost borne by a school for SFA compared with an alternative reading program (incremental out-of-pocket costs)?
- How much more time, space, and effort does a school's implementation of SFA require than what is needed for an alternative reading program, accounting for the fact that some reallocated existing resources do not have a direct cost but do have alternative uses (incremental full resource costs)?

To address the first research question, the analysis presents the out-of-pocket costs of assembling the resources a school would typically fund to implement SFA or an alternative reading program.[1] For the second research question, the analysis defines "full" cost in terms of all resources used, or the "cost ingredients," regardless of who paid for them and regardless of whether they were donated, purchased directly, or reallocated from other uses, since using resources in short supply involves an opportunity cost of what cannot be done instead with those resources — for example, existing staff and existing facilities. Analysis for this second question is not intended to represent what schools or districts actually paid in dollars, but rather to estimate, relative to the alternative reading program, the full resource cost of replication. Given the tight resource constraints facing districts nationwide during the grant period, applying a cost to resources that have alternative uses can illustrate trade-offs facing schools and districts.[2]

[1] In the context of this evaluation, some of these items were subsidized or completely covered by the i3 grant. The analysis in this chapter ignores the funding source and instead focuses on the resources required for implementation in the absence of a grant.

[2] For opportunity costs to actually exist, resources must have competing alternative uses and not be unallocated. In crowded school buildings operating under tight budgets, this assumption is often realistic.

This is one of the few studies to compare SFA program costs to alternative reading program costs, and to go beyond a school's out-of-pocket expenses to consider the full resource cost to society. In this way, the chapter provides a gauge of the relative resource requirements of SFA. If, for example, SFA's full resource cost is greater than that of the alternative reading program while the out-of-pocket costs are the same, the result would imply that SFA reading program implementation prompts schools to spend more time on or devote more space to SFA's reading program than they would for an alternative reading program. The research questions do not address cost-effectiveness, and cost differences cannot be linked directly to the impacts for reasons related to the research design.[3] A detailed analysis of resource costs in one district presents a way to understand the resource requirements for the program, which may be associated with impacts. After an introduction to this case study district, the chapter presents the answers to the research questions, providing details of the assumptions and data sources used to conduct the analyses, and concludes by testing the assumptions and results in a series of sensitivity checks.

Although the study team collected and reviewed cost data for multiple study districts, the chapter presents cost findings for one district, so that a single policy context and a single business-as-usual reading program implemented in the control group schools form the comparison. In the case study district, the control group reading program is a commonly used reading curriculum supplemented by reading intervention materials from other programs.[4] The set of items included in the full resource cost for the case district is similar to what would be included for other districts in the evaluation.

Observed differences in the number of instructional staff allocated to the reading program by the end of the school year, in amount of teacher and school leader time, and in materials costs between program and control group schools can be interpreted as incremental program costs. The chapter focuses on incremental annual per-student costs rather than incremental totals.

The study uses multiple sources of data — several rounds of surveys, Success for All Foundation (SFAF) manuals and administrative records, public data on staffing and expenditures, and interviews with five principals in the case study district specifically to inform the cost analysis — to tabulate or estimate the cost of key program features and learn about contextual factors that may have affected implementation.

[3] Ideally one would want to calculate the "yield" or increased output resulting from the program (for example, additional students diverted from special education identification) and link that to estimated incremental costs. But the research design, a cluster randomized controlled trial with schools randomized within blocks, means there is not sufficient statistical power within a single district to obtain a reliable impact estimate for a given district. As a result, even though the analysis can calculate a cost difference for a single district, that result cannot be linked to a corresponding impact. Sufficient data were not available to calculate cost differences for all districts that make up the impact estimate.

[4] Specific curriculum titles are not cited in the chapter to protect the anonymity of the district. Cost estimates were based on specific core reading and intervention curricula.

Key Findings

- Out-of-pocket costs in SFA schools were $119 more per student per year than in control group schools. This difference is driven by the additional time required for the reading program facilitator and additional training delivered by an SFAF point coach. The *total* out-of-pocket costs are $276 per student per year in SFA schools in this study and $157 in the control group schools.

- Full resource costs — all the extra time, effort, and space required to implement the program — were $227 per student per year more in SFA schools than in control group schools. This difference is driven by additional principal time associated with launching the SFA program, additional costs for the facilitator and training and professional development, and additional classroom space and teacher time associated with teaching reading and providing reading supports. *Total* resource costs were $1,811 per student per year in SFA schools in this study and $1,584 in the control group schools.

- Sensitivity tests indicate that varying the level of reading program implementation and allocation of staff and their time, in both program and control group schools, does not substantially alter the per-student, per-year difference in full resource costs.[5]

For both program and control group schools, the analysis assumes that reading programs started in 2011-2012. In this way, the analysis compares the start-up cost for SFA with that for an alternative reading program in its first year, as well as follow-up "steady-state" costs in the subsequent two years. The analysis spreads the higher start-up cost over the three-year implementation period. Tables report average annual costs based on a three-year reading program.[6]

These results suggest that schools (or districts) spend more and invest more staff time and space on SFA than they would for other programs. In the context of average per-pupil annual spending of nearly $5,650 in all districts in the evaluation sample,[7] the incremental out-of-pocket costs of a little more than $100 per student per year are relatively modest. Thus, the addi-

[5] Results from a similar size district in the evaluation, which adopted a new control group school reading program at the same time as the SFA program began, showed an out-of-pocket cost difference of $343 per student per year, driven primarily by additional materials and training costs.

[6] For schools that continue SFA beyond the initial three-year period, steady-state costs may be a more relevant measure. An analysis limited to third-year, steady-state costs shows a difference of $180 per student per year. This is $47 less than the average annual difference in the three-year estimate. The third-year-only cost is similar to the full resource cost analysis because principal and teacher time remained the same, and facilitator time increased because the case study district was able to fund full-time facilitators in all program schools. Results are in Appendix Table E.2.

[7] Average per-pupil spending is based on district-level data from the Common Core of Data, National Center for Education Statistics, 2010-2011.

tional costs per student are potentially within reach for schools and districts with some budget flexibility.

The Case District and Its Representativeness

The case study district is not an outlier on pre-SFA implementation (baseline) values of key school or district features that might affect resource allocation — such as total staffing and per-pupil instructional spending — with respect to other study districts.[8] As Chapter 7 describes, the majority of districts recruited during the Investing in Innovation (i3) scale-up grant period had three or fewer schools adopting SFA; to this end, the case district is typical, with three SFA schools and three control group schools. These six schools have demographic characteristics comparable to other schools in the evaluation sample, and their total enrollment and staffing remained steady from the year before the grant to the final year of the grant, as they did in other study districts.[9] The change in Title I funding per student in the case district was not any greater over time than it was in other districts.

The district does differ from other study districts in some respects. As Figure 6.1 shows, in terms of its ratio of students to instructional staff (not only reading teachers), the case district (District D in the figure) has the highest ratio of the five study districts, though it still falls within the recommended SFAF class size of 20 students per full-time-equivalent instructor. In terms of spending in the study schools (in program and control group schools combined), per-pupil expenditures in the case district in the final year were $3,143, reflecting a decline of about $250 per student from the year before the SFA program was implemented. (See Appendix Figure E.1.) This spending level is less than the average per-pupil spending of other evaluation districts.[10]

In terms of the reading program, the district was distinct in some ways. The district had a high concentration of English language learners and taught them during a longer reading block (four hours total) than the 90-minute reading block for general education students. This district also offered a relatively high amount of tutoring in both program and control group schools. In addition, some of the districtwide reading coach's time was allocated to coach and support SFA schools, supplementing the time of the SFAF point coach. Only one other district in the study retained this district coach structure for the duration of the grant period.[11] Despite

[8] These results are described in Appendix E.

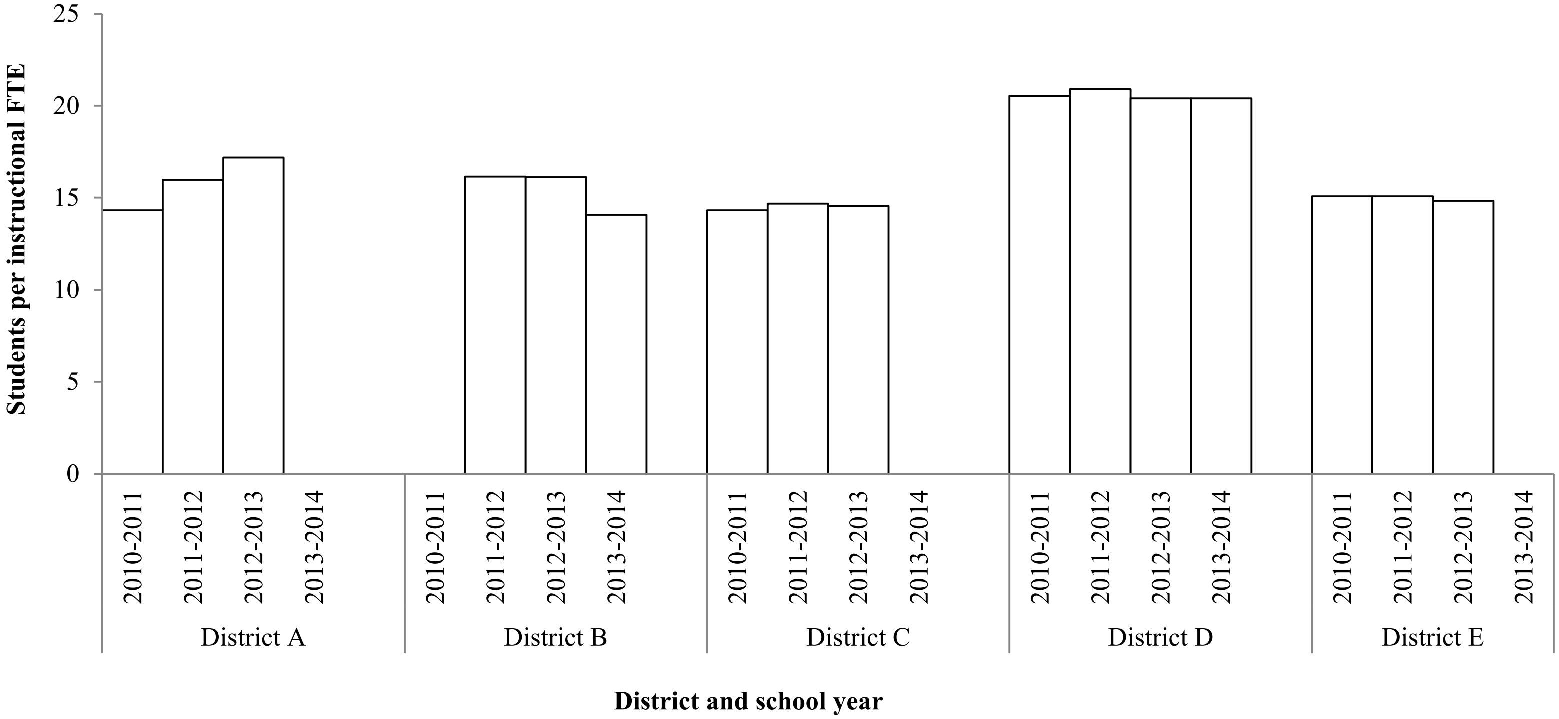
[9] Demographic characteristics in this case district differ from other districts in the state, but that is to be expected given the underresourced districts that SFAF targets.

[10] Based on the federal Common Core of Data Local Education Agency fiscal file in 2010-2011, the average spending in the other four evaluation districts is $6,111. Data for other districts in the final year of implementation were not publicly available at the time of this analysis.

[11] This district coach role is discussed more in Chapter 7 of this report, which examines SFAF's scale-up experience under the i3 grant.

Figure 6.1

Ratio of Students to Instructional Full-Time Equivalent Staff in Study Districts, from 2010-2011 to 2013-2014



SOURCES: MDRC calculations based on publicly available state data on full-time-equivalent (FTE) instructional staff and student enrollment.

NOTES: Data represent 37 study schools in five school districts. Enrollment was calculated by summing all students in kindergarten through grade 5 in all study schools in each district. Individual staff members were considered "instructional" if they directly provided instruction to students, worked in some capacity in a classroom, provided coaching to teachers, or worked on curriculum development. District D is the case district for the main cost analysis.

the local coaching and support, SFA schools in this district attained low implementation scores relative to other SFA evaluation sites. As a result of this difference, Appendix E presents costs for scenarios with more intense and complete implementation of both SFA and the control group school reading programs to test whether the results are sensitive to assumptions about different levels of implementation fidelity.

Program and control group schools appear relatively similar on key elements that drive costs, such as the number of teachers, both in the year before the SFA program began and during the program. However, the ratio of students to teachers is lower in SFA schools than in control group schools. Although schools in both the program and control groups in the district have an average of 74 core subject teachers each year,[12] program group schools have fewer students in kindergarten through grade 5, which affects per-student cost estimates.[13] In addition, the composition of the core subject faculty differs: SFA schools have a higher percentage of new teachers (those with 0 to 3 years' experience) in each year, a difference of 13 percentage points averaged across years. New teacher time costs less per training session (because of lower salaries), and new teachers did not receive as much training over the three years as existing teachers did. Table 6.1 gives a picture of resource allocation in the case district for items that will be included in the resource cost analysis. A full description of resource allocation across a broader set of resource categories is presented in Appendix E. (See Appendix Table E.1.)

Approach 1: Out-of-Pocket Costs

This section describes the method used to calculate the direct costs incurred by schools. It explains which costs were included and how they were estimated for the corresponding tables that summarize the findings.

Out-of-pocket costs are defined here as what the schools in the case study district (District D) would have spent on their own if they had to purchase the program and were not receiving a grant as part of the evaluation.[14] To estimate out-of-pocket costs in both program and control group schools, the analysis requires a set of resources to be included, an estimate of the level (or quantity) of resource use, and a price associated with each resource use. The difference in these costs between the groups represents the incremental out-of-pocket costs.

[12] A teacher was included in the analysis of the reading program if his or her assignment was listed as "Elementary Classroom," "Kindergarten Classroom," "Reading Classroom," "Mathematics," or "Communications Arts" in state records of staff assignments. In some cases mathematics teachers are cocategorized as elementary classroom teachers, so they are included in the sample.

[13] Although schools experienced considerable student mobility in certain years, on average across the three years the net change in enrollment was essentially the same in both program and control group schools, so estimated costs were not adjusted for this.

[14] Note that schools also drew on their existing resources. These are counted in the full cost estimate in the next section.

Table 6.1

Resource Allocation by Year in SFA and Control Group Schools in the Case District, Annualized for Years 1-3

Resource	Program Group	Control Group	Difference
Instructional staff (FTEs)	96.6	87.5	9.1
Items included in resource cost analysis			
Core instructional FTEs[a]	73.7	72.7	1.0
Facilitator FTEs[b]	1.9	1.1	0.9
Principal and supervisory FTEs	3.3	3.3	0.0
Percentage of reading teachers with 0-3 years of experience	33.0	19.8	13.3
Total student enrollment	1,838.3	1,947.3	-109.0
Core instructional student-to-FTE ratio	25.0	26.8	-1.8

SOURCES: MDRC calculations based on publicly available state data on school staff counts and student enrollment. Program school facilitator counts were calculated based on responses to the study's principal survey and principal interviews. The percentage of teachers with zero to three years of experience was calculated based on responses to the study's teacher survey.

NOTES: Staff counts are provided in full-time equivalent (FTE) units.

[a]A teacher is considered a core instructional FTE if the teacher's job assignment was listed as "Elementary Classroom," "Kindergarten Classroom," "Reading Classroom," "Mathematics," or "Communications Arts."

[b]The estimate assumes each control group school had a reading program facilitator equivalent to 0.36 FTE during each of the three implementation years, and both program and control group schools had no facilitator prior to Year 1.

Four key expenses are considered out-of-pocket costs in this analysis:

- After-school tutoring provided or supervised by certified teachers[15]
- Reading program facilitators
- Training (professional development and ongoing coaching) provided by SFAF coaches
- Materials and supplies

[15] While federal Title I funding helped support teacher time in program schools, state funding, especially for English language learners, supported teacher time for after-school tutoring in control group schools that applied for and received state funding.

Table 6.2 provides additional detail on the data sources, quantities, and pricing for each of these out-of-pocket expenses.

Out-of-Pocket Costs: Results

The difference in out-of-pocket costs in the case district is $119 per student per year, as shown in Table 6.3. More than half of this difference ($71 per student per year) came from the cost of training and professional development (for example, time the SFAF point coach spent training and coaching) and represents the increased number of days of training in SFA schools compared with training days in control group schools. Next, even though control group schools had reading program facilitators, and facilitators in SFA schools were not always devoting all their time to that role, the additional time that SFA facilitators did spend represented $38 per student per year. (If the SFA school-based facilitators in all program schools had been working full-time, the difference would be even larger.) Tutoring provided or supervised by certified teachers represented just $12 per student per year more, because control group schools in the case district also provided tutoring by teachers during and after school. And SFA materials cost $3 per student per year less than the core reading and intervention materials purchased by the control group schools. The total out-of-pocket expenses for SFA schools were $505,047, or $276 per student per year, and for control group schools they were $309,255, or $157 per student per year, yielding the $119 difference.

This analysis suggests that SFA represents a slight premium for training costs and facilitator time, but the program could be affordable. Relative to annual per-pupil spending in the case district of about $3,100 on average over the grant period, out-of-pocket costs of an additional $119 per year per student are likely feasible for some districts.

Sensitivity Checks on Out-of-Pocket Costs

In the case district, the control group school reading program did not actually start in all schools in 2011-2012, but for the sake of the analysis it was assumed to have done so. The sensitivity analysis looks at out-of-pocket costs in another small district (District E) that actually did start a new reading program in 2011-2012 (a different program from the one used in the case district). This analysis serves to check the assumptions in the case district about start-up costs for the control group reading program. By examining costs for another new reading program, and facilitator, tutoring, and training costs, the study team could identify what factors determine the size of the estimated per-student annual difference in costs. (The full extent of start-up costs, which may include substantial after-school hours on the part of the principal and facilitator, are not included because the analysis would have to include too many assumptions.) The two districts share some similarities: three program schools in the evaluation, similar district spending levels per pupil, and control group schools with a reading program facilitator.

Table 6.2

Data Sources for Direct Expenses for Reading Programs in the Case District

Expenses	Key Data Sources for Quantity	Data Sources, Quantities, and Pricing
Tutoring	Principal survey, NCES Schools and Staffing Survey	Both program and control group schools in the case district provided tutoring by certified teachers after school. The cost of tutoring represents the number of teachers principals said were involved in tutoring and the cost of those teachers' after-school time in terms of their hourly wage (salary plus benefits), multiplied by a lower bound of tutoring session length (41 minutes) and frequency (1-5 times a week, depending on the school response).
Reading program facilitator	Principal survey, interviews, NCES Schools and Staffing Survey	Both SFA and control group schools had reading program facilitators. SFA facilitators spent anywhere from 0.5 to 1 full-time equivalent (FTE) position to coach teachers, review data, and regroup students. Control group facilitator time was allotted for the reading block (1.5 hours) plus 1 hour for tutoring and interventions (essentially a 0.36 FTE), based on interview descriptions of this role. Control group school facilitators provided general reading or Title I intervention support, rather than support associated with the specific commercial reading program by Houghton-Mifflin. Interviews in both SFA and control group schools indicated the level of experience of facilitators, and the corresponding salary and wages were taken from the U.S. Department of Education schools and staffing survey.
Training	SFAF contracts in program group schools and interviews in control group schools	SFAF coaches visited each school for 12-19 sessions of training per year. In control group schools, principal interviews indicated that trainers or reading coaches visited schools six days per year. One additional day of training in the summer before Year 1 was included for control group school in-service time. Trainer salaries and benefits were obtained from SFAF contracts and were assumed to be the same for both program and control group schools, as prices are based largely on experience.
Materials and supplies	SFAF contracts, commercial publisher's list price	SFA school materials costs varied by school, and quantities and prices were taken directly from contracts across years. SFAF, as a nonprofit foundation, makes its chapter books, teacher lesson plans, facilitator guides, and Member Center data system and other assessment systems available to schools essentially at cost. To estimate the one-time reading curriculum and materials costs for the control group school reading program in Year 1, the study team consulted the commercial publisher's list price and prices on Amazon.com (for a national price) for student texts and teacher editions, as well as supplemental intervention materials. Student text prices were multiplied by the number of students enrolled (assuming that schools may have purchased slightly fewer books than they had students). A 25 percent discount was applied, recognizing that districts probably did not pay retail price.[a]

NOTES: The study team obtained prices from national and local administrative data and adjusted these prices to be expressed in 2012 dollars (that is, to reflect prices at the end of the first year of implementation). Prices are also discounted and geographically adjusted in order to be nationally representative. For staff positions, when only salaries were listed, the study team added a 33 percent increment for benefits, as is standard.

[a]Appendix E discusses the rationale for this discount.

Table 6.3

Out-of-Pocket Expenses and Differences, Annualized for a Three-Year Reading Program, in SFA and Control Group Schools in the Case District

	Program Group		Control Group		Difference	
Costs ($)	Total	Per Student	Total	Per Student	Total	Per Student
Certified tutoring after school[a]	38,922	21	17,622	9	21,299	12
Reading program facilitator[b]	138,105	75	71,830	37	66,275	38
Training and professional development[c]	132,908	72	2,367	1	130,540	71
Reading program materials[d]	195,112	107	217,436	110	-22,324	-3
Total	505,047	276	309,255	157	195,791	119

SOURCES: Teacher full-time equivalent (FTE) values are taken from publicly available state data. Teacher and facilitator salaries were calculated from the U.S. Department of Education schools and staffing survey. Teacher experience and facilitator experience levels were calculated based on responses to the study's teacher survey. Tutoring frequency and duration, facilitator FTE values, and control group reading programs are taken from the study's principal survey and principal interviews. Training and materials costs for program group schools are based on SFA program contracts. Reading program materials costs for control group schools were obtained from publisher websites and prior evaluations.

NOTES: Rounding may cause slight discrepancies in calculating differences. Prices are given in geographically adjusted 2012 dollars. Personnel estimates include benefits amounting to one-third of salary. See Appendix Table E.6 for a more detailed description of cost calculations and assumptions.

[a] A lower bound for tutoring frequency, for both certified and volunteer tutors, was taken from responses to the principal survey. A lower bound of tutoring session length (41 minutes) was also taken from the principal survey, since all respondents indicated that tutoring sessions in their schools lasted at least 41 minutes.

[b] The estimate assumes that all non-SFA schools have a reading program facilitator equivalent to 0.36 FTE.

[c] Estimates include point coach training and observation time.

[d] Reading program materials include teacher and student textbooks, classroom consumables, assessment materials, support and licensing fees, and intervention materials when applicable. This estimate assumes that all study schools purchased a new set of reading program materials in Year 1.

The estimated difference in out-of-pocket costs in District E is more than double the difference in the case district: an additional $343 per student per year rather than an additional $119. The larger annual per-pupil difference in District E compared with District D is driven by several factors. First, the number of students enrolled in District E was smaller than in District D, so on a per-student basis, even a similar total cost difference to District D would appear higher in District E. Second, there were differences in "business as usual" in the two districts: SFAF-developed materials did not cost District E more in absolute terms but were much greater than those required for the control group reading program in District E. Finally, the incremental cost of the reading program facilitator in SFA schools in District E was $75 per student per year, nearly double the difference in District D, and the incremental cost of training and professional development was also higher. District E offered less tutoring than District D in both program and control group schools, so there is essentially no per-pupil difference in annual tutoring cost, compared with $12 in District D.

Approach 2: Full Resource Costs

This approach takes the perspective that reallocated resources with alternative uses have costs. Even though a school might have had resources on hand, such as classrooms and staff, before adopting SFA, the use of those resources for the SFA program means the school forgoes their use for some alternative reading program or school activity. It is unlikely that these schools had "extra" resources with no alternate use; schools reported insufficient funding to implement aspects of the reading program, and teacher turnover and absenteeism were reported in the case district. This view of alternative uses is applied to both the SFA and control group schools. The difference between the cost estimates represents the *incremental* full resource cost required for SFA.

To estimate resource costs, the method builds on that used for out-of-pocket costs. In addition to using the expenses listed above, the next step is to determine the difference in resource use required for SFA and for the alternative reading program. For example: How many more (or fewer) hours of staff time do SFA schools use, compared with what control group schools use, for their reading program? What is the difference in classroom space used?

To get a gauge of the relative resource requirements of SFA and the alternative reading program, the hours any individual devotes to the respective reading programs are included, be they school-paid hours that teachers would have spent doing something else or not.[16] As a re-

[16] Levin and McEwan (2001).

sult, staff time is included in the list of resources used to estimate full resource cost. The analysis also includes staff members who were reallocated to the reading program.[17]

SFAF manuals and information accompanying district contracts with SFA determined critical items to include, and surveys and interviews identified additional elements or allocations of time that could drive costs and cost differences. The resources include the following:

- *Teachers and their time:* This includes time spent on reading instruction, after-school tutoring, and training, and other time spent on reading support structures and processes.
- *Volunteer tutors:* The resource cost reflects the time volunteers spent providing tutoring.
- *Principal time:* Because SFA is both a reading program and a reform of school practices and structures related to student progress, principals are likely to be devoting substantial time, at least in the first year, to ensure the program takes hold. In interviews, SFA school principals indicated they spent at least 10 to 12 hours per day on their job, and at least 5 to 6 of those hours were for SFA in the first year. This analysis therefore assumes that principals spend half their time on SFA, especially considering the staffing constraints and limits on facilitator time in these schools. Control group school principals are assigned a little less time for the launch of their reading program in Year 1.[18]
- *Facilities and classrooms:* The primary facilities cost relates to the number of classrooms used during the reading block and for after-school tutoring.
- *Out-of-pocket expenses* (described above):
    - After-school tutoring
    - Reading program facilitators
    - Training provided by SFAF coaches
    - Materials and supplies

[17] Even though a school may have already had a Title I-funded position or intervention coordinator on staff, reallocating that person to SFA is not without costs. As interviews with facilitators and principals pointed out, those Title I staff organized supports and performed tasks related to state testing, English language learner support, and accountability requirements that still needed to be completed. Given that schools facing budget constraints could not hire additional staff, respondents reported that facilitators were often doing the job of several people.

[18] As noted in Chapter 4, there was no statistically significant difference between SFA teachers and control group school teachers in their ratings of the extent to which their principal was involved with the school's reading program. The teacher survey items from which the rating is derived do not ask about the amount of time their principals spent in implementing and monitoring the reading program.

Table 6.4 goes into more detail about these resources and the sources from which information was drawn. National market prices are used to project their cost, and prices are discounted by 3.5 percent, as is standard.[19] The details of the pricing are discussed in Appendix E. As noted earlier, to make the burden of training and materials costs comparable between SFA and control group schools, the analysis assumes that control group schools were also starting up a new reading program.[20]

Full Resource Costs: Results

In the case district, the difference in total resource costs, as shown in Table 6.5, is $227 per student per year. Some of the largest incremental costs come from the out-of-pocket differences described earlier: the costs of training and professional development provided by SFAF coaches and the time of the reading program facilitator. Principal time made up a larger portion of the difference than teacher time. Principal time in program group schools costs more per student per year than in control group schools, because the analysis assumes principals in SFA schools devoted half their time in the first year to launching, and in subsequent years to sustaining, the reading program and associated structural and procedural changes in the school, while control group principals spent about a third of their time.

The number of teachers, including new teachers, ends up driving cost differences related to teacher time. The cost of total teacher time allocated to SFA, especially the portion of time spent on the reading block, appears as an incremental total savings for SFA schools because SFA schools have more new teachers (who cost less). However, the control group schools have more students, so on a *per-student* basis, the savings become neutral. Also note that tutoring time shows a relatively small per-student per-year difference because the control group schools in the case study district also provided tutoring. Volunteer tutors operated only in control group schools, so that resource cost is subtracted from the total difference.

Classroom space and materials made up the rest of the cost differences. Even though SFA and control group schools had a similar number of teachers instructing in classrooms, the difference comes from the addition of classroom hours for tutoring as well as the number of classrooms required to maintain a 20-to-1 student-teacher ratio, as SFAF requires for beginning

[19] Moore, Boardman, and Vining (2013).

[20] Since control group schools in the case district did not in fact adopt a new program in 2011-2012, any teacher turnover associated with bringing in a new program is not present in the control group schools. The main analysis did not adjust for turnover costs in either the SFA or control group schools. Appendix Table E.5 accounts for the number of new teachers in program and control group schools and the associated training costs, as well as the number of substitute teachers.

Table 6.4

Data Sources for Full Resource Costs for Reading Programs in the Case District

Resources	Key Data Sources for Quantity	Data Sources, Quantities, and Pricing
Teachers and their time[a]	Principal survey, teacher survey, interviews and focus groups, SFAF Snapshot, SFAF classroom counts	The number of teachers is limited to those who are main classroom teachers, as determined from job class titles in state administrative data. The median teacher experience level in each school is associated with a salary plus benefits, which determines the price of these teachers' time. Time includes: • Instruction during the reading block: 90 minutes on average for both SFA and control group schools • Tutoring after school: using the same estimates as reported for out-of-pocket expenses • Training: For both SFA and non-SFA schools, the analysis assumes three hours of meeting time (which could include after-school or in-service day sessions) and that each session serves only a fraction of all teachers.[b] For SFA schools, the analysis assumes that each session serves about one-third of teachers, or that teachers participate in every third session, based on the distribution of teachers across KinderCorner, Roots, and Wings program levels reported in SFAF administrative data for evaluation schools. For control group schools, the analysis also assumes that each training session serves one-third of teachers. • Support team participation (such as an attendance team): SFA teachers in focus groups reported spending time in at least one Solutions Team. For control group school teachers, no time for this function is included.[c]
Volunteer tutors	Principal survey	While SFA schools did not use volunteers for tutoring, control group schools used a mix of certified teachers and volunteers. The resource value of volunteer time is estimated as the equivalent of a paraprofessional salary or as volunteer time as priced by the Independent Sector.
Principal time	Principal survey, interviews	SFA principals were assumed to spend half their time (0.5 FTE) reviewing student progress, grouping students, monitoring instruction and after-school tutoring, convening Solutions Teams meetings, and engaging parents. Control group school principals were assumed to spend 2.5 hours a day (0.36 FTE) monitoring the 90-minute reading block, monitoring tutoring and interventions, and leading support team functions. Principal salary levels plus benefits were determined using experience levels reported on the principal survey.

(continued)

Table 6.4 (continued)

Resources	Key Data Source for Quantity	Data Sources, Quantities, and Pricing
Facilities and classrooms[d]	Staffing data from state administrative records	Classrooms used during the reading block and for after-school tutoring make up this cost. The analysis assumes that each teacher had his/her own classroom or its equivalent space during the reading block. In addition, if tutoring was provided, the analysis estimates the "rental" price for one classroom per hour of after-school tutoring. Tutoring may have happened in one large room (as in a computer lab), and different tutors may have worked different days of the week in that same classroom, so the analysis assumes just one classroom for tutoring. As a result, this cost may be a conservative estimate of required space.
Room for materials and facilitator		SFA requires a dedicated room where the facilitator works and stores materials to be easily available to teachers. In some schools, this may be a classroom, while in others it may be a closet. The analysis assumes each SFA school had one materials and facilitator room, and that control group schools did not have a room devoted to this purpose. It was priced as the equivalent rental price of a classroom for a full day.
All items in Table 6.2[e]		

NOTES: [a]The number of teachers is restricted to "core subject teachers": those who teach English, math, or social studies. It excludes music, art, and gym teachers, as well as paraprofessionals, aides, and substitutes. These numbers and job class titles were checked against the teacher survey responses, which provided a close but not exact corroboration.

[b]The number of proposed sessions is taken from contracts, but actual numbers may vary. Reliable data were not available to account for additional or fewer visits made to a given school based on the school's needs.

[c]When the SFAF Snapshot (implementation scoring form) indicated full implementation of Solutions Teams, the cost was based on having two teachers on each of the five teams. If Solutions Teams met once a month, that is, nine times per year, this would be 9 hours of time. When the Snapshot did not indicate full implementation, the frequency was reduced to one teacher on five teams meeting three times per year for an hour each meeting. Solutions Teams functions were addressed in control group schools, but the data provided no indication that teachers there spent time on them.

[d]The team depreciated the cost of classroom space, computers, and furniture to obtain the equivalent of a "rental price" for their use, given that these items could be used for alternative subjects or programs in the school. Some schools had laptops and/or a computer lab for a brief time. But the team could not determine the presence of this equipment consistently across schools, so these costs are excluded.

[e]Note that certified tutoring time, included in Table 6.2, is reflected in teacher time in the full resource cost tables, as explained in the "Teachers and their time" row.

reading classes. The rest of the difference comes from materials (books and teacher guides) and the dedicated room for materials and the facilitator in SFA schools.

The additional resource costs indicate that schools need to draw on more staff, effort, and space to implement the program, but to a relatively modest extent. These estimates were calculated in a specific funding and implementation environment; the resource costs could differ in other settings and times. Cost differences could be greater if the program schools implemented SFA to its full extent (for example, by employing a full-time facilitator or putting in place

Table 6.5

Resource Costs and Differences, Annualized for a Three-Year Reading Program, in SFA and Control Group Schools in the Case District

	Program Group		Control Group		Difference	
Costs ($)	Total	Per Student	Total	Per Student	Total	Per Student
Teacher time (total)[a]	954,484	519	957,936	492	-3,452	27
Teachers: reading block[b]	869,063	473	920,013	473	-50,950	0
Teachers: support teams[c]	7,144	4	0	0	7,144	4
Certified tutoring after school[d]	*38,922*	*21*	*17,622*	*9*	*21,299*	*12*
Teacher time in training	45,896	25	20,300	7	25,596	18
Reading program facilitator[e]	*138,105*	*75*	*71,830*	*37*	*66,275*	*38*
Principal time devoted to reading program	182,792	99	134,226	69	48,566	30
Volunteer tutors[f]	0	0	22,300	11	-22,300	-11
Facilities and classrooms[g]	1,703,908	927	1,681,026	864	22,882	63
Materials and facilitator room[h]	18,923	10	0	0	18,923	10
Training and professional development[i]	*132,908*	*72*	*2,367*	*1*	*130,540*	*71*
Reading program materials[j]	*195,112*	*107*	*217,436*	*110*	*-22,324*	*-3*
Total	3,326,232	1,811	3,087,121	1,584	239,111	227

(continued)

Table 6.5 (continued)

SOURCES: Teacher and principal full-time equivalent (FTE) values are taken from publicly available state data. Teacher, principal, facilitator, and volunteer salaries are calculated from the U.S. Department of Education schools and staffing survey. Teacher and facilitator experience levels were calculated based on responses to the study's teacher survey. Tutoring frequency and duration, facilitator FTE values, control group reading programs, and principal experience levels were determined based on responses to the study's principal survey. Point coach salaries were obtained from correspondence with the SFA Foundation. Facilities and classroom costs are based on square footage costs from the Center for Benefit-Cost Studies of Education Database of Educational Prices. Training and materials costs for program group schools were obtained from SFA program contracts. Reading program materials costs for control group schools were obtained from publisher websites and prior evaluations.

NOTES: Rounding may cause slight discrepancies in calculating differences. Italicized items also appear in Table 6.3. Prices are given in geographically adjusted 2012 dollars. Personnel estimates include benefits amounting to one-third of salary. See Appendix Table E.6 for a more detailed description of cost calculations and assumptions.

[a]Teacher counts reflect an estimate of core-curriculum teachers. A teacher was considered a core instructional FTE if the teacher's job assignment was listed as "Elementary Classroom," "Kindergarten Classroom," "Reading Classroom," "Mathematics," or "Communications Arts."

[b]This estimate assumes that all reading teachers spend 90 minutes per day in the reading block, except teachers of English language learners who spend 4 hours per day.

[c]The estimate assumes that no teachers in control group schools met on support teams.

[d]A lower bound for tutoring frequency, for both certified and volunteer tutors, was taken from responses to the principal survey. A lower bound of tutoring session length (41 minutes) was also taken from the principal survey, since all respondents indicated that tutoring sessions in their schools lasted at least 41 minutes.

