THE SUCCESS FOR ALL MODEL OF SCHOOL REFORM

Early Findings from the Investing in Innovation (i3) Scale-Up

Janet C. Quint
Rekha Balu
Micah DeLaurentis
Shelley Rappaport
Thomas J. Smith
Pei Zhu

October 2013

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

The Success for All Model of School Reform

Early Findings from the Investing in Innovation (i3) Scale-Up

Janet C. Quint
Rekha Balu
Micah DeLaurentis
Shelley Rappaport
Thomas J. Smith
Pei Zhu

with

Emma Alterman
Herbert Collado
Emily Pramik

October 2013

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, The George Gund 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 Web site: www.mdrc.org.

Overview

First implemented in 1987, the Success for All (SFA) school reform model combines three basic elements:

- Reading instruction that is characterized by an emphasis on phonics for beginning readers and comprehension for students at all levels, a highly structured curriculum, an emphasis on cooperative learning, across-grade ability grouping and periodic regrouping, frequent assessments, and tutoring for students who need extra help
- Whole-school improvement components that address noninstructional issues
- Strategies to secure teacher buy-in, provide school personnel with initial and ongoing training, and foster shared school leadership

Success for All was selected to receive a five-year scale-up grant under the U.S. Department of Education's first Investing in Innovation (i3) competition. This report, the first of three, examines the program's implementation and impacts in 2011-2012, the first year of operation, at 37 kindergarten through grades 5 and 6 (K-5 and K-6) schools in five school districts that agreed to be part of the scale-up evaluation: 19 "program group" schools were randomly selected to operate SFA, and 18 "control group" schools did not receive the intervention. Program and control group schools were very similar at the start of the study. The analysis compares the experiences of school staff as well as the reading performance of a cohort of kindergarten students who remained in SFA schools throughout the year (and therefore received the maximum "dosage" of the program) with those of their counterparts in the control group schools.

Key Findings

- While teachers in the SFA schools initially expressed concerns about implementing this new, complex, and demanding initiative, by the end of the first year, many teachers were beginning to feel more comfortable with the program.
- Almost all the program group schools had reached a satisfactory level of early implementation as determined by the Success for All Foundation, the nonprofit organization that provides materials, training, and support to schools operating the reform. Yet there was also ample room for schools to implement additional program elements and to refine the elements that they had put in place.
- Reading instruction in the two sets of schools was found to differ in key ways.
- Kindergartners in the SFA schools scored significantly higher than their control group counterparts on one of two standardized measures of early reading. The impact on this measure seems to be robust across a range of demographic and socioeconomic subgroups, as well as across students with different levels of literacy skills at baseline.

Subsequent reports will examine the reading skills of these students as they progress through first and second grades and will also measure the reading skills of students in the upper elementary grades.

Contents

Appendix

List of Exhibits

Table

Table

Table

Figure

Preface

Improving reading instruction and student reading achievement has been a major focus of federal, state, and local education programs in recent decades because reading is a foundational skill for all academic success. Though progress has been made in understanding the characteristics of effective reading instruction, student reading achievement has improved only gradually over the past 20 years, and black and Hispanic students continue to lag behind their white counterparts.

A continuing issue has been how to secure wider take-up by schools and school districts of instructional approaches that have been shown to be effective. One such approach is Success for All (SFA). Starting in the 1980s, SFA's developers, Robert Slavin and Nancy Madden, developed a reading program — including curriculum materials and teacher professional development — that emphasizes both phonics and comprehension and that makes extensive use of cooperative learning techniques. In the following years, they and their colleagues built a substantial research record demonstrating SFA's positive effects on students' reading achievement.

The U.S Department of Education's Investing in Innovation (i3) program has created an opportunity to expand programs that have previously been shown to be effective and to test their continued effectiveness at scale and in new settings. SFA's solid evidentiary base positioned it to win one of the first scale-up grants awarded under the i3 competition. This report is the first of three that examine the implementation and impacts of the i3 scale-up of SFA.

This fresh look at SFA will address important new questions about this much-studied initiative. As schools' approaches to reading instruction continue to evolve, does the SFA model remain substantially different from other reading programs used in the participating districts, or have the strategies used by those schools begun to look more like SFA? How do teachers and principals respond to SFA in a world of high-stakes testing and school accountability? Can schools implement SFA with the needed fidelity? And, finally, does SFA continue to produce better reading outcomes for students than other programs? This report provides early, encouraging answers based on the first year of program implementation. Later reports on this project will continue the analysis through two more years of program implementation and will follow children into later grades.

Gordon L. Berlin
President

Acknowledgments

This report reflects the efforts of a great many people. First and foremost, we want to acknowledge the principals, research liaisons, and teachers at the schools that participated in this study of the Success for All program. Their assistance and cooperation were vital for providing the rich and detailed information on which this report is based. Staff members in the central offices of the five school districts where the study took place provided us with critical student records data.

At the Success for All Foundation, Nancy Madden, Dan Anderson, Sharon Fox, and Jill Hanson were unfailingly responsive to our information requests. Nancy Madden and Robert Slavin provided helpful critiques of earlier drafts.

Joseph Jenkins, Professor Emeritus at the University of Washington, helped us assess the advantages and drawbacks of several assessment measures. Brian Rowan and Robin Jacob at the University of Michigan provided useful insights on the study's instructional logs and their analysis.

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 Matthew Au worked with DIR to ensure that these data were collected on schedule and with minimal disruption to school operations.

Also 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. Marjorie Singer, Mariella Quimoyoz-Cruz, Ann Fenton, and Amor Aquino helped resolve a number of issues that arose.

Fred Doolittle ensured that the team received both material and moral support at every juncture. He, along with Margaret Bald, Jean Grossman, Robert Ivry, and Leigh Parise, carefully reviewed earlier drafts of the report and made comments that improved the final product. Mario Flecha was his usual unflappable self in handling a variety of administrative and other tasks. Elizabeth Nelson assisted with the coding of the qualitative data.

Robert Weber edited the report with his customary mix of sensitivity and punctiliousness, and Stephanie Cowell prepared the report for publication.

The Authors

Executive Summary

First implemented in 1987, Success for All (SFA) is one of the best-known and most thoroughly evaluated school reform models. It combines three basic elements:

- Reading instruction that emphasizes phonics for beginning readers and comprehension for students at all levels, and that is characterized by a highly structured curriculum, an emphasis on cooperative learning, across-grade ability grouping and periodic regrouping, frequent assessments, and tutoring for students who need extra help
- Whole-school improvement components that address noninstructional issues that can affect student learning, such as behavior, attendance, and parental involvement
- A set of strategies for securing teacher buy-in, providing school personnel with initial training and ongoing professional development, and fostering shared leadership in schools

Previous evaluations of Success for All showed positive effects on students' reading performance. The strength of the program's evidentiary base was critical to the selection in 2010 of the Success for All Foundation (SFAF) as one of four recipients of five-year scale-up grants awarded under the U.S. Department of Education's first Investing in Innovation (i3) competition. SFAF is the nonprofit organization that provides materials, training, and support to schools implementing the intervention. The i3 grant called for SFAF to expand its operations to 1,100 additional schools over the five-year period and for MDRC to conduct an independent evaluation of the implementation and impacts of that expansion.

Further study of the initiative is especially important for two reasons. First, over the decade since SFA was rigorously studied, the program model has continued to evolve, with greater emphasis being placed on the use of engaging technology in the classroom. Second, many school reading programs have also modified their practices, strengthening their teaching of phonics and incorporating additional instructional supports for students who are not making adequate progress in the classroom. These developments leave open the question of whether SFA continues to lead the early reading field.

This report, the first of three, examines the program's implementation and impacts in 2011-2012, the first year of operation. Thirty-seven kindergarten through grades 5 and 6 (K-5 and K-6) schools in five school districts agreed to be part of the scale-up evaluation; 19 "program group" schools were randomly selected to operate SFA, and 18 "control group" schools did not receive the intervention. The analysis compares the experiences of school staff as well as the reading performance of a cohort of kindergarten students in the SFA schools with those of their counterparts in the control group schools. Subsequent reports will examine the reading skills of these students as they progress through first and second grades and will also measure the reading skills of students in the upper elementary grades.

Data sources for the report include principal surveys, teacher surveys, and teacher-completed logs describing reading instruction, which were administered at all schools; "School Achievement Snapshot" forms completed by SFAF coaches to report on the extent of program implementation; assessments administered to kindergartners at the beginning and end of the school year; and administrative records obtained from the districts. During spring 2012 site visits, researchers also conducted interviews with principals at SFA schools and control group schools as well as interviews with SFA facilitators and focus groups with teachers at SFA schools.

In brief, this report finds that while teachers initially expressed concerns about implementing this new, complex, and demanding initiative, by the end of the first year, almost all the program schools had reached what SFAF considers a satisfactory level of early implementation, and many teachers were beginning to feel more comfortable with the program. Reading instruction in SFA schools was found to differ in key ways from instruction in the control group schools. Finally, kindergartners in the SFA schools scored significantly higher than their control group counterparts on one of two standardized measures of early reading.

Characteristics of Participating Schools and Students

- **The 19 program group schools were similar to the 18 control group schools on all school-level characteristics at baseline, although the 37 evaluation schools were not fully representative of all schools participating in SFA's i3 scale-up.**

As expected, random assignment produced two groups of schools that, at the outset of the demonstration, had similar characteristics. The evaluation schools tended to be larger than other schools participating in the SFA scale-up and to serve more Hispanic students — not surprising, given the location of the majority of the evaluation schools in districts within 200 miles of the U.S. border with Mexico.

- **While on most student-level characteristics students in the program and control group schools were statistically indistinguishable, students in SFA schools were significantly more likely than those in control group schools to be English language learners, and control group students had**

slightly higher scores, on average, on one of two measures of early reading skills.

Randomization in the research design ensures that any baseline differences between the program group and the control group are themselves random — that is, due to chance. Nonetheless, these chance differences can be statistically significant when the samples are very large. The impact analysis controls statistically for these baseline differences.

- **Mobility was fairly high in the study schools. About 10 percent of kindergarten students who were enrolled in fall 2011 (the baseline point for the study) left the study schools over the course of the school year, and about 11 percent of students who were enrolled in these schools in spring 2012 had transferred during the year from a school that was not in the study.**

Mobility patterns were similar in program and control group schools. The main analysis sample for which impacts are measured consists of students who did not move either in or out of the study schools. It comprises 2,567 students who were enrolled in fall 2011 in regular (not special education) classes and who could be tested in English, remained in these schools in the spring, and had valid scores on at least one of the two standardized reading measures used in the evaluation. This group of students had the best chance of receiving the full amount of the SFA program during the year.

Implementing the Initiative

- **The adoption, summer workshops, and professional development processes prescribed by SFAF were generally followed, although some teachers voiced concerns about each of these sets of activities.**

At all the study schools, at least 75 percent of the teachers voted to adopt SFA, although, in retrospect, some teachers reported that they were given limited information beforehand and that the decision was rushed. In a similar vein, teachers reported that workshops held before the school year began and professional development from SFA coaches did not fully prepare them for the day-to-day experience of teaching in an SFA classroom.

- **Teachers and facilitators at the SFA schools frequently reported that insufficient staff made it difficult for the schools to do everything that they were expected to do. Teachers also voiced concerns about implementing some aspects of the program model.**

Although 15 of the 19 schools had SFA facilitators who were supposed to be available full time, many of these individuals had to divide their efforts between the program and other non-SFA responsibilities. Many schools lacked the staff needed to put in place the daily tutoring for struggling students that is called for by the program model. A mismatch between the number of students identified for instruction at a certain level and the number of teachers prepared to teach at that level sometimes complicated the regrouping process, making for too many students at some levels and too few at others (so that students at somewhat different levels had to be grouped together). And staffing challenges also meant that about half the schools did not put in place the committees charged with implementing the whole-school aspects of SFA.

Features of the program model also posed implementation difficulties. Some teachers complained that the highly structured curriculum stifled their creativity. They also feared that classes were moving too quickly for struggling students, and some distrusted SFAF's reassurances that students who did not grasp the material the first time around would have opportunities to relearn it at a later point. Finally, school staff complained that SFA's data system was complicated and demanding.

- **Such issues notwithstanding, by the end of the first year, all but one of the study schools were deemed to have met SFAF's standards for adequate first-year implementation, although there was also considerable room for improving the breadth and depth of that implementation.**

On average, the 19 study schools were judged to have put in place 85 percent of the items on the school Snapshot that describe program features that SFAF considers to have the highest priority for first-year implementation. Similarly, they were judged to have put in place 79 percent of all the features whose implementation was measured during the first year.

However, the standard for what constitutes faithful implementation of the SFA model changes as the program rolls out. Only two-thirds of all the Snapshot items were rated during the first year, mostly because SFAF does not expect schools to implement the remaining items until the second year of program operations. The Snapshot ratings indicate that all schools could improve their implementation of SFA, not only by putting in place additional program features but also by improving the depth and quality of features already in place.

- **The SFA schools varied a good deal in their implementation ratings, with higher ratings generally being found at the schools that had more experienced teachers.**

In collaboration with SFAF, the researchers used the Snapshot ratings to create a scoring system — a more refined measure of implementation that accounts for the extent of implementation by weighting key practices and taking into account the proportion of classrooms demonstrating a given practice. On average, schools earned just over half the maximum possible score for the items rated in 2011-2012 (55.8 of a maximum possible score of 105). The lowest-scoring school achieved just 40 percent of the maximum possible score, while the highest-scoring school achieved 74 percent of the maximum score. The higher-scoring schools not only pursued more practices but also implemented those practices in more classrooms. The schools in the top quartile of the scale measuring fidelity of implementation had teachers with two years' more experience, on average, than schools in the bottom three quartiles.

- **Teachers and principals agreed that SFA benefited their schools.**

Despite the issues that they faced in implementing the program, in response to a survey item, 71.4 percent of the teachers expressed agreement with the statement "Overall, your school has benefited from the SFA program." Principals were unanimous in agreeing with this statement. Moreover, by the end of the first year, many teachers reported in focus groups that they felt more comfortable with the program.

Instructional and Other Characteristics of SFA Schools and Control Group Schools

- **The reading programs used in the control group schools appear to cover similar content as SFA, and all programs provide similar kinds of materials.**

The majority of the control group schools taught reading/language arts using commonly used basal programs available from leading educational publishers. Like SFA, these cover phonics, phonemic awareness, vocabulary, fluency, and reading comprehension. Like SFA, too, these materials include a teacher's manual, reading selections for students, assessments, and strategies for assisting struggling readers.

- **Teachers in the SFA schools received more professional development in reading and rated it more highly than did teachers in control group schools.**

Teachers in the SFA schools were more likely to report having received professional development in reading instruction, and on a greater number of reading-related topics, than their counterparts in the control group schools. These patterns held whether SFA teachers were compared with teachers in all control group schools or only with teachers at those control group schools that also adopted new reading programs. In general, SFA teachers also rated the professional development that they received as more helpful than did control group teachers.

- **Several factors appear to have differentiated reading instruction in the SFA and control group schools.**

Cooperative learning and cross-grade ability grouping and periodic regrouping are key features of the SFA instructional model, and teacher survey responses make it clear that these practices were much more common in program group schools than in control group schools. Furthermore, instructional logs completed by the teachers indicate that early-grade reading instruction in SFA schools was much more likely to center on comprehension and vocabulary; teachers in control group schools, in contrast, were more likely to emphasize spelling. Finally, teachers in SFA schools were much less likely than those in control group schools to say that they changed parts of the reading program that they disliked or with which they disagreed.

- **There are no significant differences between program and control group schools along other dimensions of reading instruction that SFA deems important.**

In both program and control group schools, the average length of the reading period was approximately an hour and forty minutes. Teachers in both sets of schools gave their principals virtually identical ratings on a scale measuring instructional leadership in reading (giving the principals ratings that were somewhat higher than a neutral midpoint). The use of data to check on students' progress was equally common in the two groups of schools. Finally, while a higher proportion of program group school principals than control group school principals reported that staff members provided tutoring to students needing extra assistance, the difference was not statistically significant. (As previously noted, staffing issues made it difficult to implement the tutoring component at a number of SFA schools.)

- **There are no significant differences between SFA and control group schools in the extent to which the schools had staff members charged with improving attendance and behavior, securing parental and community support, or undertaking other whole-school improvement activities not directly related to instruction.**

While principal surveys indicate that SFA schools were generally more likely than control group schools to have an individual or a group of people who were responsible for carrying out various activities associated with non-instructional whole-school reforms, the differences were not statistically significant. Again, this may be because the committees charged with implementing the whole-school improvement aspects of SFA were not fully operational at a number of schools.

Early Program Impacts

- **By the end of the first implementation year, SFA produced a positive and statistically significant impact on one of the two reading outcomes measured for the main sample of kindergarten students who remained in their study schools for the whole school year and who had maximum possible exposure to the program.**

Both the Woodcock-Johnson Letter-Word Identification test and the Woodcock-Johnson Word Attack test measure phoneme awareness and decoding. The Letter-Word Identification test asks the student to read real words of increasing complexity, while the Word Attack test has the student apply phonic/decoding skills to nonsense words. The program impact on the Woodcock-Johnson Word Attack score is 0.55 raw score point, or 0.18 standard deviation in effect size. (Program and control group students had similar scores on the second measure, the Woodcock-Johnson Letter-Word Identification test.)