[e]The estimate assumes that all non-SFA schools have a reading program facilitator equivalent to 0.36 FTE.

[f]No volunteer tutors were reported in program group schools.

[g]This estimate assumes that each teacher has his/her own classroom. This figure includes the cost of a classroom for after-school tutoring by certified teachers.

[h]This estimate assumes that each SFA school had one room for the program facilitator and reading program materials and that control group schools did not have a room devoted to this purpose.

[i]Estimates include point coach training and observation time.

[j]Reading program materials include teacher and student textbooks, classroom consumables, assessment materials, support and licensing fees, and intervention materials when applicable. This estimate assumes that all study schools purchased a new set of reading program materials in Year 1.

more Solutions Teams). Differences could also be smaller if control group schools had more staff associated with their programs. The next section on sensitivity checks explores these alternative scenarios.

Sensitivity Checks on Full Resource Costs

One can imagine that the estimated resource costs in the case district might be driven by unique situations related to the particular level of implementation or the way in which staff time and counts were estimated. To check the case district assumptions, the study team varied the resource quantities based on different assumptions about staffing and the level of implementation.

The first set of checks relates to the number of staff members and their time allocation. The analysis uses a different estimation of costs for existing items, such as teacher experience and training time, as well as an estimation of additional costs for the instructional support staff, such as instructional aides and paraprofessionals; the district reading coach (who also served part-time as a local SFA coach); and substitute teachers. The last item is included because of the relatively high rate of chronic absenteeism in the case district (more than a third of teachers were absent for 10 or more days a year). Paying replacement teachers can be seen as an additional cost to maintain the same staffing level.

None of the staff-related changes alter the per-student, per-year cost difference substantially. (See Appendix Table E.4 for details.) Some of these checks do not alter the estimated cost difference, while others increase the costs by up to $43 per student per year.

The second set of checks relates to the level of implementation. Program group schools were compared with control group schools in three scenarios, with both schools implementing their respective reading programs at a low, moderate, or high level. Because the analysis compares implementation at the same level in both schools, the amount of principal time, facilitator time, and teacher time is assumed to be similar in both program and control group schools. The resource cost differences in the two scenarios are similar (between $191 and $206 more per student per year in SFA schools compared with control group schools). Moreover, each scenario's resource cost difference is less than the estimated difference of $227 per student per year shown in Table 6.5. (See Appendix Table E.5 for results for each scenario and Appendix E for an explanation of the results.) Appendix Table E.6 enumerates the exact assumptions used for the resource cost analysis and each implementation scenario.

Limitations

As with any cost analysis, there are aspects of implementation to which costs could not be assigned. The analysis does not account for smaller expenses that are likely not driving the costs of the program, such as the cost of photocopying assessments for the periodic regrouping tests, nor does it account for other aspects of the program, such as parent engagement, because reliable data were not available. As noted earlier, it does not reflect full start-up costs due to limited data availability regarding time use after school.

This analysis also is not able to accurately account for the costs of assembling the full range of reading materials, all included in SFA, that may be required in control group schools. For example, control group schools incur coordination costs related to buying intervention curricula that are separate from their core curricula; SFA schools have interventions built into the reading program materials. In addition, control group schools may have to install and coordinate separate data systems for screening all students and identifying those students who need attention, while SFA schools can rely on the Member Center monitoring system.

In interviews, principals cite qualitative aspects of the program, such as better staff coordination, better training of new teachers, and better data systems, but there is no clear way to assign costs to these features.

Staff turnover was cited by principals as a barrier to implementation and smooth operations. The district did not provide counts of leavers and joiners and the experience or salary levels of those in each group, information that would be needed in order to estimate the cost associated with each staff member's length of service during the school year. This could affect the accuracy of the cost estimates, given that personnel costs dominate costs in a school or program.

Discussion and Conclusion

It is worth noting that the implementation in SFA and control group schools observed in this study occurred in the context of a specific funding environment, as mentioned in Chapter 1. In a more favorable funding situation, overall expenditures might be higher if schools are able to hire more teachers and provide more tutoring. But schools in earlier eras of SFA implementation also faced constraints in funding a full-time school-based facilitator.

To put this analysis in context, and to see whether program costs have changed over time as the SFA program has evolved, it is useful to compare results from the case district to an earlier cost study of SFA by Borman and Hewes in 2002.[21] The studies differ in several ways. First, the prior study did not leverage its access to data on matched control group schools to calculate the incremental or added cost of SFA relative to the counterfactual condition (or "business as usual"), as is done in this chapter. That cost study provided the level of out-of-pocket costs for five program sites, relied exclusively on SFAF administrative data, and assumed that the program training and materials costs were the same for all sites, rather than accounting for the different amounts of materials and training required relative to school size (as this analysis

[21] Borman and Hewes (2002).

did).[22] The prior study did not estimate resource costs for issues such as additional time spent by staff beyond the regular school day.

The incremental resource costs matter for this evaluation, because "business as usual" has changed substantially since the Borman and Hewes study. Widespread adoption of Response to Intervention practices took hold in many elementary schools, including those in the case district and in the broader study. As a result, many control group schools reported tutoring and intervention structures similar to those of SFA. Thus, even if the 2002 study had estimated the incremental cost of SFA, it would have reflected a different reading program and policy context than the current study.

The feasible comparison between that earlier study and the case district relates to annual out-of-pocket costs. In order to make a fair comparison, the study team made some adjustments to the Borman and Hewes calculations, because their study included several costs that differ from the case district costs. The first adjustment was to make tutoring time comparable. The prior study assumed that certified reading tutors were full-time equivalents, suggesting that tutoring occurred seven hours a day, five days a week. By contrast, this evaluation accounts only for the after-school tutoring time spent by certified teachers (on average, just one after-school hour for three days per week). The adjustment reduced tutoring time for each tutor to one period or hour a day from seven hours a day, to conservatively estimate dedicated tutoring time. Second, the earlier study accounted for a full-time-equivalent family support staff member, a position that did not exist in the current evaluation districts. That position was eliminated from the prior study estimate in this adjustment. Both studies estimated that facilitators in some schools worked less than full time on SFA.

The adjusted comparison between the earlier study and the one in this chapter shows that total annual out-of-pocket cost *per SFA school* increased over time. In the analysis presented in this chapter, the average annual out-of-pocket cost per SFA school is about $168,348.[23] With the previously described adjustments to the 2002 study, the average annual out-of-pocket

[22] Although Table 2 in the Borman and Hewes article is titled "Program Ingredients and Costs of Success for All by School and by Year," it presents information on only the elements that this chapter calls "out-of-pocket" expenses — tutors, facilitator, materials, and training — plus family support staff, a position that no longer exists on its own. That table does not provide a full accounting of resources relative to time, space, or opportunity costs.

[23] The out-of-pocket levels in both studies for the materials and training are based on data from SFAF contracts. The out-of-pocket levels for after-school tutoring and the facilitator are estimated costs based on survey data for this study and on SFAF administrative data for the Borman and Hewes study. Total costs were calculated across three years for each school; then the study team calculated the average cost per year per school. In the Borman and Hewes study, the average was across five schools. In the case district in this evaluation, the average was across three schools.

cost per SFA school would have been about $107,111, unadjusted for inflation.[24] Annual *per-student* out-of-pocket costs for each school, a more policy-relevant measure, were not presented in the Borman and Hewes study.[25]

In sum, the SFA program direct costs to schools have increased over time as the program has offered more services to schools related to data tracking and analysis (as described in Chapter 3). When compared with different staffing scenarios, or direct costs in another district, the out-of-pocket costs per student are relatively stable. The incremental per-student annual out-of-pocket and resource costs represented by SFA over other reading programs from 2011-2012 through 2013-2014 are modest and represent potentially reasonable costs for many districts.

[24] In 2012 dollars, the per-school cost is $142,811. However, this number is slightly misleading. Teacher salaries and benefits have lagged behind inflation, especially considering budget cuts that held down salaries in the years before 2012. According to the U.S. National Center for Education Statistics, teacher pay was higher in 1999-2000 (the period for the prior study) than in 2011-2012 or 2012-2013 (in 2012-2013 dollars). Given that salaries for teachers and trainers make up more than half the costs, it is not clear how best to adjust earlier estimates over time. With the original Borman and Hewes assumptions, the estimated annual out-of-pocket cost was about $319,490 per SFA school in 2000 dollars (or $415,337 in 2012 dollars).

[25] The study presented per-pupil annual expenditures, which are not necessarily equivalent to out-of-pocket costs.

Chapter 7

The Scale-Up of Success for All

This chapter addresses three important dimensions of scaling a mature program: (i) *outreach strategies*: what methods the program used to approach more schools and districts; (ii) *scope and efficiency*: whether and how the program reached its expansion targets and achieved operational efficiency; and (iii) *support mechanisms and implementation outcomes*: among schools that adopted the program, what supports were received and what challenges arose, and whether the program model was implemented as intended.

Given that the expansion of Success for All (SFA) occurred on the heels of the Great Recession, and at a time when more schools had evidence-based reading programs already in place (as noted in Chapter 1), this chapter also studies the interaction between contextual and program factors, to describe what facilitated or hindered the expansion process and investigate whether some program features were less appealing to implement in tough financial circumstances. The Success for All Foundation (SFAF) initially planned to recruit 1,100 new schools in five years. The chapter explores the relative success SFAF encountered in each scale-up dimension during the period funded by the U.S. Department of Education's Investing in Innovation (i3) grant (September 2010 through September 2015).[1]

Several factors motivate the study of scale-up efforts. First, the evidence-based policy movement includes government- and foundation-funded initiatives that seek to expand the scale of programs with proven effectiveness.[2] Funders and program developers alike are thinking about effective strategies to expand their programs to new sites and/or new populations. Second, researchers and program developers are interested in whether expanding to new sites requires a shift in implementation strategies.[3] Third, the public management question of setting public targets with large publicly funded initiatives, such as i3, makes this expansion of SFA different from its earlier expansion during the 1990s. Such targets, and the accompanying reporting and measurement of progress by a federal funding agency, can create an environment of public pressure and accountability that distinguishes scale-up under these conditions from expansion using only an organization's own financing.[4]

[1] The analysis in this chapter reflects data through May 2015 and therefore does not include completed recruitment and expansion data for the final year of the grant.

[2] Other organizations doing such work include the Coalition for Evidence-Based Policy and the National Center on Scaling Up Effective Schools.

[3] Quint et al. (2005).

[4] Educational institutions change their practices in response to accountability targets, such as markers of academic progress, discipline, and graduation (Desimone 2013; Hamilton et al. 2007).

Research Questions

Research questions for the chapter correspond to the three dimensions of scaling up that stem from the motivating factors discussed above:

Outreach Strategies

- What was SFAF's plan to approach districts and schools?
- To what extent did the outreach modes and incentives that SFAF designed help recruit new schools? To what extent did schools decline to consider the program, given economic circumstances?
- To what extent did characteristics of schools or districts approached during the i3 period differ from those approached before the i3 grant began?

Scope and Efficiency

- To what extent did SFA expand?
- Did SFAF increase the geographic concentration of schools, as its leaders had hoped?

Support Mechanisms and Implementation Outcomes

- What approaches did SFAF use to support schools that adopted the program?
- Did scale-up sites achieve fidelity comparable to that at evaluation sites?
- What factors facilitated or impeded SFA's ability to scale up and schools' interest in adopting or maintaining the program?

The scale-up logic model (Figure 7.1) shows the mechanisms for recruitment and outreach in the second column; expansion goals and scope are identified in the third column, with a target number of schools to be recruited in each grant year; and the last column includes mechanisms for support.

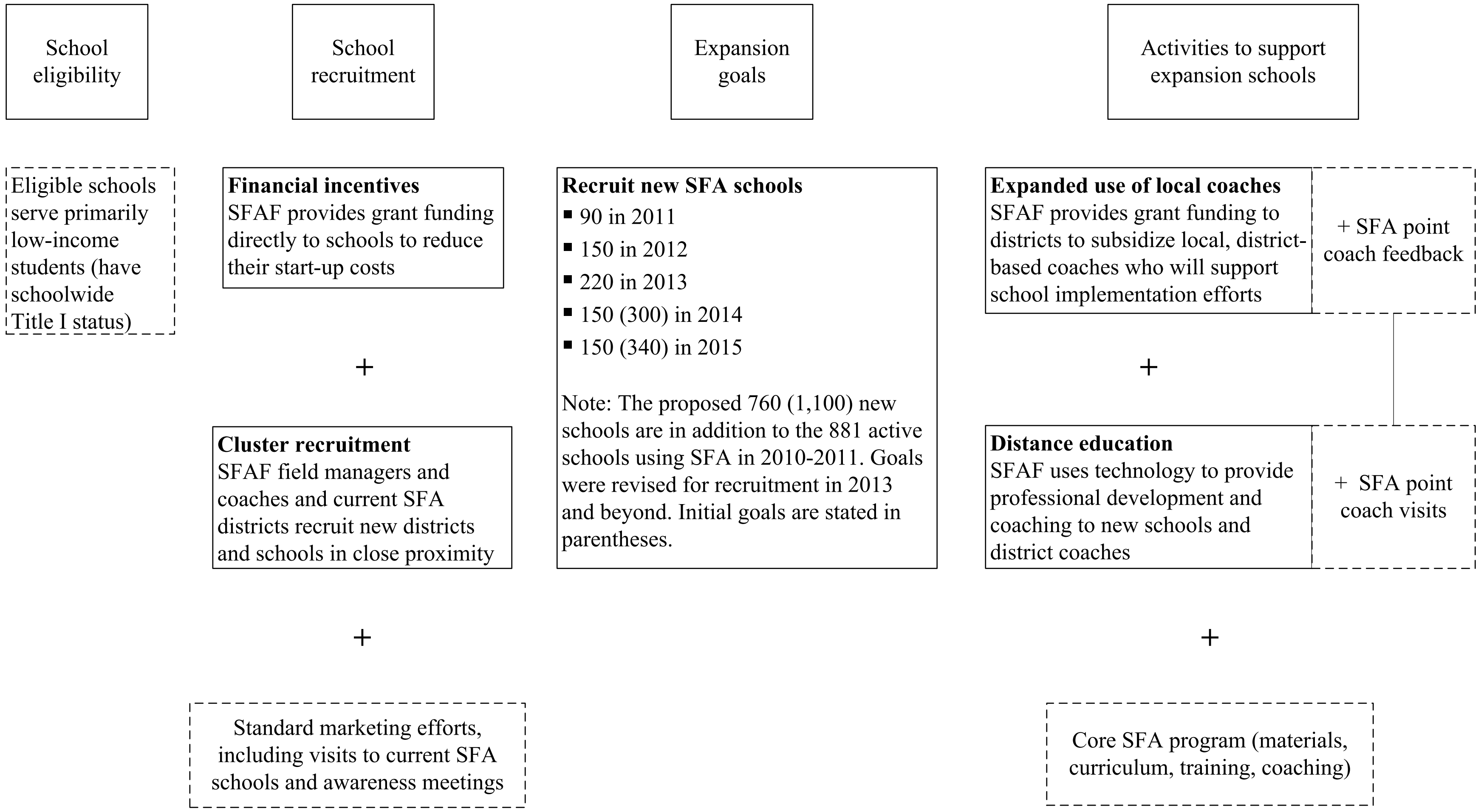
Key Findings

Outreach Strategies

- SFAF approached districts that had slightly lower per-pupil spending, schools that had slightly higher percentages of black students, and a larger percentage of schools located in the South, compared with schools using SFA immediately before the grant period.

Figure 7.1

Logic Model for SFA Scale-Up



School eligibility

Eligible schools serve primarily low-income students (have schoolwide Title I status)

School recruitment

Financial incentives
SFAF provides grant funding directly to schools to reduce their start-up costs

+

Cluster recruitment
SFAF field managers and coaches and current SFA districts recruit new districts and schools in close proximity

+

Standard marketing efforts, including visits to current SFA schools and awareness meetings

Expansion goals

Recruit new SFA schools

- 90 in 2011
- 150 in 2012
- 220 in 2013
- 150 (300) in 2014
- 150 (340) in 2015

Note: The proposed 760 (1,100) new schools are in addition to the 881 active schools using SFA in 2010-2011. Goals were revised for recruitment in 2013 and beyond. Initial goals are stated in parentheses.

Activities to support expansion schools

Expanded use of local coaches
SFAF provides grant funding to districts to subsidize local, district-based coaches who will support school implementation efforts

+ SFA point coach feedback

+

Distance education
SFAF uses technology to provide professional development and coaching to new schools and district coaches

+ SFA point coach visits

+

Core SFA program (materials, curriculum, training, coaching)

NOTE: New services or program elements are in boxes with full lines; standard elements are in dotted lines.

Scope and Efficiency

- Even during tight financial times for school districts, SFA expanded to nearly 450 schools and 146 districts by 2014-2015.[5]

- SFAF did not achieve as much geographic concentration within districts or among neighboring districts as it had anticipated: In the large majority of school districts involved in the i3 scale-up, three or fewer schools operated the program.

Support Mechanisms and Implementation Outcomes

- The hiring and use of local coaches (district employees who supported SFA implementation) to increase local support to schools and help achieve economies of scale succeeded in some districts, but not in as many districts as SFAF had hoped, either because the coaches were not hired or because their efforts did not have the desired effects. Just over half the scale-up schools had access to a local coach as well as an SFA "point coach" (an SFAF staff member).

- Schools in the scale-up sample achieved lower third-year implementation scores than schools in the evaluation sample did. Yet among the scale-up sample, schools without a local coach achieved scores similar to those of schools that did have a coach.

Potential Limiting Factors

- Schools reported feeling pinched by the recession and contraction in state revenues and federal Title I revenues during the grant period.

- The expansion of alternatives to SFA, or at least the presence of some of the whole-school structures and processes that SFA includes in its program (such as Response to Intervention frameworks), meant more "competition" for SFAF when it began this current round of marketing and expansion efforts.

This chapter takes an observational and descriptive approach to discussing the expansion. It is able to assess the extent to which the expansion unfolded as SFAF had anticipated and what got in the way, based on information from interviews with adopting and nonadopting schools, program coaches, and executives; SFAF administrative records; and public data. While the quantitative results rely heavily on SFAF records of recruitment and adoption and on contracts with districts, analysis was conducted independently of the foundation. The study was not

[5] MDRC's estimate, based on Common Core of Data information, is that the program served about 276,000 students through the grant's fourth year. SFAF estimates that at least 95 additional schools will be recruited through Year 5 and will add an additional 34,000 students to total number of students served.

designed to address whether SFAF was more or less successful at expansion than any other reading or instructional program being scaled up during the same time period. Nor can the chapter address whether SFAF would have been more or less successful with a different grant amount to offer schools.[6]

Data presented on the total number of schools recruited and students served, and the analyses and tabulations contained in this chapter, reflect counts and estimates reported by SFAF through Year 4 of the five-year i3 grant. At that point, 447 schools had begun operating the program as part of the scale-up.

Outreach: Recruitment Strategies

This section describes SFAF's planned recruitment strategies, use of financial incentives, and strategy shifts in response to recruitment progress.

During the 1990s, SFAF had added nearly 400 schools a year and had operated nearly 1,500 schools simultaneously, so its initial i3 expansion target of 150 to 300 new schools each year seemed possible. The i3 scale-up recruitment strategy was similar to what SFAF had used earlier, but with a financial incentive. The financial incentive (in the recruitment column of the logic model, Figure 7.1) was supposed to encourage and facilitate adoption. Based on past success and the promise of the subsidy, SFAF planned to recruit 1,100 schools under the i3 grant.

SFAF's recruitment message in the past typically emphasized the program's elements (such as tutoring), SFAF's own services (such as ongoing coaching over a three-year period), and evidence about the program's effectiveness in improving student performance. The organization also encouraged candidate schools to visit schools currently operating SFA, to see it in practice. For the i3 period, SFAF continued with these messages, and also emphasized the partial subsidy funded by the i3 grant as an incentive to adopt.

The organization planned to approach districts it had previously worked with and to recruit new districts. SFAF advertised in trade publications, used a marketing firm, and tried webinars to describe the program to potentially interested districts. The decision to focus on districts before approaching individual schools was in the interest of obtaining buy-in from the superintendent and aligning with district policy and reading achievement goals. One principal described this sort of process: "District leadership supported it. We were able to get our principals to buy into it right away — they were excited about it. Teachers were a bit of a different sell, but they got into it." SFAF approached at least 195 distinct districts. But the district approach was not yielding the numbers it desired, so midway through the grant, the organization decided to

[6] Determining the effect of the grant amount on recruitment would have required that schools be randomly assigned to receive different amounts; this was not a primary question of the research.

approach individual schools as well, with the recommendation that their staff visit schools already running the program. One principal said, “I sent all of the teachers to Steubenville district to observe SFA in action. I talked my superintendent into releasing money.”

The incentive SFAF had planned to offer was a partial subsidy, typically about $50,000. It usually covered materials and training costs, while a school’s aid from Title I, the federal funding stream designated for schools serving low-income students, would support the facilitator to provide ongoing professional development and support, as was the case in prior recruitment efforts. “In the past, we marketed the data” — improved outcomes from the program — said Nancy Madden, SFAF’s president. But for the 2010-2014 expansion period, she said, “we marketed the start-up grant, which was very attractive to people,” given some of the financial difficulties schools experienced. SFAF also planned to recruit schools that did not need the subsidy, reasoning that some schools would have sufficient school staff and/or other funding sources to support changes to their reading program.

The number of new schools recruited fell short of SFAF’s expectations. Schools cited cost as a significant barrier to adoption of the full program, even with the subsidy. SFAF had expected that only 50 percent of schools implementing the full kindergarten through fifth grade program would require a grant for adoption. Instead, 64 percent of full implementers required the grant (176 of 276 fully implementing schools), which meant that the grant did not go as far as SFAF had expected.

Shifts in Strategy

Altering its strategy, SFAF allowed schools to adopt components of the program, such as adopting only Reading Roots, for early grades, or only KinderCorner, for kindergarten students, rather than the whole program. The hope was that schools implementing one component of the program would adopt the rest of the program when funds became available. A school leader in Colorado said: “We loved the SFA program but the cost, even with a grant, was just too much for our district. We did adopt it for the Preschool Curiosity Corner, which has a much healthier budget.” Another school respondent noted that the school had “not gone to [implementing in] the other grade levels because we did not have the funds.” About 38 percent of schools overall adopted SFA for only some grades. About one-third of schools that adopted the program in 2011-2012 did so partially, and by 2013-2014 and 2014-2015, half of recruited schools adopted SFA in this way.

Some districts and schools decided not to participate, even with permitted adaptations. Between 9 and 23 districts per year decided against adopting. The study team approached a sample of these nonadopting districts or schools for interviews.[7] Interviewees at 9 of the 14 schools whose responses could be coded cited cost as the primary barrier. In some cases, too, district informants said they were satisfied enough with the supports their current programs provided (which typically included additional materials for struggling readers), and that schools were already providing tutoring and conducting frequent monitoring of student progress.

For the full sample of adopting and nonadopting districts, publicly available demographic data show that districts that never adopted SFA during the grant period tended to have a higher percentage of white students (46 percent compared with 33 percent among adopters) and were located more frequently in the Northeast and Midwest.[8] The adopting and nonadopting districts were similar to each other in terms of per-pupil spending, urban status, proportion of students with special needs, and student-to-staff ratio. But a test of differences across all district characteristics (rather than each individual characteristic) shows that the samples differ.[9] These differences may be more pronounced for individual schools that did and did not adopt, but spending and revenue data are not available for all states at the school level. (See Appendix Table F.1 for details.)[10]

Another shift in recruitment strategy occurred during Year 3 of the grant. Despite a relatively strong start recruiting schools in Year 1, the organization encountered difficulties in the next two years.[11] As Figure 7.2 shows, recruitment dipped to 57 schools in Year 3. In reaction to the slower-than-anticipated progress on recruitment, which was due in part to the recession and to other reading programs available to schools, SFAF and the i3 office conferred midway through the grant period. They agreed to revise initial targets for total school recruitment. Goals were revised downward to 760 from 1,100, and the number of schools targeted for recruitment

[7] Initially the study team planned for a random sample of 20 percent of nonadopters. However, many of these sites had so much staff turnover that no one at the district or school who had been involved in the adoption decision remained to be interviewed. The study team turned its focus to schools or districts where staff members were available, and as a result, the sample cannot be considered representative of all nonadopters.

[8] There is usually more within-district than between-district variation in school characteristics, so a comparison of district characteristics may mask school-level differences.

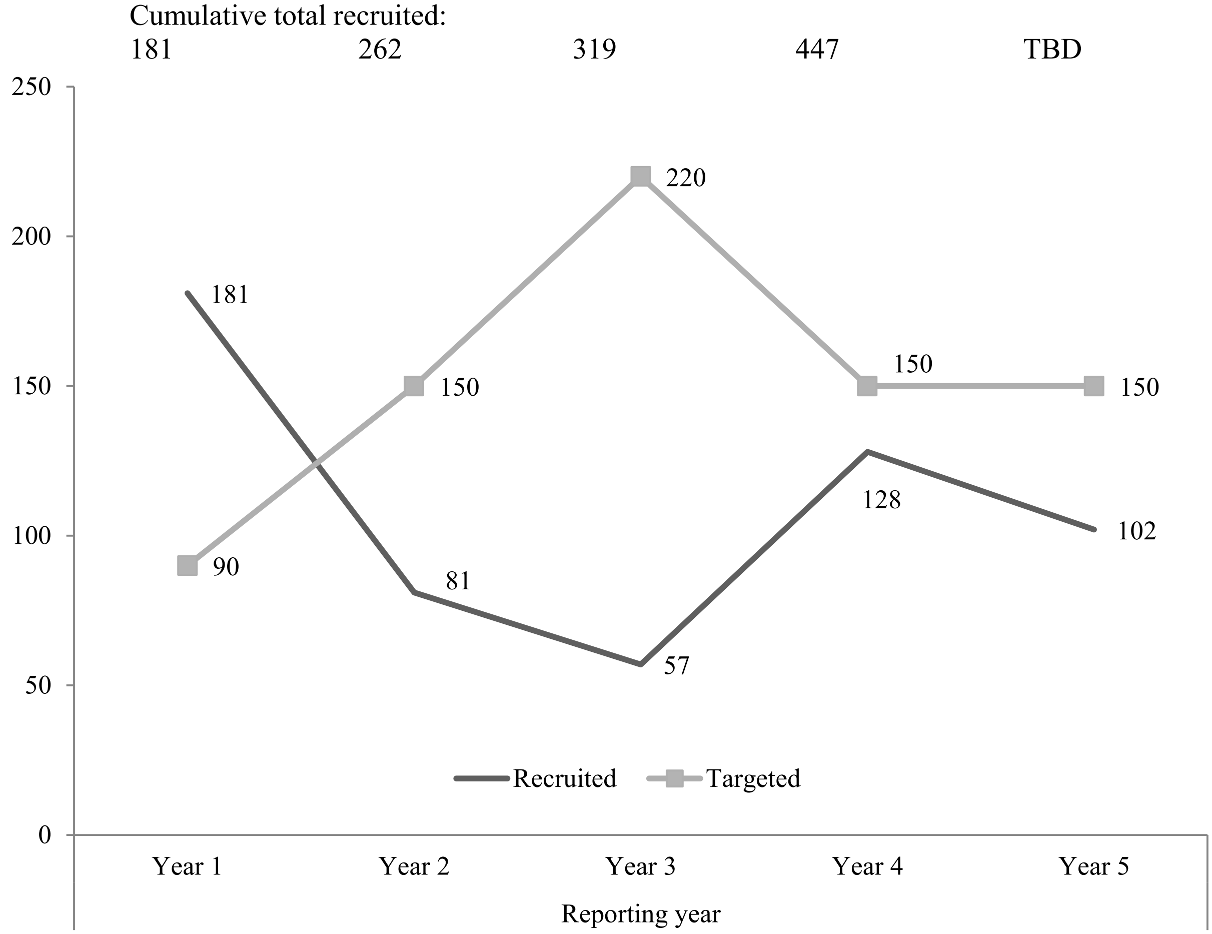
[9] In addition to testing for differences in each variable, an F-test for all district-level variables was conducted to see whether there were any overall differences in baseline district characteristics between the two groups. This test was based on a logistic regression, predicting sample status with the measured district-level baseline characteristics. The overall differences between samples were significant at $p < 0.001$.

[10] Because per-pupil expenditures are not readily available at the school level, but rather at the district level, it is not feasible to compare the exact difference in school-level resources for these samples.

[11] SFAF had begun recruiting schools in 2010, so when it received the grant in late 2010, it could count both 2010 and 2011 schools toward the Year 1 goal.

Figure 7.2

Number of Schools Targeted for SFA Scale-Up and Number Recruited



SOURCE: SFAF report on schools.

NOTES: SFAF received the Investing in Innovation (i3) grant in the fall of 2010. It had already begun recruiting and continued recruiting during the 2011 year. This explains why the total number recruited is greater than targeted in Year 1. Recruitment goals for Year 4 and Year 5 initially were 300 and 340, respectively, but were revised downward to 150 in the third year of the grant. The total number of schools includes the 19 program schools included in the evaluation sample. The number of schools recruited for Year 5 reflects the total as of May 31, 2015. TBD = to be determined.

in the last two years was adjusted to 150 from 300 in Year 4 and to 150 from 340 in Year 5, as shown in the logic model (Figure 7.1).[12]

Despite the difficult economic circumstances schools faced, SFAF was able to expand into 447 schools by 2014-2015 by offering a subsidy and allowing partial adoption. "To be hon-

[12] The last group of schools will start in the 2015-2016 school year.

est, what a lot of i3 funding enabled us to do was continue expansion in an otherwise catastrophic situation," said Robert Slavin, SFAF chairman of the board. Nonetheless, it appears almost certain that SFAF will fall short of its revised goal of 760 schools by the end of the grant period.

Scope and Efficiency: Expansion Strategies

This section describes specific expansion patterns and the extent of geographic concentration. The initial plan was that recruiting schools in a geographic cluster and having district coaches would ultimately reduce the number of times SFAF coaches would visit schools, thereby helping SFAF to achieve operational efficiency.

As noted above, after four years of recruiting, a total of 447 schools in 146 districts were operating SFA. By comparison, in the period immediately before the i3 grant began, SFAF was operating 868 schools in 313 districts. The four strategies described below built on the organization's prior operations and active locations, in part to achieve "cluster recruitment" or geographic concentration (listed in the recruitment column of the logic model, Figure 7.1). None of the four strategies dominated the expansion effort.

1. *Adding new districts, ideally near existing districts:* The organization added 76 districts in addition to having 70 existing districts continue into the i3 period.

2. *Adding new schools in existing districts:* SFAF added 116 new schools in existing districts (26 percent of total recruited schools).

3. *Adding new schools in new districts later in the grant period:* SFAF added 171 new schools this way (38 percent of recruited schools).

4. *Rerecruiting schools* that were active immediately before the grant period in existing districts, to continue with SFA in other grades or for a longer period. The organization rerecruited 160 schools (36 percent of schools).

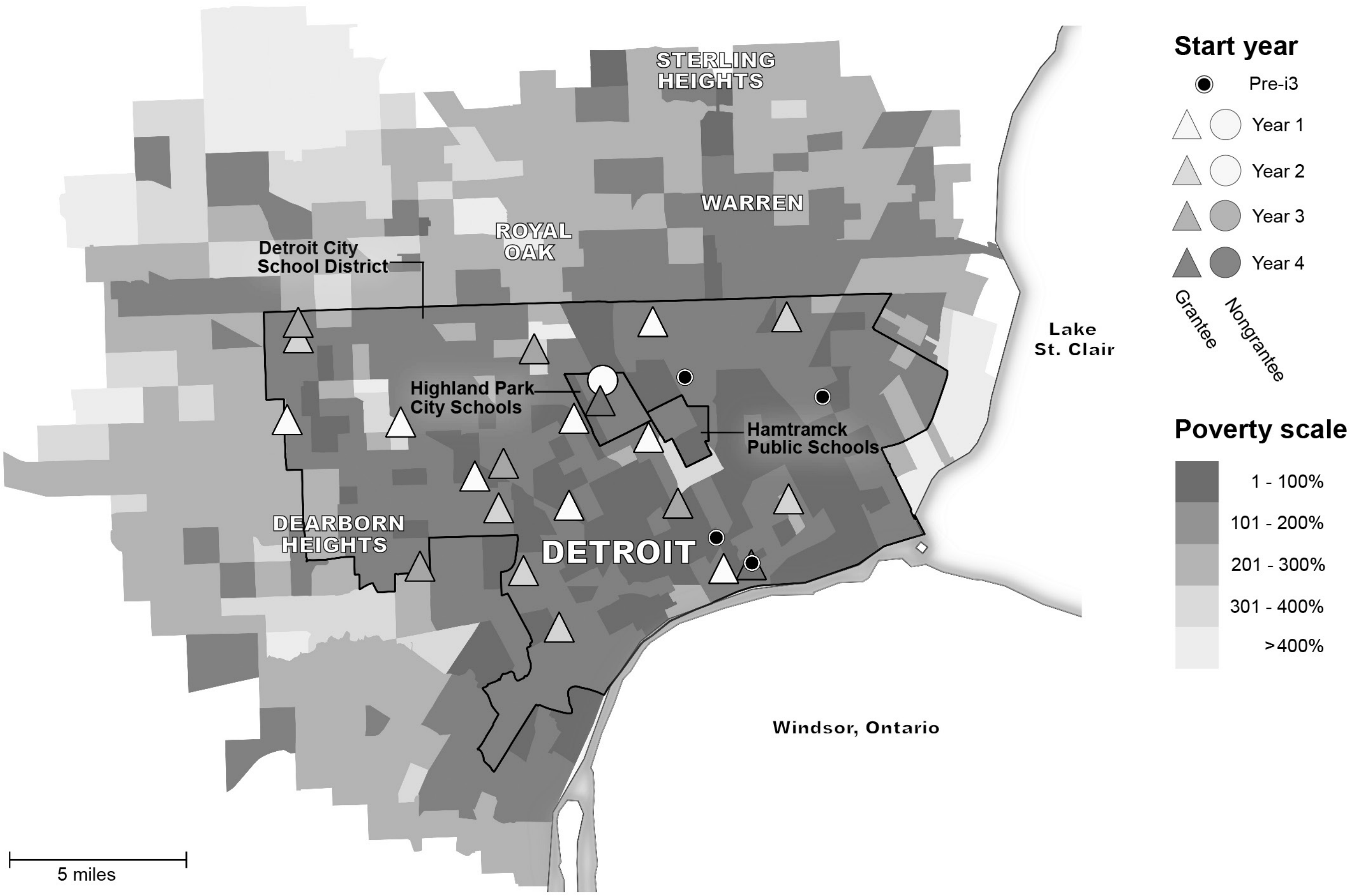
To illustrate the different expansion patterns described above, maps in Figures 7.3, 7.4, and 7.5 show examples of the approaches.

Gains and Losses

The 76 *new* districts included large urban districts such as Detroit City School District, as well as smaller districts with just one to five new schools. In Figure 7.3, Detroit represents strategies (1) and (3): It was a new district that added new schools in each year of the grant peri-

Figure 7.3

SFA Scale-Up Recruitment Pattern, by Grant Receipt, Start Year, and Neighborhood Poverty Level: Detroit



SOURCE: Income and poverty data are from the 2010 U.S. Census.