- **The program impact on the Word Attack score seems to be robust across a range of demographic and socioeconomic subgroups as well as across students with different levels of literacy skills at baseline.**

Positive and statistically significant impacts were found for male students, female students, black students, and Hispanic students, students in poverty (as defined by each district), non-English language learners, and students not in special education. (The program's impact on English language learners is positive but not statistically significant; the sample size for this subgroup is quite small.) The program's impact also does not differ for students with more or fewer literacy skills measured at baseline.

- **The positive findings on the Word Attack measure are consistent with findings in a previous study of SFA and are on a par with the impacts of other prominent school reform measures.**

Borman et al.'s study of SFA found a significant positive effect on the Word Attack measure for kindergarten students after one year of program implementation (and no impact on the Letter-Word Identification measure). Furthermore, the effect size registered in the present study is similar in magnitude to those of other major reforms.[1]

[1] Borman et al. used a meta-analysis to show that the effect size of 29 of the most widely deployed comprehensive school reforms ranged between 0.09 and 0.15 standard deviation (Borman, Hewes, Overman, and Brown, "Comprehensive School Reform and Achievement: A Meta-Analysis," *Review of Educational Research* 73: 125-230 [2003]). Similarly, the Tennessee Student-Teacher Ratio (STAR) experiment found that reducing early-grade classes in elementary schools from their standard size of 22 to 26 students to only 13 to 17

(continued)

The results of this study are encouraging. They are also preliminary, for a number of reasons. First, students were tested after only one year of exposure to the SFA program. Second, the measures used for kindergartners are less precise than those for older students, and they test phonetic skills; what ultimately matters for reading is comprehension, and this will not be assessed until students are slightly older. Third, teachers are likely to be able to implement SFA in their classrooms more easily and more smoothly in subsequent years than in this first year. Finally, it is anticipated that a number of program elements not now in place in the SFA schools — especially tutoring — will be added over time.

students significantly increased average student reading performance by 0.11 to 0.22 standard deviation in effect size (Nye, Hedges, and Konstantopoulos, "The Long-Term Effects of Small Classes: A Five-Year Follow-Up of the Tennessee Class Size Experiment," *Educational Evaluation and Policy Analysis* 21: 127-142 [1999]).

Chapter 1

Introduction

Success for All (SFA) is one of the best-known and most thoroughly evaluated school reform models. First implemented in 1987 and focused on ensuring that every child learns to read well in the elementary grades, it combines three basic elements:

- Reading instruction that emphasizes phonics for beginning readers and comprehension for students at all levels, and that is characterized by a highly specified curriculum, an emphasis on cooperative learning, across-grade ability grouping and periodic regrouping, frequent assessments, and tutoring for students who need extra help

- Whole-school improvement components that address noninstructional issues that can affect student learning, such as behavior, attendance, and parental involvement

- A set of strategies for securing teacher buy-in, providing school personnel with initial training and ongoing professional development, and fostering shared leadership in schools.

Previous evaluations, both experimental and nonexperimental, showed that, on standardized reading tests, students in SFA classrooms outperformed students receiving other kinds of reading instruction.[1] The strength of the program's evidentiary base was critical to the selection in 2010 of the Success for All Foundation as one of only four recipients of five-year scale-up grants awarded under the U.S. Department of Education's first Investing in Innovation (i3) competition. SFAF is the nonprofit organization that provides materials, training, and support to schools implementing the intervention. The i3 grant called for SFAF to expand its operations to 1,100 additional schools over the five-year period and for MDRC — a nonprofit, nonpartisan education and social policy research organization — to conduct an evaluation of the implementation and impacts of that expansion.

[1] The most salient of these evaluations is a three-year longitudinal cluster randomized experiment in which 35 Title I 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 3-5 group serving as a control group for the K-2 schools. Children in the K-2 schools scored significantly higher than their counterparts in the grades 3-5 schools on three scales from the Woodcock Reading Mastery Test. Impacts grew over time as the children progressed from kindergarten to second grade. (See Borman et al., 2007.) In other large-scale studies, results for students in SFA schools have outstripped those for students in matched comparison schools. (See, for example, Rowan, Correnti, Miller, and Camburn, 2009.)

Further evaluation of SFA is especially important for two reasons. First, the program model has continued to evolve over time, with greater emphasis being placed on the use of engaging technology in the classroom and on the deployment of school district personnel who are trained by SFAF to provide professional development services and technical assistance to schools along with SFAF coaches. Second, many school reading programs have also modified their practices since the earlier SFA evaluations were conducted. In particular, they have strengthened the teaching of phonics, and, like SFA, they have incorporated increasingly intensive instructional supports for students who are not making adequate progress (an approach commonly referred to as "Response to Intervention"). All these developments leave open the question of whether SFA continues to lead the early reading field.

The i3 evaluation of SFA employs an experimental design, in which 37 schools in five school districts that are participating 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, while the 18 control group schools did not get the intervention. The analysis compares the experiences of adults and the performance of students in the two groups of schools.

This is the first of three reports from that evaluation. It examines the implementation of SFA and its effects on student learning during the 2011-2012 school year, the first year that the program was put in place. The report considers SFA's implementation across all the grades in the 19 program group schools. The impact analysis, in contrast, centers on a group of students who entered kindergarten in the 37 study schools in fall 2011 and whose reading skills were assessed in spring 2012.[2]

The report uses quantitative and qualitative data from a wide variety of sources.[3] Through teacher and student surveys, implementation summaries completed by SFAF staff, teachers' instructional logs, interviews and focus groups conducted in the course of site visits with school personnel, school district databases, and individual assessments of students' reading skills, it addresses five main questions:

1. Are the SFA and control group schools that are participating in the i3 evaluation similar to each other and to the other schools receiving SFA under the i3 grant that are not part of the evaluation?

2. What was involved in putting the program in place, and how did school personnel respond?

[2] Subsequent reports will track the literacy growth of this group of students as they advance through first and second grades and will also measure the reading skills of students in grades 3 through 5.

[3] Appendix A describes these data sources, their purposes, and response rates.

3. To what extent were SFA's features implemented during the program's first year, and what factors were associated with more complete implementation?

4. How distinct were the program group schools and the control group schools in various aspects of school functioning?

5. Did SFA produce impacts on students' early reading skills?

The report finds, in brief, that while the treatment schools, as expected, struggled with many aspects of implementing a new and decidedly different approach to reading instruction and schoolwide collaboration, by the spring of 2012, the large majority of schools had met SFAF's expectations for what first-year schools should accomplish. While SFA schools and control group schools resembled each other in a number of respects, they differed in some important fundamentals of reading instruction. Finally, on one of two measures of an early reading skill (decoding ability), kindergartners in the SFA schools scored significantly higher than their counterparts in the control group schools.

These findings must be regarded as merely suggestive of the program's capacity to produce substantial impacts on student achievement, for several reasons. First, SFAF recognizes that it takes several years for the program to be fully implemented, and some of the components of a mature program were not yet in place during the first year. Second, it also takes time for teachers to implement lessons with skill and ease. Third, the students for whom impacts are measured will have received only one year of exposure to the program; to the extent that impacts are cumulative, this will not be reflected in the program's first year. Fourth, it is difficult to measure reading skills at this early age. Finally, while ability to decode words is an essential aspect of learning to read, ultimately what is at stake is students' ability to understand and make meaning of what they have read; comprehension will be measured in later years, but not at this early stage. Thus, while the findings presented here may whet readers' appetites for learning about SFA's implementation and impacts in the years to come, they do not necessarily provide a foretaste of future findings.

The remaining sections of this chapter describe the SFA program, the theory of change that underlies it, and the contents of this report.