NOTE: Poverty scale is calculated as median household income relative to the national poverty threshold, $22,315. Poverty rates are displayed in census tracts within four miles of an SFA school.

od. However, all but one of the schools in that district were recruited with the financial incentive, an example of how the grant was used for more schools than expected (triangles indicate a grantee school and darkly shaded triangles indicate later years). The 70 *existing* districts also included large urban and smaller suburban districts. In Figure 7.4, Atlanta demonstrates strategies (4) and (1): It was an existing district in which schools were rerecruited (see circles beneath black dots, which indicate new schools in the same location as previously operating schools) and where a nearby district recruited all schools with a grant (shown with triangles). Figure 7.5 shows an example of the expansion SFAF leaders hoped for, with all four strategies in play: Steubenville, Ohio, and its neighboring districts. The district had previously operated SFA, it added new schools to the ones that had previously operated the program, nearby districts that had not operated SFA were recruited, and these districts added new schools in each of the i3 grant years.[13] This combination was rare.[14]

What is not evident in the maps is where SFAF lost the geographic concentration it had had immediately before the i3 period. For example, Kansas City, Missouri, had 33 schools active just before the i3 period and no schools during the grant period. Similarly, Reading and Harrisburg, Pennsylvania, had about 20 schools active in the pre-i3 period and none in the grant period.[15] In Massachusetts, five districts previously operating SFA did not continue with the program; for example, Lawrence had 16 schools in the pre-i3 period and none after, and neighboring districts also stopped operating SFA.[16] The loss of some schools is to be expected; however, the loss of entire districts in which none of the schools were rerecruited meant that SFAF had to identify new districts and recruit elsewhere. In such cases, SFAF was not able to achieve the savings in staff time that comes from recruiting within existing districts and benefiting from existing relationships.

The upshot of the gains and losses is that SFAF maintained a similar geographic concentration of schools within districts during the grant period as before. The proportion of districts that had just one or two schools was similar both before and during the grant — 77 percent and 70 percent, respectively. The proportion of districts with five or more schools also stayed similar: 14 percent and 17 percent, respectively. About 54 percent of schools were recruited in a

[13] Notably, too, the region had a mix of schools adopting with and without a grant (triangles and circles).

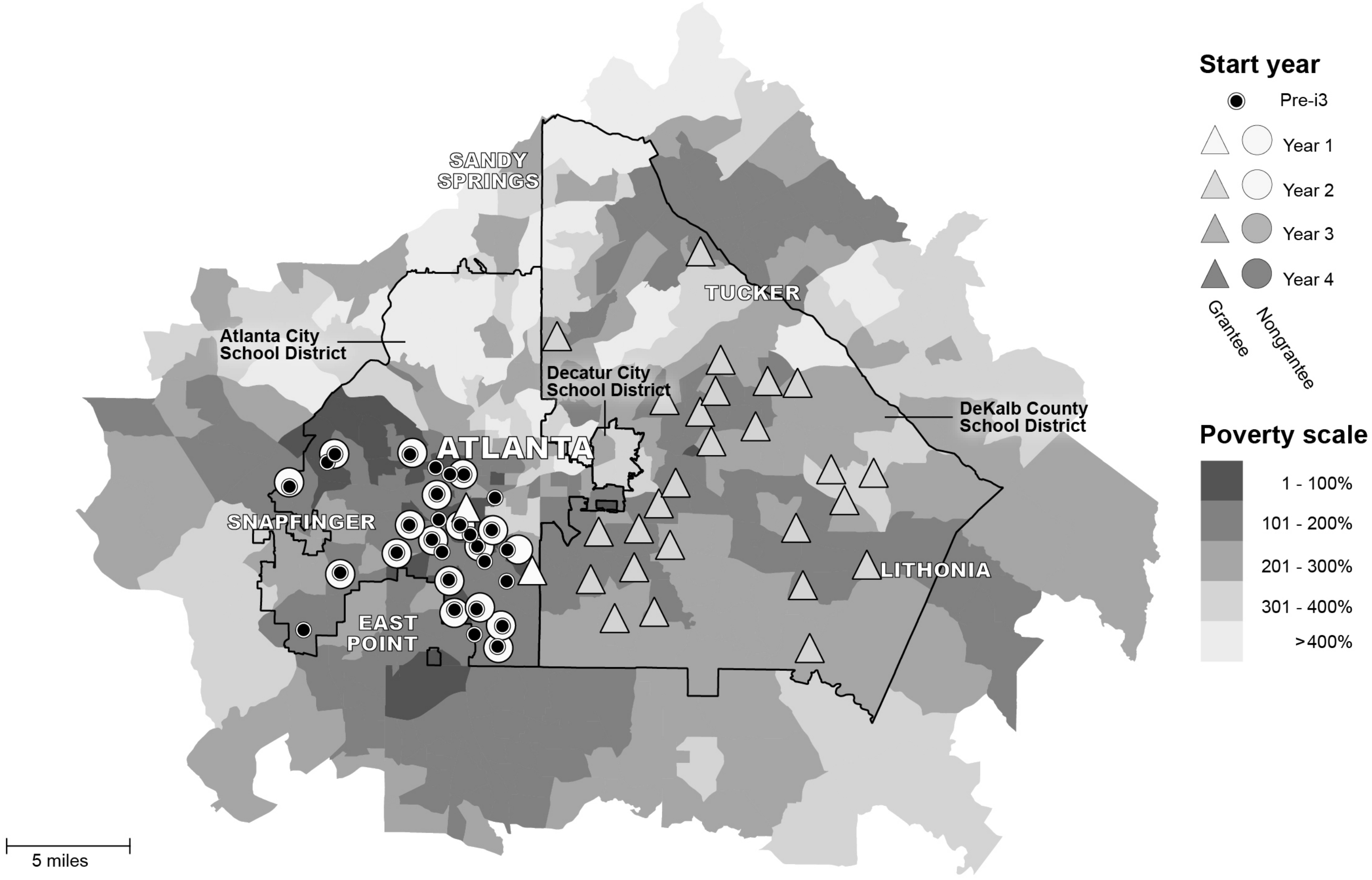
[14] Maps do not show exact district boundaries and zoom out in an effort to protect individual school identities.

[15] However, six pre-i3 districts in Pennsylvania were retained during the i3 period. Four states (Hawaii, Montana, North Dakota, and Wisconsin) had no new schools adopting SFA in the i3 period, and in both Alaska and Minnesota only one new school across previously operating districts joined SFA.

[16] Appendix Table F.3 shows a comparison of schools operating SFA immediately before the i3 grant period and those adopting it during the i3 period. Schools in the i3 period differed across a set of typically examined characteristics, as well as in terms of per-pupil spending. Data about the length of SFA implementation in the pre-grant period were not readily available for all schools.

Figure 7.4

SFA Scale-Up Recruitment Pattern, by Grant Receipt, Start Year, and Neighborhood Poverty Level: Atlanta

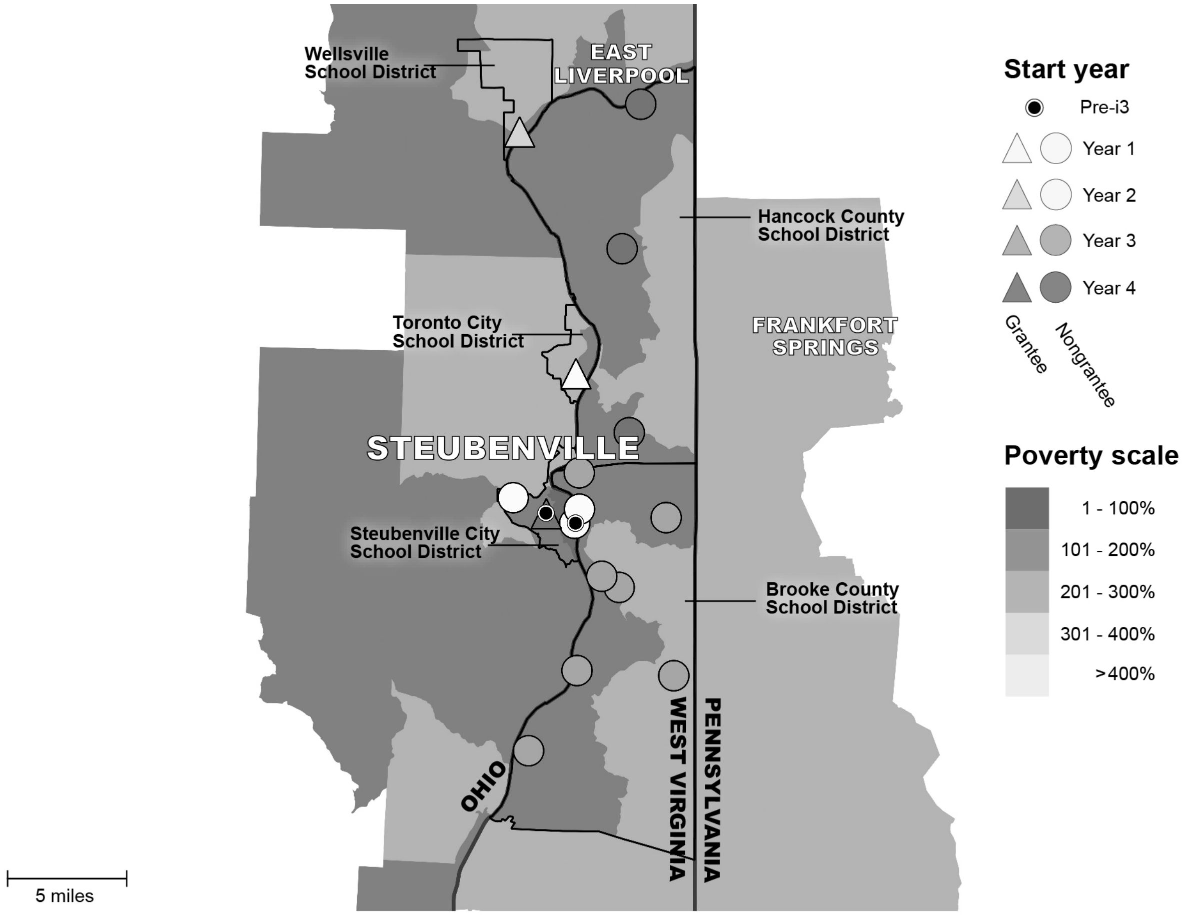


SOURCE: Income and poverty data are from the 2010 U.S. Census.

NOTE: Poverty scale is calculated as median household income relative to the national poverty threshold, $22,315. Poverty rates are displayed in census tracts within four miles of an SFA school.

Figure 7.5

SFA Scale-Up Recruitment Pattern, by Grant Receipt, Start Year, and Neighborhood Poverty Level: Steubenville



SOURCE: Income and poverty data are from the 2010 U.S. Census.

NOTE: Poverty scale is calculated as median household income relative to the national poverty threshold, $22,315. Poverty rates are displayed in census tracts within four miles of an SFA school.

"cluster," or near other schools. However, the cluster recruitment goal of enlisting neighboring *districts* to work together and create regions across districts was not realized. "We find districts are competitive and don't like to share resources. So the concept of reducing costs across district lines didn't work," said Nancy Madden, SFAF's president.

School Support Mechanisms

This section describes the support mechanisms SFAF used for schools that adopted the program, whether scale-up schools achieved fidelity to the program comparable to that of evaluation schools, and factors that may have limited both recruitment and implementation.

Initially, SFAF designed approaches it hoped would allow the organization to support schools on a larger scale more readily and more efficiently. These included providing remote coaching and support, developing and sharing a digital library, and identifying and hiring local coaches. SFAF developed a suite of remote tools to increase its support of schools that adopted the program and made videos demonstrating different components of the program. SFAF coaches joined most teacher support meetings by phone, in between their school visits. The goal was to use SFAF coaches' time on site to address schoolwide systems, and address classroom instructional practices more often via remote, online support.[17] The local coach aspect of the project operated for three of the intended five years. The next section discusses how this decentralization of the coaching function worked during its trial, why it was curtailed, and what replaced it.

Plans for Local Coaches

To create more of a local support system, SFAF had proposed to districts the idea of appointing a local coach. This would be a district position to help schools to implement SFA with fidelity. A coach based locally and funded by the district, SFAF leaders reasoned, would have more knowledge of the local context and could spend more time with the schools rather than traveling. In addition, local coaches and exemplar schools could help recruit neighboring institutions to adopt the program. In the first year, the local coach would shadow the point coach from SFAF headquarters and would learn key aspects of the program. In subsequent years, SFAF hoped, the point coach would be able to transfer coaching and support responsibilities to the local coach. In this way, SFAF would increase its ability to serve more schools, while a number of schools within a district using SFA could form a network, with someone who knew the schools' particular needs available to model effective instruction.

[17] The study does not have data on usage other than access to the website, so it is not possible to report on frequency or content of online support.

SFAF worked with 11 to 15 districts, of the 146 districts recruited during the grant period, to appoint and train a local coach, and about 40 percent of scale-up schools had a local coach at some point. These coaches reported that they were making progress with schools in terms of launching Solutions Teams, improving data use, and modeling instruction. Reported one district coach:

> I tried to work with [on-site] facilitators on being stronger coaches. My role with the principals has changed [over time]. I do a lot more walk-throughs, observations, a lot more data analysis. . . . Last year and this year, we pull reports from the SFA site, target interventions based on the data, with the principal present at this conversation. Data analysis is a big thing that has changed in the past two years.

Another local coach noted progress in both systems and instructional areas of SFA:

> We had two component leads each at [the] Kinder, Roots, Wings [levels]. . . . That's where my coaching time happened this year. [We were] building their capacity [and] started to get some momentum. We did the same thing with the Solutions committees. Teachers step[ped] up to be chairs of these committees. We really made some headway earlier this year.

Challenges Associated with the Local Coach Strategy

The use of local coaches also had a number of problems, however. First, it was difficult for districts to identify someone with the appropriate skills and time who could handle the responsibility. Coaches who had experience running schools or serving as a reading coach were preferred, but they did not get to the point of proficiency at which the SFAF coaches could completely transfer responsibilities to them, according to SFAF leaders. Second, local coaches often had to juggle coaching SFA schools with other district responsibilities, including supporting other Title I programs. According to SFAF contracts that allocated time for local coaches to spend on SFA, one coach spent at least 46 percent of her time on non-SFA duties by Year 3. And in some cases, local coaches reported filling gaps in school staffing, such as performing some of the school-based facilitator's duties because the facilitator herself was playing multiple roles at the school. Third, districts faced challenges to fund the position, given the budget cuts to nonessential positions that districts were making during the grant period. The coach structure did not last for the duration of the grant period.

The promise of the decentralized coaching model for reducing costs also was premised on operating in districts or geographic areas with a substantial concentration of SFA schools. As the previous section noted, not enough of the schools that joined were geographically clustered in this manner for the model to encourage districts to appoint local coaches. Because of this, neither the districts nor SFAF achieved the increased efficiency that was desired.

Given the challenges SFAF had in recruiting its target number of schools and the challenges schools faced in funding this local coach position, in Year 3 of the five-year grant, the federal i3 office required SFAF to discontinue funding for it.[18] Because so many schools required financial support to adopt SFA, the i3 office, according to chairman of the board Robert Slavin, wanted SFAF to redirect its resources toward school recruitment and away from support of existing SFA schools.

Without the local coach structure, point coach visits and on-site team visits continued, along with remote, online support. The SFAF coaches were not able to reduce the number of visits they made to sites, so the coaching and support aspect of SFAF's expansion strategy did not achieve the anticipated efficiency. "Our intention was to shift to digital support. We find that increases quality of implementation. [But we] haven't been able to reduce on-site support, which is interesting," said Nancy Madden of SFAF.

How Well Did the Scale-Up Sites Implement the Program?

This section describes to what degree scale-up sites implemented the full program, and whether scale-up sites outside the evaluation achieved fidelity comparable to that of the 19 evaluation sites described in Chapter 3. It concludes by describing factors that facilitated or impeded SFAF's ability to scale up the program.

Before discussing the quality of implementation, it is worth noting that about one-fourth of scale-up sites that initially agreed to implement the program for three years dropped the program before then. As Figure 7.6 shows, 100 schools started SFA by 2011-2012 or 2012-2013 but discontinued the program early; their average length of implementation was 1.35 years.[19] In comparison, all 19 SFA schools in the evaluation sample also enrolled in 2011-2012 and completed all three years, perhaps because the program was provided to them for free.

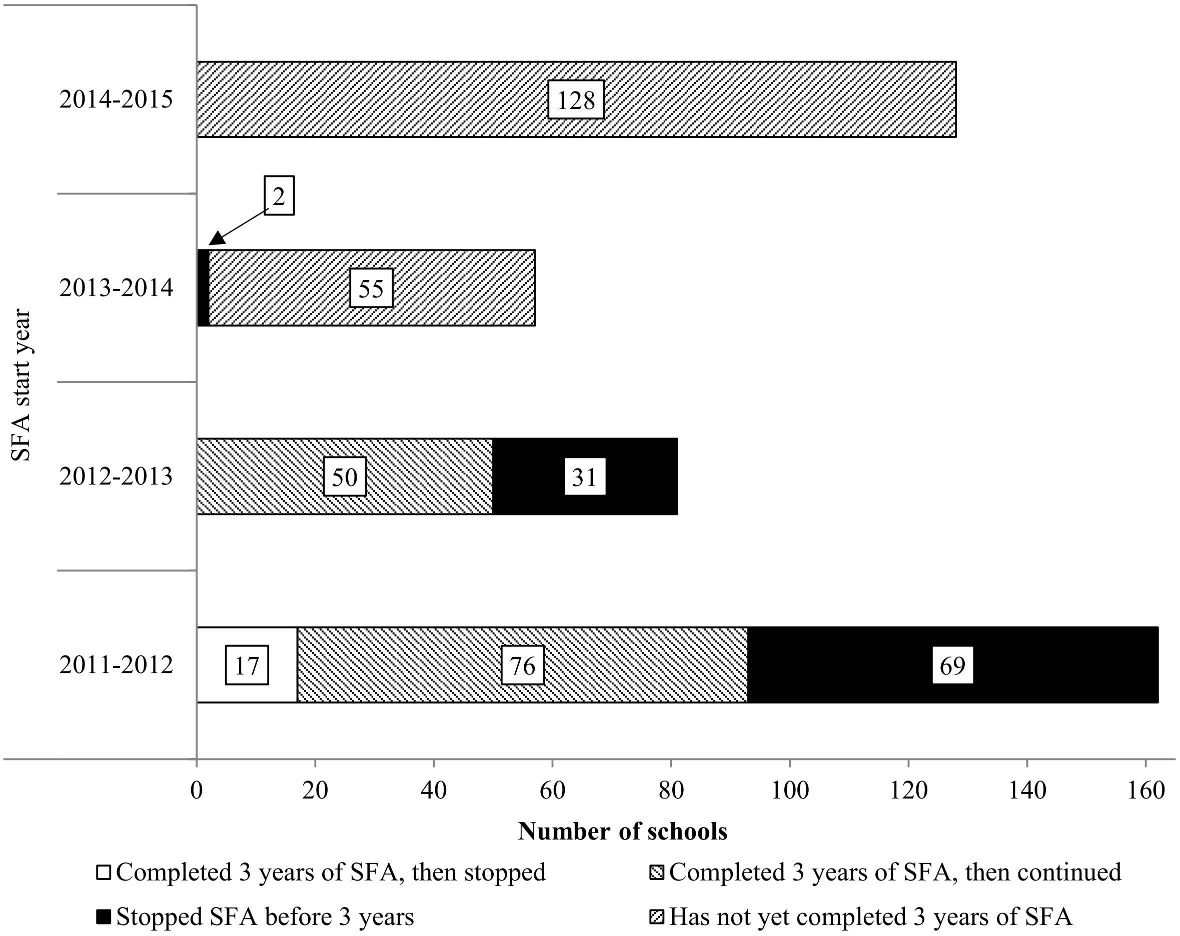
Using information from the School Achievement Snapshot forms completed for the schools that continued to operate SFA for three years, the quality of implementation at the scale-up schools can be gauged and compared with implementation at the evaluation sample schools. Among scale-up and evaluation sample schools that were comparable in a number of respects — they began in the same year, implemented the program in all grades, completed

[18] Districts that could fund the coach position on their own retained the local coach. At least six districts eliminated the district coach position due to funding cuts in their district or lack of i3 funding support.

[19] The study team was not able to reach an informant at these schools to determine the reasons for early withdrawal, often because of staff turnover. Implementation score (Snapshot) data were not available for schools that did not complete the three years, so it is not possible to determine whether schools that exited before completion scored lower, and if so, whether that was related to the decision to stop the program. Of 162 schools that started the program in 2011-2012, 69 schools did not complete it. An additional 31 schools that started in 2012-2013 did not complete it.

Figure 7.6

Number of Scale-Up Schools by Completion Status at the End of 2014-2015, by SFA Start Year



SOURCE: SFAF report on schools.

NOTES: Under the i3 grant, schools could receive subsidized implementation assistance for three years. Therefore, completion refers to whether or not a school implemented SFA for three years. Only schools that started in 2011-2012 or 2012-2013 could be verified as having completed three years of SFA. The SFAF report on schools data set does not contain SFA end dates later than 2013-2014, because information was not available to the foundation at the time this report was written. It is also unknown whether schools are continuing with SFA after 2014-2015. Therefore, the category "Has not yet completed 3 years of SFA" includes schools that started SFA after 2012-2013 and completed one or two years of SFA through 2014-2015, but whose status with respect to SFA after 2014-2015 is unknown.

This table does not include the 19 evaluation schools that started in 2011-2012, because they all completed three years but under supervision.

The counts represent schools only through the fourth year of the grant.

three years, and received a grant — the analysis calculated scores from the 2013-2014 Snapshot, which corresponds to the third year of the program for both samples.[20] The research team examined fidelity scores on the components described in Chapter 3: continuous improvement, noninstructional components, and the use of challenging instructional processes. The team also considered whether schools in the scale-up sample had received essential training, had a full-time facilitator, had a principal who was judged to be fully involved, and had access to a district coach — all factors hypothesized to be related to greater implementation quality.

Comparing the evaluation sample and scale-up schools with similar characteristics, the analysis found the following:

- At a higher proportion of scale-up schools than evaluation sample schools, all staff received essential training.
- A higher proportion of scale-up schools than evaluation sample schools had the capacity to provide tutoring.
- A lower proportion of scale-up schools than evaluation schools had access to a local coach (about 65 percent and 80 percent, respectively).
- A comparable percentage had a fully involved principal (89 percent in both samples) and full-time facilitator (84 percent of scale-up schools and 89 percent of evaluation schools).
- For schools in the scale-up sample, SFAF coaches focused on the program's structural aspects while they were on site and provided instructional support via remote coaching, whereas evaluation schools received on-site support from SFAF coaches on all aspects of the program.

On average, the schools in the scale-up sample attained lower overall implementation scores than the evaluation sample did.[21] Both scale-up sample and evaluation sample schools met the criterion for overall adequacy of implementation, defined as 80 percent of schools attaining at least 50 percent of the maximum implementation score, with 81 percent of scale-up and 89 percent of evaluation schools reaching this standard. In terms of scores and components, however, scale-up sample schools attained 66 percent of the maximum possible score, compared with 75 percent for the evaluation sample. In particular, scores on noninstructional com-

[20] As previously noted, scale-up schools did not receive as large a grant as evaluation sites did. Appendix Figure F.2 describes how the study team arrived at this analytic sample for the scale-up sample (37 schools) to ensure comparability with the evaluation sample (19 schools).

[21] See Chapter 3 for a description of the scoring approach.

ponents as well as challenging instruction were significantly lower for the scale-up schools.[22] It is notable that schools that had a district coach as well as an SFAF coach did not attain significantly higher scores than schools with only an SFAF coach.[23]

Because this comparison of samples is limited to schools that completed three years of implementation, inferences should be made with caution. The generalizability may be limited to schools that completed the program. Furthermore, implementation scores were available only for scale-up schools that started in 2011-2012 or 2012-2013, which may have faced different constraints and therefore implemented the program differently from schools that started later.

Limiting Factors: The Economic and Policy Environment

A combination of policy and economic changes coincided with SFAF's expansion, creating a set of barriers that made this expansion different from earlier iterations. The Great Recession shrank federal and state revenues, affecting Title I funds. As noted in Chapter 1, short-term federal stimulus funds,[24] intended to limit the effects of the recession from 2009-2011, had expired by the time SFA approached schools for its first year of recruitment. Many districts eliminated teaching and instructional support positions in anticipation of the expiration of funding. State budgets approved late in the school year, and federal and state funding uncertainty, limited schools' ability to counter staff turnover or other staffing instability. This meant that SFAF was recruiting schools to take on a labor-intensive program at precisely the time when schools felt short-staffed.

In addition, in Year 3 of the grant (fiscal year 2013), Title I funding decreased from \$14.52 billion to \$13.76 billion, affecting all of the nation's approximately 33,000 Title I schools, the very schools that SFAF recruits.[25] Based on reports on the Great Recession and

[22] Scale-up schools attained 63 percent of the maximum score on noninstructional components and challenging instruction, compared with 79 percent and 73 percent, respectively, for the evaluation schools. One hypothesis might be that evaluation sites focused on implementation more because they were being studied by the research team with ongoing data collection, while schools in the larger scale-up sample did not receive as much support or observation.

[23] Fidelity, or implementation ratings, could vary by whether the scale-up mechanisms SFAF designed were in place for certain schools. However, the analysis finds no significant difference between schools that did or did not share the following characteristics: scale-up cohort (whether the school began implementing in 2011-2012, as did schools in the evaluation sample); presence of a local district coach for that school; grant/subsidy receipt; and cluster recruitment (whether the school was recruited as part of a cluster or not). On these dimensions, schools on average attained implementation scores between 55 percent and 66 percent of the maximum possible score, but the differences were not statistically significant. This is true for the 37 scale-up schools used in comparison with the evaluation sample, as well as for the broader sample of 300 scale-up schools with fidelity scores.

[24] U.S. Department of Education (2009).

[25] U.S. Department of Education (2015b).

study team interviews,[26] budget cuts prompted schools to use their funds in different ways. Title I is typically supposed to be used for staff who provide supplementary support services for students in need, not for core instruction. Schools that operated SFA in the past typically used Title I funds to support the full-time facilitator position. Now some schools, facing cuts in their general revenues, started to use Title I to support core teaching staff time. This limited the funds available to support a full-time facilitator. "Finding a facilitator was a difficult cost issue. It ended up increasing the cost of the program (for schools)," said Nancy Madden. While the school-based facilitator had always been part of the SFA model, in the postrecession environment, shifting support costs to the school decreased demand for the program.

In addition to the funding challenges, many schools felt a sense of reform overload. With demands of Common Core-related curriculum alignment and teacher training starting in some states in 2012, some schools reported limited capacity to take on a new program. For example, the principal of a school in New Mexico that opted not to adopt SFA reported that the school's number one priority was alignment with the Common Core State Standards. This was echoed by some informants at other schools as well, though it was not the primary reason they cited for not adopting the program. "It was just not a very receptive marketplace," said Madden.

Conclusion

The story of scaling up SFA was largely about contextual factors creating barriers to expansion. In the i3 period, these contextual changes included the advent of the Common Core State Standards (policy), the recession's aftermath effect on revenues and curtailed staffing (economy), and the presence of increased competition from commercial reading programs that include some version of reading intervention supports, such as the RtI reading framework (market). Moreover, major sources of federal funding that schools regularly received during the study period, such as Title I, did not require schools to adopt a program with proven success at boosting early reading outcomes.[27] In the absence of such requirements, local decision-makers seemed to address immediate priorities (staying within the budget) rather than longer-term potential (a possible increase in student performance).

On the one hand, in the face of these competing reforms and considerable economic challenges, SFAF succeeded in reaching a substantial number of new schools. It accomplished this by using a financial incentive to encourage adoption (delivering coaching and comprehensive materials at reduced cost) and expanding to new districts and schools while maintaining relationships with existing districts and schools.

[26] Oliff and Leachman (2011).

[27] Legislation under consideration in 2015 may require that Title I and School Improvement Grant recipients adopt proven evidence-based programs, such as Success for All.

On the other hand, despite SFA's considerable expansion — to 447 schools by the end of the fourth year of the demonstration and, according to SFAF, to an estimated 540 schools by the grant's end — the expansion was less extensive than the organization had hoped, based on expectations from its major expansion in the 1990s. This shortfall occurred despite rerecruiting schools that had previously operated the program and recruiting schools that implemented the program for fewer grade levels than the program was intended to serve and over a shorter time period.

It is impossible to determine whether other strategies might have eased recruitment. For example, SFAF required schools to support the full-time on-site facilitator's position (as they had done in past expansions as well). Had SFAF itself (via the i3 grant) footed this bill, more schools might have been able to adopt the program. But doing this would have entailed a trade-off: Under this hypothetical approach, SFAF would have had to spend more money per school, so the total number of schools that the organization proposed to include in the i3 scale-up would have had to be reduced. And a revised strategy to fund the facilitator position would not have addressed another problem the organization confronted: overcoming schools' reform fatigue.

The chairman of the SFAF board, Robert Slavin, cited some lessons. It could be easier to expand a more discrete program, as opposed to a whole-school program that requires buy-in at multiple levels (from district leadership to teachers) and substantial upfront investments such as a full-time program facilitator. Relative to what schools were already doing to support struggling readers, SFA did not appear distinct enough to some schools unfamiliar with the program to warrant switching, according to interviews with nonadopting districts and schools. To this end, SFAF leaders said they learned that word of mouth, rather than specific expansion staff or a marketing strategy, may still be the best way to engage new schools unfamiliar with the program.

The mismatch between the expansion targets and actual recruitment may reflect the difficulty any program faces in accurately forecasting and anticipating how multiple changes in context may converge. The ambitious targets may also reflect a paradox of seeking public funding for an expansion: To present a compelling case for funding, programs set targets high; yet regardless of whether targets are set too low or too high, the programs may be judged by whether they meet their initial forecast, not by how well they adapt to new contexts. The SFA experience suggests that both programs seeking to expand and the funders of these efforts need to take context more fully into account in considering expansion targets and strategies.

Chapter 8

Looking Backward, Looking Forward

When Success for All (SFA) was first introduced in the late 1980s, the program was unique in its combination of program elements: a strong emphasis on phonics, extensive use of cooperative learning methods, ability grouping of students across grade levels, inclusion of whole-school and family support components, and heavy reliance on data-driven decision-making. A number of earlier studies point to the effectiveness of SFA's approach to the teaching of beginning readers. Today, an observer in an SFA classroom would continue to see reading instruction that looks quite different from instruction in non-SFA schools: Teachers follow highly structured lesson plans, while students work in pairs or small groups during almost every reading class. Still, over time, non-SFA schools have increasingly come to resemble schools operating SFA ("SFA schools"), adopting a phonics-based approach to beginning reading, for example, and incorporating outreach to students' families. With the narrowing of the gap between SFA schools and other schools, it is important to assess whether SFA continues to make a difference vis-à-vis other reading programs.[1]

At the outset of this study of Success for All, funded under the U.S. Department of Education's Investing in Innovation (i3) competition, the evaluators were required to state their "confirmatory" question — the answer to which would be used to judge the effectiveness of the intervention in turning around low-performing schools. The confirmatory question guiding this evaluation is:

> *Compared with non-SFA schools, what is the impact of SFA on students' reading scores in the areas of alphabetics, comprehension, and fluency at the end of the third year following the start of SFA?*

Other analyses — of impacts for subgroups of this main sample, for students with differing amounts of exposure to the program, for students in other grades — are considered "exploratory," that is, of secondary importance. The sample used to answer the confirmatory question consists of second-graders at the 19 program group and 18 control group schools who were enrolled at these schools from kindergarten on.[2] In the program group schools, these are the students with maximum exposure to the program; they learned to read "the SFA way."

[1] See, for example, Lemons, Fuchs, Gilbert, and Fuchs (2014) for another consideration of the changing counterfactual in educational research. Please see Chapter 4 for additional information on how SFA and non-SFA schools are similar to or different from each other.

[2] To be more precise: As previously noted, these students were included in the primary analysis sample if they had both baseline test scores and at least one valid spring test score for each of the three years of the study.

For this sample, the evaluation yielded one statistically significant impact, on one of two measures of students' phonetic skills. Knowledge of phonics is an indisputably important element of reading, but it is also acquired relatively early. By the time students are in second grade, they should ideally be reading with some degree of fluency and should understand what they read. Students in SFA schools did about the same as those in control group schools on two tests measuring these more advanced reading skills. This is not to say that SFA didn't work to inculcate these skills — only that, for the sample as a whole, it did not work better than other reading programs.

This is one instance, however, in which the exploratory findings may have as much importance for policymakers and practitioners as the confirmatory results. The subgroup findings indicate that second-grade students in SFA schools who entered kindergarten knowing few letters and words did significantly better than their control group counterparts on both of the measures of phonics skills and were able to read words with greater fluency. Their scores on a measure of comprehension were also higher, although the impact was not large enough to meet established standards of statistical significance. The program did not have comparable effects on students who began kindergarten with higher baseline skills.

It is possible to speculate about, but not to say with certainty, what aspects of SFA made the program more effective for lower-skilled students. What is most distinctive about SFA remains its instructional core. SFA's instructional videos may teach phonics to beginning readers in a particularly engaging, lively way. And it may be, too, that the SFA practice of grouping students with lower skills in a single classroom allows teachers to focus on their needs more effectively than would be the case in more heterogeneous classrooms, where greater differentiation of instruction would be required.

The impact results suggest that, for schools whose students arrive in kindergarten with some early literacy skills — knowing a number of letters of the alphabet and recognizing some words (either through sounding them out or by sight) — the choice of reading program may not matter much. By the end of second grade, these students are likely to be able to learn basic reading skills whatever reading program their schools adopt, especially as most reading programs have come to place a greater emphasis on phonemic awareness and phonics in the early grades.

But for schools in which many students start out behind their age peers in these literacy skills, the reading program does matter — and Success for All continues to beat the competition. The subgroup finding may be especially important for Title I schools serving large numbers of low-income students — the very schools that SFA targets — given the high concentration of poor readers among students eligible for free- or reduced-price lunches.[3]

In debating whether to adopt or continue with Success for All (or any reading program), schools and school districts will obviously need to take into consideration the cost of the intervention. As Chapter 6 shows, SFA's incremental cost, when compared with the cost of the business-as-usual reading program in a representative district, is relatively modest: an additional $119 per student per year in direct expenses, or $227 per student per year in terms of total resources needed to mount the program.

As Chapter 3 documents, schools got better at operating SFA over the years, and the large majority of schools implemented the program with fidelity. Thus, the SFA model can be considered to have received a fair test. Nonetheless, there remained room for improvement, and it's reasonable to speculate that if implementation had been stronger, effects might have been larger. The Student Achievement Snapshot, teacher and principal surveys, interviews, and other data point to a number of issues during the third year of operations at the evaluation sites: insufficient training for new staff, reduced adherence to the model (perhaps implying a need for stronger monitoring), and incomplete implementation of the program's tutoring intervention.

Training (and retraining) would appear to be especially important when there is considerable staff turnover. The evaluation does not have a reliable measure of teacher turnover, but as reported in Chapter 4, in Year 3, 24 percent of program group teachers but only 12 percent of control group teachers were new to their schools.[4] According to the teacher survey, however, almost a third of the new SFA teachers reported either that they had not received professional development on how to implement their school's reading program or that the professional development they did get was not helpful to them. Although this figure is much lower than the comparable proportion for new teachers in control group schools (46 percent), the need for extensive and high-quality professional development is arguably greater when the reading program is as highly structured, fast-paced, and demanding as SFA.

The large influx of teachers without previous exposure to SFA raises questions about their ability to implement the program with fidelity. The survey data indicate that, far more than in previous years, SFA teachers made their own changes to the program. In the third year, over half the teachers (54 percent) acknowledged that they changed parts of the reading program that, in their opinion, do not work for students.[5] The Success for All Foundation (SFAF) en-

[3] As indicated in Chapter 5, the scores of students in grades 3-5 in the evaluation schools on the Gates-MacGinitie test ranged between the 26th and the 30th percentiles nationally.

[4] The difference was apparent in three of the five study districts, including the one with the largest number of schools.

[5] In the first and second years, the proportions were 26 percent and 45 percent, respectively. While the wording of the item in the first two years was somewhat different — teachers were asked whether they

courages teachers, once they have mastered the program and understand it thoroughly, to change it in ways that will better meet their students' needs. But it was teachers who were new to their schools, rather than those who had been there longer, who were more likely to say that they modified the program (67 percent and 52 percent, respectively). And teachers who were new to teaching altogether were even more likely to alter the SFA lesson plans (74 percent and 53 percent).