The SFA Program and Its Theory of Change

Table 1.1 lays out the key features associated with each of the three main elements of the SFA program: the instructional model, the noninstructional (whole-school) components, and the implementation strategies. The key features include both structures (for example, a 90-minute reading block and a group of staff members whose mission is to improve relationships with

Table 1.1

Key Elements of the Success for All Program

The instructional model

- A K-6 reading program with three levels:
  - KinderCorner (kindergarten)
  - Reading Roots (usually first grade – beginning readers)
  - Reading Wings (usually second grade and up)
- An emphasis on phonics in the lower levels and on vocabulary and comprehension at all levels
- A 90-minute reading period
- "Scripted" lesson plans that lay out timed activities and language for teachers to use in presenting them
- Instruction that is rapidly paced, uses technology, and employs cooperative learning in pairs and small groups
- Cross-grade ability grouping for reading, with students frequently leaving their regular classroom to receive reading instruction from another teacher ("walking to read")
- Frequent use of data to monitor student learning
- Quarterly assessments to measure students' progress toward grade-level standards and to regroup students into the highest levels at which they can be successful ("aggressive placement")
- A team of staff members charged with fostering instructional improvement efforts
- Computerized small-group tutoring and individual tutoring for students who need extra assistance

Whole-school improvement features

- A "Leading for Success" continuous improvement model whose key elements include distributed leadership, quarterly review of student achievement data, and the harnessing of school resources to meet specified achievement goals
- A Leadership team (including the principal, SFA facilitator, and Schoolwide Solutions coordinator, among others) that provides vision, direction, and monitoring
- Leading for Success teams that include:
  - Instructional component teams of teachers for each level (KinderCorner, Roots, Wings)
  - "Solutions" teams of teachers and other staff members charged with:
    - Improving student attendance
    - Developing appropriate interventions (academic, behavioral, health-related, social, and attendance-related) for particular students with learning difficulties
    - Putting in place "Getting Along Together," a schoolwide program for social skills development and conflict resolution, as well as other behavioral interventions
    - Increasing family involvement
    - Engaging community businesses and institutions to support the school

Implementation strategies

- An adoption process that includes a presentation on the program followed by a teacher referendum
- Designation of a school staff member as the program facilitator
- Initial training of school leaders, program facilitators, and teachers
- Delivery of SFA curricular and other materials
- Ongoing professional development supplied by "coaches" (SFAF employees or district employees trained by SFAF) and by school-based program facilitators

students' families) and processes (for example, using data to fine-tune instruction and regroup students and providing ongoing professional development). The program's multifaceted nature and the interconnectedness of SFA components help to explain why it is likely to take several years for full and high-quality implementation to occur.

Figure 1.1 depicts the theory of change guiding the program and the evaluation. (For simplicity and clarity, the diagram has been stripped down to essential elements more or less as they unfold in chronological order.) Strategies for implementing the program are introduced to participating schools, which then put into place SFA's instructional model and its whole-school improvement components. The operationalization of these program elements leads to intermediate changes in teaching and learning and in the school environment more generally. The SFA instructional model results in improved classroom instruction for all students, in greater individualization of instruction, and in teachers' greater confidence in their ability to help all students achieve success. The whole-school improvement components benefit individual students (for example, by securing eyeglasses for students who need them or by enlisting parents in support of their children's learning) as well as the broader school environment (for example, by creating a more orderly environment and by engaging teachers in more collaborative efforts). Finally, these intermediate changes produce changes in student outcomes: greater engagement and improved attendance and behavior, higher levels of achievement, and steady academic progress through the elementary grades.

The Contents of This Report

The structure of this report largely follows the SFA program's theory of change. After this introductory chapter, Chapter 2 examines the process of selecting sites for the i3 demonstration in general and for the evaluation in particular, as well as the characteristics of the study schools and other schools participating in the i3 scale-up.

Chapter 3 examines the *implementation strategies* outlined in the leftmost column of the theory of change (Figure 1.1) and discusses teachers' responses to these strategies.

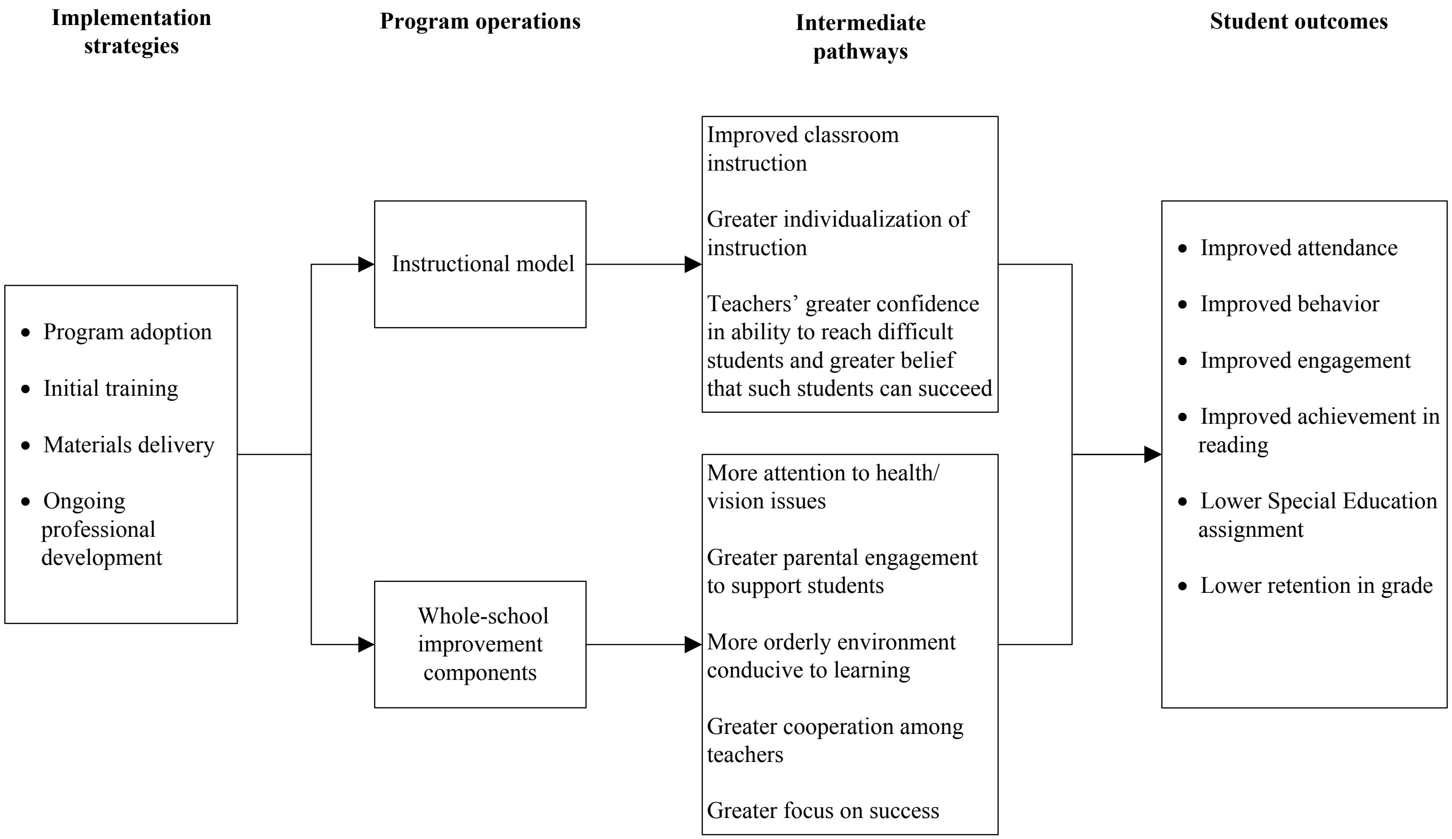
Chapter 4 deals with the second column of Figure 1.1 — *program and school operations* — and considers the extent to which program components were put in place, the factors contributing to greater or lesser implementation, and broader issues associated with implementation fidelity.

Chapter 5 uses data from the principal and teacher surveys to examine ways in which SFA schools and control group schools were similar or different along various dimensions

The Success for All Evaluation

Figure 1.1

The Success for All Theory of Change



Implementation strategies

- Program adoption
- Initial training
- Materials delivery
- Ongoing professional development

Program operations

Instructional model

Whole-school improvement components

Intermediate pathways

Improved classroom instruction

Greater individualization of instruction

Teachers’ greater confidence in ability to reach difficult students and greater belief that such students can succeed

More attention to health/ vision issues

Greater parental engagement to support students

More orderly environment conducive to learning

Greater cooperation among teachers

Greater focus on success

Student outcomes

- Improved attendance
- Improved behavior
- Improved engagement
- Improved achievement in reading
- Lower Special Education assignment
- Lower retention in grade

related to implementation strategies (especially with respect to professional development), instructional and noninstructional aspects of program operations, and the *intermediate pathways* hypothesized to lead to program impacts.

Chapter 6 assesses SFA's impact on *student outcomes* (shown in the rightmost column of Figure 1.1).

Finally, Chapter 7 reflects on what has been learned to date and lays out additional questions to be addressed in future reports.

Chapter 2

Schools and Students in the Success for All Evaluation

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

Key Findings

- The majority of the 37 study schools are located in large or midsize cities in the Northeast, South, and West; serve 550 students, on average; and enroll students who are primarily Hispanic and low-income.

- Compared both with other SFA schools in the i3 scale-up and with a national sample of schools serving low-income students, the study schools are more concentrated geographically, are larger, and serve more Hispanic students.

- About 10 percent of kindergarten students who were enrolled in fall 2011 (the baseline point for the study) left the study schools altogether over the course of the year. Conversely, about 11 percent of students who were enrolled in spring 2012 and whose reading skills were assessed at that point had transferred into their new school from a school that was not in the study.

- Random assignment produced two groups of schools that, at baseline, were very similar on school-level characteristics.

- At baseline, students in SFA schools were statistically indistinguishable from students in control group schools in most respects. The students in SFA schools were, however, significantly more likely than those in control group schools to be English language learners, and control group students had slightly higher scores, on average, on one of two measures of early reading skills.

Recruitment and Random Assignment

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 puts in place the SFA program, or a control group, which implements 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 cost, and schools willing to participate in the evaluation received the program gratis.[1] Each study school had to meet certain eligibility criteria: (1) it had to serve students from kindergarten through fifth grade; (2) at least 40 percent of the students at the school had to be eligible for the free and reduced-price lunch program; (3) the school had to be willing to participate in a random assignment experiment; (4) it had to identify a school staff member to serve as the SFA facilitator; and (5) at least 75 percent of its teachers had to vote to adopt the SFA program.

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 from 4 to 17, producing a total sample of 37 schools. In spring 2011, the schools within each district were assigned randomly to program or control conditions. The 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 and Student Samples

Table 2.2 shows the average characteristics of the 37 schools in the study sample and how the study schools compared with all 67 schools in the scale-up sample and with a national sample of elementary schools that serve students in kindergarten (K) through grade 5 and in which at least 40 percent of enrolled students are eligible for free and reduced-price lunch.[2]

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

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

The Success for All Evaluation

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
Estellita District	4	2	2
Poplar City	17	9	8
Rivers County	4	2	2
Seaboard Boroughs	6	3	3
Southwest City	6	3	3
Number of schools	37	19	18

SOURCE: Success for All evaluation data.

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

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. The average school enrolls about 550 students. The majority of students in the study schools are Hispanic and are eligible for free and reduced-price lunch.

Compared with the scale-up sample, the study schools differ in their geographical location; more than half are in the South, and none are in the Midwest. The study schools are also larger than the scale-up schools on average, in terms of both total student enrollment and the number of full-time employees in the school. Finally, the study schools have more Hispanic students and fewer black (non-Hispanic) students than the schools in the scale-up sample. This reflects the fact that three of the five study districts are located relatively close to the U.S. border with Mexico.[3]

Similar patterns exist when the study schools are compared with a national sample of elementary schools that met the demographic eligibility criteria. Overall, the study schools differ from the national sample in most measures reported in Table 2.2. Specifically, the study

[3] Chapter 6 explores the effects of SFA on Hispanic students, particularly those who were instructed in Spanish.

Table 2.2

Selected Characteristics of the Study Schools, Schools in the SFA Scale-Up, and the National Population of Similar Schools (2010-2011)

	Study Sample	Schools in Scale-Up Sample	National Population of Elementary Schools with Grades K-5 and Low-Income Students
Geographic region (% of schools)			
Northeast	16.2	11.9	13.1
South	67.6	41.8 ***	24.6 ***
Midwest	0.0	26.9 ***	24.8 ***
West	16.2	19.4	37.5 ***
Urbanicity (% of schools)			
Large or midsize city	62.2	53.7	29.1 ***
Urban fringe and large town	21.6	26.9	40.0 **
Small town and rural area	16.2	19.4	30.9 *
Title I status (% of schools)	100.0	98.5	94.7
Free and reduced-price lunch (school average % of students)	56.8	62.6	68.3 ***
Race/ethnicity (school average % of students)			
White	13.8	17.2	41.8 ***
Black	22.6	43.7 ***	18.4
Hispanic	61.8	33.9 ***	31.4 ***
Asian	0.7	1.3 *	4.1 **
Other	1.1	3.9 *	4.2 ***
Male (school average % of students)	51.3	52.0	51.4
Total school enrollment	546.5	450.3 ***	456.1 ***
Number of full-time-equivalent teachers (all grades)	32.2	27.4 ***	27.5 **
Number of schools	37	67	6,047

SOURCE: 2010-2011 Common Core of Data (CCD).

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

F-tests were applied to categorical characteristics of geographic region and urbanicity for the pair-wise comparisons. For geographic region, the comparisons between the study sample and the scale-up sample yield statistically significant differences (p-value less than 0.001), and, for urbanicity, the comparison does not yield statistically siginificant differences (p-value equal to 0.534). The differences between the study sample and the national population of schools with grades K-5 and at least 40 percent low-income students are statistically significant (p-values less than 0.001 for both variables).

To examine if there is any systematic difference between the the study sample and the scale-up sample, an F-test was conducted in a regression model controling all school characteristics reported in this table. The p-value of the test is equal to 0.491. A similar test for systematic difference between the study sample and the national population also produced a p-value of less than 0.001.

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.

schools are more likely to be located in the South and in urban areas and to have more low-income and more Hispanic students. The study schools also tend to have larger enrollments than the national sample of schools.

Once schools were randomly assigned, all the kindergarten students in these schools who were in regular classes (that is, were not in separate special education classes) in the fall of the 2011-2012 school year and who could be tested in English (all except three students) were included in the *baseline sample.* This sample comprises 2,956 students across the 37 schools.

The analysis of SFA's impacts focuses on a *main analysis sample* (or *main sample*) of 2,568 students who were present in the study schools in the fall and spring of the school year and who had valid spring test scores.[4] The sample excludes students who moved out of a study school between the fall of 2011 and the following spring — about 10 percent of all students enrolled in these schools in the fall. Students in the main sample had the best chance of receiving the full amount of the SFA program during the year.

A third sample, the *spring sample,* consists of the 2,897 students who were enrolled in the schools in spring 2012 and who had valid spring assessment scores. Like the main sample, this sample differs from the baseline sample in that it excludes "out-movers" as well as students who remained in a study school but did not have a valid spring test score. Unlike the main sample, however, it includes students who moved into a study school from a school that was not in the study during the same time period. Such "in-movers" constitute about 11 percent of the spring sample.[5]

Results for the spring sample show how the implementation of SFA affects the average performance level of all students in the schools. While the spring sample is not the primary sample for the impact analysis, it is nonetheless of interest because it takes into account student mobility, a phenomenon over which school administrators exercise little or no control and that is especially widespread in schools serving low-income students.[6]

In addition, a small proportion of students in the main sample were instructed primarily in Spanish during their kindergarten year and were tested in Spanish as well as English at both baseline and follow-up. This sample forms the *subsample of students with Spanish test scores,* and it is analyzed separately as a subgroup of the main sample. The study also explores the

[4] Students with at least one valid reading test score from the spring are included in the sample. As a result, the actual numbers of students used for the impact analysis for the two tests can differ slightly.

[5] Both the main sample and the spring sample include a small number of students in the study schools who were ineligible for testing or who otherwise went untested in fall 2012 but had at least one valid test score the following spring.

[6] For a detailed description of student mobility during the year among students in the program and control groups, see Appendix B.

program effects on various subgroups defined by student baseline characteristics, such as race/ethnicity, gender, poverty status, special education status, and English language learner status.

Equivalence of Baseline Characteristics of Program Group and Control Group Schools and Students

The purpose of random assignment is to produce a program group and a control group that are statistically equivalent on all characteristics of these schools at the start of the study. If the two groups are indeed equivalent at the outset, and if any attrition from the sample over the course of the study is balanced across groups, one can be reasonably confident that any differences that are later found in outcomes between the two groups 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 characteristics at the beginning of the study. There were no statistically significant differences in any school-level baseline characteristics between program and control groups.[7] 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 school characteristics at baseline between the two groups of schools.[8] The results indicate no such differences in school characteristics.

Using the demographic data received from students' district records, as well as baseline test scores, Table 2.4 presents selected characteristics of students in the baseline sample. On average, these students were five and a half years old as of September 1, 2011; a majority of the students were Hispanic and were poor, as defined by their school districts. About 7.5 percent were special education students.

For the baseline sample, differences between students in the program group and those in the control group on most characteristics are not statistically significant. Two notable exceptions are that the program group had a higher percentage of English language learners and that control group students registered slightly higher average scores on one of the two baseline test scores (the Woodcock-Johnson Letter-Word Identification test). Both differences are statistically significant at the 10 percent level. Randomization ensures that the program group and control group start out similar to each other at baseline and that any differences between the two groups are themselves randomly distributed — that is, any differences are due to chance. However, these chance differences can be statistically significant even when

[7] In this report, the term "statistically significant" applies to differences that had a 10 percent probability or less of having arisen by chance.

[8] 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.

The Success for All Evaluation

Table 2.3

Selected Characteristics of the Study Schools, by Program or Control 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	NA
Students eligible for free and 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: 2010-11 Common Core of Data (CCD); district-provided state reading test data, 2010-2011; state reading test records, 2010-2011.

NOTES: The estimated differences for school-level data are regression-adjusted using ordinary least squares regressions, 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 difference between the treatment and control groups, an F-test was calculated for the full sample of 37 schools in a regression model controlling the following variables: indicators of random assignment strata and all school characteristics reported in this table. The p-value of the test is 0.999.

the samples are large. An F-test was conducted across all the baseline variables as a group and points to an overall statistically significant difference between program and control group students in the baseline sample.[9]

[9] This test was also based on a logistic regression, predicting program status with the measured student-vel aseline characteristics. The P-value for the F-test is less than 0.01.