It may well be that those teachers who do not adequately grasp the reasons why the SFA reading program is structured as it is may feel entitled to change it. But unless the changes they make are just as effective as SFA's scripted curriculum, extensive departures from the script may reduce the program's impacts. While the SFA facilitators at program schools have much to do, they may need to give special attention to monitoring the reading instruction of teachers who are new to their schools and for whom SFA may pose special challenges.

In addition to not understanding the program, a second reason that teachers may have deviated from the program more in the third and final year of the i3 demonstration is that they were uncertain whether their schools would continue with SFA after the demonstration ended. The evaluation did not probe this issue. It is notable, however, that morale was considerably lower in SFA schools than in control group schools: According to the teacher survey, only 47 percent of the SFA teachers, compared with 74 percent of the control group teachers, agreed that teacher morale at their school had been high since the start of the school year.[6]

Along with the findings about high turnover and incomplete adherence to the program, the implementation analysis also indicates that straitened economic circumstances prevented SFA's tutoring component from being implemented at ten of the program schools in Year 3. According to interviews with principals and to their survey responses, these schools could not afford to pay teacher-tutors for the extra hours of work that the tutoring component required.[7] As a consequence, many lower-skilled students who might have benefited from tutoring that was closely aligned with their regular reading program did without it. Some of these students received extra assistance through Response to Intervention initiatives at their schools, but so did their counterparts at the control group schools; in fact, according to principal reports, the pro-

changed parts of the reading program that they didn't like or disagreed with — the overall end result is a sharp increase in the proportion of teachers who reported modifying the program as they saw fit.

In qualitative interviews with teachers of English language learners (ELLs) in several schools in one district, the teachers said that they generously supplemented the SFA ELL curriculum or replaced it with the materials that were used in the control group schools in the district. Some teachers also complained that they found errors in the SFA products, perhaps because the curriculum used terms unfamiliar to speakers of Mexican Spanish.

[6] This question was not asked on the previous teacher surveys, so whether teacher morale declined over time cannot be ascertained.

[7] The amount of tutoring provided in the district where the cost analysis was conducted was exceptional in this regard.

portions of second-graders who received tutoring during the 2013-2014 school year were statistically indistinguishable in program group and control group schools (16 percent and 21 percent, respectively). Had more students received SFA tutoring, as called for by the program model, it seems reasonable to think that they would have registered higher scores on the assessments used to measure program impacts.

Along with implementation issues, one aspect of SFA pedagogy that may bear closer scrutiny and possible modification is instruction related to reading comprehension. The program had no effect on this outcome for the primary analysis sample: Second-graders in SFA schools and in control group schools registered similar scores on the comprehension measure. The program made more of a difference for students in the low-skilled subgroup, but this difference was not large enough to be statistically significant.

The i3 evaluation does not provide the sole or definitive word on this score. Earlier well-regarded studies, both experimental and quasi-experimental, did find that SFA improved comprehension among early-grade students. In particular, the findings of this study stand in contrast to those from a randomized controlled trial conducted by Borman et al., which found a statistically significant positive impact on comprehension for students who, like those in the i3 study, were exposed to the program for three years.[8] However, neither Borman nor the present study reports positive and statistically significant program impacts for students in the upper elementary grades. Upper-grade students in the SFA i3 study schools did not score higher than their control group counterparts on the Gates-MacGinitie test, which measures reading comprehension and vocabulary. This study did not find effects one way or the other on state reading tests, which also have a comprehension focus, for grades 3 and 5; for grade 4, the impact was negative and statistically significant at the 10 percent level.[9]

As the findings from instructional logs presented in Chapter 4 indicate, reading lessons in SFA schools and in control group schools did not differ significantly in the extent to which

[8] See Borman et al. (2005a, 2005b, 2007). In addition, the What Works Clearinghouse (2009) reports on six quasi-experimental studies of SFA that measured comprehension as an outcome. Five of these registered impacts with effect sizes of 0.15 or more. The studies had relatively small samples, and none of the comprehension effects was found to be statistically significant when analyzed at the cluster level.

[9] Because of the Borman study design (schools were randomly assigned to receive SFA either in kindergarten through grade 2 or in grades 3 through 5), students in grades 3 through 5 in that study were not exposed to the program in the earlier grades. The researchers therefore conclude that the program may not be beneficial to students who do not receive SFA instruction before third grade. (See Hanselman and Borman 2013.) In this study, however, third-graders in the SFA schools got the program in both first and second grades, while fourth-graders began SFA instruction in second grade. Unless SFA instruction needs to begin in kindergarten to produce impacts on comprehension, other explanations for the absence of impacts in this domain would appear to be in order. In any event, Borman has suggested that comparisons with his study are of limited usefulness because of differences in the study settings and samples (the majority of students in the Borman study were African-American, while in this study the majority of students are Hispanic) (email communication, Geoffrey Borman to Janet Quint, April 20, 2015).

they emphasized vocabulary, activation of students' prior knowledge, story structure, analysis and synthesis of information, and other elements of comprehension. Students in SFA classes were more likely to be asked to discuss text, and presumably to demonstrate their understanding of it. On the other hand, students in control group classes were more likely to engage in writing and in the analysis of word structure (including prefixes, suffices, and contractions) — activities that can also boost understanding. Teaching comprehension effectively is unquestionably a challenge.[10] But it may well be that SFA will need to develop a more sustained and distinctive approach if it is to achieve more robust impacts in this area, especially because comprehension becomes of paramount importance as students progress through elementary school and beyond.

In an economic climate characterized by budgetary cutbacks that forced many school districts to cut staff and restrict program offerings, the Success for All Foundation did not meet the ambitious expansion goals it set for itself. Nonetheless, the program was able to reach some 276,000 students in 447 i3 scale-up schools through the fourth grant year, and SFAF anticipates enlisting another 95 schools and 34,000 students by the end of the grant. While not all these schools could afford to implement SFA in all grades or to continue it for the entire demonstration period, the i3 scale-up has heightened the program's prominence on the educational landscape.

As the three-year evaluation drew to a close, three of the five participating school districts opted to continue with Success for All, retaining it in eight program group schools and expanding it to four control group schools as well. (Of the two districts that chose to end the program, one did so despite the fact that teachers in six of the district's nine SFA schools had voted to keep it.) Clearly, administrators, principals, and others in these districts found something of value in the program. This report suggests that they were right in their appraisal. With additional professional development, closer monitoring, more tutoring, and the enhancement of SFA's comprehension-focused curriculum components, Success for All might make an even bigger difference, and for more students, than it already does.

[10] In recognition of the fact that changes in classroom reading instruction have not resulted in anticipated improvements in reading comprehension, the Institute of Education Sciences within the U.S. Department of Education funded the Reading for Understanding Research Initiative to guide the development of new interventions in this area. (See Douglas and Albro 2014.)

Appendix A

Data Sources and Response Rates for the Success for All Evaluation

Appendix Table A.1 examines the sources of data used in the implementation and impact analyses and shows the response rate for data from each source for the 2013-2014 school year. For all sources of data, the desired goal was to achieve the highest response rates possible, so that the results would be fully representative of schools, principals, leaders, teachers, and students in the evaluation. As the table shows, the Success for All Foundation supplied School Achievement Snapshot ratings for all program group schools. Scores on each follow-up test used in the impact analysis were available for at least 94 percent of all second-grade students, depending on the specific test. Lower response rates were obtained for principal and teacher surveys, teacher logs, and (in the program group schools only) Success for All (SFA) facilitator interviews and teacher focus groups, but for each of these data sources, the response rate was 75 percent or higher.

Appendix Table A.2 shows that response rates in program group and control group schools did not differ significantly for any data source.

The data used in the cost and scale-up analyses are discussed separately in those chapters.

Appendix Table A.1

Data Sources and Overall Response Rates, 2013-2014

Instrument and Purpose	Number Targeted	Number of Respondents	Response Rate (%)
Principal survey			
Survey administered to all principals at both program and control group schools. Program group surveys also included questions about SFA. The survey provides information about the school's reading program, professional development, and school practices and supports. Additionally, it describes the launch and implementation of SFA in program group schools.	37	28	75.7
Teacher survey[a]			
Survey administered to all reading teachers at both program and control group schools. Program group surveys also included questions about SFA. The survey provides information about the school's reading program, professional development, and school practices and supports. Additionally, it describes the launch and implementation of SFA in program group schools.	704	530	75.3
School visit data			
Principal interviews: Interviews with both program and control group principals to learn about the SFA adoption process, school context, and implementation of the reading program.	37	29	78.4
Facilitator interviews: Interviews with the SFA facilitator at program group schools to learn about his or her duties and the SFA implementation story.	19	15	78.9
School Achievement Snapshot			
Evaluations created by SFA and filled out by an SFA coach who visited the school during each quarter to determine implementation levels of SFA components.	19	19	100.0
Teacher logs[b]			
Logs of teaching practices filled out by both program and control group teachers. The logs track the classroom practices of a group of randomly selected students over the course of a school day. The logs are used to highlight differences between program and control group classroom practices.	2,242	1,771	79.0

(continued)

Appendix Table A.1 (continued)

Instrument and Purpose	Number Targeted	Number of Respondents	Response Rate (%)
Impact analysis follow-up tests			
Woodcock-Johnson Word Attack test was administered to all sample students in spring 2014. Test scores serve as an outcome variable in the impact estimation model.	3,049	2,907	95.3
Woodcock-Johnson Letter Word Identification test was administered to all sample students in spring 2014. Spanish versions of the tests were administered to students without English mastery. Test scores serve as an outcome variable in the impact estimation model.	3,049	2,902	95.2
Woodcock-Johnson Passage Comprehension test was administered to all sample students in spring 2014. Spanish versions of the tests were administered to students without English mastery. Test scores serve as an outcome variable in the impact estimation model.	3,049	2,894	94.9
Test of Word Reading Efficiency was administered to all sample students in spring 2014. Test scores serve as an outcome variable in the impact estimation model.	3,049	2,873	94.2
District records Demographic and state testing information from each of the five districts for each student in the study. These data are used as covariates in the impact estimation model.	3,049	2,914	95.6

NOTES: [a]Teacher surveys were received from 31 of 37 schools; 4 program group schools and 2 control group schools did not return surveys.

[b]Log response rates were calculated based on the number of logs distributed to a given teacher, which was typically eight logs. The statistical test was computed at the level of logs, and it tests whether the experimental status of the school to which a teacher belonged affected the probability that the teacher would return a completed log.

Appendix Table A.2

Data Sources and Response Rates, by Program or Control Group Status, 2013-2014

	Program Group			Control Group			P-Value of
	Number	Number of	Response	Number	Number of	Response	Response Rate
Instrument	Targeted	Respondents	Rate (%)	Targeted	Respondents	Rate (%)	Difference[a]
Principal survey	19	14	73.7	18	14	77.8	0.779
Teacher survey[b]	407	297	73.0	297	233	78.5	0.696
School visit data							
Principal interviews	19	14	73.7	18	15	83.3	0.779
Facilitator interviews	19	15	78.9	—	—	—	—
Teacher focus groups	19	16	84.2	—	—	—	—
School Achievement Snapshot	19	19	100.0	—	—	—	—
Teacher logs[c]	1,203	981	81.5	1,039	790	76.0	0.723
Impact analysis follow-up tests							
Woodcock-Johnson Letter-Word Identification	1,632	1,553	95.2	1,417	1,349	95.2	0.887
Woodcock-Johnson Word Attack	1,632	1,557	95.4	1,417	1,350	95.3	0.711
Woodcock-Johnson Passage Comprehension	1,632	1,549	94.9	1,417	1,345	94.9	0.781
Test of Word Reading Efficiency	1,632	1,537	94.2	1,417	1,336	94.3	0.871
District records	1,632	1,565	95.9	1,417	1,349	95.2	0.650

(continued)

Appendix Table A.2 (continued)

NOTES: See Appendix Table A.1 for instrument descriptions. A two-tailed t-test was applied to differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a] Some measures were intended only for the program group; therefore, it is not possible to test the difference in response rates between the program and control groups.

[b] Teacher surveys were received from 31 of 37 schools; 4 program group schools and 2 control group schools did not return surveys.

[c] Log response rates were calculated based on the number of logs distributed to a given teacher, which was typically eight logs. The statistical test was computed at the level of logs, and it tests whether the experimental status of the school to which a teacher belonged affected the probability that the teacher would return a completed log.

Appendix B

Supplementary Tables for Chapter 2

Appendix Table B.1

Selected Baseline Characteristics of Students in the Full Baseline Sample

	Program Group	Control Group	Estimated Difference	P-Value for Estimated Difference
Age (years)	5.5	5.5	0.0	0.591
Students in poverty (%)	88.0	88.6	-0.5	0.795
Race/ethnicity (%)				
White	12.9	13.5	-0.6	0.704
Black	20.5	19.3	1.2	0.794
Hispanic	63.2	64.9	-1.7	0.727
Asian	1.6	0.9	0.7	0.441
Other	1.8	1.2	0.6	0.239
Male (%)	50.9	49.3	1.6	0.428
English language learners (%)	23.2	16.9	6.2	0.069 *
Special education status (%)	7.7	7.5	0.2	0.883
Peabody Picture Vocabulary Test, standard score	88.8	89.7	-0.9	0.442
Woodcock-Johnson Letter-Word Identification test, raw score	10.0	10.7	-0.7	0.083 *
Number of students	1,542	1,414		

SOURCES: MDRC calculations based on baseline test scores on the Peabody Picture Vocabulary Test (PPVT) and the Woodcock-Johnson Letter-Word Identification (WJLWI) test, administered to the baseline student sample in fall of the 2011-2012 school year. Student records data collected from the five districts in the study sample were also used.

NOTES: The full baseline sample includes all students eligible for baseline testing in the fall of 2011. The values for the 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 differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Appendix Table B.2

Selected Baseline Characteristics of Students in the Baseline Analysis Sample

	Program Group	Control Group	Estimated Difference	P-Value for Estimated Difference
Age (years)	5.5	5.5	0.0	0.895
Students in poverty (%)	87.4	88.6	-1.3	0.541
Race/ethnicity (%)				
White	12.5	12.6	-0.1	0.965
Black	18.9	17.5	1.4	0.764
Hispanic	65.4	67.2	-1.8	0.678
Asian	1.5	0.9	0.6	0.602
Other	1.7	1.3	0.4	0.493
Male (%)	49.6	49.0	0.6	0.820
English language learners (%)	25.8	20.3	5.5	0.163
Special education status (%)	6.2	6.4	-0.3	0.826
Peabody Picture Vocabulary Test				
Standard score	92.3	92.6	-0.3	0.848
Percentile equivalent	30	32		
Woodcock-Johnson Letter-Word Identification test, raw score	10.6	11.1	-0.5	0.288
Number of students	1,468	1,363		

SOURCES: MDRC calculations based on baseline test scores on the Peabody Picture Vocabulary Test (PPVT) and the Woodcock-Johnson Letter-Word Identification (WJLWI) test, administered to the baseline student sample in fall 2011. Student records data collected from the five districts in the study sample were also used.

NOTES: The baseline analysis sample includes all students with valid baseline tests on both the PPVT and WJLWI, administered in the fall of 2011.

A two tailed t-test was applied to differences between the program and control groups. Although there was no significant difference on any individual baseline characteristic, there was a statistically significant difference in the joint distribution of these baseline characteristics. This is based on a logistic regression predicting program status from student-level baseline. The p-value for the F-test is < 0.0001.

The values for the 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.

Appendix Table B.3

Selected Baseline Characteristics of Students in the Primary Analysis Sample and Sample Out-Movers

	Primary Analysis Sample	Sample Out-Movers	Estimated Difference	P-Value for Estimated Difference	
Age (years)	5.5	5.5	0.0	0.528	
Students in poverty (%)	89.0	90.4	-1.4	0.181	
Race/ethnicity (%)					
White	14.1	15.3	-1.2	0.153	
Black	14.2	16.4	-2.2	0.021	**
Hispanic	69.1	65.3	3.8	0.001	***
Asian	1.0	1.5	-0.5	0.271	
Other	1.5	1.6	-0.1	0.825	
Male (%)	49.2	52.0	-2.8	0.137	
English language learners (%)	25.6	18.0	7.6	0.000	***
Special education status (%)	6.0	9.6	-3.6	0.000	***
Peabody Picture Vocabulary Test, standard score	92.4	90.6	1.8	0.000	***
Woodcock-Johnson Letter-Word Identification test, raw score	10.7	9.6	1.1	0.000	***
Number of students	1,635	1,321			

SOURCES: MDRC calculations based on baseline test scores on the Peabody Picture Vocabulary Test (PPVT) and the Woodcock-Johnson Letter-Word Identification (WJLWI) test, administered to the baseline student sample in fall of the 2011-2012 school year. Student records data collected from the five districts in the study sample were also used.

NOTES: Out-movers are students in the full baseline sample who became ineligible for membership in the primary analysis sample. To be in the primary analysis sample, a student must have valid scores on both tests administered at baseline in the fall of 2011 and at least one valid score from each follow-up test administration period, occurring annually in the spring.

The values for the primary analysis sample are the weighted average of the observed district means for students in this sample (using number of program group schools in each district as weight). The values for out-movers in the next column are the regression-adjusted means using the observed distribution of the main analysis sample across blocks as the basis of the adjustment.

Appendix Table B.4

Selected Baseline Characteristics of Sample Out-Movers, by Program or Control Group Status

	Program Group	Control Group	Estimated Difference	P-Value for Estimated Difference	
Age (years)	5.5	5.5	0.0	0.598	
Students in poverty (%)	89.7	89.7	0.0	0.982	
Race/ethnicity (%)					
White	13.4	14.1	-0.6	0.759	
Black	23.0	21.3	1.7	0.716	
Hispanic	59.9	62.5	-2.6	0.612	
Asian	1.8	1.0	0.8	0.374	
Other	1.8	0.9	0.9	0.262	
Male (%)	52.9	49.8	3.1	0.325	
English language learners (%)	19.2	11.7	7.4	0.017	**
Special education status (%)	11.0	9.2	1.7	0.459	
Peabody Picture Vocabulary Test, standard score	83.5	85.4	-1.9	0.086	*
Woodcock-Johnson Letter-Word Identification test, raw score	8.8	9.8	-1.0	0.038	**
Number of students	688	633			

SOURCES: MDRC calculations based on baseline test scores on the Peabody Picture Vocabulary Test (PPVT) and the Woodcock-Johnson Letter-Word Identification (WJLWI) test, administered to the baseline student sample in fall of the 2011-2012 school year. Student records data collected from the five districts in the study sample were also used.

NOTES: Out-movers are students in the full baseline sample who became ineligible for membership in the primary analysis sample. To be in the primary analysis sample, a student must have valid scores on both tests administered at baseline in the fall of 2011 and at least one valid score from each follow-up test administration period, occurring annually in the spring.

The values for the 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) who became out-movers. 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.

Appendix Table B.5

Selected Characteristics of the 2013-2014 Auxiliary Student Sample, by Program or Control Group Status

	Program Group	Control Group	Estimated Difference	P-Value for Estimated Difference	
Grade 3					
Age (years)[a]	8.6	8.6	0.0	0.253	
Students in poverty (%)	89.5	91.3	-1.8	0.383	
Race/ethnicity (%)					
White	13.9	15.2	-1.3	0.422	
Black	20.8	18.0	2.8	0.488	
Hispanic	63.2	64.6	-1.4	0.733	
Asian	0.8	0.4	0.3	0.482	
Male (%)	51.3	52.5	-1.2	0.635	
English language learners (%)	22.5	20.3	2.2	0.618	
Special education status (%)	10.4	12.1	-1.8	0.275	
Mean proficient on 2010-2011 state test (%)[b]	73.0	72.1	0.9	0.742	
Grade 4					
Age (years)[a]	9.7	9.7	0.0	0.883	
Students in poverty (%)	89.0	91.5	-2.5	0.183	
Race/ethnicity (%)					
White	14.5	14.0	0.5	0.730	
Black	20.6	19.6	1.1	0.781	
Hispanic	62.3	64.6	-2.4	0.556	
Asian	1.0	1.2	-0.2	0.803	
Male (%)	49.7	50.1	-0.3	0.873	
English language learners (%)	19.4	13.2	6.2	0.061	*
Special education status (%)	10.3	9.9	0.4	0.776	
Mean proficient on 2010-2011 state test (%)[b]	68.9	66.7	2.2	0.451	

(continued)

Appendix Table B.5 (continued)

	Program Group	Control Group	Estimated Difference	P-Value for Estimated Difference
Grade 5				
Age (years)[a]	10.7	10.7	0.0	0.362
Students in poverty (%)	87.9	90.5	-2.6	0.204
Race/ethnicity (%)				
White	14.3	14.0	0.3	0.769
Black	20.1	19.3	0.8	0.819
Hispanic	63.3	64.2	-0.9	0.818
Asian	1.0	1.1	-0.1	0.788
Male (%)	49.5	50.7	-1.2	0.616
English language learners (%)	16.1	12.9	3.3	0.240
Special education status (%)	12.9	12.2	0.7	0.711
Mean proficient on 2010-2011 state test (%)[b]	65.2	65.0	0.2	0.941

SOURCES: Student records data collected from the five districts in the study sample for the 2013-2014 school year.

NOTES: The auxiliary student sample is defined as the set of students who were present in grades 3, 4, and 5 in the sample schools in the 2013-2014 school year who have either a valid state test score or Gates-MacGinitie Reading Test score on the vocabulary or reading comprehension subtest.

In each grade level, approximately 99 percent of the analysis sample had demographic data, and there was no difference in data availability for program and control group students. Therefore, the sample sizes presented below are for the total auxiliary student sample and differ only slightly from the numbers used in the calculations above.

The sample of third-grade students ranges from 2,572 students (1,311 in the program group and 1,261 in the control group) to 2,841 students (1,450 in the program group and 1,391 in the control group).

The sample of fourth-grade students ranges from 2,604 students (1,329 in the program group and 1,275 in the control group) to 2,656 students (1,361 in the program group and 1,295 in the control group).

The sample of fifth-grade students ranges from 2,752 students (1,459 in the program group and 1,293 in the control group) to 2,789 students (1,478 in the program group and 1,311 in the control group).

Due to data availability, the number of observations examined in the table varies by characteristics. 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 for the program group are the weighted averages 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 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.

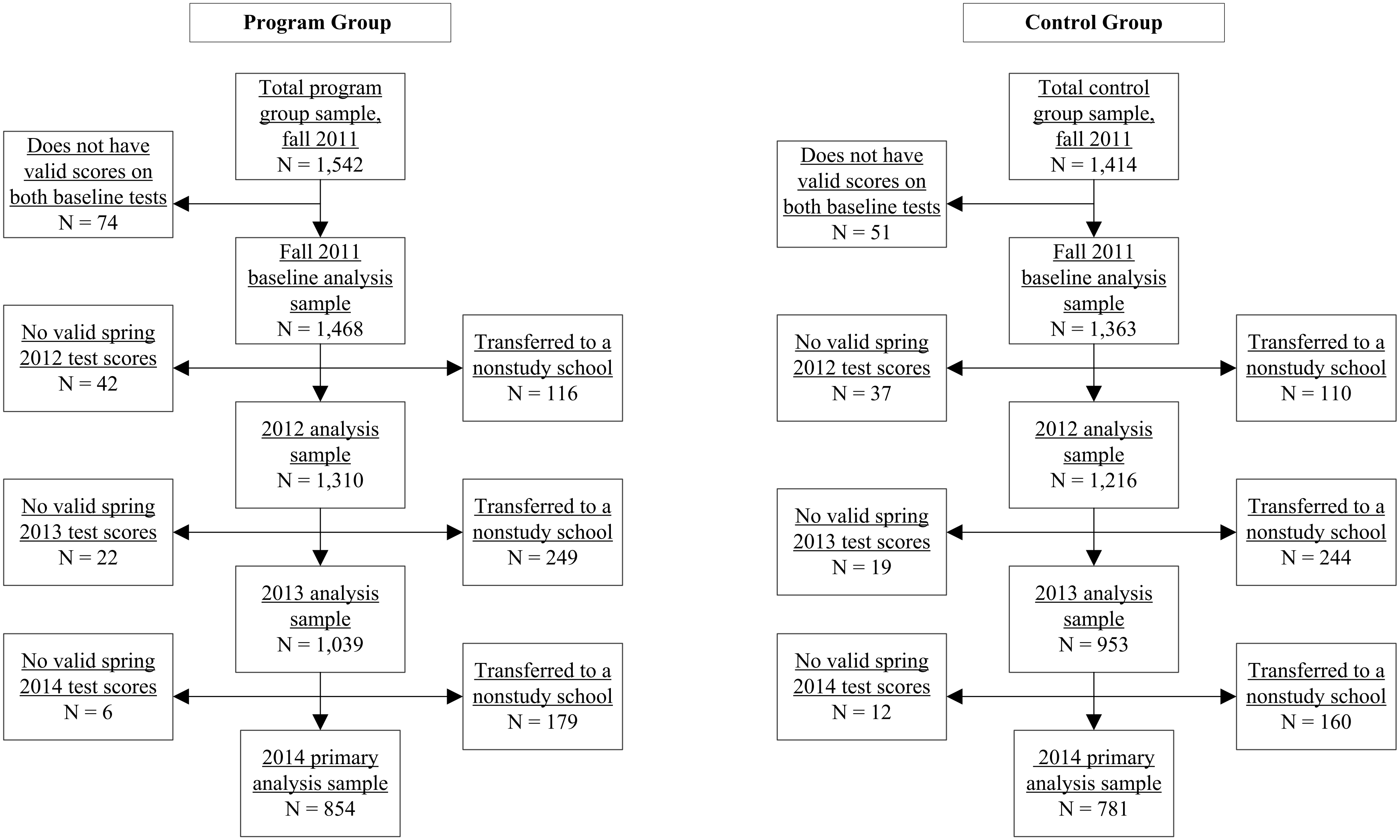
A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a] Age is calculated as the age (in years) of a student as of September 1, 2012.

[b] This variable, measured at school level for each grade, shows the percentage of students in 2010-2011 who were proficient on the state reading test. The number of observations by grade level is 37 in each case, which is the number of study schools.

Appendix Figure B.1

Tracking the Formation of the Primary Analysis Sample


Program Group
Total program group sample, fall 2011 N = 1,542
Does not have valid scores on both baseline tests N = 74
Fall 2011 baseline analysis sample N = 1,468
No valid spring 2012 test scores N = 42
Transferred to a nonstudy school N = 116
2012 analysis sample N = 1,310
No valid spring 2013 test scores N = 22
Transferred to a nonstudy school N = 249
2013 analysis sample N = 1,039
No valid spring 2014 test scores N = 6
Transferred to a nonstudy school N = 179
2014 primary analysis sample N = 854
Control Group
Total control group sample, fall 2011 N = 1,414
Does not have valid scores on both baseline tests N = 51
Fall 2011 baseline analysis sample N = 1,363
No valid spring 2012 test scores N = 37
Transferred to a nonstudy school N = 110
2012 analysis sample N = 1,216
No valid spring 2013 test scores N = 19
Transferred to a nonstudy school N = 244
2013 analysis sample N = 953
No valid spring 2014 test scores N = 12
Transferred to a nonstudy school N = 160
2014 primary analysis sample N = 781

Appendix C

Supplementary Table for Chapter 4

Appendix Table C.1

SFA-Control Group Comparisons on Teacher Perceptions of Professional Development (Implementation Year 2013-2014)

	Program Group	Control Group	Estimated Difference	P-Value
Percentage of teachers who agreed that they received helpful professional development in:				
Learning how to implement their school’s reading program properly	77.9	67.2	10.7	0.083 *
Learning new techniques for reading instruction	72.5	73.9	-1.4	0.831
Learning how to teach students at different reading levels or in different reading groups	53.0	68.3	-15.3	0.072 *
Developing strategies to better meet the needs of the reading students who struggle the most	58.1	58.0	0.1	0.986
Using classroom materials, including technology, to improve reading instruction	63.3	64.5	-1.2	0.865
Learning how to better use the time allocated to reading instruction	65.8	55.5	10.3	0.211
Learning how to implement cooperative learning techniques among students	79.1	61.5	17.6	0.008 ***

SOURCES: Spring 2014 teacher surveys.

NOTES: Items on the teacher surveys 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 and dividing by the total number of respondents to that item.

The means reported for teacher survey items are means of school means. First, means are taken within each school at the teacher level, then the mean across school means is taken. This was done to prevent overweighting schools with more teachers.

A two-tailed t-test was applied to differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

Completed surveys were received from 15 out of 19 SFA schools and 16 out of 18 control group schools. Completed surveys were received from 297 teachers at SFA schools and 233 teachers at control group schools.

Response rates for all teacher survey items presented were above 72 percent for both SFA and control group teachers. The response rates for these items are lower than response rates presented in other tables because item responses were counted only if a teacher received professional development in the relevant area.

Appendix D

Impact Estimation Model and Other Analytic Issues

Estimation Model

The basic model for the Success for All (SFA) impact analyses 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, with each district's estimate weighted in proportion to the number of program schools from each district in the sample. Findings in this report therefore represent the impact on student performance in the average program school within the study districts. The results do not necessarily reflect what the program effect would be in the wider population of districts from which districts participating in the study were selected.

Specifically, a two-level hierarchical model with students nested within schools is used for impact estimations for student-level outcomes reported in Chapter 5 of the report:

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

where

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

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

T_k = one if school *k* is assigned to receive the SFA 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*

μ_k , ε_{ik} = 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, since students are not associated with a specific reading teacher in the SFA model. The model is estimated as a two-level hierarchical model with the MIXED procedure in SAS.

Similarly, a one-level ordinary least squares (OLS) model is used to estimate the impacts on outcomes that are measured at school level: special education identification and retention rate.

$$Y_k = \sum_m \gamma_{0m} D_{mk} + \sum_m \gamma_{1m} T_k D_{mk} \ + \gamma_2 Y_{-1k} + \mu_k \ ,$$

where

Y_k = outcome for school k

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

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

Y_{-1k} = pre-program outcome measure for school k

μ_k = school-level random error, assumed to be independently and identically distributed.

In both models, the weighted average γ_1 (weighted by the number of program schools in each district, or block) of the estimated γ_{1m} coefficients for the five districts is the estimated program effect on the average program 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 both models are fixed effects models instead of random effects models. This is 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 program status determined by the status of the school in which he or she was enrolled at the time of the baseline tests.

Other Analytic Issues

Covariate Selection

The principles and rules for choosing covariates in the impact model 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.

For this reason, both student-level and school-average baseline Peabody Picture Vocabulary Test and Woodcock-Johnson Letter-Word Identification test scores are included in the impact model. They are closely related to the outcome and can potentially explain a fair amount of variation in the outcomes, which could lead to improved precision of the impact estimation.

- If theory and prior empirical evidence do not provide enough guidance for choosing covariates, use the following method to select appropriate covariates: (1) Rerandomize the sample schools to create pseudo program and control groups. (2) Run impact model using this pseudo-program indicator. (3) Add potential covariates to the impact model one by one, keeping only those that reduce the standard error of the pseudo-program effect.

Using this procedure, the study team further identified the students' English language learner (ELL) status, special education (SPED) status, age, and gender as covariates for the impact model.

For the impact estimation of school-level outcomes, the preprogram measure (for SPED and retention) is included as a covariate in the model because it closely relates to the outcome and because including more covariates might reduce the degrees of freedom of the estimation, which would lead to reduced statistical precision.

Treatment of Missing Values

Students with missing outcomes were dropped from the impact analyses for which they lacked data. In cases of missing covariate measures, the missing data were replaced with zeros, and a dichotomous variable indicating the missing status of a given covariate for each observation was added to the impact analysis model. This approach is chosen because it is straightforward to implement and because it is unlikely to create bias in impact estimates in an experimental setting.[1]

Multiple Hypothesis Testing

Following the What Works Clearinghouse guideline (version 3, 2014), we apply the Benjamini-Hochberg procedure within each of the outcome domains to adjust for the p-values, to reduce the risk of drawing inappropriate conclusions about the impact of SFA on the basis of statistically significant results that may occur by chance alone.[2]

[1] 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) for a detailed discussion on this issue.

[2] Benjamini and Hochberg (1995); What Works Clearinghouse (2014).

This procedure works in the following way:

1. Order the p-values in ascending order.
2. Let m equal the number of hypotheses to be tested.
3. Let q equal the desired false discovery rate (FDR).
4. Reject hypothesis H_i if $p(i) \leq i/m * q$.

For the report, there are two confirmatory tests for the alphabetics domain: one for program impact on the Woodcock-Johnson Word Attack (WA) score, the other for program impact on the Woodcock-Johnson Letter-Word Identification score. The following unadjusted p-values were obtained from the impact estimation model:

$$\{0.022, 0.243\}.$$

If we want to limit the FDR to 0.05 or less, we would compare the first p-value (0.022) to

$$1/2 * (0.05) = 0.025.$$

Since $0.022 < 0.025$, the first null hypothesis (for WA score) can be rejected at the 0.05 level. This result indicates that the positive impact on the WA score is significant at the 0.05 level after the Benjamini-Hochberg adjustment.

Statistical Precision Based on Sample Used

A common way to convey a study's statistical power is through the minimum detectable effect size (MDES). Formally, the MDES is the smallest true program impact, scaled as an effect size, that can be detected with a reasonable degree of power (in this case, 80 percent) for a given level of statistical significance (in this case, 5 percent for a two-tailed test). The numbers of students and schools in the sample are crucial factors that determine the precision with which impacts can be estimated, in order to accept or reject with confidence the hypothesis that the program had no effect. In general, larger sample sizes provide more precise impact estimates.

Appendix Tables D.1 and D.2 present the minimum detectable effect sizes for the impact estimates reported in Chapter 5. The MDES values in this table are based on the number of students and schools used in the actual impact estimation and the standard errors of the estimated impact of actual assignment to intervention. Hence, the values in the tables represent the actual precision of the analyses. The tables show that, across the grades, the study is equipped to detect impacts on reading achievement that are as small as 0.17 to 0.18 for the primary analysis sample and that range from 0.11 to 0.25 across grades for the auxiliary sample (numbers are

Appendix Table D.1

Realized Minimum Detectable Effect Sizes for the Primary Analysis Sample, Related Subgroups of the Primary Analysis Sample, and the Spring Analysis Sample (Implementation Year 2013-2014)

	WJLWI	WJWA	TOWRE	WJPC
Full primary analysis sample	0.17	0.18	0.18	0.17
Subgroups defined by baseline WJLWI				
Lower-performing students	0.27	0.26	0.32	0.27
Higher-performing students	0.20	0.22	0.16	0.16
Subgroups defined by baseline PPVT				
Lower-performing students	0.28	0.26	0.29	0.25
Higher-performing students	0.19	0.21	0.17	0.18
Students tested in both English and Spanish[a]	0.75	0.61	0.81	0.74
Black	0.56	0.67	0.65	0.57
White	0.98	1.18	0.66	1.07
Hispanic	0.31	0.31	0.30	0.26
Female	0.25	0.27	0.27	0.26
Male	0.18	0.19	0.18	0.17
Special education	0.59	0.67	0.75	0.57
Non-special education	0.18	0.19	0.19	0.18
English language learner	0.64	0.55	0.67	0.52
Non-English language learner	0.15	0.16	0.17	0.18
Poverty status	0.19	0.19	0.19	0.18
Non-poverty status	0.64	0.69	0.64	0.54
Spring analysis sample	0.18	0.17	0.17	0.15

SOURCES: Baseline Peabody Picture Vocabulary Test (PPVT) and Woodcock-Johnson Letter-Word Identification (WJLWI) test, administered to the baseline student sample in fall of the 2011-2012 school year, as well as the WJLWI test (Spring 2014), the Woodcock-Johnson Word Attack (WJWA) test (Spring 2014), the Test of Word Reading Efficiency (TOWRE) (Spring 2014), the Woodcock-Johnson Passage Comprehension (WJPC) test (Spring 2014), and student records data collected from the five districts in the study sample.

NOTES: The "primary analysis sample" consists of students from 37 schools (19 program group schools and 18 control group schools) and includes any student who had at least one valid spring test score in each of the three implementation years and who had valid scores on the fall baseline 2011 Peabody Picture Vocabulary Test and fall baseline 2011 Woodcock-Johnson Letter-Word Identification test.