The Success for All Evaluation

Table 2.4

Selected Characteristics of the Student Baseline Sample, by Program or Control Status (Fall of School Year 2011-2012)

	Program Group	Control Group	Estimated Difference	P-Value for Estimated Difference
Age (years)[a]	5.5	5.5	0.0	0.531
Students in poverty (%)	88.1	88.5	-0.5	0.826
Race/ethnicity (%)				
White	12.9	13.5	-0.6	0.702
Black	20.4	19.4	1.0	0.819
Hispanic	63.3	64.9	-1.6	0.742
Asian	1.6	0.9	0.7	0.435
Other	1.8	1.2	0.6	0.243
Male (%)	50.8	49.3	1.5	0.460
English language learners (%)	23.2	17.0	6.2	0.071 *
Special education status (%)	7.8	7.6	0.2	0.888
Peabody Picture Vocabulary Test				
Standard score	91.4	92.2	-0.9	0.442
Percentile equivalent[b]	27	30	–	–
Woodcock-Johnson Letter-Word Identification test				
Standard score	93.3	95.0	-1.8	0.065 *
Percentile equivalent	32	37	–	–

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.

NOTES: “Student baseline sample” is defined as the set of students who were present in the sample schools in fall of 2011. This sample consists of 2,956 students across the 37 schools.

2,835 students had valid fall 2011 PPVT test scores: 1,468 are in the program group, and 1,367 are in the control group.

2,850 students had valid fall WJLWI test scores; 1,481 are in the program group, and 1,369 are 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 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 group students, an F-test was conducted for the full sample of 37 schools in a regression model controlling the following variables: indicators of random assignment strata, all student characteristics reported in this table, and corresponding missing indicators. The p-value of the test is less than 0.001.

[a] Age is calculated as the age (in years) of a student as of September 1, 2011.

[b] The percentile equivalent corresponds to the mean standard score for a given study group.

Chapter 3

Launching Success for All

Chapter 3 considers the initial stages of implementing Success for All in the 19 program group schools, beginning with the decision in spring 2011 to adopt SFA and the ensuing steps taken to help ensure that well-functioning interventions would be put into place. This description of the adoption process and the professional development and other activities that followed helps set the context for the implementation findings presented in Chapters 4 and 5.

The chapter draws on interviews with teachers, principals, and facilitators in the SFA schools and on principal and teacher surveys to describe these processes and staff members' responses to them. It should be stressed that both interviews and surveys were completed in spring 2012, often well after the events that they describe and in light of subsequent experience with the SFA program. Caution is therefore needed in interpreting these findings — particularly the interview data, which reflect subjective recall of events that took place nearly one year earlier.[1]

Key Findings

- The recruitment and adoption process prescribed by SFA was generally followed, although teachers voiced concerns about having enough information and time to make an informed choice.
- SFA facilitators who worked in the schools had strong and relevant backgrounds — as instructional specialists, reading coaches, and/or administrators — but a majority also had responsibilities outside of their SFA duties.
- Training and professional development unfolded largely as planned, although some teachers criticized both the content and the timing of some components and questioned the supportiveness of SFA "point coaches" during the school year.

[1] As documented in Appendix A, response rates to the principal and teacher surveys are very high, so that there is little doubt about the representativeness of the findings. In contrast, researchers generally conducted one or two focus groups with teachers in each school, and the focus groups generally included four to six teachers. There is no reason to doubt, but also no way to be sure, that the views of focus group participants resembled those of other teachers in their schools.

- The SFA curricular materials found general acceptance among teachers, with one significant exception: Teachers who taught SFA in Spanish in the schools with a large bilingual population complained that the Spanish-language materials arrived over the course of the year rather than at the outset, making planning difficult and requiring them at times to create their own materials. The Spanish version of SFA also lacked the multiplicity of videos that children and teachers in SFA's English-language classrooms found appealing.

- Many teachers reported that, over the course of the year, they became more comfortable with SFA's materials and procedures.

Adopting Success for All

The typical SFA implementation regimen that follows a school's decision to adopt the program requires the school to appoint a full-time SFA facilitator; have school staff attend summer trainings ("leadership training" for principals and facilitators and separate introductory workshops for teachers); distribute SFA curricular materials when they arrive at the school; and work on an ongoing basis with the coach assigned to the school — the "point coach" in SFA parlance. The coaches are employees of the Success for All Foundation (SFAF) or, in some cases, are employees of the local school district who have been trained by SFAF, to support implementation and ensure fidelity to the SFA model. Figure 3.1 shows the timeline of several key activities: initial SFAF recruitment presentations to the study schools, the schools' vote on whether or not to adopt the program, preimplementation training sessions conducted in the summer before the start of the school year, and follow-up training over the course of the year.

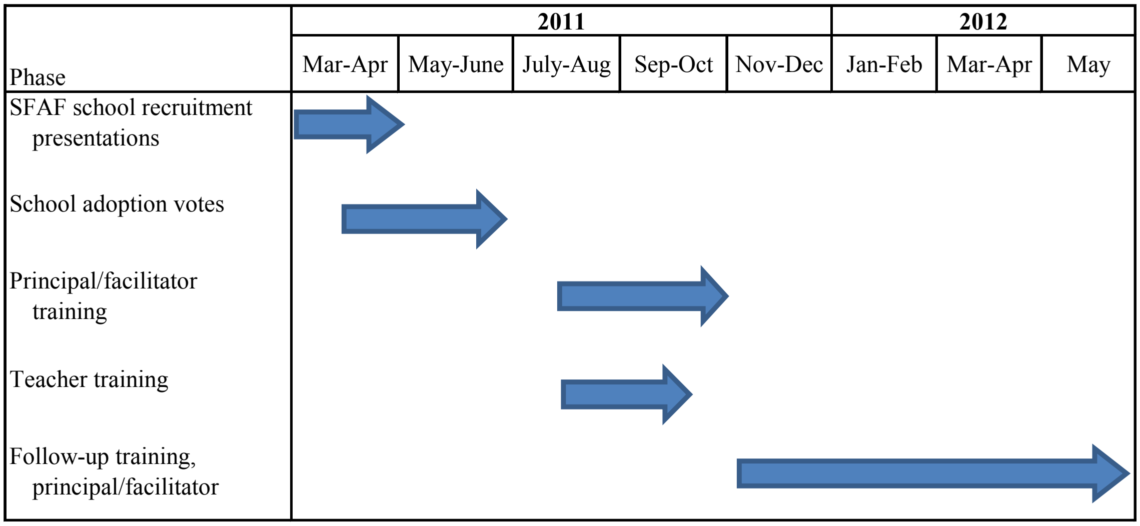
SFAF recruited study sites in two principal ways. First, it distributed information about the grant opportunities nationally through a variety of education-related outlets. Second, it drew on current and previous partnerships with schools and districts. About half the schools in the study were recruited through each method.

Next steps included telephone discussions to deepen understanding of the SFA model; the arranging of visits to existing SFA schools; presentations to district and school staff to build awareness; and, finally, a vote about whether to participate on the part of school staff. The process of random assignment was explained prior to the vote, so that school personnel knew that their school would have a 50 percent chance of being designated a control group school and, if so, would not be allowed to implement SFA for the three years of the study.

All the schools in the evaluation did, in fact, vote to adopt SFA. Their decisions to do so did not evidently stem from strong dissatisfaction with their existing reading programs. The first

Figure 3.1

Major Phases of SFA Adoption and Professional Development for Study Group Schools



SOURCE: Success for All Foundation.

portion of Table 3.1 shows support for the prior year's reading program. Over two-thirds of the teachers felt that the reading program they had been using was effective, a perception shared (to a lesser extent) by their principals.

Presumably, then, the decisions to adopt the program had much to do with the persuasiveness of SFA's presentations. Teachers were generally impressed by these presentations and were inclined to vote for SFA, despite the fact that the presentations were often fairly brief — as little as 15 minutes long, according to one teacher. SFA's emphasis on the use of data and its history of success appeared to make the program a wise choice for their schools. School personnel also reported that they were more likely to feel positive about SFA in advance of the vote if a teacher or principal in the building had previous SFA experience. In over one-third of schools, a teacher or principal had had a positive experience with SFA in the past and thus encouraged the staff to adopt the program.

Teachers in the program group schools later voiced three main concerns about the adoption process. The first concerned the timing of the vote. Many schools held the vote immediately following the SFAF presentation. This gave teachers little time to digest the information and no time to do independent research on the program to further their understanding.