The minimum detectable effect sizes are calculated based on realized standard errors from the impact estimation for the primary analysis sample, subgroups of that sample, and the spring analysis sample.

[a] The minimum detectable effect sizes for the Spanish-language versions of the Woodcock-Johnson exams are as follows: BATLWI (Spanish version of WJLWI) = 0.99; BATWA (Spanish version of WJWA) = 0.79; BATPC (Spanish version of WJPC) = 0.94.

Appendix Table D.2

Realized Minimum Detectable Effect Sizes for the Auxiliary Analysis Sample, by Grade (Implementation Year 2013-2014)

	Gates-MacGinitie			State Reading
	Vocabulary	Comprehension	Total	Test
Grade 3	0.22	0.21	0.22	0.22
Grade 4	0.17	0.18	0.18	0.19
Grade 5	0.19	0.25	0.23	0.11

SOURCES: Gates-MacGinitie Reading Comprehension and Vocabulary subtests (Spring 2014) and student state testing records collected from the five districts in the study sample.

NOTES: The minimum detectable effect sizes are calculated based on realized standard errors from the impact estimation for the auxiliary analysis sample.

The "auxiliary analysis sample" is defined as the set of students who were present in grades 3, 4, or 5 in the sample schools in the 2013-2014 school year and who have state testing scores or vocabulary or reading comprehension subtest scores from the Gates-MacGinitie Reading Test.

The sample of third-grade students ranges from 2,572 students (1,311 in the program group and 1,261 in the control group) to 2,841 students (1,450 in the program group and 1,391 in the control group).

The sample of fourth-grade students ranges from 2,604 students (1,329 in the program group and 1,275 in the control group) to 2,656 students (1,361 in the program group and 1,295 in the control group).

The sample of fifth-grade students ranges from 2,752 students (1,459 in the program group and 1,293 in the control group) to 2,789 students (1,478 in the program group and 1,311 in the control group).

expressed in effect-size units).[3] These numbers are very close to the MDES of 0.15 that was targeted at the planning stage of the study.

Note that the study was not designed to detect subgroup or differential subgroup effects, and all such analyses reported in Chapter 5 are exploratory and only for the purpose of generating hypotheses.

[3] Note that an estimated impact smaller than the MDES can still be found to be statistically significant. This is because the calculation of the MDES incorporates not only the probability of making a Type I error (that is, concluding that there is an impact when, in fact, there is not) but also the probability of making a Type II error (that is, concluding that there is no impact when, in fact, the program was effective).

Effect Size Calculation

The impact tables in Chapter 5 provide the estimated effect size and p-value for each impact estimate. The effect size indicates the magnitude of the estimated effect, calculated as a proportion of the standard deviation of the outcome measure for the control group. The standard deviations used for the effect size calculations for the primary analysis sample are 8.81 for the Woodcock-Johnson Letter-Word Identification test; 6.81 for the Woodcock-Johnson Word Attack test; 15.82 for the Test of Word Reading Efficiency; and 4.83 for the Woodcock-Johnson Passage Comprehension test.

For the auxiliary sample, the effect size calculations use the following standard deviation numbers:

- Gates-MacGinitie Comprehension test: 39.27 for Grade 3, 38.11 for Grade 4, and 34.40 for Grade 5
- Gates-MacGinitie Vocabulary test: 41.87 for Grade 3, 36.51 for Grade 4, and 33.78 for Grade 5
- Gates-MacGinitie Total Score: 36.67 for Grade 3, 33.65 for Grade 4, and 30.59 for Grade 5
- State Reading Achievement test: The scores are standardized within each district already, so the standard deviation is around 1 for this measure.

Issues with Outcome Measures

This section discusses how the analysis dealt with issues with some outcome measures used in Chapter 5. It explains why raw scores were used rather than standard scores. It also reports the data sources for special education and retention rates, how these rates were constructed, and some irregularities in the data as reported.

Use of Raw Scores for Outcomes

Raw test scores are typically converted to more easily interpretable measures, such as standard scores and percentile ranks. Such measures allow a student's score to be compared with a distribution of scores obtained for a norming sample, which is selected to be representative of the full population of students of a comparable age. A student's percentile rank, for example, is based on the percentage of students in the norming sample who received a raw score at or below the student's raw score. Therefore, for scaled measures to be meaningful the norming sample has to be up to date.

When examining the distribution of standard scores for students in the SFA control group, the research team found that these students were, on the whole, performing substantially better on both outcome measures than the “average” student as defined by the norming sample. This was highly unexpected given that the schools in this study serve economically disadvantaged students who, on average, perform beneath national academic standards.

Investigating the issue further, the research team learned that the Woodcock-Johnson tests were normed between 1996 and 1999. In 2005, the test publishers made adjustments to the weights assigned to demographic groups in the norming sample based on new U.S. census data, but no new students were actually tested. Because reading instruction for kindergartners has greatly changed since 1999, with far more emphasis on explicit instruction in letter-sound identification, comparing standard scores for students in the SFA study schools with the out-of-date norming sample would be misleading.

Construction of School-Level Special Education Measures

Special education (SPED) status applies to students who are so categorized because of disabilities; a subcategory of special education includes students who are classified as having specific learning disabilities (SLDs). The study team requested that counts of special education students be tabulated from student enrollment in Individualized Education Programs (IEPs) in the spring of a given school year, but that classification and declassification counts be cumulative, so that they represent the total number of classification changes throughout the school year, as of the spring. The five school districts in the study sample provided the following grade-level counts of SPED students for the spring of 2011 (baseline year) and the spring of 2014 (third follow-up year):

- Total number of students with an IEP in the spring of a given school year
- Total number of students with SLDs in the spring of a given school year
- Students who were *newly classified* as having an IEP in a given school year
- Students who were *newly classified* as having SLDs in a given school year
- Students who previously had an IEP and were *declassified* in a given school year
- Students who previously had SLDs and were *declassified* in a given school year
- Total number of students enrolled in the springs of 2011 and 2014[4]

For each special education measure, the collected spring special education counts were divided by spring grade-level enrollment counts in a given school. Thus, the outcome measures used in the impact analysis are the percentage of spring-semester students in a given school and grade with a particular special education designation.

[4] All but one of the districts provided enrollment data for the spring of 2011. For the district that did not, the research team used enrollment data provided by the state department of education.

There are some irregularities in the district-reported SPED counts, most likely due to districts having different reporting conventions:

- One district reported a large drop in SPED declassification rate in one grade. In this district, the fifth-grade IEP declassification rate in study schools declined from 19.7 percent to 3.8 percent during the period from 2011 to 2014. The district could not provide an explanation for this change. The decline was evident in control group schools as well as in program group schools.

- There are cases in two districts in which the newly identified IEP students outnumbered the total IEP students in a given school, grade, or year. In one of the districts, this may be because the district includes home-schooled and private school students in their newly identified IEP counts but not in their total IEP counts. It may also be that some newly identified students left the school before year-end IEP total counts were tabulated.

- A fair number of schools reported zero counts for one or more of the SPED measures. While some of these may reflect true zero counts, in other cases, the value of zero might have been used erroneously when the counts were actually missing. Appendix Table D.3 looks at the prevalence of nonzero values reported by both SFA and non-SFA schools by grade.

- Even though districts were requested to report the cumulative total for each measure for the entire school year, there could be cases where districts provided cumulative counts for only a portion of the school year. Since the information was submitted by districts, however, the same reporting rules would have been applied to both the program and control group schools and therefore would not bias the impact estimates reported in Chapter 5.

Construction of School-Level Retention Rates

The five school districts also provided grade retention counts for the spring of 2011 and the spring of 2014. Retention rates for each grade level in each school were calculated by dividing spring retention counts by spring enrollment counts. One district did not provide enrollment or retention data for the spring of 2011; for this district, the school-by-grade-level retention rates were calculated using data from the state department of education.

As with the SPED data, a fair number of schools reported zero counts for one or more of the retention measures. Some of them could reflect true zero counts, but in other cases, the value of zero may have been used erroneously for missing counts. Appendix Table D.4 looks at the prevalence of nonzero retention counts.

Appendix Table D.3

Percentage of Schools Reporting Nonzero Counts for Special Education and Specific Learning Disability Identification, by Program Status and Grade (Implementation Year 2013-2014)

	All Special Education Categories		Specific Learning Disability (SLD)	
	Percentage of Schools Reporting Nonzero Student Counts		Percentage of Schools Reporting Nonzero Student Counts	
	Program	Control	Program	Control
Kindergarten				
Total identification	94.74	88.89	15.79	27.78
New identification	78.95	61.11	10.53	22.22
Declassification	52.63	38.89	5.26	5.56
Grade 1				
Total identification	100.00	100.00	57.89	55.56
New identification	78.95	72.22	47.37	22.22
Declassification	57.89	38.89	26.32	5.56
Grade 2				
Total identification	100.00	100.00	78.95	50.00
New identification	78.95	72.22	63.16	44.44
Declassification	57.89	50.00	36.84	22.22
Grade 3				
Total identification	100.00	100.00	94.44	83.33
New identification	83.33	72.22	61.11	50.00
Declassification	50.00	44.44	38.89	16.67
Grade 4				
Total identification	100.00	100.00	100.00	94.12
New identification	61.11	64.71	50.00	58.82
Declassification	27.78	41.18	11.11	17.65
Grade 5				
Total identification	100.00	100.00	94.44	94.12
New identification	77.78	64.71	66.67	29.41
Declassification	38.89	35.29	22.22	5.88
Number of schools	19	18		

SOURCE: MDRC calculations based on special education records collected from the five districts in the study sample.

Appendix Table D.4

Percentage of Schools Reporting Nonzero Counts for Students Retained, by Program Status and Grade (Implementation Year 2013-2014)

	Percentage of Schools Reporting Nonzero Student Counts	
Grade	Program	Control
Kindergarten	26.32	50.00
Grade 1	73.68	72.22
Grade 2	57.89	83.33
Grade 3	66.67	72.22
Grade 4	38.89	35.29
Grade 5	72.22	64.71
Number of schools	19	18

SOURCE: MDRC calculations based on student retention and enrollment data collected from the five districts in the study sample.

Additional Impact Findings

This section provides results on additional or complementary impact analyses based on different analytic approaches or different analysis samples.

Additional Analysis for Impact Variation by Baseline Reading Performance Level

Analysis reported in Chapter 5 (Table 5.3) assesses potential differential effects of SFA by looking at SFA's impacts on subgroups of students defined by their baseline scores. Alternatively, one could assess this issue by looking at linear interactions between baseline student test scores and program status. Specifically, the study team reestimated the student impact model, including the main effect of the program, baseline student test scores, and the interaction between the two as covariates in the model. The estimated coefficient for the interaction terms provides an indication of whether the difference between the program group's and the control group's reading achievement levels depends on students' baseline achievement level. A statistically significant and positive (or negative) estimate would indicate that students with higher baseline test scores benefited more (or less) from the program.[5]

Because there are two baseline tests in this study, three different models are used to explore the differential effects of the two baseline tests both separately and together. Model 1 uses only the baseline score on the Peabody Picture Vocabulary Test (PPVT) and its interaction with the program indicator in the regression; Model 2 uses only the baseline score on the Woodcock-Johnson Letter-Word Identification (WJLWI) test and its interaction with the program indicator in the regression; and Model 3 uses both baseline test scores and their interactions with the program indicator in the regression.

Appendix Table D.5 reports the estimated coefficients for the interaction terms for all three models. The results are consistent with findings reported in the chapter. Of particular note are the following:

- Results from Model 1 show that SFA's impacts on the four outcomes do not seem to vary by students' baseline PPVT scores.
- Negative and significant coefficient estimates from Model 2 indicate that, for all four outcome measures, the magnitude of the SFA impacts decreases for students with higher baseline WJLWI scores.
- Results from Model 3 show that, when both baseline scores and their interactions with the program indicator are in the model, there is a significant and negative relationship between impact size and students' baseline WJLWI scores, and a significant and positive relation between impact size and students' baseline PPVT scores.

The linear patterns between impacts and students' baseline reading scores can also be illustrated graphically. Appendix Figures D.1 and D.2 present scatter plots of student outcomes against the two baseline test scores: Figure D.1 has four graphs showing the relationship between each of the four outcome measures and students' baseline PPVT scores; Figure D.2 shows the same graphs based on students' baseline WJLWI scores. The scatter plots use different symbols for the program group and control group students, and separate trend lines for the program and control groups are added to make the relationship more visible. In general, the patterns observed from these figures corroborate the findings from the subgroup analysis and the linear interaction analysis.

[5] The model implicitly assumes a linear relationship between baseline student test scores and the program effect.

Appendix Table D.5

Interaction of Baseline Test Scores and the Effects of SFA for the Primary Analysis Sample (Implementation Year 2013-2014)

	Baseline Test Score Interaction Effect			
Standardized Outcome	Estimate	Effect Size	Standard Error	P-Value
Model 1: Interaction with fall baseline PPVT[a] score				
WJLWI[b]	0.02	0.00	0.03	0.464
WJWA[c]	0.01	0.00	0.02	0.803
TOWRE[d]	-0.01	-0.00	0.05	0.900
WJPC[e]	0.01	0.00	0.01	0.353
Model 2: Interaction with fall baseline WJLWI score				
WJLWI	-0.17	-0.02	0.07	0.010 ***
WJWA	-0.14	-0.02	0.05	0.008 ***
TOWRE	-0.36	-0.02	0.12	0.004 ***
WJPC	-0.08	-0.02	0.04	0.028 **
Model 3: Interaction with fall baseline WJLWI and PPVT score				
WJLWI				
Interaction with PPVT	0.08	0.01	0.03	0.005 ***
Interaction with WJLWI	-0.26	-0.03	0.08	0.001 ***
WJWA				
Interaction with PPVT	0.05	0.01	0.02	0.030 **
Interaction with WJLWI	-0.20	-0.03	0.06	0.001 ***
TOWRE				
Interaction with PPVT	0.10	0.01	0.05	0.061 *
Interaction with WJLWI	-0.47	-0.03	0.14	0.001 ***
WJPC				
Interaction with PPVT	0.04	0.01	0.01	0.008 ***
Interaction with WJLWI	-0.11	-0.02	0.04	0.004 ***

(continued)

Impact Findings for Subgroups of the Primary Analysis Sample

Appendix Table D.6 presents detailed findings for each of the subgroups of the primary analysis sample reported in Table 5.4. Students are grouped by race or ethnicity, gender, special education status, English language learner status, and poverty status.

A small subset of the students in the primary sample spoke Spanish when they entered kindergarten and therefore might have had difficulty receiving instruction in English at school. Different districts have different policies for such students. In one of the districts in the sample, these students were provided reading instruction primarily in Spanish. For these students, the study team tested their reading achievement in both English and Spanish at follow-up.

Appendix Table D.5 (continued)

SOURCES: Baseline Peabody Picture Vocabulary Test (PPVT) and Woodcock-Johnson Letter-Word Identification (WJLWI) test, administered to the baseline student sample in fall of the 2011-2012 school year, as well as the Woodcock-Johnson Letter-Word Identification test (Spring 2014), Woodcock-Johnson Word Attack test (Spring 2014), the Test of Word Reading Efficiency (Spring 2014), the Woodcock-Johnson Passage Comprehension test (Spring 2014), and student records data collected from the five districts in the study sample.

NOTES: The student sample size for the Woodcock-Johnson Letter-Word Identification test is 1,631 students (851 in the program group and 780 in the control group).

The student sample size for the Woodcock-Johnson Word Attack test is 1,635 students (854 in the program group and 781 in the control group).

The student sample size for the Test of Word Reading Efficiency is 1,625 students (847 in the program group and 778 in the control group).

The student sample size for the Woodcock-Johnson Passage Comprehension test is 1,625 students (848 in the program group and 777 in the control group).

The impact analyses for reading achievement were conducted using raw scores. 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, as well as interaction terms between baseline scores and program indicator. Only the estimated coefficient for the interaction term in each model is reported here.

Effect sizes were computed using the full control group's standard deviations for the respective measures. The control group standard deviations are as follows: 8.81 for the Woodcock-Johnson Letter-Word Identification Test, 6.81 for the Woodcock-Johnson Word Attack test, 15.82 for the Test of Word Reading Efficiency, and 4.83 for the Woodcock-Johnson Passage Comprehension test.

A two-tailed t-test was applied to differences between program and control groups. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a] Peabody Picture Vocabulary Test (Form B).

[b] Woodcock-Johnson Letter-Word Identification test (Form B).

[c] Woodcock-Johnson Word Attack test (Form B).

[d] Test of Word Reading Efficiency.

[e] Woodcock-Johnson Passage Comprehension test (Form B).

Appendix Table D.7 provides impact estimates on both the English and the Spanish versions of the four tests for this group of students. SFA produced statistically significant impacts on only one outcome for this group. Note that, given its sample size, this subgroup has limited statistical power to detect significant findings. Even though some of the estimates reported here, especially those for the English tests, are quite large, it is hard to know whether the results are driven by the uncertainty and noise in the estimation due to small sample size or by a true strong impact. In addition, the analysis is confined to one district in the sample, therefore confounding the district effect with the program effect.

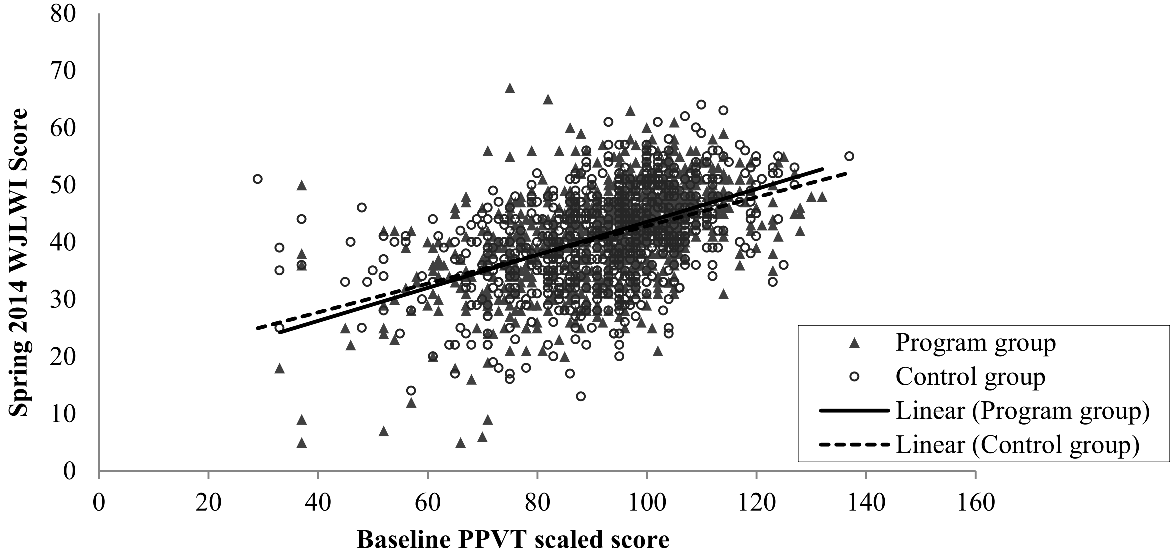
Impact Findings for All School-Level Special Education Outcomes

Appendix Table D.8 supplements Table 5.7 and reports findings for all school-level special education outcomes, including the total identification, new identification, and declassification rates across kindergarten through grade 5 for all SPED categories, as well as for the SLD category in particular.

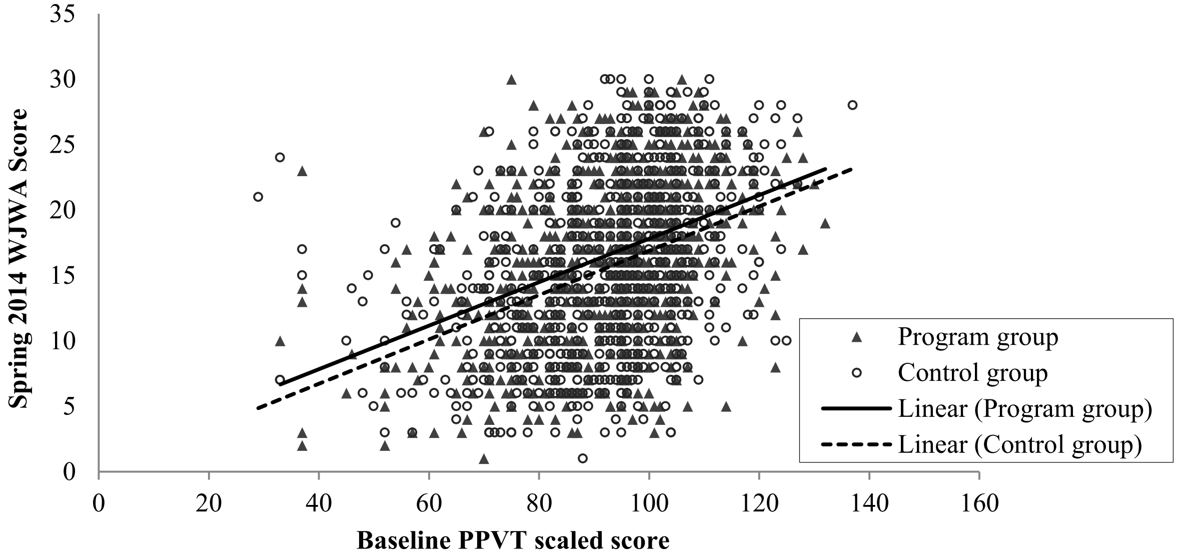
Appendix Figure D.1

Linear Relationship Between Reading Outcome and Baseline PPVT Scores, by Program or Control Group Status

Woodcock-Johnson Letter-Word Identification



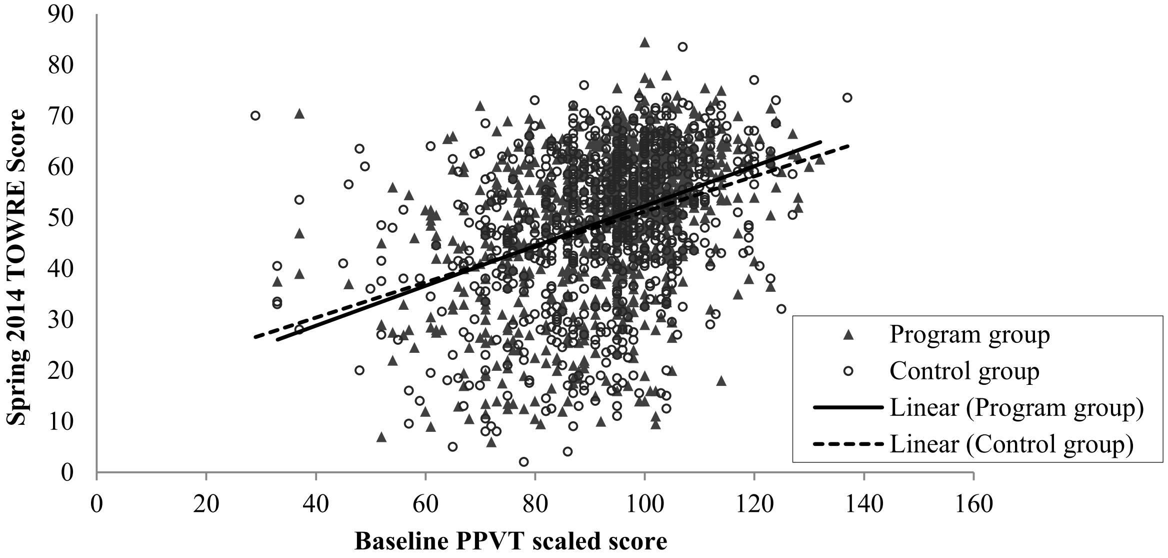
Woodcock-Johnson Word Attack



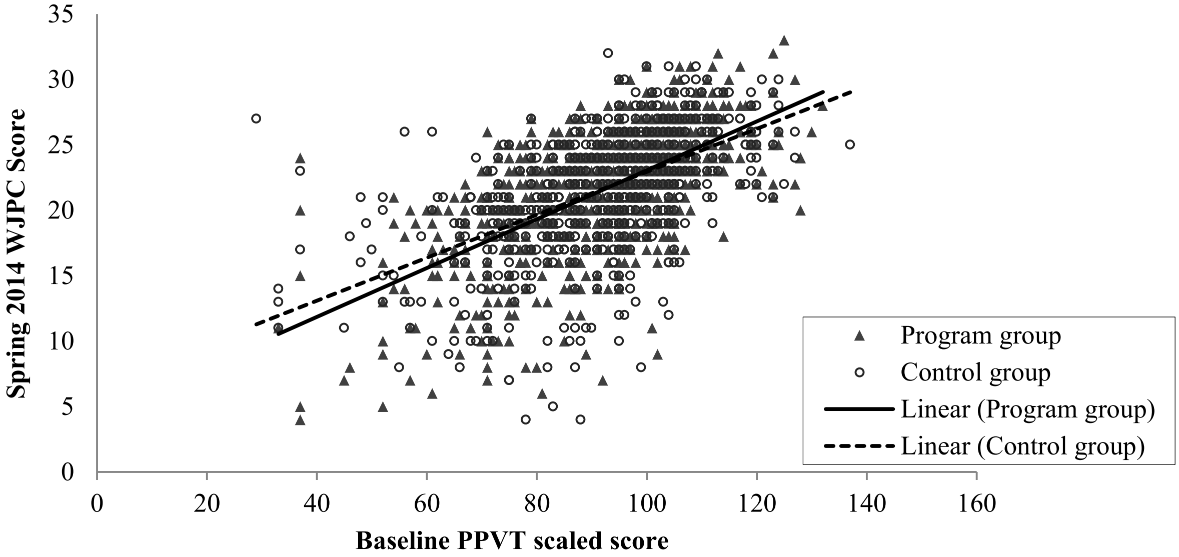
(continued)

Appendix Figure D.1 (continued)

Test of Word Reading Efficiency



Woodcock-Johnson Passage Comprehension

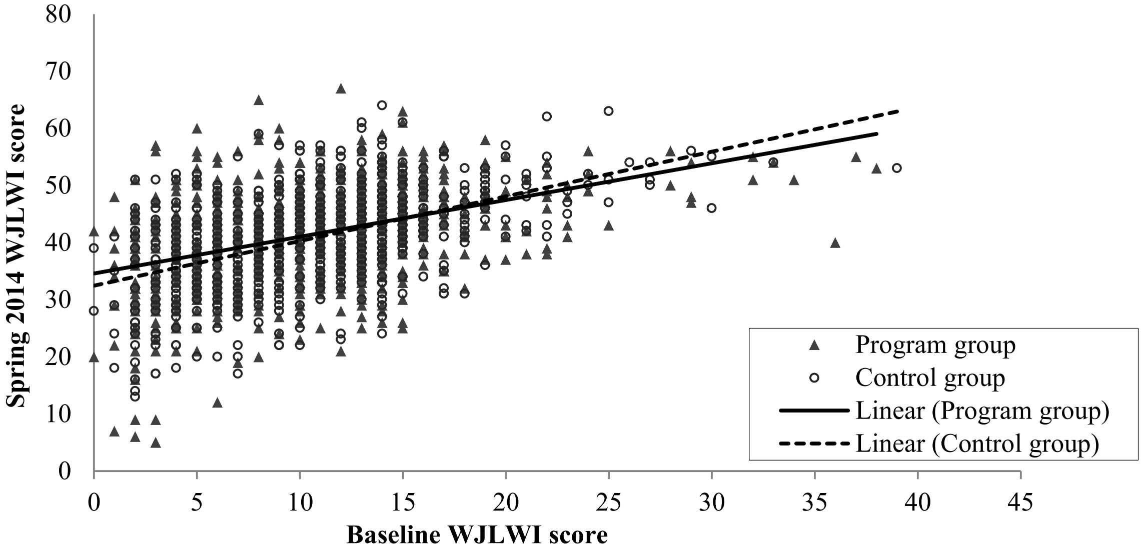


SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2014), Woodcock-Johnson Word Attack test (Spring 2014), Woodcock-Johnson Passage Comprehension test (Spring 2014), Test of Word Reading Efficiency (Spring 2014), and Peabody Picture Vocabulary Test (Fall 2011).

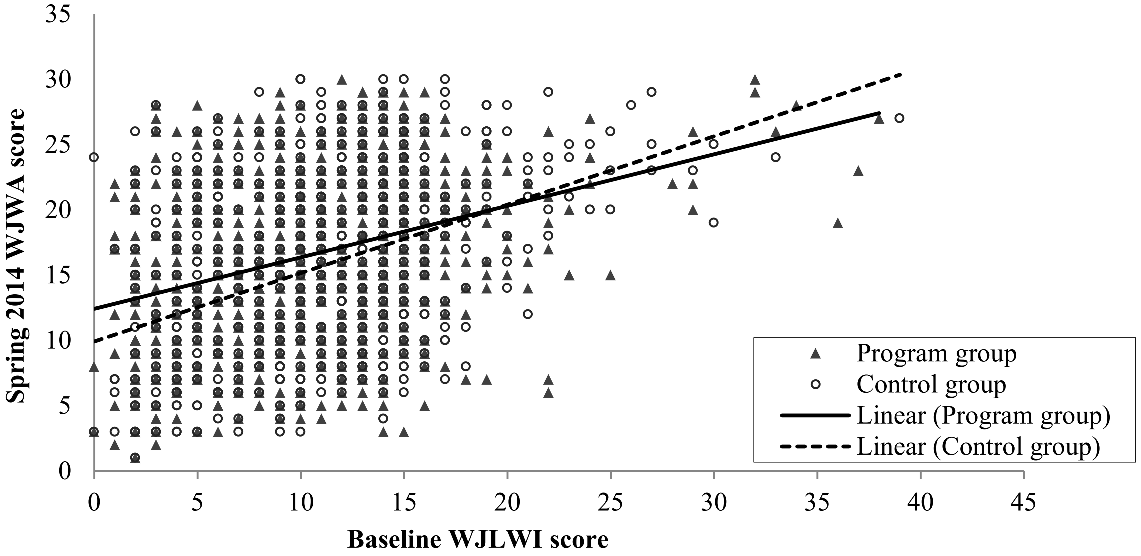
Appendix Figure D.2

Linear Relationship Between Reading Outcome and Baseline WJLWI Test Scores, by Program or Control Group Status

Woodcock-Johnson Letter-Word Identification



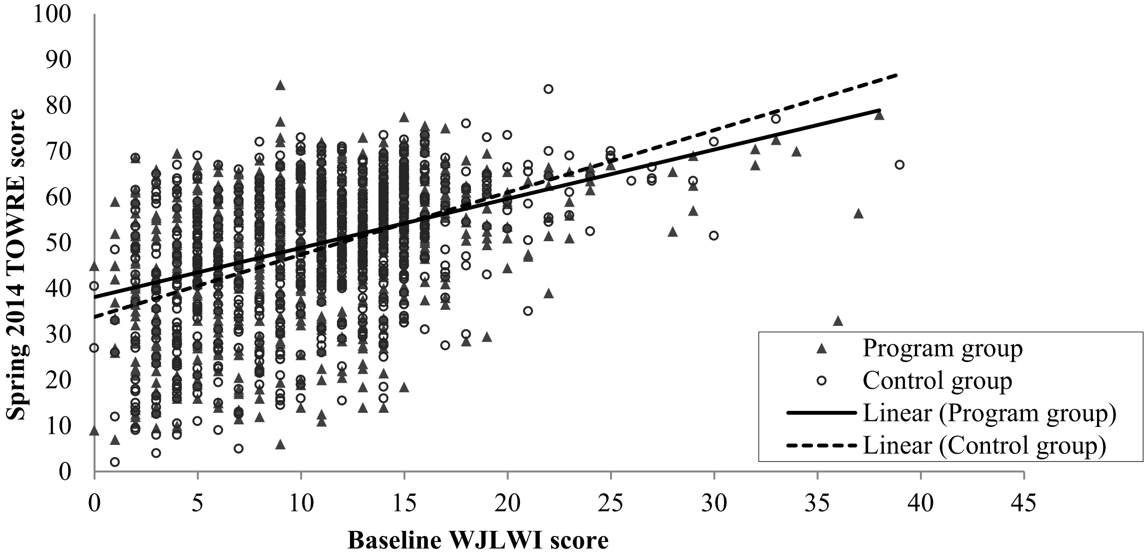
Woodcock-Johnson Word Attack



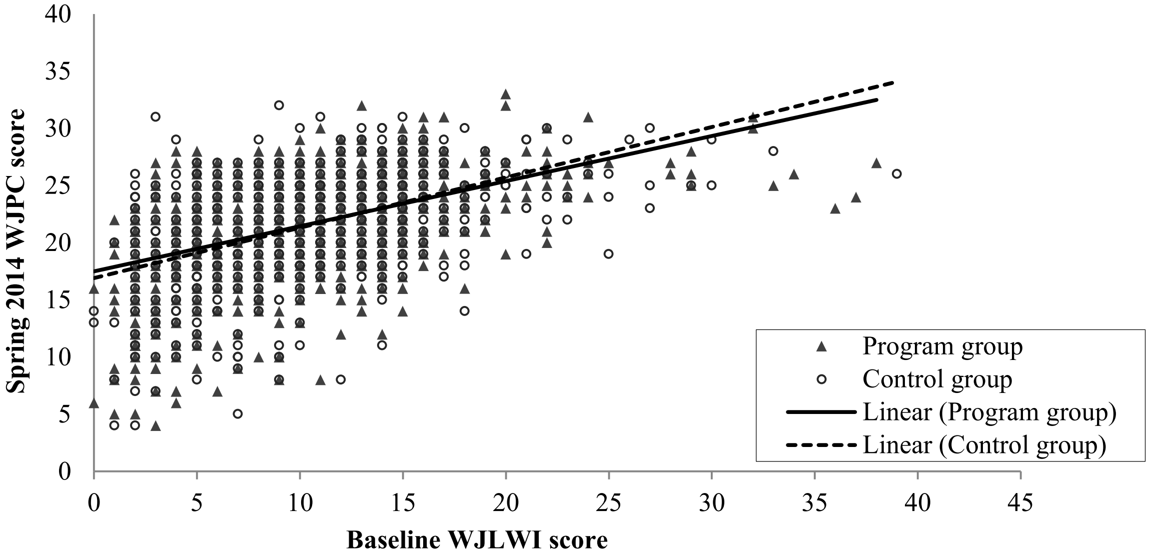
(continued)

Appendix Figure D.2 (continued)

Test of Word Reading Efficiency



Woodcock-Johnson Passage Comprehension



SOURCES: Woodcock-Johnson Letter-Word Identification test (Fall 2011 and Spring 2014), Woodcock-Johnson Word Attack test (Spring 2014), Woodcock-Johnson Passage Comprehension test (Spring 2014), and Test of Word Reading Efficiency (Spring 2014).

Appendix Table D.6

Impact of SFA on Average Second-Grade Reading Achievement for Subgroups of the Primary Analysis Sample (Implementation Year 2013-2014)

				Estimated		Number in	Number in
Subgroup	Program	Control	Estimated	Impact		Program	Control
and Outcome	Group	Group	Impact	Effect Size	P-Value	Group	Group
Black							
WJLWI[a]	42.00	41.24	0.76	0.09	0.647	132	98
WJWA[b]	17.02	16.32	0.69	0.10	0.652	134	98
TOWRE[c]	52.40	49.73	2.67	0.17	0.441	133	98
WJPC[d]	22.77	22.18	0.59	0.12	0.526	133	98
White							
WJLWI	43.24	38.98	4.26	0.48	0.155	125	103
WJWA	17.52	15.30	2.22	0.33	0.403	125	103
TOWRE	51.14	50.84	0.30	0.02	0.929	125	102
WJPC	23.37	20.54	2.83	0.59	0.118	125	103
Hispanic							
WJLWI	40.08	39.94	0.14	0.02	0.884	569	557
WJWA	15.86	15.46	0.39	0.06	0.585	570	558
TOWRE	48.02	48.43	-0.41	-0.03	0.804	564	556
WJPC	20.99	21.04	-0.05	-0.01	0.914	565	554
Female							
WJLWI	41.37	40.33	1.03	0.12	0.191	437	394
WJWA	16.35	15.15	1.20	0.18	0.067 *	437	394
TOWRE	49.92	48.52	1.40	0.09	0.348	434	394
WJPC	21.68	21.49	0.19	0.04	0.661	435	391
Male							
WJLWI	40.98	40.73	0.25	0.03	0.645	414	386
WJWA	16.47	15.62	0.86	0.13	0.062 *	417	387
TOWRE	48.91	48.21	0.70	0.04	0.490	413	384
WJPC	21.47	21.30	0.17	0.04	0.552	413	386
Special education							
WJLWI	37.31	36.92	0.39	0.04	0.830	49	48
WJWA	14.37	13.03	1.34	0.20	0.399	49	49
TOWRE	43.19	46.32	-3.14	-0.20	0.443	47	48
WJPC	19.51	19.74	-0.23	-0.05	0.807	49	48

(continued)

Appendix Table D.6 (continued)

Subgroup and Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value		Number in Program Group	Number in Control Group
Non-special education								
WJLWI	41.45	40.69	0.76	0.09	0.186		802	732
WJWA	16.57	15.51	1.06	0.16	0.025	**	805	732
TOWRE	49.84	48.34	1.50	0.10	0.151		800	730
WJPC	21.72	21.47	0.25	0.05	0.394		799	729
English language learner								
WJLWI	36.48	36.35	0.13	0.01	0.946		236	181
WJWA	13.92	13.26	0.66	0.10	0.604		237	181
TOWRE	44.22	42.74	1.48	0.09	0.685		232	180
WJPC	18.50	18.81	-0.31	-0.06	0.713		234	180
Non-English language learner								
WJLWI	42.38	41.58	0.80	0.09	0.096	*	615	599
WJWA	17.30	15.95	1.35	0.20	0.002	***	617	600
TOWRE	50.78	49.56	1.22	0.08	0.187		615	598
WJPC	22.38	22.20	0.18	0.04	0.552		614	597
Poverty status								
WJLWI	40.89	40.21	0.68	0.08	0.249		748	703
WJWA	16.26	15.23	1.03	0.15	0.030	**	751	704
TOWRE	48.96	47.88	1.08	0.07	0.300		744	701
WJPC	21.39	21.23	0.16	0.03	0.595		745	700
Non-poverty status								
WJLWI	44.55	43.80	0.74	0.08	0.699		103	77
WJWA	18.08	17.12	0.96	0.14	0.550		103	77
TOWRE	53.85	53.68	0.17	0.01	0.961		103	77
WJPC	24.40	23.23	1.17	0.24	0.195		103	77

(continued)

Appendix Table D.6 (continued)

SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2014), Woodcock-Johnson Word Attack test (Spring 2014), Woodcock-Johnson Passage Comprehension test (Spring 2014), Test of Word Reading Efficiency (Spring 2014), and student records data collected from the five districts in the study sample.

NOTES: The impact analyses for reading achievement were conducted using raw scores. 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.

Due to small sample sizes, estimates could not be computed for race/ethnicity groups other than white, black, and Hispanic.

Effect sizes were computed using the full control group's standard deviations for the respective measures. The control group standard deviations are as follows: 8.81 for the Woodcock-Johnson Letter Word Identification Test, 6.81 for the Woodcock-Johnson Word Attack test, 15.82 for the Test of Word Reading Efficiency, and 4.83 for the Woodcock-Johnson Passage Comprehension test.

A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

[a]Woodcock-Johnson Letter-Word Identification test.

[b]Woodcock-Johnson Word Attack test.

[c]Test of Word Reading Efficiency.

[d]Woodcock-Johnson Passage Comprehension test.

Test Score Trends and Impact Findings for a Consistent Student Sample

Figure 5.2 and the top panel of Figure 5.3 show the test score trends and impact findings for three implementation years, based on the analysis sample for each year. Recall that, for each implementation year, the analysis sample is defined as all students with at least one valid baseline test score and at least one valid outcome test score from spring of that school year. Therefore, the analysis sample differs from year to year. To assess whether the different impact findings are caused by changes in the sample composition or true changes in the program effects, the study team also examines the trends and impacts by implementation year on a consistent sample. This sample consists of students who were in the study schools for all three implementation years. Appendix Figures D.3 and D.4 show patterns that are consistent with those observed in Figures 5.2 and 5.3. This indicates that the changes in the impact findings over the years are unlikely to be caused by changing student composition in the sample.

Appendix Table D.7

Impact of SFA on Average Second-Grade Reading Achievement for the Spanish Test Analysis Sample (Implementation Year 2013-2014)

Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value	
Woodcock-Johnson Letter-Word Identification	33.83	31.45	2.38	0.28	0.283	
Woodcock-Johnson Word Attack	11.67	10.32	1.35	0.21	0.302	
Woodcock-Johnson Passage Comprehension	17.15	16.20	0.95	0.20	0.417	
Test of Word Reading Efficiency	41.15	33.61	7.54	0.48	0.099	*
BATLWI (Spanish version of WJLWI)	43.44	43.55	-0.12	-0.01	0.973	
BATWA (Spanish version of WJWA)	23.35	22.86	0.49	0.09	0.741	
BATPC (Spanish version of WJPC)	22.38	22.46	-0.08	-0.02	0.956	
Number of schools: 14	8	6				

SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2014), Woodcock-Johnson Word Attack test (Spring 2014), Woodcock-Johnson Passage Comprehension test (Spring 2014), Test of Word Reading Efficiency (Spring 2014), Spanish versions of the same tests (Spring 2014), and student records data collected from one district in the study sample.

NOTES: The "Spanish test analysis sample" includes students in the primary analysis sample who had at least one valid score on the Spanish-language versions of the Woodcock-Johnson exams, and it consists of students from 14 schools (8 program group schools and 6 control group schools) from one school district.

The student sample size for the WJLWI test is 181 students (117 in the program group and 64 in the control group).The student sample size for the WJWA test is 181 students (117 in the program group and 64 in the control group). The student sample size for the WJPC test is 180 students (117 in the program group and 63 in the control group). The student sample size for the Test of Word Reading Efficiency is 176 students (112 in the program group and 64 in the control group). The student sample size for the BATLWI, BATWA, and BATPC tests is 181 students (117 in the program group and 64 in the control group).

The impact analyses for reading achievement were conducted using raw scores. 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.

Effect sizes were calculated using the full control group's standard deviation for the respective measures. The control group standard deviations are as follows: WJLWI: 8.66, WJWA: 6.38, WJPC: 4.67, TOWRE: 15.86, BATLWI: 10.71, BATWA: 5.65, and BATPC: 4.74.

Appendix Table D.8

Impact of SFA on Special Education and Specific Learning Disability Identification and Declassification Rates by Grade (Implementation Year 2013-2014)

	All Special Education Categories				Specific Learning Disability (SLD) Category			
Outcome	Program Group (%)	Control Group (%)	Estimated Impact (%)	P-Value	Program Group (%)	Control Group (%)	Estimated Impact (%)	P-Value
Kindergarten								
Total identification rate	9.03	7.03	2.00	0.215	0.15	0.38	-0.23	0.096 *
New identification rate	4.94	2.55	2.39	0.146	0.18	0.40	-0.22	0.237
Declassification rate	1.16	1.46	-0.31	0.531	0.11	0.07	0.03	0.809
Grade 1								
Total identification rate	9.81	10.10	-0.29	0.879	1.20	0.95	0.24	0.549
New identification rate	2.69	2.70	-0.01	0.990	0.80	0.32	0.47	0.139
Declassification rate	1.34	0.62	0.72	0.057 *	0.56	0.14	0.42	0.201
Grade 2								
Total identification rate	12.08	10.63	1.46	0.284	2.65	2.53	0.12	0.854
New identification rate	3.23	2.57	0.66	0.459	1.30	1.51	-0.21	0.655
Declassification rate	1.57	1.47	0.10	0.849	0.75	0.46	0.28	0.308

(continued)

Appendix Table D.8 (continued)

	All Special Education Categories				Specific Learning Disability (SLD) Category				
Outcome	Program Group (%)	Control Group (%)	Estimated Impact (%)	P-Value	Program Group (%)	Control Group (%)	Estimated Impact (%)	P-Value	
Grade 3									
Total identification rate	12.73	12.50	0.23	0.914	5.72	5.17	0.55	0.650	
New identification rate	2.88	3.12	-0.24	0.781	1.45	1.77	-0.32	0.682	
Declassification rate	1.27	1.18	0.09	0.887	0.83	0.34	0.49	0.158	
Grade 4									
Total identification rate	13.54	13.23	0.31	0.840	6.92	5.53	1.38	0.144	
New identification rate	2.22	2.85	-0.63	0.478	1.41	1.52	-0.11	0.834	
Declassification rate	0.79	0.86	-0.07	0.888	0.26	0.35	-0.08	0.782	
Grade 5									
Total identification rate	14.04	15.83	-1.79	0.351	8.01	5.80	2.21	0.130	
New identification rate	1.81	1.71	0.11	0.865	1.31	0.78	0.53	0.356	
Declassification rate	1.02	1.19	-0.17	0.810	0.47	-0.12	0.59	0.012	**
Number of schools: 37	19	18			19	18			

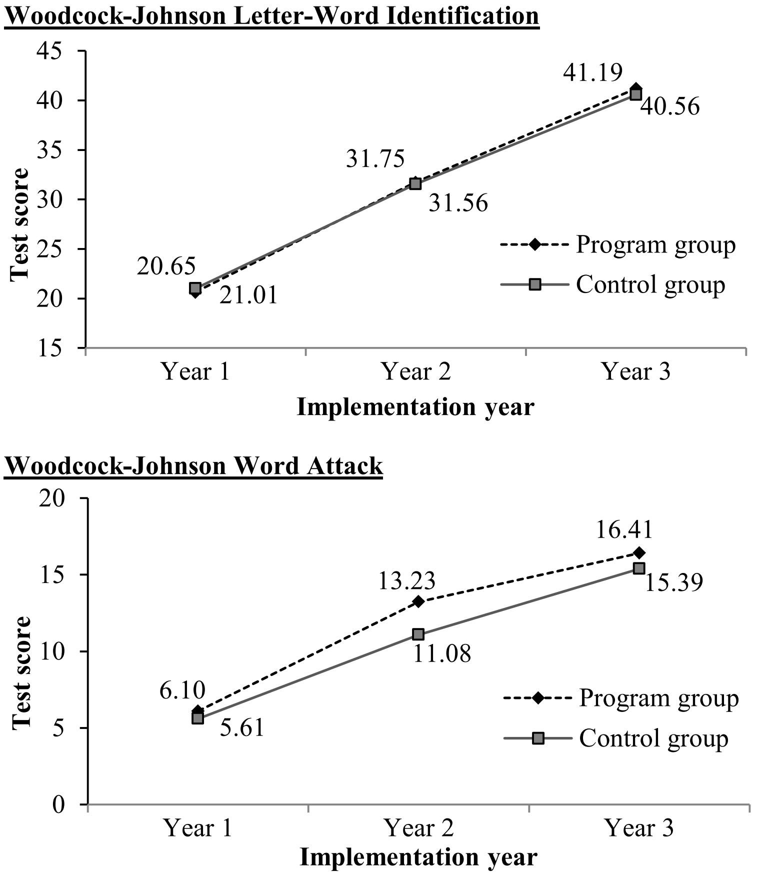
SOURCE: MDRC calculations based on special education records collected from the five districts in the study sample.

NOTES: The estimated impacts are based on an ordinary least squares (OLS) model with school-level data, controlling for random assignment block and school-level preprogram outcome measures. The program group and control group columns display regression-adjusted mean outcomes for each group. Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Appendix Figure D.3

Mean Test Scores for the Consistent Primary Analysis Sample over Time, by Program or Control Group Status



Woodcock-Johnson Letter-Word Identification

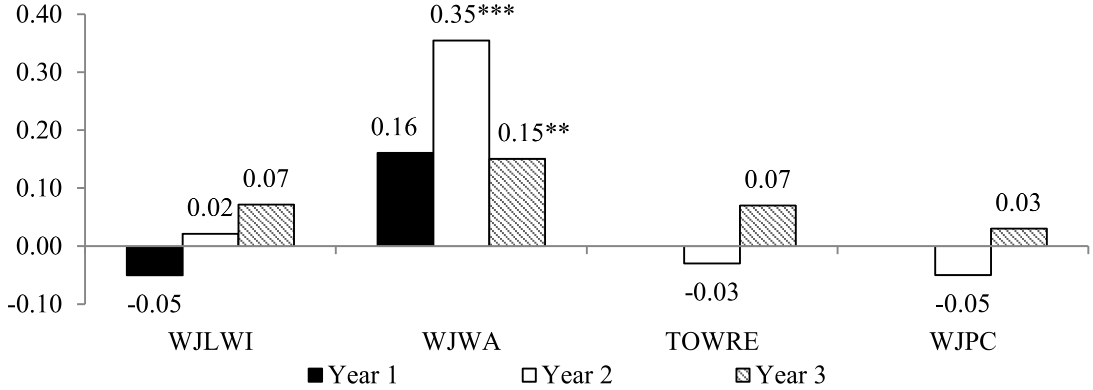
Woodcock-Johnson Word Attack

SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2012-2014), Woodcock-Johnson Word Attack test (Spring 2012-2014), and student records data collected from the five districts in the study sample.

NOTES: The "consistent primary analysis sample" consists of students from 37 schools (19 program group schools and 18 control group schools) and is limited to students in the 2013-2014 primary analysis sample; that is, those students who were present in study schools in 2013-2014 who had at least one valid spring test score in each of the three implementation years and who had valid scores on the fall baseline 2011 Peabody Picture Vocabulary Test and fall baseline 2011 Woodcock-Johnson Letter-Word Identification test.

Appendix Figure D.4

Impact of SFA on Average Reading Achievement for the Consistent Primary Analysis Sample, by Implementation Year



SOURCES: Woodcock-Johnson Letter-Word Identification (WJLWI) test (Spring 2012-2014), Woodcock-Johnson Word Attack (WJWA) test (Spring 2012-2014), Woodcock-Johnson Passage Comprehension (WJPC) test (Spring 2012-2014), Test of Word Reading Efficiency (TOWRE) (Spring 2012-2014), and student records data collected from the five districts in the study sample.

NOTES: The student sample size for the Woodcock-Johnson Letter-Word Identification test ranges from 1,623 students (852 in the program group and 771 in the control group) in Year 2 to 1,631 students (851 in the program group and 780 in the control group) in Year 3. The student sample size for the Woodcock-Johnson Word Attack test ranges from 1,628 students (851 in the program group and 777 in the control group) in Year 2 to 1,635 students (854 in the program group and 781 in the control group) in Year 3. The student sample size for the Test of Word Reading Efficiency ranges from 1,578 students (823 in the program group and 755 in the control group) in Year 2 to 1,625 students (847 in the program group and 778 in the control group) in Year 3. The student sample size for the Woodcock-Johnson Passage Comprehension test is 1,625 in both years (848 to 851 students in the program group and 774 to 777 students in the control group).

Students were tested using both Form A and Form B of the Test of Word Reading Efficiency. The scores reported above represent the average.

The impact analyses for reading achievement were conducted using raw scores. 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.

Effect sizes were computed using the full control group's standard deviations for the respective measures. A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Appendix E

Methodology of the Cost Analysis

Data Collection

The Success for All (SFA) evaluation team requested school-level cost and personnel data from the five evaluation districts for the three study years. The data requested included costs of reading curricula and other reading materials and technology; professional development days and cost; full-time equivalent (FTE) counts of personnel, by job category and experience level, as well as salary and benefits expenditures; and school revenues from federal Title I funds, state funding, and private grants to support the reading program.

Three of the five districts were able to fulfill part of this data request, providing personnel counts and salary expenditures; only one district provided FTE information. For all districts, state department of education websites supplied FTE data, student enrollment counts, and expenditures data of varying completeness. Figure E.1 shows the average enrollment and per-pupil expenditures for schools in each district, where available.

District D was selected for a focused case study because it offered the most complete and detailed set of cost information. District D was the only evaluation district in a state that makes available FTE, enrollment, and expenditures data at the school level and for all study years. In addition, the publicly available expenditures data for District D disaggregated overall school expenditures into separate categories used in the cost analysis, including expenditures on operations and instructional supplies. This combined set of data allowed the MDRC research team to construct a more accurate and nuanced picture of costs incurred by schools during the study years.

Determining the Representativeness of the Study District

The study team used data from the National Center for Education Statistics' 2010-2011 Common Core of Data (CCD) to determine District D's comparability with other districts in its state and with other districts in the evaluation (unless otherwise noted).[1] Below is a brief summary of tests conducted.

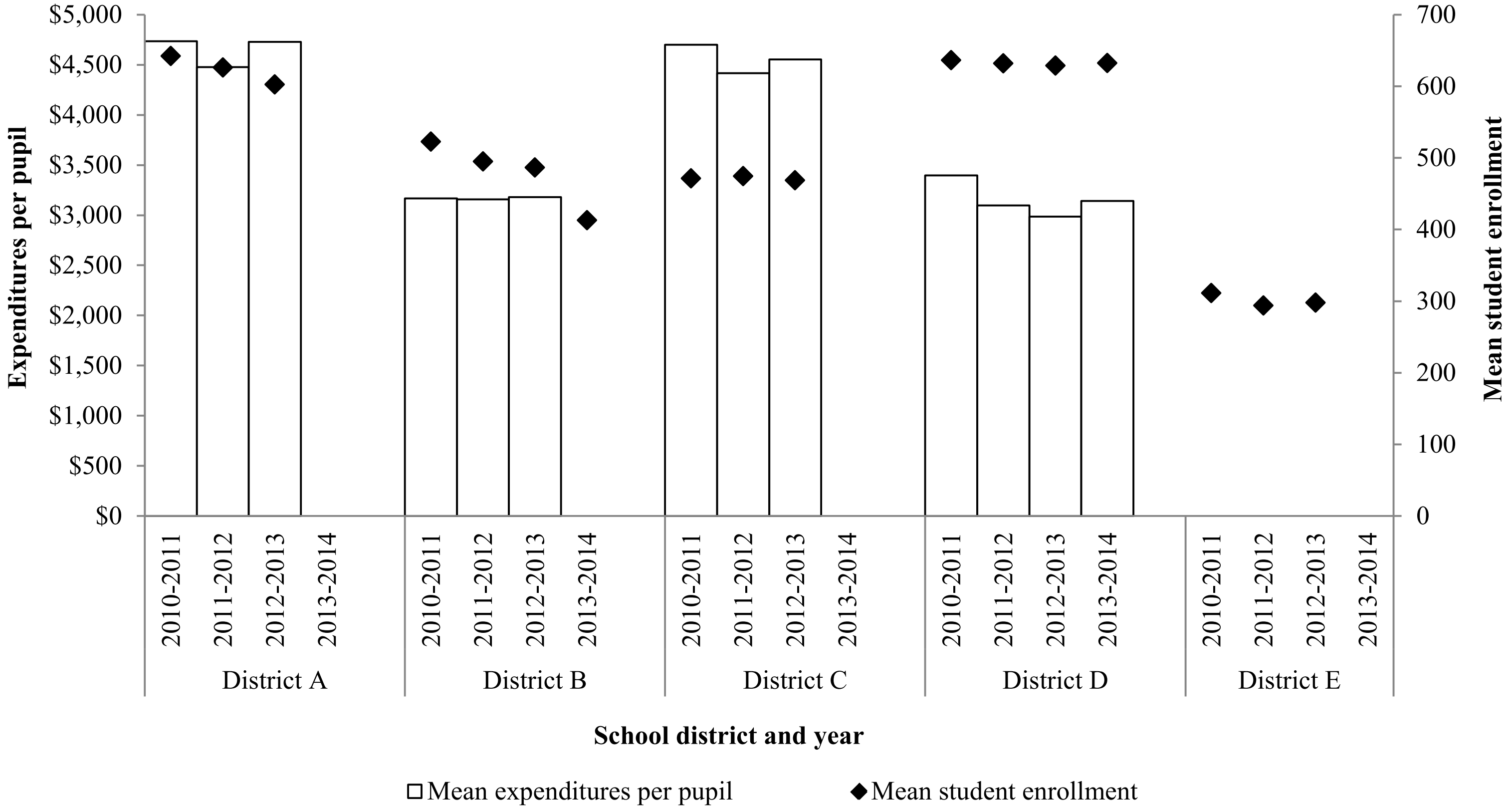
Comparing District D with All Other Districts in the State

- A t-test of whether per-pupil instructional spending in District D differed from the statewide average showed no significant differences ($p = 0.94$). This comparison uses state administrative data for each school.

[1] National Center for Education Statistics (2015).

Appendix Figure E.1

Average Student Enrollment and Per-Pupil Expenditures Per School, by Study District, from 2010-2011 to 2013-2014



SOURCES: Publicly available state administrative data on school-level school expenditures and student enrollment only for schools in the evaluation sample. Expenditures data were not available for District E.

NOTES: Enrollment was calculated by summing all students in kindergarten through grade 5 in all study schools in each district. Per-pupil expenditures were calculated by dividing total instructional expenditures by total student enrollment.

- A t-test of whether the school-level student-teacher ratio differed between schools in District D and all other schools in the state showed no significant differences (p = 0.74).

- An F-test of whether the schools in the district differed on a set of student characteristics (race variables, percentage of students qualified for free or reduced-price lunch, Title I status) showed significant differences (p < 0.01). This is to be expected, given the kinds of underresourced schools SFA aims to serve.

Comparing District D with All Other Districts in the Evaluation

- To compare per-pupil instructional spending across states for a test statistic, the study team would need a multistate data source, but the Common Core of Data and other federal sources provide instructional spending data only at the district level, not the school level. This would mean comparing the value for just one district (which has no variance) against the average of other study districts (which does have a variance); therefore, a comparison would not yield a meaningful test result.

- A t-test of whether the school-level student-teacher ratio differed showed that schools in District D had significantly higher student-teacher ratios than schools in the other districts (p < 0.01).

- An F-test of whether schools in District D differed from schools in other districts on a set of student characteristics (race variables, percentage of students qualified for free or reduced-price lunch, Title I status) showed no significant differences (p > 0.99).

Methodological Approach

The resource cost analysis presented in Chapter 6 draws on the "cost ingredients" method described by Levin and McEwan,[2] assigning quantities and prices to each component resource necessary to implement Success for All and the alternative reading programs in control group schools. The estimation of quantities is described in Chapter 6. This section provides more detail on the sources and estimation of prices as well as further information about quantity assumptions.

[2] Levin and McEwan (2001).

Prices

Price Adjustments

- *Time adjustments:* All costs are expressed in 2012 dollars, to be consistent with the start of SFA implementation in the 2011-2012 school year. Prices are adjusted using the Bureau of Labor Statistics' Consumer Price Index for All Urban Consumers (CPI-U).
- *Regional adjustments:* All costs calculated using national price estimates are converted to reflect relative cost of living in the case study district, using the Bureau of Economic Analysis's state-level regional price parity index for 2012.
- *Intertemporal discounting:* All costs are expressed in terms of their present value in the first year of the program. All costs incurred in Years 2 and 3 of the program are discounted assuming an interest rate of 3.5 percent, as is conventional standard.[3]
- *Benefit costs for personnel:* All personnel are assumed to receive benefits valued at one-third of their annual salary.

Price Estimates

Almost all personnel prices come from the Database of Educational Resource Prices from the Center for Benefit-Cost Studies of Education (CBCSE),[4] which pools cost data from multiple sources. Teacher, principal, facilitator, and volunteer tutor prices all came from the U.S. Department of Education's 2011 School and Staffing Survey (SASS),[5] as reported in the CBCSE's database.

Teachers. Teacher salary data were taken from the 2011 SASS national estimates for public school teachers with a bachelor's degree. For the main analysis, the median teacher experience level was taken from responses to the teacher survey administered as part of the broader SFA evaluation described in this report, and the corresponding salary was applied as the price.[6]

[3] Moore, Boardman, and Vining (2013).

[4] Center for Benefit-Cost Studies of Education (2015).

[5] National Center for Education Statistics (2011).

[6] The 2011 SASS reports teacher salaries within the following experience groupings: fewer than 2 years, 2 years, 3 years, 4 years, 5 years, 6-9 years, 10-14 years, 15-19 years, 20-24 years, 25-29 years, 30-34 years, and 35 or more years.

Facilitators. SFA facilitator experience level was taken from interview notes, and the corresponding SASS salary was then applied as the price. Facilitators at control group schools were assumed to have the same range of experience (10 to 14 years).

Principals. Principal experience level was taken from responses to the principal survey administered as part of the broader study, and the corresponding SASS salary was applied as the price.[7]

Classrooms and space. Classroom and furniture costs came from the Peter Li Education Group 2013 Annual Report, as reported in the CBCSE's database. The costs presented use CBCSE's estimates for cost per square foot and square footage required per student to calculate the price of an individual unfurnished classroom. The following standard table was used to calculate the cost of an individual classroom.[8] These calculations were applied to both SFA and control group schools.

1	Cost per square foot	$204.79
2	Average square footage recommended	1,029
3	Overall cost of classroom (row 1 * row 2)	$210,728.91
4	Estimated useful life in years	40
5	Cost of depreciation per year of use (row 3 / row 4)	$5,268.22
6	Multiply undepreciated amount by interest rate of 3.5% ((row 3 – row 5) * 0.035)	$7,191.12
7	Depreciation plus forgone interest (row 5 + row 6) = annual cost	$12,459.35

Furniture. The following calculations show how the annual cost of furnishings was calculated, using the CBCSE's cost-per-square-foot estimate. These calculations were applied to both SFA and control group schools.

1	Cost per square foot	$100.87
2	Average square footage recommended	1,029
3	Overall cost of classroom (row 1 * row 2)	$103,795.23
4	Estimated useful life in years	15
5	Cost of depreciation per year of use (row 3 / row 4)	$6,919.68
6	Multiply undepreciated amount by interest rate of 3.5% ((row 3 – row 5) * 0.035)	$3,390.64
7	Depreciation plus foregone interest (row 5 + row 6) = annual cost	$10,310.33

[7] SASS reports principal salaries within the following groupings: fewer than 3 years, 3-9 years, more than 9 years.

[8] Levin and McEwan (2001, p. 67), Simon (2011, p. 102).

Training costs. The case study district had a district reading coach who provided ongoing coaching during the school year, a position analogous to the Success for All Foundation (SFAF) point coach. When such coaches were interviewed across districts, they reported having about six years of teaching experience. Thus, trainers in control group schools were assumed to receive the salary equivalent of a teacher with six to nine years of experience, based on SASS.

Materials costs. Control group school reading programs were identified using responses to the principal surveys and interviews administered by the study during the 2011-2012 school year. The MDRC cost analysis compares the purchase price of these alternative reading program materials with the cost of the SFA program materials, assuming adoption during the 2011-2012 school year. In reality, in District D, the actual adoption year of the control group school reading programs varied. Quantities for reading program materials include both core curriculum materials and assessment and intervention program materials, when the latter were reported.

Estimates for the price of control group school reading materials were obtained from publisher websites for all but one school in District D. For that school, the cost of reading program materials was estimated from a combination of current publisher prices and per-student cost estimates provided by a study reviewed by the What Works Clearinghouse.[9] Because publisher websites provide the full retail cost of individual materials, and districts often do not pay the full retail cost, the study team applied a 25 percent discount to core curriculum and assessment material prices to account for potential wholesale or institutional discounts.[10] This discount was not applied to intervention materials or to cost estimates obtained from prior evaluations.

The school-level cost of student and teacher materials was estimated by multiplying the cost of individual materials by the number of students or teachers in the school. The cost of intervention materials for students was estimated using the number of students receiving tutoring, as reported in the study-administered principal survey.

Quantities

Because available data were incomplete, a number of assumptions about ingredient quantities were needed for the main resource cost analysis. The following list shows which assumptions were required for each resource.[11]

[9] What Works Clearinghouse (2013).

[10] The study team found in a cursory search of trade magazines that volume discounts ran the gamut from 10 percent to 60 percent.

[11] See Appendix Table E.6 for the actual quantities corresponding to each of these assumptions.

- *Teacher time on support teams:* Number of teachers on each team, frequency of meetings, length of meetings
- *Teacher time on certified tutoring:* Exact frequency of sessions, exact length of sessions
- *Teacher time in training (in control group schools):* Number of training/professional development sessions
- *Facilitator:* Share of FTE devoted to reading program
- *Principal time:* Share of FTE devoted to reading program
- *Materials room:* Share of room devoted to this purpose
- *Training costs (in control group schools):* Number of training/professional development sessions
- *Materials:* Wholesale discount applied to published prices

In the absence of time diaries showing how facilitators actually allocated their time, quantities were assigned based on self-reporting during interviews or based on principals' reported assignment of facilitator time. If an SFA facilitator reported working full-time on the program, she was counted as spending all her time on SFA. Control group school facilitators were assumed to spend 2.5 hours out of a 7-hour day on their school's reading program, including monitoring and coaching during the 90-minute reading block and providing support and/or tutoring services for 1 hour.

Another question related to quantity was whether to include support staff in addition to teaching staff. To this end, the study team examined different staff categories and whether the number of staff in each category presented differences at baseline (even before SFA implementation began). Appendix Table E.1, a comprehensive companion to Table 6.1, presents differences in staffing allocation as well as state-provided costs of supplies.

At baseline, program group schools spent less on supplies and had a lower ratio of students to "core subject" teachers: 25 to 1 in program group schools and 27 to 1 in control group schools.[12] This difference is driven not by a different number of core subject teachers, but by program group schools having fewer students than control group schools. The real difference in resource allocation appears to come from other staff categories: Program group schools had about 8 more non-core-subject instructional staff members, 27 more members of the instructional support staff, and 7 fewer members of the professional support staff across the three

[12] A teacher was considered "core" if his or her assignment was listed as "Elementary Classroom," "Kindergarten Classroom," "Reading Classroom," "Mathematics," or "Communications Arts" in state records of staff assignments. In some cases mathematics teachers are cocategorized as elementary classroom teachers, so they are included in the sample.

Appendix Table E.1

Resource Allocation by Year in Program and Control Group Schools in District D

	Program Group		Control Group		P-C Differences	
Resource	Annualized Years 1-3	Annualized Minus Baseline	Annualized Years 1-3	Annualized Minus Baseline	Between Annualized Years 1-3	Between Baselines
Instructional staff (FTEs)	96.6	0.6	87.5	-2.4	9.1	6.1
Items included in resource costs						
Core instructional FTEs[a]	73.7	-0.3	72.7	-1.3	1.0	0.0
Facilitator FTEs[b]	1.9	—	1.1	—	0.9	—
Principal and supervisory FTEs	3.3	0.3	3.3	0.3	0.0	0.0
Percentage of reading teachers with 0-3 years of experience	33.0		19.8		13.3	
Total enrollment	1,838.3	-12.7	1,947.3	-20.7	-109.0	-117.0
Core instructional student-to-FTE ratio	25.0	-0.1	26.8	0.2	-1.8	-1.6
Other costs						
Noncore instructional FTEs	22.9	0.9	14.8	-1.1	8.1	6.1
English language learner instructional FTEs	1.0	-1.0	1.3	-0.2	-0.3	0.5
Special education instructional FTEs	11.0	2.0	3.2	0.2	7.8	6.0
Other instructional FTEs	10.9	-0.1	10.3	-1.1	0.6	-0.4
Professional support FTEs[c]	20.1	-8.0	26.9	4.5	-6.7	5.8
English language learner instructional FTEs	0.0	0.0	0.0	0.0	0.0	0.0
Special education instructional FTEs	1.0	-1.0	0.7	0.7	0.3	2.0
Instructional support FTEs[d]	41.4	3.6	14.4	-0.3	27.1	23.1
English language learner instructional FTEs	0.0	0.0	0.0	0.0	0.0	0.0
Special education instructional FTEs	26.0	1.1	2.2	0.5	23.8	23.2
Instructional supplies ($)[e]	152,334.28	-58,931.10	145,536.46	-126,880.13	6,797.82	-61,151.21

(continued)

Appendix Table E.1 (continued)

SOURCES: All personnel, supplies, and student enrollment information was obtained from publicly available state data. Net change in student enrollment was estimated from the study's student testing data. SFA school facilitator counts were taken from responses to the study's principal survey and principal interviews. The percentage of teachers with 0-3 years of experience was taken from responses to the study's teacher survey.

NOTES: Staff counts are provided in full-time-equivalent (FTE) units.

[a] A teacher is considered a core instructional FTE if the teacher's job assignment was listed as "Elementary Classroom," "Kindergarten Classroom," "Reading Classroom," "Mathematics," or "Communications Arts."

[b] The estimate assumes that all non-SFA schools have a reading program facilitator equivalent to 0.36 FTE in each implementation year, and that both program and control group schools had no facilitator prior to Year 1.

[c] Professional support staff includes staff with the following job titles: coach, community services representative/resource worker, guidance counselor, librarian, library assistant, media specialist, nurse, occupational therapist, physical therapist, prevention coordinator, psychologist, resource teacher, social worker, speech/language therapist, and support specialist.

[d] Instructional support staff refers to instructional aides, interventionists, and FTEs devoted to professional development.

[e] Instructional supplies expenditures include supplies for all subject areas, not just the reading program.

schools.[13] Because it could not be determined to what extent these staff members were involved in reading instruction, they are excluded from the primary analysis presented in the chapter.

Supplementary Analysis: Estimating the Cost to Schools of Continuing with SFA

The estimates presented in the main text reflect the cost required to launch SFA in a district by averaging the program's costs for the first three years of implementation. These costs are expected to decline over time, eventually reaching a consistent level required to maintain the program's operation. A consideration of continuation costs is useful if schools implement SFA beyond the initial three years, as three of the five study districts chose to do.

Third-Year Costs

Appendix Table E.2 shows full resource costs incurred in the third year of the program only. The third-year cost provides an estimate for the resource commitment required to maintain the operation of SFA after the start-up period. This cost difference of $180 per student is lower than the $227 annualized full resource cost difference presented in Table 6.5, because none of the start-up costs incurred in the first two years of the program are reflected in this estimate.

SFA Costs over Time

Appendix Figure E.2 shows how reading program costs in SFA and non-SFA schools changed during the three years of the study, with a decline, as expected, after Year 1 for both groups. Control group schools' costs increased from Year 2 to Year 3, largely because of a higher proportion of more experienced teachers, who receive higher salaries. SFA school costs also modestly increased from Year 2 to Year 3, primarily due to additional facilitator time.

[13] Instructional support staff includes instructional aides, paraprofessionals, intervention specialists, and professional development providers. Professional support staff includes coaches, community service representatives, guidance counselors, librarians, media specialists, nurses, occupational/physical therapists, prevention coordinators, psychologists, resource teachers, social workers, speech and language therapists, and support specialists.

Appendix Table E.2

Steady-State (Third-Year) Resource Costs and Differences in Program and Control Group Schools in District D

	Program Group		Control Group		Difference	
Costs ($)	Total	Per Student	Total	Per Student	Total	Per Student
Teacher time (total)[a]	940,601	511	1,058,985	542	-118,384	-31
Teachers: reading block[b]	860,042	467	1,023,346	524	-163,304	-57
Teachers: support teams[c]	8,394	5	0	0	8,394	5
Certified tutoring after school[d]	40,268	22	14,881	8	*25,387*	*14*
Teacher time in training	40,290	22	20,757	11	19,533	11
Reading program facilitator[e]	165,603	90	74,315	38	*91,288*	*52*
Principal time devoted to reading program	183,958	100	138,869	71	45,089	29
Volunteer tutors[f]	0	0	20,087	10	-20,087	-10
Facilities and classrooms[g]	1,723,524	937	1,747,149	895	-23,625	42
Materials and facilitator room[h]	19,640	11	0	0	19,640	11
Training and professional development[i]	110,799	60	2,324	1	*108,475*	*59*
Reading program materials[j]	53,403	29	0	0	*53,403*	*29*
Total	3,197,528	1,738	3,041,729	1,557	155,799	180

(continued)

Appendix Table E.2 (continued)

SOURCES: Teacher and principal full-time equivalent (FTE) values are taken from publicly available state data. Teacher, principal, facilitator, and volunteer salaries are calculated from the U.S. Department of Education schools and staffing survey. Teacher and facilitator experience levels are based on responses to the study's teacher survey. Tutoring frequency and duration, volunteer tutor counts, facilitator FTE values, control group reading programs, and principal experience levels are based on responses to the study's principal survey. Point coach salaries were obtained from correspondence with the SFA Foundation. Facilities and classroom costs are based on square footage costs from the Center for Benefit-Cost Studies of Education Database of Educational Prices. Training and materials costs for program group schools were obtained from SFA program contracts. Reading program materials costs for control group schools were obtained from publisher websites and prior evaluations. Training costs for control group schools were estimated based on interview responses. Substitute teacher counts and days required were calculated based on records of chronically absent teachers published by the U.S. Department of Education's Office of Civil Rights.