Table 3.1

SFA Teachers' and Principals' Survey Responses Related to Program Adoption

	Mean Survey Scores	
	Principals	Teachers
Support for prior year's reading program		
Percentage of respondents who agree or strongly agree that the prior year's reading program was effective[a]	61.1	68.9
Percentage of principals who agree or strongly agree that their teachers believed the prior year's reading program was effective[b]	72.2	–
Adequacy of information needed to make adoption decision		
Percentage of teachers who agree or strongly agree that they had adequate information to decide to adopt the program[c]		38.7
Percentage of principals who agree or strongly agree that their teachers felt they had enough information to decide to adopt the program[d]	61.1	
Teachers' attitudes about likely efficacy of SFA		
Percentage of teachers who agree or strongly agree that they were skeptical about whether SFA would improve reading outcomes for students for the 2011-2012 school year[e]		63.7
Percentage of principals who agree or strongly agree that teachers at their school were skeptical about whether SFA would improve reading outcomes for students for the 2011-2012 school year[f]	61.1	

SOURCES: Spring 2012 teacher survey and spring 2012 principal survey.

NOTES: On the spring 2012 teacher survey, teachers were asked to indicate their level of agreement with each item reported, where 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree. The percentages who agreed or strongly agreed, as presented above, were calculated by first taking the percentage of teachers at a given school who selected 3 or 4 for a given item, where the denominator was the number of nonmissing responses.The mean of all such school-level means was taken to produce the percentages presented above. The mean across school means is used so as not to give more weight to schools with more teachers or vice versa.

[a] SFA means and standard deviations for this item: principal survey mean = 2.78, principal survey standard deviation = 0.878, teacher survey mean = 2.77, teacher survey standard deviation = 0.199.

[b] SFA mean and standard deviation for this item: mean = 2.89, standard deviation = 0.68.

[c] SFA mean and standard deviation for this item: mean = 2.16, standard deviation = 0.43.

[d] SFA mean and standard deviation for this item: mean = 2.78, standard deviation = 1.11.

[e] SFA mean and standard deviation for this item: mean = 2.73, standard deviation = 0.22.

[f] SFA mean and standard deviation for this item: mean = 2.72, standard deviation = 0.83.

A second concern was that while most schools did conduct a secret, anonymous vote, not all did. Teachers in one school reported that their school required them to write their names on their ballots and to include their reasoning if they voted against SFA. Another school held an open vote during a meeting, where those in favor simply raised their hands. In schools where the vote was not anonymous, some teachers reported feeling pressured into voting for SFA.

Finally, regarding the adequacy of information needed to make an adoption decision (the middle part of Table 3.1), the survey results indicate that the majority of teachers felt that they had not received enough information before the vote. Similarly, some focus group participants reported feeling underinformed or misinformed by the presentation about the realities of SFA. After having experienced SFA in the school, treatment school teachers pointed out what they saw to be inaccuracies and inconsistent information in the original presentation. Teachers at several schools reported that, at the time of the vote, they had not fully understood that SFA is more than a reading curriculum and that it incorporates whole-school reform elements as well. Teachers also noted that they had had misconceptions about the amount of work necessary to implement the program properly. Thus, while teachers voted for SFA in all the schools in the study, their attitudes toward the adoption experience were mixed.[2]

Principals, on the other hand, reported a much more positive perspective about the adequacy of information before the vote. They had learned of the opportunity to have SFA from their districts; in some instances, specific schools were asked by district administrators or by school board members to apply to be a part of SFA. They understood that they themselves had to register interest in the program before it could be introduced to their schools.

Many principals were impressed by the data systems and professional development that SFA was expected to bring. One principal explained: "From my understanding of SFA, I thought really it would benefit my teachers beyond reading. I thought it would just make them better practitioners of good teaching practices." In turn, principals were encouraging in discussing SFA with their teachers. Most principals recalled the adoption decision as though SFA was a clearly positive choice, despite the different perception voiced by many teachers.

Appointing the SFA Facilitator

A pivotal task in all the SFA schools was appointment of an SFA facilitator, who would be part of the school's staff, manage SFA materials, interact with teachers and SFAF coaches, and

[2] As Table 3.1 also shows, teachers — however impressed by SFAF's presentations they may have been — nonetheless reported a marked degree of skepticism about SFA's likely capacity to help students in their schools during the program's first year. Almost two-thirds harbored some doubt that SFA would improve reading outcomes. Principals seemed aware of the teachers' skepticism (again shown in Table 3.1).

ensure that the program was being implemented with fidelity. In general, the qualifications of the SFA facilitators were strong. All had taught. Many had backgrounds as reading coordinators or instructional specialists. A handful had done facilitation work previously, and some had administrative experience as well.

While postings for these positions drew a number of responses, schools usually gave preference to a qualified applicant who already was on staff at the school. One facilitator put it this way: "I had developed a real relationship with the teachers, and having been in a coaching position, I thought I would be more successful as their facilitator because we've established that relationship already working with them and coaching them."[3]

Most facilitators were in place by the beginning of the school year, and most had been able to attend an introductory conference for school leaders before starting.[4] It should be noted that the job was, from SFA's vantage point, expected to be full time. In fact, the majority of facilitators had other responsibilities within the school, which somewhat limited the attention that they could give to their SFA responsibilities.

Summer Training

In order to help schools begin implementing the program, SFAF provided training and support to school leaders and staff. The first training session took place in Baltimore, during the summer preceding initial implementation, at the New Leaders Conference. For those few leaders unable to make it to that session, a second leadership conference was held in the fall, and later training sessions were held as well (Figure 3.1).

The leadership conference lasted for a week and was geared toward principals, facilitators, and (sometimes) district staff, in order to prepare them for their new roles as SFA school leaders. At the conference, participants were introduced to the SFA components and to the strategies needed to implement them successfully. They were also introduced to the SFA data system (known as "Member Center") and the role that it was to play both to support reading and as part of the school's continuous improvement process.

SFA consultants were expected to visit the schools and provide training before the beginning of the school year. On their first day on campus, the SFA staff met with the Leadership

[3] While teachers generally gave their facilitators positive reviews, this was especially true when the facilitator had previously served as a reading specialist or reading teacher. Principals also relied on the facilitator to promote teachers' buy-in to SFA. One principal noted: "One of the things that has facilitated it is the strong reading background of the facilitator. . . . And so she had a huge buy-in and saw the benefits to the program. And so she's very articulate in expressing that and sharing that with the staff."

[4] One facilitator did not participate in any of the leadership conferences.

Team. That day was to be devoted to team-building, reviewing and establishing realistic school goals, establishing the Leading for Success model, and preparing to present to the faculty in the remaining days of the visit.

On the second day, all faculty members — not just reading teachers — were invited to participate in introductory workshops. They got an overview of how the SFA components fit together, received training on cooperative learning and the Getting Along Together classroom management component, and were introduced to the Leading for Success structures. During the following two days, reading staff were separated into groups by reading level — KinderCorner (for kindergarten students), Reading Roots (for beginning readers), or Reading Wings (usually for students in second grade and higher) — and by language (in districts where there was to be bilingual instruction). There they were introduced to the instructional methods, materials, data tools, and procedures and routines that they would be using during the school year.[5]

Teachers and facilitators alike reported finding the amount of material that these initial sessions covered a bit overwhelming. A related issue was that the workshops sometimes took place several weeks before school started, which led teachers to complain that they did not have what they had learned fresh in their minds. An added complication was that there were no materials, such as teachers' guides, available to teachers for study between the summer training and the beginning of the school year. And the training sometimes happened before some teachers knew what level of SFA (Reading Roots or Reading Wings) they would be teaching at school's start. Teachers might have attended one training level only to find themselves assigned to teaching the other.[6]

Finally, the initial training was also the first time that teachers came to grips with certain realities of SFA instruction. Many had not previously realized that instruction was so highly structured. Others were left confused about just how "scripted" it was intended to be. Teachers also felt that, despite the scripting, they were uncertain how much latitude they had in presenting it.[7] There also were concerns about the rapid instructional pacing that the curriculum required. Teachers worried that a class might be expected to move on to a new activity before some students had fully absorbed the material in the current one. (SFA's expectation was that

[5] This overview of training is drawn from the SFA facilitators' guide (Leading for Success Team, 2012). Also see Table 1.1, in Chapter 1, which briefly describes the program components discussed here.

[6] Similarly, when teachers switched from one reading level to another during the school year, they sometimes felt that they had been given a task for which they were unprepared. While some teachers said that they got about 45 minutes of training before transitioning from one level to the next, others were surprised to hear that this training might have been available to them.

[7] Some teachers said that, in early SFA professional development sessions, they were told that they had to follow the manual strictly, while others got the message that they had some freedom in presenting SFA, as long as they followed the program's broad strictures.

those students would "catch on" when the material was revisited in a later lesson.) Chapter 4 discusses this issue in detail.