NOTES: Rounding may cause slight discrepancies in calculating differences. Italicized items also appear in Table 6.3. All prices are given in geographically adjusted 2012 dollars. All personnel estimates include benefits costs amounting to one-third of salary. See Appendix Table E.6 for a more detailed description of cost calculations and assumptions.

[a] Teacher counts reflect an estimate of core-curriculum teachers. A teacher was considered a core instructional FTE if the teacher's job assignment was listed as "Elementary Classroom," "Kindergarten Classroom," "Reading Classroom," "Mathematics," or "Communications Arts."

[b] This estimate assumes that all reading teachers spend 90 minutes per day in the reading block, except teachers of English language learners who spend 4 hours per day.

[c] The estimate assumes that no teachers in control group schools met on support teams.

[d] A lower bound for tutoring frequency, for both certified and volunteer tutors, was taken from responses to the principal survey. A lower bound of tutoring session length (41 minutes) was also taken from the principal survey, since all respondents indicated that tutoring sessions in their schools lasted at least 41 minutes.

[e] The estimate assumes that all non-SFA schools have a reading program facilitator equivalent to 0.36 FTE.

[f] No volunteer tutors were reported in program group schools.

[g] This estimate assumes that each teacher has his/her own classroom. This figure includes the cost of a classroom for after-school tutoring by certified teachers.

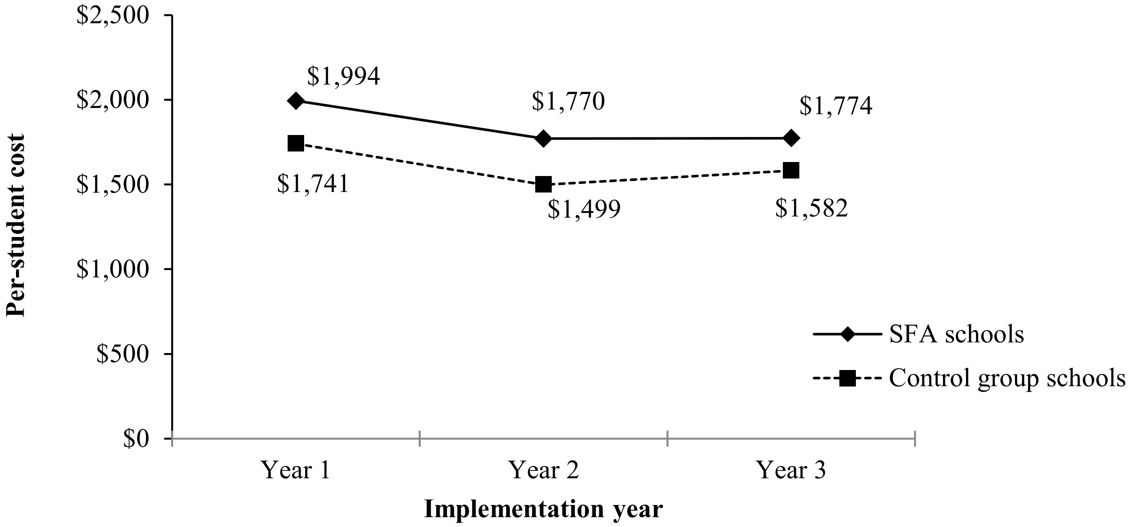
[h] This estimate assumes that each SFA school had one room for reading program materials and the use of the program facilitator, and that control group schools did not have a room devoted to this purpose.

[i] Estimates include point coach training and observation time.

[j] Reading program materials include teacher and student textbooks, classroom consumables, assessment materials, support and licensing fees, and intervention materials when applicable. This estimate assumes that all study schools purchased a new set of reading program materials in Year 1.

Appendix Figure E.2

Per-Student Costs of the SFA and Control Group School Reading Programs over Time in District D



SOURCES: MDRC calculations based on publicly available state data on school staff counts, enrollment, and expenditures; survey responses; SFA contracts and correspondence with the SFA Foundation; and estimated materials costs.

NOTE: See Appendix Table E.6 for a detailed description of cost calculations and assumptions.

Sensitivity Checks

Check on Out-of-Pocket Costs: Alternative Reading Program

Appendix Table E.3 shows the higher relative out-of-pocket costs incurred by SFA schools in District E, an evaluation district that actually did adopt a new reading program in 2011-2012 (as District D is assumed to have done for the purposes of analysis). Like District D, District E is small (with fewer than 20 elementary schools), with three SFA schools and three control group schools. Along with these similarities, the districts had some key differences:

- The six schools included in the study in District E had significantly lower student enrollment than the six study schools in District D.[14] This difference led to higher per-student costs in District E.

- Both program and control group schools offered less tutoring in District E than in District D. Only one school (an SFA school) reported offering any tutoring, and that school reported that each session lasted just 21 to 30 minutes. In District D, all schools responding to the principal survey indicated that tutoring occurred and that each session lasted at least 41 minutes.

- Facilitators in SFA schools in District E allocated more time to SFA than their counterparts did in District D.

- On a per-student basis, SFA schools in District E spent more on training and materials than SFA schools in District D, while control group schools in District E spent much less on materials than the control group schools in District D. The lower enrollment in District E schools led to a larger per-student cost of materials in SFA schools, despite slightly lower total spending on SFA materials than in District D.

These factors contributed to a larger out-of-pocket difference between program and control group schools in District E. The per-student out-of-pocket cost estimate in District E is $343, $224 larger than in District D (as presented in Table 6.3). Differences in reading program materials alone accounted for $151 of the larger per-student difference in District E. This sensitivity check shows how the per-student out-of-pocket cost differential observed in District D would increase in a district with lower enrollment and a less expensive control group school reading program.

Check on Full Resource Costs: Alternative Staffing Assumptions

Appendix Table E.4 calculates how the estimated difference in full resource costs would change as a result of six modifications of staffing cost assumptions presented in Table 6.5.

[14] SFA schools in District E had an average enrollment of 1,010 during the three study years, and control group schools in District E had an average enrollment of 779. In District D, SFA schools had an average enrollment of 1,838 and control group schools had an average enrollment of 1,947.

Appendix Table E.3

Out-of-Pocket Expenses and Differences, Annualized for a Three-Year Reading Program, in SFA and Control Group Schools in District E

	Program Group		Control Group		Difference	
Costs ($)	Total	Per Student	Total	Per Student	Total	Per Student
Certified tutoring after school[a]	3,144	3	0	0	3,144	3
Reading program facilitator[b]	169,375	168	72,848	94	96,527	75
Training and professional development[c]	119,012	120	2,382	3	116,630	117
Reading program materials[d]	182,367	185	29,395	37	152,972	148
Total	473,897	477	104,625	134	369,273	343

SOURCES: Teacher and facilitator salaries are calculated from the U.S. Department of Education schools and staffing survey. Teacher experience and facilitator experience levels were calculated based on responses to the study's teacher survey. Training and materials costs for program group schools are based on SFA program contracts. Training costs for control group schools were estimated based on interview responses.

NOTES: Rounding may cause slight discrepancies when calculating differences. All prices are given in geographically adjusted 2012 dollars. All personnel estimates include benefits costs amounting to one-third of salary.

[a] A lower bound for tutoring frequency, for both certified and volunteer tutors, was taken from responses to the principal survey. A lower bound of tutoring session length (41 minutes) was also taken from the principal survey, since all respondents indicated that tutoring sessions in their schools lasted at least 41 minutes.

[b] The estimate assumes that all non-SFA schools have a reading program facilitator equivalent to 0.36 FTE.

[c] Estimates include point coach training and observation time.

[d] Reading program materials include teacher and student textbooks, classroom consumables, assessment materials, support and licensing fees, and intervention materials when applicable. All study schools in this district purchased a new set of reading program materials in Year 1.

Using weighted average for teacher experience. Applying this change would increase the per-student annualized total resource cost difference by $2/year. Appendix Figure E.3 shows that teacher experience varied between program and control group schools in the case district, as SFA schools had a higher proportion of novice teachers than control group schools. This difference increased over time, as SFA schools hired increasing numbers of novice teachers.[15] To account for this difference, this sensitivity check replaces the median experience level used in Table 6.5 with a weighted average of teacher experience for each school and year.

[15] It is worth noting that the dramatic decrease in total number of teachers in control group schools is likely due to survey attrition, not due to an actual exodus from these schools.

Appendix Table E.4

Sensitivity Checks: Changes in Total Annualized Cost and Per-Student Annualized Cost, for Select Changes to Staffing Assumptions in District D

	Program Group Cost ($)		Control Group Cost ($)		Difference		Relevant Cost Difference in Table 6.5		Change to Total Annualized Cost Difference with Sensitivity Check Applied	
Sensitivity Check	Total	Per Student	Total	Per Student	Total	Per Student	Total	Per Student	Total	Per Student
Items included in resource cost analysis[a]										
Weighted average for teacher experience	930,105	506	927,788	477	2,317	29	-3,452	27	5,769	2
Reduce training time to 1 hour/session:										
Teacher time[b]	15,293	8	6,641	3	8,652	5	25,596	18	-16,944	-13
Control group school direct training costs[c]	132,908	72	789	0	132,119	72	130,540	71	1,578	1
Increase control group teacher time on support teams from 0 to 15 hours per year[d]	7,144	4	1,926	1	5,218	3	7,144	4	-1,926	-1
Additional costs										
Instructional support staff[e]	79,869	43	63,039	32	16,830	11	—	—	16,830	11
District coach (program and control group schools)	83,117	45	4,207	2	78,910	43	—	—	78,910	43
Substitute teachers	65,061	35	45,652	23	19,409	12	—	—	19,409	12

(continued)

Appendix Table E.4 (continued)

SOURCES: Teacher full-time equivalent (FTE) values, instructional support staff FTE and salary values, and substitute teacher salaries are taken from publicly available state and district data. Teacher salaries are calculated from the U.S. Department of Education schools and staffing survey. Teacher experience levels were calculated based on responses to the study's teacher survey. Point coach and district coach costs were obtained from correspondence with the SFA Foundation. Training costs for program group schools were obtained from SFA program contracts. Training costs for control group schools were estimated based on interview responses. Substitute teacher counts and days required were calculated based on records of chronically absent teachers published by the U.S. Department of Education's Office of Civil Rights.

NOTES: Rounding may cause slight discrepancies in calculating differences. Prices are given in geographically adjusted 2012 dollars. Personnel estimates include benefits amounting to one-third of salary. See Appendix Table E.6 for a more detailed description of cost calculations and assumptions.

[a] Annualized cost analysis results are available in Table 6.5.

[b] This sensitivity check reduces the length of each training session from 3 hours to 1 hour for all schools.

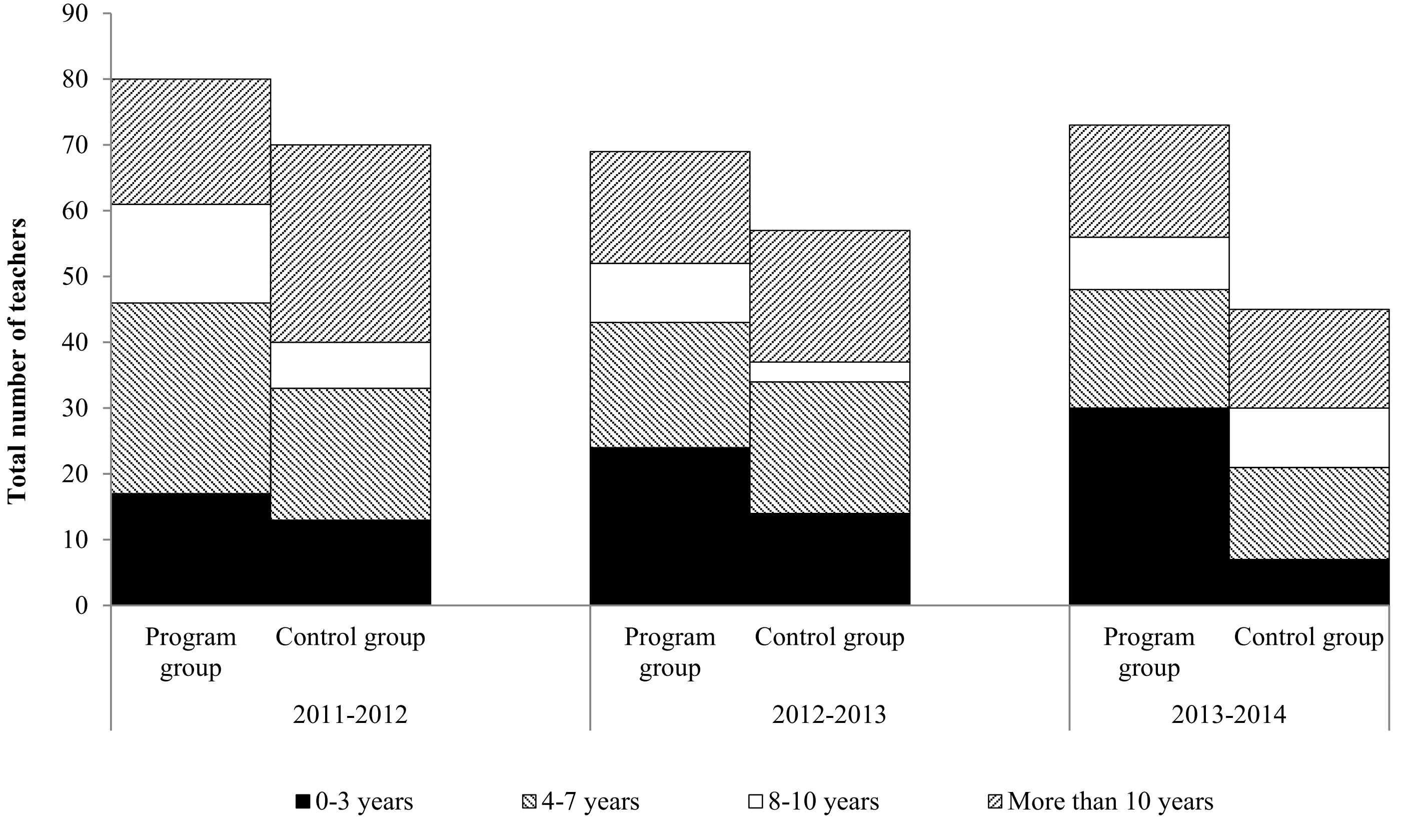
[c] This sensitivity check reports the savings for control group schools if the trainer's time was reduced from 3 hours per session to 1 hour per session. Actual SFA training costs remain the same in this analysis.

[d] This check adjusts values for the control group only; program group values are taken from the cost analysis as reported in Table 6.5.

[e] This count includes English language learner support staff but not special education support staff. Before the implementation of SFA in this district, there was a large and persistent difference in special education staffing between program group and control group schools, so these differences in staffing could not be attributed to the program.

Appendix Figure E.3

Reading Teacher Experience Levels in Study District D, by School Program Status and Year



SOURCE: Teacher survey responses.

NOTE: This figure shows experience levels for all reading teachers in grades 1-5 who responded to the study's teacher survey in a given year.

Reducing the length of each training session to 1 hour. The effect on teacher time reduces the per-student cost difference by $13 per year. The effect of reducing trainer time (direct training costs) increases the cost difference by a negligible $1 per year.

In Table 6.5, each training session is assumed to last 3 hours. This sensitivity check is shown on two separate rows in Appendix Table E.4; the first row shows less teacher time spent in training for both program and control group teachers, and the second row shows the decrease in control group schools' direct training costs. Direct training costs that schools would pay to SFA could differ based on how long training sessions last, but data available to the study team could not be disaggregated this way, so total training costs are reported in the main tables.

Increasing control group time on support teams from 0 to 15 hours per year. Applying this change produces a negligible decrease of $1 per year in estimated cost difference.

In Table 6.5, the analysis assumes that control group school teachers spend no time on support teams. Note that 15 hours per year is also the amount assumed for SFA schools for which the Snapshot rating form did not indicate full implementation of Solutions Teams.

Adding instructional support staff. Including this support staff increases the per-student cost difference by $11 per year.

Schools often use support staff members, including English language learner support and special education support staff, to deliver reading interventions in elementary schools. The sensitivity check does not include special education staff, however. A very large special education support staff in place at one program group school before SFA was implemented contributed to a significant and persistent difference in the number of support staff members used in SFA schools compared with control group schools. The inclusion of special education staff in the sensitivity check would produce an exaggerated estimate of the difference between SFA schools and control group schools in support staff deployment and would attribute previously existing differences to the introduction of SFA.

Adding in district coach. Including the district coach's time spent with both program and control group schools produces the largest change of all the sensitivity checks in Appendix Table E.4, increasing the annual per-student resource cost difference by $43.

District D had a district coach, or a locally appointed staff person who supplemented the work of the SFA Foundation point coach. Estimates for the district coach's time and salary were taken from SFAF records. Because district coaches were often existing district employees, districts allocated a portion of the coaches' time to SFA. For allocated SFA time, the coach's recorded time was divided evenly among the three SFA schools in District D. Of the remaining one-third of their time, about 10 percent was assumed to be spent supporting reading in the three control group schools (given that the coach was likely to be supporting other schools outside the study).

Adding in substitute teachers. Adding in substitute teachers increases the annual per-student resource cost by $12.

Substitute teacher counts were estimated using data from the U.S. Department of Education's Office of Civil Rights (OCR) for the first year.[16] The share of teachers who were chronically absent in 2011-2012 was calculated from the OCR's publicly available data and then applied to staff counts taken from state administrative data. Because the data are reported as the number of teachers absent 10 or more days, the study team had to make some assumptions. Half of all chronically absent teachers (absent 10 or more days) were assumed to require a substitute for 10 days, while the other half were assumed to require a substitute for 20 days. The number of teachers absent fewer than 10 days per year is not known, so the analysis does not estimate the cost of a substitute for these teachers. Therefore this cost may be a lower bound. To assign a price to this expense, the study uses the wage rate for a daily substitute provided by the school district.

Discussion of Staffing-Related Sensitivity Checks

Adding the district coach's time resulted in the largest change to the per-student total resource cost difference in Table 6.5. Including this cost accounts for the ongoing professional development that occurred in control group schools in District D throughout the course of the study. The other staffing checks did not change the main cost difference by more than $12 per student.

Check on Full Resource Costs: Implementation Intensity Scenarios

Appendix Table E.5 shows how the estimated full resource cost difference between SFA schools and control group schools differs under low, moderate, and high levels of implementation for both SFA and control group programs. In other words, in the low-intensity scenario, both SFA schools and control group schools are assumed to implement relatively little of their required program, while in the high-intensity scenario, both SFA schools and control group schools are assumed to implement their respective reading programs to the fullest extent.

For the majority of resources, the quantities assumed in the main resource cost estimate (Table 6.5) are identical to the quantities in one of the three implementation scenarios. For these

[16] U.S. Department of Education (2015a).

Appendix Table E.5

Sensitivity Checks: Differences in Annualized Three-Year Reading Program Costs in SFA and Control Group Schools in District D, by Implementation Scenario

	Low Implementation		Moderate Implementation		High Implementation	
Costs ($)	Total Difference	Per Student Difference	Total Difference	Per Student Difference	Total Difference	Per Student Difference
Teacher time (total)[a]	51,667	59	39,615	53	117,934	97
Teachers: reading block[b]	-50,950	0	-50,950	0	-50,950	0
Teachers: support teams	7,144	4	7,689	4	6,794	4
Certified tutoring after-school	*21,299*	*12*	*14,439*	*9*	*3,780*	*5*
Teacher time in training	76,789	44	71,195	41	159,996	89
Reading program facilitator	*86,151*	*47*	*66,715*	*38*	*11,403*	*12*
Principal time devoted to reading program	11,025	9	10,392	10	12,251	12
Volunteer tutors	-22,300	-11	-59,918	-31	-117,920	-60
Facilities and classrooms[c]	5,940	16	11,587	32	22,882	63
Materials and facilitator room[d]	18,954	10	15,163	8	0	1
Training and professional development[e]	*130,540*	*71*	*152,438*	*83*	*152,438*	*83*
Reading program materials[f]	*-22,324*	*-3*	*-22,324*	*-3*	*-22,324*	*-3*
Total	259,654	199	213,667	191	176,664	206

(continued)

Appendix Table E.5 (continued)

SOURCES: Teacher and principal full-time equivalent (FTE) values are taken from publicly available state data. Teacher, principal, facilitator, and volunteer salaries are calculated from the U.S. Department of Education schools and staffing survey. Teacher and facilitator experience levels were calculated based on responses to the study's teacher survey. Tutoring frequency and duration, facilitator FTE values, control group reading programs, and principal experience levels were determined based on responses to the study's principal survey. Point coach salaries were obtained from correspondence with the SFA Foundation. Facilities and classroom costs are based on square footage costs from the Center for Benefit-Cost Studies of Education Database of Educational Prices. Training and materials costs for program group schools were obtained from SFA program contracts. Training costs for control group schools were estimated based on interview responses. Reading program materials costs for control group schools were obtained from publisher websites and prior evaluations.

NOTES: This table shows the estimated incremental full resource costs of SFA given "low," "moderate," and "high" levels of reading program implementation in both SFA and non-SFA schools. Italicized items also appear in Table 6.3. Prices are given in geographically adjusted 2012 dollars. Personnel estimates include benefits amounting to one-third of salary. See Appendix Table E.6 for a more detailed description of cost calculations and assumptions behind low, moderate, and high implementation scenarios.

Rounding may cause slight discrepancies in calculating differences.

[a]Teacher counts reflect an estimate of core-curriculum teachers. A teacher was considered a core instructional FTE if the teacher's job assignment was listed as "Elementary Classroom," "Kindergarten Classroom," "Reading Classroom," "Mathematics," or "Communications Arts."

[b]All estimates assume reading teachers spend 90 minutes per day in the reading block, except teachers of English langauge learners, who spend 4 hours per day.

[c]All estimates include the cost of a classroom for after-school tutoring by certified teachers.

[d]This estimate assumes that each SFA school had one room for reading program materials and the use of the program facilitator. The space allocated for this use in non-SFA schools varies across scenarios.

[e]Non-SFA training costs assume that the trainer's annual salary is $50,000.

[f]Reading program materials include teacher and student textbooks, classroom consumables, assessment materials, support and licensing fees, and intervention materials when applicable. This estimate assumes that all study schools purchased a new set of reading program materials in Year 1.

ingredients, the other two implementation scenarios represent sensitivity checks on the quantities used in the main cost estimate.[17] For a complete description of assumptions associated with each scenario and the full resource cost analysis, see Appendix Table E.6.

Quantities for teacher time spent in the reading block and reading program materials cost are identical in the main ingredients estimate and in all three implementation scenarios. The reading block is assumed to be fixed at 90 minutes regardless of implementation level. SFA materials costs are documented in SFAF contracts based on the number of students, rather than the level of implementation, and therefore should not be expected to vary. Control group schools' materials costs are also estimated consistently across implementation levels; all three scenarios assume adequate materials were purchased in the control group schools, given student enrollment figures and the number of students reported as receiving tutoring, which could determine purchases of intervention material.

Principal time was estimated in a fundamentally different way in the implementation scenarios than in the main cost estimates. In the main cost figure, principal time is estimated as a share of FTE. All SFA school principals are assumed to be devoting half their time to SFA, and principals at all control group schools are assumed to be devoting 2.5 hours (90 minutes for the reading block and an additional hour for general support) in a 7-hour school day to the school's reading program. In short, the main estimates assume that every SFA school requires 0.5 FTE of principal time and every control group school requires 0.36 (2.5/7) FTE of principal time.

Rather than assign a fraction of a principal FTE to each school, the implementation scenarios estimate each component of principal time separately. Varying quantities for each of these activities are assigned to SFA school and control group school principals across the scenarios.[18]

One consequence of this approach is a dramatic decrease in the estimated difference in principal time between SFA schools and control group schools. Since each comparison refers to the same implementation level in both groups, the differences in principal time spent on the reading program are relatively small in all three scenarios. Table 6.5 shows the differences in reported principal time in District D. In Table 6.5, SFA principals are estimated to spend slightly

[17] Estimates for the following ingredients are identical in the main estimate and the "low" scenario: teacher support teams (Solutions Teams in SFA), teachers' certified tutoring, teacher time in training, volunteer tutors, and training and professional development. Estimates for reading program facilitator and facilities and classrooms are identical in the main analysis and the "moderate" scenario.

[18] The following principal activities are estimated in each scenario: reading block (does not vary; 90 minutes for each scenario), time spent on Solutions Teams (support teams), tutoring observations, rehiring teachers, reallocating students to reading groups, hiring substitute teachers.

Appendix Table E.6

Inputs and Assumptions for Implementation Sensitivity Checks on Full Resource Cost Analysis

		Sensitivity Check Scenarios		
Ingredient	Full Resource Cost Analysis[a]	Low Implementation	Medium Implementation	High Implementation
Teacher time				
Reading block	Analysis assumes 90 minutes for non-ELL teachers and 4 hours for ELL teachers. The number of teachers is taken from public state data.	Same as full resource cost analysis.	Same as full resource cost analysis.	Same as full resource cost analysis.
Support teams	Program group: If the SFAF Snapshot indicates fully implemented Solutions Teams, then analysis assumes 5 teams met 9 times per year for 1 hour per session and includes 2 teachers per team. Otherwise, the analysis assumes 5 teams met 3 times per year for 1 hour per session and includes 1 teacher per team. Control group: No support team time is assigned.	Same as full resource cost analysis.	Program group: For schools where the SFAF Snapshot does not indicate fully implemented Solutions Teams, the analysis now assumes teams met 5 times per year.	Program group: For schools where the SFAF Snapshot does not indicate fully implemented Solutions Teams, the analysis now assumes teams met 9 times per year. Control group: The analysis now assumes 5 teams met 3 times per year for 1 hour per session and includes 1 teacher per team.

(continued)

Appendix Table E.6 (continued)

		Sensitivity Check Scenarios		
Ingredient	Full Resource Cost Analysis[a]	Low Implementation	Medium Implementation	High Implementation
Teacher time				
Certified tutoring	The number of certified tutors is taken from responses to the study's principal survey. All sessions are assumed to last 41 minutes. The number of sessions per week represents the lowest possible value indicated by responses to the principal survey.	Same as full resource cost analysis.	Number of minutes per session is increased to 61 minutes. Unless the principal survey indicates daily tutoring, tutoring is assumed to occur 2 days per week. If no principal survey was completed, tutoring is assumed to occur 1 day per week.	Number of minutes per session is increased to 81 minutes. Unless the principal survey indicates daily tutoring, tutoring is now assumed to occur 3 days per week. If no principal survey was completed, tutoring is assumed to occur 2 days per week.
Training time	Program group: The number of training sessions is taken from SFA program contracts. Control group: Analysis assumes 6 training days per year, plus 1 day in summer prior to year 1. Both groups: Analysis assumes all sessions last 3 hours, and 1/3 of all core teachers participate in any given training session.	Same as full resource cost analysis.	Analysis assumes an additional 4 sessions of training for new teachers in Years 2 and 3 in all schools.[b]	Analysis assumes an additional 4 sessions of training for new teachers in Years 2 and 3 in all schools.[b] Program group: Session time increased to 5 hours.

(continued)

Appendix Table E.6 (continued)

		Sensitivity Check Scenarios		
Ingredient	Full Resource Cost Analysis[a]	Low Implementation	Medium Implementation	High Implementation
Facilitator	Program group: Facilitator FTE counts are taken from responses to the study's principal survey and interviews. Control group: Analysis assumes 0.36 facilitator FTE at each school, based on principal interview responses.	Control group: Analysis assumes 0.1 facilitator FTE at each school.	Same as full resource cost analysis.	All schools are assumed to have 1 full-time facilitator.
Principal time	Program group: Analysis assumes 0.5 principal FTE devoted to the reading program. Control group: Analysis assumes 0.36 principal FTE devoted to the reading program.	For all schools, principal time includes a 90-minute reading block, 45 Solutions Team hours, 45 hours rehiring, 30 hours reallocating students (3 hours for each of 6 reallocations), and time hiring substitute teachers based on the number of chronically absent[c] teachers (5.5 hours per teacher).	For all schools, principal time includes a 90-minute reading block, 45 Solutions Team hours, 9 hours rehiring, 40 hours reallocating students (5 hours per reallocation), time for hiring substitute teachers (5.5 hours per teacher), and 200 hours observing tutoring in Year 1 and 100 hours in Years 2-3.	For all schools, principal time includes a 90-minute reading block, 45 Solutions Team hours, 0 hours rehiring, 50 hours reallocating students (7 hours per reallocation), time for hiring substitute teachers (5.5 hours per teacher), and 400 hours observing tutoring in Year 1 and 200 hours in Years 2-3.

(continued)

Appendix Table E.6 (continued)

		Sensitivity Check Scenarios		
Ingredient	Full Resource Cost Analysis[a]	Low Implementation	Medium Implementation	High Implementation
Volunteer tutors	The number of tutors is based on responses to the study's principal survey. No volunteer tutors were reported in program group schools. All sessions are assumed to last 41 minutes. The number of sessions per week represents the lowest possible value indicated by responses to the study's principal survey.	Same as full resource cost analysis.	Number of minutes per session is increased to 61. Unless the principal survey indicates daily tutoring, tutoring is assumed to occur 2 days per week. If no principal survey was completed, tutoring is assumed to occur 1 day per week.	Number of minutes per session is increased to 81. Unless the principal survey indicates daily tutoring, tutoring is now assumed to occur 3 days per week. If no principal survey was completed, tutoring is assumed to occur 2 days per week.
Facilities/ classrooms	Analysis assumes 1 teacher per classroom.	Analysis assumes 4 teachers per classroom.	Analysis assumes 2 teachers per classroom.	Same as full resource cost analysis.
Materials and facilitator room	Program group: Analysis assumes each school has 1 dedicated room for materials. Control group: No materials room is assumed.	Same as full resource cost analysis.	Control group: Analysis assumes each school has 0.2 room dedicated for materials.	All schools are assumed to have 1 dedicated room for materials.
Training	Program group: The number of training days was taken from SFA program contracts. Control group: The analysis assumes 6 days per year, plus 1 additional training day in the summer before Year 1.	Same as full resource cost analysis.	Program group: Analysis assumes an additional 4 sessions of training for new teachers in Years 2 and 3.[d] Control group: 1 additional session assumed in Years 2 and 3.	Same as moderate implementation.

(continued)

Appendix Table E.6 (continued)

		Sensitivity Check Scenarios		
Ingredient	Full Resource Cost Analysis[a]	Low Implementation	Medium Implementation	High Implementation
Reading program materials	Program group: Materials costs are taken from SFA program contracts. Control group: Materials costs assume an initial purchase in the 2011-2012 school year. Prices were obtained from publisher websites and prior evaluations.	Same as full resource cost analysis.	Same as full resource cost analysis.	Same as full resource cost analysis.

NOTES: Staff counts are provided in full-time equivalent (FTE) units. ELL = English language learner.

[a] The results of the full resource cost analysis are shown in Table 6.5.

[b] The number of new teachers in each school was calculated by applying the proportion of new teachers reported in the study's teacher survey to state totals of core instructional teachers at each school.

[c] A teacher is considered chronically absent if he or she is absent for 10 or more days in a school year. Counts of chronically absent teachers were obtained from data from the U.S. Department of Education's Office of Civil Rights.

[d] The cost of an additional session is estimated at $2,700, and was calculated by dividing total yearly training costs by number of sessions per year in each school's SFA program contract.

more time on SFA than they would in the “high” scenario, and control group principals are assumed to spend slightly more time on the reading program than they would in the “moderate” scenario. As a result, the range of total principal time cost differences in the three implementation scenarios ($10,392 to $12,251) are all significantly lower than the $48,566 difference presented in the main figure, which explains why the estimated per-student cost differences for the three scenarios are all lower than the total estimated per-student cost difference in the main figure.

Interpretation and Discussion of Implementation Scenarios

The implementation scenarios lend themselves to very different interpretations, depending on whether one focuses on total cost differences or per-student cost differences. The total cost differences show a steep decline in the relative expense of SFA as one moves from the low-intensity to the high-intensity scenario. On the other hand, the per-student cost differences are relatively consistent across the three scenarios. The discrepancy between total and per-student differences can be explained by the lower student enrollment figures in SFA schools in the case district. As such, the per-student cost estimates are more relevant for assessing the actual demands of SFA compared with control group reading programs.

The similar per-student costs in these three scenarios suggest that the full resource cost is not significantly sensitive to changes in the relevant assumptions, when they are applied to both SFA schools and control group schools. However, the range of yearly per-student full resource costs in the scenarios ($191 to $206) is slightly lower than the estimated $227 per-student in Table 6.5. This decrease should be expected, since each of the scenarios compares the cost of similar levels of implementation across both groups, while the main estimate presented in Table 6.5 reflects the starker difference in implementation actually observed in District D.

Appendix F

Methodology of the Scale-Up Analysis

Data Sources

Success for All Foundation Administrative Data

The Success for All Foundation (SFAF) provided different sources of data. For comparing samples and mapping schools by SFA adoption year and grant status, the study team used SFAF data on the name and location of schools implementing the program before or during the Investing in Innovation (i3) grant period, the year of adoption, and whether the school received an i3 subsidy. SFAF's data on start and end dates were used to produce the figure on years of implementation of the program.

To assess adequacy of program implementation, the study team used SFAF's School Achievement Snapshot, a 99-item form created for SFAF point coaches to identify areas of strength and note areas where improvement is needed. Using 67 items from the Snapshot, MDRC created a scale to measure the extent of implementation. (Snapshot items related to student engagement were excluded, because these are better seen as early outcomes of implementation.)

The Common Core of Data: School- and District-Level Data Sets

The Common Core of Data (CCD) from the National Center for Education Statistics was used to compare demographic and fiscal characteristics of schools and districts in different school samples.[1] Most of these comparisons appear in Appendix Tables F.1 through F.5. Separate district-level data sets provide fiscal and demographic data. The demographic district files were used to get data on the numbers and proportions of English language learners and students in special education. There were no analogous school-level variables. The fiscal district files were used to obtain per-pupil expenditures.

Addressing Missing Data on School Characteristics

In order to perform the comparisons described above, the study team had to merge SFAF's administrative data with CCD. But because SFAF provided school names rather than the numerical identifiers used in the CCD, the merge was imperfect. Most of the schools were matched using school name, district name, and state name. Eighty-eight out of 868 adopting schools in the pre-i3 period (about 10 percent) and 26 out of 447 adopting schools during the i3 period (about 6 percent) could not be matched to school-level CCD data sets. Nineteen out of

[1] These data are publicly available; see National Center for Education Statistics (2015).

Appendix Table F.1

Selected Characteristics of Districts That Did and Did Not Adopt SFA During the Grant Period

District Characteristics	Nonadopter Sample	Adopter Sample	Estimated Difference	P-Value	
Geographic region (% of districts)					
Northeast	23.33	12.38	10.95	0.021	**
South	16.67	41.67	-25.00	0.000	***
Midwest	41.67	22.62	19.05	0.001	***
West	18.33	23.33	-5.00	0.388	
Urbanicity (% of districts)					
Large or midsize city	24.59	33.02	-8.43	0.188	
Urban fringe or large town	29.51	31.83	-2.32	0.716	
Small town or rural area	45.90	35.15	10.75	0.104	
Race/ethnicity (school average % of students)					
White	46.49	33.14	13.35	0.004	***
Black	30.79	36.98	-6.19	0.206	
Hispanic	16.52	20.93	-4.41	0.234	
Asian	1.84	2.00	-0.16	0.808	
Other	4.37	6.95	-2.58	0.291	
Male (school average % of students)	51.28	51.23	0.04	0.863	
Percentage of students who are English language learners	4.93	7.73	-2.79	0.033	**
Percentage of students with special education status	13.93	13.90	0.03	0.967	
Student-to-FTE ratio	15.49	15.47	0.01	0.978	
Per-student expenditures ($)	14,118.76	13,426.67	692.10	0.276	
Number of districts	55	146			

(continued)

Appendix Table F.1 (continued)

SOURCES: Common Core of Data (CCD) district-level fiscal and demographic data; SFAF report on schools.

NOTES: Analysis was conducted on district-level variables and weighted by the number of schools in the district that adopted SFA or did not adopt SFA.

The adopter sample includes any district in which at least one school adopted SFA. The nonadopter sample includes any district in which at least one school decided against adopting SFA. Thirteen districts in the nonadopter sample also had schools that adopted, and therefore appear in the adopter sample as well.

Variable missing rates ranged between about 6 percent and 16 percent for districts that adopted SFA. For this sample, English language learner status was missing for about 11 percent of cases, special education status was missing for about 13 percent of cases, race and gender were missing for about 11 percent of cases, and student-to-full-time-equivalent (FTE) ratio was missing for about 16 percent of cases.

Variable missing rates ranged between about 6 percent and 9 percent for districts that did not adopt SFA.

Student-to-FTE ratio is based on district-level data in this table. The total number of full-time teachers from district-level data was used to calculate this ratio.

A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

An overall F-test was conducted to determine whether the samples are statistically different when all variables are examined simultaneously. The test uses logistic regression to predict sample membership using all variables presented in this table. The p-value of the F-test is < 0.0001, indicating with very high probability that the samples are different.

313 adopting districts during the pre-i3 period and 6 out of 146 adopting districts during the i3 period could not be matched to district-level CCD data sets. Three out of 55 districts that did not adopt SFA for any schools were missing district-level CCD data.

Missing CCD variables presented in the table notes for Appendix Tables F.1 through F.5 are expressed as percentages of the entire sample of schools or districts, and so include schools or districts that could not be merged to CCD data sets at all, as well as those that could be matched but that were missing data on the relevant variables.

Especially when the p-values of the differences are less than 0.01, as many are, missing data will not substantially alter the results. For example, in Appendix Table F.5, about 7 percent of the nonevaluation sample is missing the variable showing the percentage of Hispanic students enrolled (no evaluation schools are missing this variable). Even if all the nonevaluation schools missing this variable had 100 percent Hispanic students, the estimated difference would not change by more than 3 percentage points.

When comparing samples, however, the p-values of interest pertain to the overall F-test comparing differences across a set of characteristics, rather than the p-values for differences in any individual characteristic. This allows for a clearer takeaway on whether samples differ.

Appendix Table F.2

Selected Characteristics of Schools, by Grant Status

School Characteristics	Awarded Direct School Grant	Not Awarded Direct School Grant	Estimated Difference	P-Value
Geographic region (% of schools)				
Northeast	13.85	13.92	-0.08	0.981
South	48.21	40.76	7.45	0.121
Midwest	23.59	21.94	1.65	0.685
West	14.36	23.53	-9.17	0.016 **
Urbanicity (% of schools)				
Large or midsize city	41.97	23.79	18.18	0.000 ***
Urban fringe or large town	33.68	30.40	3.28	0.473
Small town or rural area	24.35	45.81	-21.46	0.000 ***
Title I status (% of schools)	99.38	89.37	10.00	0.000 ***
Free or reduced-price lunch (school average % of students)	76.94	68.48	8.46	0.001 ***
Race/ethnicity (school average % of students)				
White	21.93	39.04	-17.11	0.000 ***
Black	50.16	28.65	21.51	0.000 ***
Hispanic	22.80	20.65	2.15	0.449
Asian	1.66	1.88	-0.22	0.704
Other	3.45	9.79	-6.34	0.000 ***
Male (school average % of students)	51.68	51.83	-0.15	0.674
Number of grades served	7.51	6.35	1.16	0.000 ***
Number of schools	197	250		

SOURCES: 2011-2012 Common Core of Data (CCD) school-level data; SFAF report on schools.

NOTES: Missing rates were under 10 percent for all but Title I status, which was missing for about 18 percent of cases. Free/reduced-price lunch status was missing for 9.6 percent of cases.

A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

An overall F-test was conducted to determine whether the samples are statistically different when all variables are examined simultaneously. The test uses logistic regression to predict sample membership using all variables presented in this table. The p-value of the F-test is < 0.0001, indicating with very high probability that the samples are different.

Appendix Table F.3

Selected Characteristics of Schools Operating SFA Before the Grant Period and During the Grant Period

	Grant Period	Pre-Grant Period	Estimated		
School Characteristics	Sample	Sample	Difference	P-Value	
School level					
Geographic region (% of schools)					
Northeast	13.89	26.74	-12.85	0.000	***
South	44.11	25.69	18.42	0.000	***
Midwest	22.69	20.49	2.20	0.362	
West	19.40	27.08	-7.68	0.002	***
Urbanicity (% of schools)					
Large or midsize city	32.14	21.27	10.87	0.000	***
Urban fringe or large town	31.90	39.17	-7.27	0.013	**
Small town or rural area	35.95	39.56	-3.61	0.222	
Title I status (% of schools)	93.73	93.01	0.72	0.652	
Free or reduced-price lunch					
(school average % of students)	72.33	72.25	0.09	0.950	
Race/ethnicity (school average % of students)					
White	31.12	32.84	-1.72	0.402	
Black	38.60	27.17	11.44	0.000	***
Hispanic	21.65	26.97	-5.32	0.004	***
Asian	1.77	3.54	-1.77	0.005	***
Other	6.86	8.10	-1.25	0.365	
Male (school average % of students)	51.76	50.86	0.90	0.000	***
Student-to-FTE ratio	16.08	15.10	0.98	0.003	***
Number of schools: 1,315	447	868			
District level					
Percentage of students who are English language learners	7.73	10.07	-2.34	0.001	***
Percentage of students with special education status	13.90	13.95	-0.05	0.875	
Number of grades served	6.88	6.70	0.18	0.247	
Per-student expenditures ($)	13,426.67	14,832.82	-1,406.15	0.00	***
Number of districts: 459	146	313			

(continued)

Appendix Table F.3 (continued)

SOURCES: Common Core of Data (CCD) school-level demographic data and district-level fiscal and demographic data; SFAF report on schools.

NOTES: SFAF received the Investing in Innovation (i3) grant in the fall of 2010 and began recruiting for the 2011 year. The final year of recruiting ends on Sept. 15, 2015. This table reflects data from the 2011-2012 CCD for the grant period schools, and from the 2009-2010 CCD for the pre-i3 grant schools. The district characteristic on per-pupil expenditures comes from the 2010-2011 fiscal file. For the grant period sample, free/reduced-price lunch status was obtained from the 2012-2013 CCD, because unreliable values were obtained from the 2011-2012 data. For the pre-i3 grant sample, free/reduced-price lunch status was obtained from the 2009-2010 CCD.

For the pre-i3 grant sample, variable missing rates ranged from about 0.4 percent to about 13 percent. Free/reduced-price lunch status was missing for about 13 percent of cases. For variables pertaining to race and gender, about 11 percent of cases had missing values.

For the i3 grant sample, missing rates were under 10 percent for all but Title I status, which was missing for about 18 percent of cases. Free/reduced-price lunch status was missing for 9.6 percent of cases.

A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

An overall F-test was conducted to determine whether the samples are statistically different when all variables are examined simultaneously. The test uses logistic regression to predict sample membership using all variables presented in this table. The p-value of the F-test is < 0.0001, indicating with very high probability that the samples are different.

Addressing Missing Geographic Values for Maps

Latitudes and longitudes were obtained from the CCD. Because some schools could not be matched to the CCD, school address data and geographic coordinates were not available. In some such cases, however, the associated district was matched to the CCD even though the school was not. In these cases, latitude and longitude were imputed by taking a random school in that district and using its latitude and longitude; all schools with missing data in a given district were given the same imputed coordinates. This mode of imputation allowed the detection of district-level geographic clustering patterns. Location was imputed for 52 schools: 4 out of 447 from the i3 grant period, and 48 out of 868 from the pre-i3 grant period. In other cases, both district and school identifiers were missing; therefore, 22 schools in the i3 period and 40 from the pre-i3 period do not appear in the map at all. Some missing schools were actually after-school centers or other entities that do not appear in the CCD, and therefore even a district match was not possible. Unless all of these schools were characterized by just one of the four outreach strategies discussed in Chapter 7, it is unlikely that these missing data change the distribution or overall pattern in a meaningful way.

Appendix Table F.4

Selected Characteristics of Schools in the SFA Scale-Up Analytic Sample and the Evaluation Sample

School Characteristics	Evaluation Sample	Scale-Up Analytic Sample	Estimated Difference	P-Value	
Geographic region (% of schools)					
Northeast	15.79	8.11	7.68	0.388	
South	68.42	16.22	52.20	0.000	***
Midwest	0.00	54.05	-54.05	0.000	***
West	15.79	21.62	-5.83	0.611	
Urbanicity (% of schools)					
Large or midsize city	63.16	35.14	28.02	0.047	**
Urban fringe or large town	21.05	37.84	-16.79	0.210	
Small town or rural area	15.79	27.03	-11.24	0.355	
Title I status (% of schools)	100.00	97.14	2.86	0.467	
Free or reduced-price lunch					
(school average % of students)	72.17	67.55	4.62	0.605	
Race/ethnicity (school average % of students)					
White	12.83	27.30	-14.47	0.102	
Black	22.26	41.76	-19.50	0.063	*
Hispanic	62.65	22.16	40.49	0.000	***
Asian	0.88	1.73	-0.84	0.132	
Other	1.38	7.06	-5.68	0.125	
Male (school average % of students)	50.99	51.15	-0.17	0.844	
Number of grades served	7.37	7.70	-0.33	0.524	
Number of schools	19	37			

SOURCE: 2011-2012 Common Core of Data (CCD).

NOTES: The total sample size is 56 schools for all variables except for Title I status: 2 schools had missing data for this variable.

The scale-up analytic sample meets the same criteria as the evaluation sample: schools that started in 2011-2012, completed three years of implementation for grades K-5, received a grant, and had a local coach. See Appendix Figure F.2 for the sample selection process.

Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

An overall F-test was conducted to determine whether the samples are statistically different when all variables are examined simultaneously. The test uses logistic regression to predict sample membership using all variables presented in this table. The p-value of the F-test is < 0.0001, indicating with very high probability that the samples are different.

Appendix Table F.5

Selected Characteristics of Schools in the Full SFA Scale-Up Sample and the Evaluation Sample

	Evaluation	Full Scale-Up	Estimated	
School Characteristics	Sample	Sample	Difference	P-Value
Geographic region (% of schools)				
Northeast	15.79	13.80	1.99	0.807
South	68.42	43.00	25.43	0.029 **
Midwest	0.00	23.73	-23.73	0.016 **
West	15.79	19.57	-3.78	0.685
Urbanicity (% of schools)				
Large or midsize city	63.16	30.67	32.48	0.003 ***
Urban fringe or large town	21.05	32.42	-11.37	0.300
Small town or rural area	15.79	36.91	-21.12	0.061 *
Title I status (% of schools)	100.00	93.39	6.61	0.248
Free or reduced-price lunch (school average % of students)	72.17	72.34	-0.17	0.977
Race/ethnicity (school average % of students)				
White	12.83	32.00	-19.17	0.019 **
Black	22.26	39.39	-17.13	0.058 *
Hispanic	62.65	19.68	42.97	0.000 ***
Asian	0.88	1.82	-0.94	0.498
Other	1.38	7.12	-5.74	0.188
Male (school average % of students)	50.99	51.79	-0.81	0.336
Number of grades served	7.37	6.86	0.51	0.374
Number of schools	19	428		

SOURCE: 2011-2012 Common Core of Data (CCD).

NOTES: No data were missing for the evaluation schools. For the 428 schools in the full scale-up sample, most variables had a missing rate between 3 percent and 10 percent. For the Title I status variable, 19 percent of cases were missing data. About 10 percent of cases were missing data for free/reduced-price lunch status.

Rounding may cause slight discrepancies in calculating sums and differences.

A two-tailed t-test was applied to the impact estimate. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

An overall F-test was conducted to determine whether the samples are statistically different when all variables are examined simultaneously. The test uses logistic regression to predict sample membership using all variables presented in this table. The p-value of the F-test is < 0.0001, indicating with very high probability that the samples are different.

Calculating Cumulative Students Served

The number of students served in a given year, as shown in Appendix Figure F.1, was determined only for the grade levels in which SFA was instituted in a school. Enrollment counts were obtained from the CCD, and missing grade-level enrollments were imputed based on median enrollment counts at schools serving the same grade levels from the 2011-2012 CCD. The enrollment numbers in the CCD represent enrollment in the fall — as of October 1 — of a given school year.

The number of students served in a given school during the first year it implemented SFA is defined as grade-level fall enrollment for all grades plus the number of students who transferred into one of these grade levels after October 1 of the same school year.

To estimate the number of students served, the study team used the fall enrollment count of the lowest grade level served and added an estimate of the number of transfers, and repeated these calculations for all years in which SFA was implemented at that school. The transfer rate for all nonevaluation schools is assumed to be the same as that observed in the evaluation sample. For all grades other than the lowest grade level, the study team estimates the number of new students in the spring by dividing the total number of students in the spring by the number of students in the fall. In the CCD for the 2011-2012 school year, this is 112.5 percent. Therefore, the number of new students entering by the end of spring will be about 12.5 percent of the fall enrollment in our estimate.

To estimate 2013-2014 and 2014-2015 data, 2012-2013 CCD enrollment data were used with the same process.

Below is an example for a school serving kindergarten through grade 5 (K-5).

Year 1

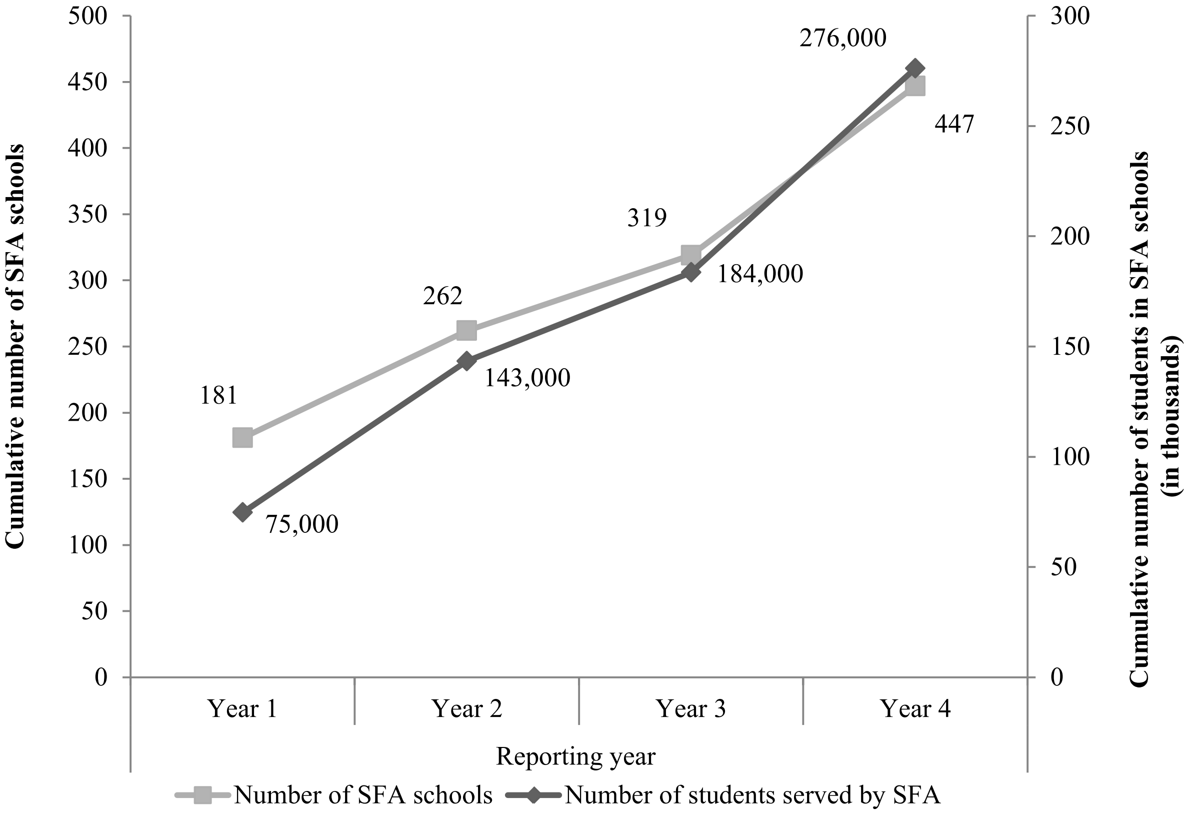
The number of students enrolled in the fall of 2013 is 50 + 42 + 44 + 39 + 38 + 42 = 255. The study team estimates that the number of new students who transfer into K-5 (from fall to the end of spring) is about 12.5 percent of the total fall enrollment. Therefore, the total number of students served at this school in Year 1 is estimated to be 1.125 * 255 = 286.9.

Year 2

The number of new students served by SFA (those who were not enrolled at all in the prior year, but enroll at some point in Year 2) equals the number of kindergartners in the fall (assumed to remain at 50) plus new kindergartners who transfer into the school after the fall (12.5 percent of 50 = 6.25) plus any new students in grades 1-5 who have entered at any time during Year 2 of this school's implementation of SFA.

Appendix Figure F.1

Cumulative SFA Schools and Students, by Reporting Year



SOURCES: SFAF report on schools, 2010-2011, 2011-2012, 2012-2013 Common Core of Data (CCD) school-level data.

NOTES: The number of SFA schools corresponds to the primary Y-axis on the left side of the chart, and the number of students served corresponds to the secondary Y-axis on the right.

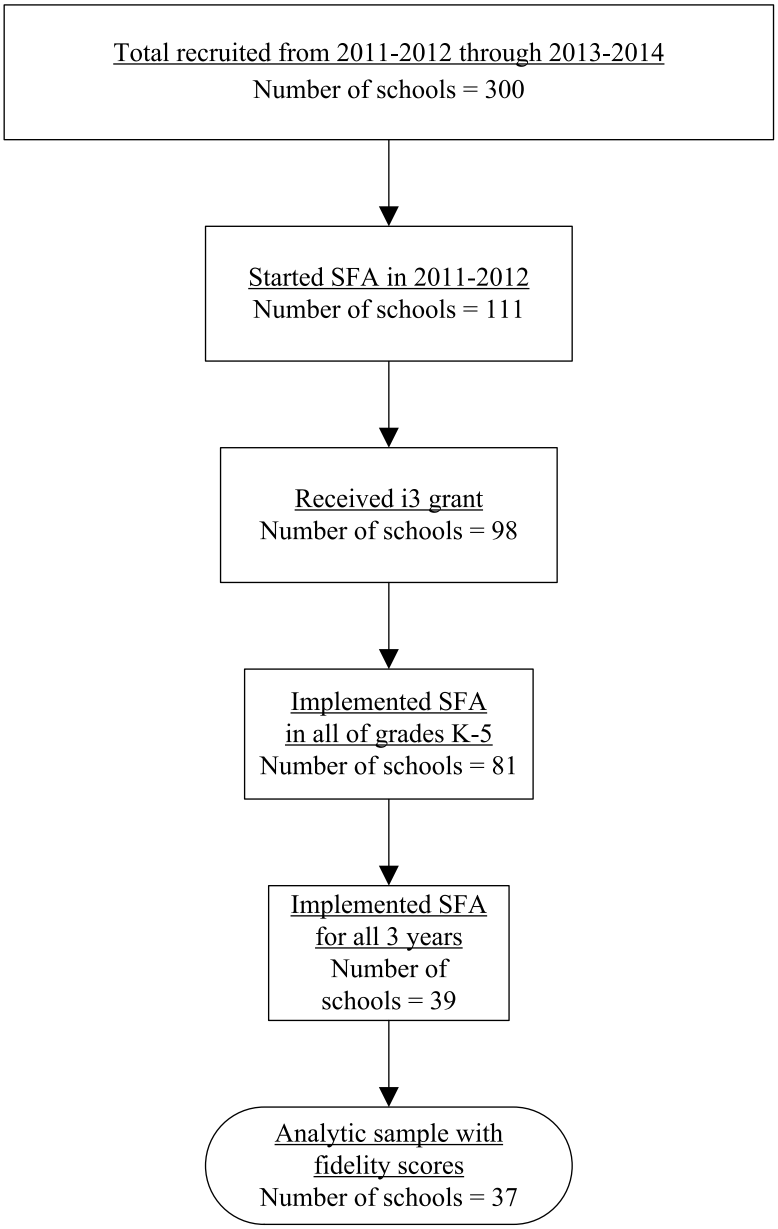
The number of students served is estimated using only the grades for which SFA is being implemented in a school.

The number of students served was calculated using CCD data. At the time of this analysis, CCD data were not available past the 2012-2013 school year. Therefore, to estimate the number of new students served in 2013-2014 and 2014-2015, 2012-2013 data were used. When CCD data were not available for a given school, the number of students served was imputed by using the median number of students by grade level for all sample schools that were not missing CCD 2011-2012 enrollment data and that had nonmissing and nonzero enrollments in the relevant grade. An imputed value was given to 309 grades from 100 schools, out of a total of 2,399 grades from 443 schools used in the calculations. Four schools were missing grade-level data and were not used in the calculations at all.

The cumulative number of schools tracks the number of SFA schools that were ever in operation since the recruitment period corresponding with the scale-up grant. This includes schools that stopped using SFA; therefore, the total presented for a given year does not equal the number of schools in operation during that year. The same is true for the cumulative total of students served for a given year.

Appendix Figure F.2

Scale-Up Sample Selection Process to Identify Nonevaluation Schools Comparable to the Evaluation Sample



SOURCES: Data provided by Success for All Foundation; characteristics defined by MDRC.

The number of transfers in a whole year (from spring to spring, or fall to fall) is estimated to be 25 percent of the prior year's enrollment of the same population. Therefore, the number of new students at fall enrollment in Year 2 will be about 25 percent of the total fall enrollment in grades 1-5 in the prior year (205 in 2013). However, students who transferred within-year in Year 1 (12.5 percent of the grades 1-5) are already counted. Therefore, the students not yet counted who enter before fall enrollment counts is 25 percent – 12.5 percent = 12.5 percent of the prior year's enrollment: in this case, 12.5 percent of 205, or 25.625. Moreover, the number of within-year transfers in grades 1-5 will also be about 12.5 percent of the Year 2 fall enrollment in grades 1-5, also 25.625.

Summing up, the total number of additional students in Year 2 is:

New fall grade 1-5 enrollees and within-year grade 1-5 transfers	2 * 25.625 = 51.25
+ new cohort of kindergartners	50.00
+ within-year kindergarten transfers	6.25
Total	=107.50

In Years 1 and 2, the total served by SFA is 286.9 students in Year 1 + 107.5 students in Year 2 = 394.4 students.

This method is applied for subsequent years until the end date of the SFA program in a school. If a school did not have an end date for SFA, it is assumed to serve students in 2014-2015.

References

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society* 57, 1: 289-300.

Black, Alison Rebeck, Fred Doolittle, Pei Zhu, Rebecca Unterman, and Jean Baldwin Grossman. 2008. *The Evaluation of Enhanced Academic Instruction in After-School Programs: Findings After the First Year of Implementation* (NCEE 2008-4021). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Borman, Geoffrey D., and Jerome V. D'Agostino. 1996. "Title I and Student Achievement: A Meta-Analysis of Federal Evaluation Results." *Educational Evaluation and Policy Analysis* 18, 4: 309-326.

Borman, Geoffrey D., and Gina M. Hewes. 2002. "The Long-Term Effects and Cost-Effectiveness of Success for All." *Educational Evaluation and Policy Analysis* 24, 4: 243-266.

Borman, Geoffrey D., Gina M. Hewes, Laura T. Overman, and Shelly Brown. 2003. "Comprehensive School Reform and Achievement: A Meta-Analysis." *Review of Educational Research* 73, 2: 125-230.

Borman, Geoffrey D., Robert E. Slavin, Alan Cheung, Anne M. Chamberlain, Nancy A. Madden, and Bette Chambers. 2005a. "Success for All: First-Year Results from the National Randomized Field Trial." *Educational Evaluation and Policy Analysis* 27, 1: 1-22.

Borman, Geoffrey D., Robert E. Slavin, Alan C. K. Cheung, Anne M. Chamberlain, Nancy A. Madden, and Bette Chambers. 2005b. "The National Randomized Field Trial of Success for All: Second-Year Outcomes." *American Educational Research Journal* 42, 4: 673-696.

Borman, Geoffrey D., Robert E. Slavin, Alan Cheung, Anne M. Chamberlain, Nancy A. Madden, and Bette Chambers. 2006. "School-Level Factors in Comprehensive School Reform." Pages 219-246 in Daniel K. Aladjem and Kathryn M. Borman (eds.), *Examining Comprehensive School Reform*. Washington, DC: Urban Institute Press.

Borman, Geoffrey D., Robert E. Slavin, Alan C. K. Cheung, Anne M. Chamberlain, Nancy A. Madden, and Bette Chambers. 2007. "Final Reading Outcomes of the National Randomized Field Trial of Success for All." *American Educational Research Journal* 44, 3: 701-731.

Bradley, M. C., Tamara Daley, Marjorie Levin, Fran O'Reilly, Amanda Parsad, Anne Robertson, and Alan Werner. 2011. *IDEA National Assessment Implementation Study: Final Report* (NCEE 2011-4027). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Center for Benefit-Cost Studies of Education. 2015. "Database of Educational Resource Prices." Website: http://cbcse.org/cost-resources/.

Center on Response to Intervention. 2015. "Gates-MacGinitie Reading Tests (GMRT)." American Institutes for Research. Website: http://www.rti4success.org/gates-macginitie-reading-tests-gmrt-reading#rel.

Coburn, Cynthia E., P. David Pearson, and Sarah Woulfin. 2011. "Reading Policy in the Era of Accountability." Pages 561-593 in Michael L. Kamil, P. David Pearson, Elizabeth Moje, and Peter Afflerbach (eds.), *Handbook of Reading Research*, vol. IV. New York: Routledge.

Common Core State Standards Initiative. 2015. "Preparing America's Students for Success." Website: http://www.corestandards.org/.

Correnti, Richard, and Brian Rowan. 2007. "Opening Up the Black Box: Literacy Instruction in Schools Participating in Three Comprehensive School Reform Programs." *American Educational Research Journal* 44, 2: 298-339.

Dee, Thomas S., and Brian A. Jacob. 2010. "The Impact of No Child Left Behind on Students, Teachers, and Schools." With comments by Caroline M. Hoxby and Helen F. Ladd. *Brookings Papers on Economic Activity* (Fall): 149-207.

Desimone, Laura M. 2013. "Reform Before NCLB." *Phi Delta Kappan* 94, 8: 59.

Dianda, Marcella, and John Flaherty. 1995. "Effects of Success for All on the Reading Achievement of First Graders in California Bilingual Programs." Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Douglas, Karen M., and Elizabeth R. Albro. 2014. "The Progress and Promise of the Reading for Understanding Research Initiative." *Educational Psychology Review* 26, 3: 341-355.

Dunn, Douglas M., and Lloyd M. Dunn. 2007. *Peabody Picture Vocabulary Test: Manual*. Minneapolis: Pearson Assessments.

Garet, Michael S., Andrew J. Wayne, Fran Stancavage, James Taylor, Kirk Walters, Mengli Song, Seth Brown, Steven Hurlburt, Pei Zhu, and Susan Sepanik. 2010. *Middle School Mathematics Professional Development Impact Study: Findings After the First Year of Implementation* (NCEE 2010-4009). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gordon, Tracy. 2012. "State and Local Budgets and the Great Recession." Brookings Institution. Website: www.brookings.edu.

Hamilton, Laura S., Brian M. Stecher, Julie A. Marsh, Jennifer Sloan McCombs, Abby Robyn, Jennifer Lin Russell, Scott Naftel, and Heather Barney. 2007. *Standards-Based Accountability Under No Child Left Behind: Experiences of Teachers and Administrators in Three States.* Santa Monica, CA: Rand Corporation.

Hanselman, Paul, and Geoffrey D. Borman. 2013. "The Impact of Success for All on Reading Achievement in Grades 3-5: Does Intervening During the Later Elementary Grades Produce the Same Benefits as Intervening Early?" *Educational Evaluation and Policy Analysis* 35, 2: 237-251.

Herlihy, Corinne, James Kemple, Howard Bloom, Pei Zhu, and Gordon Berlin. 2009. *Understanding Reading First*. New York: MDRC.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey. 2007. *Empirical Benchmarks for Interpreting Effect Sizes in Research*. New York: MDRC.

Konstantopoulos, Spyros, and Larry V. Hedges. 2004. "Meta-Analysis." Pages 281-297 in David W. Kaplan (ed.), *Handbook of Quantitative Methodology for the Social Sciences*. New York: Sage.

Lemons, Christopher J., Douglas Fuchs, Jennifer K. Gilbert, and Lynn S. Fuchs. 2014. "Evidence-Based Practices in a Changing World: Reconsidering the Counterfactual in Education Research." *Educational Researcher* 43, 5: 242-252. doi: 10.3102/0013189X14539189.

Levin, Henry M., and Patrick J. McEwan. 2001. *Cost-Effectiveness Analysis: Methods and Applications.* Thousand Oaks, CA: Sage.

Madden, Nancy A., Robert E. Slavin, Nancy L. Karweit, Lawrence J. Dolan, and Barbara A. Wasik. 1993. "Success for All: Longitudinal Effects of a Restructuring Program for Inner-City Elementary Schools." *American Educational Research Journal* 30, 1: 123-148.

Mather, Nancy, and Richard W. Woodcock. 2001. *Woodcock-Johnson III Tests of Achievement Examiner's Manual.* Itasca, IL: Riverside Publishing.

McGrew, Kevin S., Frederick A. Schrank, and Richard W. Woodcock. 2007. *Woodcock-Johnson III Normative Update Technical Manual.* Rolling Meadows, IL: Riverside Publishing.

Moore, Mark A., Anthony E. Boardman, and Aidan R. Vining. 2013. "More Appropriate Discounting: The Rate of Social Time Preference and the Value of the Social Discount Rate." *Journal of Benefit-Cost Analysis* 4, 01: 1-16.

National Assessment of Educational Progress. 2013. "The Nation's Report Card." Website: http://www.nationsreportcard.gov/reading_math_2013/.

National Bureau of Economic Research. 2010. *Business Cycle Dating Committee*. Press release. Cambridge, MA: National Bureau of Economic Research. Website: http://www.nber.org/cycles/sept2010.html.

National Center for Education Statistics. 2011. "Schools and Staffing Survey (SASS)." U.S. Department of Education. Website: https://nces.ed.gov/surveys/sass/.

National Center for Education Statistics. 2013. "Digest of Education Statistics." U.S. Department of Education. Website: https://nces.ed.gov/programs/digest/index.asp.

National Center for Education Statistics. 2015. "Common Core of Data (CCD)." U.S. Department of Education. Website: https://nces.ed.gov/ccd/.

National Reading Panel. 2000. *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction.* Reports of the Subgroups. Washington, DC: National Institute of Child Health and Human Development, National Institutes of Health.

Nye, Barbara, Larry V. Hedges, and Spyros Konstantopoulos. 1999. "The Long-Term Effects of Small Classes: A Five-Year Follow-Up of the Tennessee Class Size Experiment." *Educational Evaluation and Policy Analysis* 21, 2: 127-142.

Oliff, Phil, and Michael Leachman. 2011. "New School Year Brings Steep Cuts in State Funding for Schools." Center on Budget and Policy Priorities. Updated October 7. Website: www.cbpp.org.

Oliff, Phil, Chris Mai, and Michael Leachman. 2012. "New School Year Brings More Cuts In State Funding for Schools." Center on Budget and Policy Priorities. Updated September 4. Website: www.cbpp.org.

PRO-ED. 2012. "TOWRE-2 Test of Word Reading Efficiency - Second Edition, Complete Kit." Website: http://www.proedinc.com/.

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.

Quint, Janet C., Rekha Balu, Micah DeLaurentis, Shelley Rappaport, Thomas J. Smith, and Pei Zhu. 2013. *The Success for All Model of School Reform: Early Findings from the Investing in Innovation (i3) Scale-Up*. New York: MDRC.

Quint, Janet C., Rekha Balu, Micah DeLaurentis, Shelley Rappaport, Thomas J. Smith, and Pei Zhu. 2014. *The Success for All Model of School Reform: Interim Findings from the Investing in Innovation (i3) Scale-Up*. New York: MDRC.

Quint, Janet, Howard S. Bloom, Alison Rebeck Black, Lefleur Stephens, and Theresa M. Akey. 2005. *The Challenge of Scaling Up Educational Reform: Findings and Lessons from First Things First*. New York: MDRC.

Riverside Publishing. 2011. "Gates-MacGinitie Reading Tests® (GMRT®) Fourth Edition, Forms S and T - Paper-Pencil." Houghton Mifflin Harcourt. Website: http://www.riversidepublishing.com/.

Ross, Steven M., Marty Alberg, and Mary McNelis. 1997. *Evaluation of Elementary School School-Wide Programs: Clover Park School District, Year 1: 1996-97.* Memphis, TN: Center for Research in Educational Policy.

Ross, Steven M., and Jason Casey. 1998. *Longitudinal Study of Student Literacy Achievement in Different Title I School-Wide Programs In Ft. Wayne Community Schools, Year 2: First Grade Results*. Memphis, TN: Center for Research in Educational Policy.

Ross, Steven M., Mary McNelis, Tracey Lewis, and Steven Loomis. 1998. *Evaluation of Success for All Programs: Little Rock School District, Year 1: 1997-1998*. Memphis, TN: Center for Research in Educational Policy.

Rowan, Brian, Eric Camburn, and Richard Correnti. 2004. "Using Teacher Logs to Measure the Enacted Curriculum: A Study of Literacy Teaching in Third-Grade Classrooms." *Elementary School Journal* 105, 1: 75-101.

Rowan, Brian, Richard Correnti, Robert J. Miller, and Eric M. Camburn. 2009. *School Improvement by Design: Lessons from a Study of Comprehensive School Reform Programs*. Philadelphia: Consortium for Policy Research in Education.

Simon, Jessica. 2011. "A Cost-Effectiveness Analysis of Early Literacy Interventions." PhD diss., Columbia University.

Smith, L. J., S. M. Ross, A. Faulks, J. Casey, M. Shapiro, and B. Johnson. 1993. *1991-1992 Ft.Wayne, Indiana SFA Results*. Memphis, TN: Center for Research in Educational Policy.

U.S. Department of Education. 2009. "American Recovery and Reinvestment Act of 2009: Title I, Part A Funds for Grants to Local Education Agencies." Website: http://www2.ed.gov/policy/gen/leg/recovery/factsheet/title-i.html.

U.S. Department of Education. 2015a. "Civil Rights Data Collection." Website: http://ocrdata.ed.gov/.

U.S. Department of Education. 2015b. "Department of Education Budget Tables." Website: http://www2.ed.gov/about/overview/budget/tables.html.

U.S. Executive Office of the President. 2011. "Teacher Jobs at Risk." Website: www.whitehouse.gov.

Wendling, Barbara J., Fredrick A. Schrank, and Ara J. Schmitt. 2007. *Educational Interventions Related to the Woodcock-Johnson III Tests of Achievement* (Assessment Service Bulletin No. 8). Rolling Meadows, IL: Riverside Publishing.

What Works Clearinghouse. 2009. *WWC Intervention Report: Success for All®*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.

What Works Clearinghouse. 2013. "Adolescent Literacy intervention report: LANGUAGE!®" Institute of Education Sciences, U.S. Department of Education. Website: http://whatworks.ed.gov.

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

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 City and Oakland, California, 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 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.