In retrospect, many teachers reported that the summer workshops, while a satisfactory overview of the SFA curriculum, did not adequately expose them to what using SFA would be like in classroom practice and did not really prepare them for the day-to-day experience of teaching in an SFA classroom. As a result, they often found themselves overwhelmed and unsure of what they were supposed to do when they actually began using the SFA materials. To one teacher, it seemed as though they were expected to "magically teach" SFA. One facilitator described her teachers' experience:

> We had, as they put it, the sales pitch that everything was going to be there for them, and so, of course, they went ahead and they voted yes. I think that was a disappointment [later] when they got their training. It's, like, "Wait a minute. What happened? What do we do?"

The SFA Materials

SFA is a highly detailed and structured program; its curricular materials are voluminous, and they were unfamiliar to many of the schools in the study. An initial logistical challenge for facilitators and teachers was simply getting the materials housed and organized — a task complicated by the late arrival of the materials in about a quarter of the study schools. SFA facilitators were tasked with cataloging, organizing, and shelving the materials. In a few schools where no SFA facilitator had yet been appointed, other SFA staff or coaches pitched in to get the materials into usable order.

While the volume of materials was itself challenging for all teachers, additional issues concerned SFA's Spanish-language materials. At 8 of the 19 program group schools — schools with large concentrations of students whose primary language was Spanish and that were located in districts that permitted bilingual instruction — some students received reading instruction using these materials. Administrators and teachers complained that the Spanish materials, unlike the English materials, arrived in installments over the course of the year, making it difficult for teachers to plan ahead and sometimes requiring them to create materials on their own while waiting for the next shipment. One principal described her frustration to an interviewer: "If you're going to sell a program for all students, make sure that all material is ready for all students. Bilingual teachers should not have to wait." Furthermore, videos were not yet available for the bilingual classrooms. As a teacher noted: "They [students who receive instruction in English] have beautiful videos that depict the vocabulary, the sounds, everything in an incentive way, whereas the bilingual students don't have that video. So I do everything by hand." A facilitator at a different school also commented that SFA's bilingual materials did not meet quite the same standards as the English materials. Some minor errors were noted (and subsequently corrected); a complicating factor was that the SFA materials used some Spanish words and expressions that were different from the Spanish terms used by students of Mexican-born parents.

Ongoing Training

Throughout the school year, the SFA coaches visited the program group schools. During these visits, they spent the majority of their time with the Leadership Team, helping them review progress on implementing the SFA requirements. They did classroom walk-throughs and provided feedback on what was going well and what could be improved. Together with the SFA facilitator, the coaches also reviewed the school's data and used it to discuss how the school could strengthen its efforts. They answered questions and helped resolve problems with Member Center (the data system) and with specific program components. The coaches also worked with the Schoolwide Solutions teams and sometimes met with the teachers — grouped by reading level or individually — to provide specific feedback.

Principals and facilitators were satisfied with the training and support that they received from SFA coaches. They felt that their specific questions and concerns were addressed during the coaches' site visits. The majority of school leaders found the coaches to be attentive, very helpful, quick to respond to questions asked by phone or e-mail, knowledgeable, and tactful about sensing when and how to present particular information.

By contrast, teachers — especially in the early going — were generally less positive. Their main concern was that neither the introductory sessions nor the periodic visits by the coaches adequately prepared them for teaching an SFA class. Some indicated that coaches were not "upfront with the teachers" and did not clearly tell them: "This is what's required from you. . . . This is what you're going to do."

A majority of teachers — and a number of the facilitators and principals on the teachers' behalf — requested that coaches model what a properly paced lesson would look like. Coaches repeatedly responded, however, that such modeling was something that they could not do. Some explained that they did not know the students in the class and, therefore, could not teach it. Infrequently, teachers were shown videos of a lesson, but these were generally not perceived to be helpful to a teacher who was just beginning to teach SFA reading. One teacher opined that students in the training video were apparently already familiar with the SFA system, adding that, with her own students, "it looked nothing like that."

Teachers also reported that they got conflicting messages and guidance from the introductory presentations and the coaches. As one teacher noted: "They [initially] told us not to deviate from the script, and every time she's met with me, she tells me to go off the script. It's, like, I thought I was supposed to stay with the script." Different teachers at the same schools sometimes got different instructions, depending on which SFA staff member they had interacted with. For instance, one teacher said that she was told to post the grades online, but another teacher was told that she did not need to do this.

Principals generally had a more measured perspective. While they agreed that they had heard of some teachers reporting conflicting messages, they also could understand why that might happen. At one school, for example, a principal noted that the teachers had complained to the coach that students could not get through an assignment in the allotted time. The coach suggested that teachers at first have students answer only three questions out of the four that were presented. Later, once students were more familiar with completing this type of assignment, the coach directed the teachers to require students to answer all the questions. The principal opined that the coach's suggestions were meant to ease teachers and students into a set routine, even as teachers interpreted the advice as "waffling."

Teachers had several other issues with the ongoing coaching and support that they received. They commonly griped about coaches who, they felt, were overly critical during site visits. One teacher noted that "people from SFA come over and give their observations of what they thought we should do and how to improve, and it's, like, 'You didn't do this' or 'That should have happened' — just giving their opinions about what we could've done differently." Several teachers interpreted coaches' critiques as negative comments on their performance as teachers.

On this point, principals again felt that they could see both sides of the issue. On the one hand, they conceded that teachers might earlier have gotten clearer directions about what they needed to do once in the classroom. On the other hand, they believed that teachers had not always grasped that SFA required a real mental shift on their part and that the coaches were there to help them implement the program with fidelity, not to say, "Gotcha!" One principal noted:

> One of the things that we had to have people understand and realize is that it is [the coaches'] job to tell you, "Hey, you're taking a 30-minute section of this program and you're making it 45 minutes," and what makes the program is really sticking to the time frame and all of these different things. It was not anything personally against a teacher; it was, "We want correct implementation of the program."

Teachers' and other staff members' opinions about the professional development that they received seemed to improve during the course of the first year. One principal described getting a lot of initial pushback from teachers, with some even threatening to transfer outside the district. The principal was happy to report that "a lot of those individuals I know have been won over."

It appears that a mix of things brought about this turnaround. For one, school staff began to notice improvements in their students, and this boosted morale and encouraged teachers to stick with SFA. And teacher survey results — despite the often-negative opinions expressed in interviews — paint a fairly complimentary assessment of the training and professional development. The responses shown in Table 3.2 indicate that nearly three-quarters of the teachers believed that the professional development that they received did provide them with useful guidance in implementing SFA (though they had more mixed opinions about the utility of the professional development in providing them with strategies to help struggling students).

Some teachers were not sure whether their increasingly positive views over time were due to the coach or to their own growing comfort and familiarity with SFA. As one teacher noted: "The more that [coaches] came around, [the more they] figured out what we were trying to ask, or maybe we figured out the answers on our own. I don't know." By the end of the school year, having experienced SFA and acquired some of the skills needed for its implementation in their classrooms, many teachers were beginning to look forward to the coming school year.

The Success for All Evaluation

Table 3.2

SFA Teachers' Perceptions of the Value of Professional Development

	Mean Survey Scores
Percentage of teachers who agree or strongly agree that SFA professional development helped in learning how to implement the reading program properly[a]	73.9
Percentage of teachers who agree or strongly agree that SFA professional development helped in learning new techniques for reading instruction[b]	70.2
Percentage of teachers who agree or strongly agree that SFA professional development helped in developing strategies to better meet needs of struggling students[c]	55.8

SOURCE: Spring 2012 teacher survey.

NOTES: On the spring 2012 teacher survey, teachers were asked to indicate their level of agreement with each item reported, where 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree. The percentages who agreed or strongly agreed, as presented above, were calculated by first taking the percentage of teachers at a given school who selected 3 or 4 for a given item, where the denominator was the number of nonmissing responses. The mean of all such school-level means was taken to produce the percentages presented above. The mean across school means is used so as not to give more weight to schools with more teachers or vice versa.

[a] SFA mean and standard deviation for this item: mean = 2.80, standard deviation = 0.26.

[b] SFA mean and standard deviation for this item: mean = 2.76, standard deviation = 0.33.

[c] SFA mean and standard deviation for this item: mean = 2.57, standard deviation = 0.33.

Chapter 4

Putting Success for All in Place

Chapter 4 describes the Success for All (SFA) program in operation — the next stage in the program's theory of change depicted in Chapter 1 (Figure 1.1), which guides the SFA program and evaluation. This chapter addresses basic research questions about the fidelity of program implementation:

1. To what extent was the program implemented as the Success for All Foundation (SFAF) intended?
2. How much did SFA implementation vary among program group schools?
3. What were major facilitators and barriers to implementation?

Answers to the first two questions draw on the School Achievement Snapshot, a rubric used by the SFAF coaches to assess progress in program implementation. (See Appendix Figure E.1.) The evaluation uses this instrument to derive two measures of implementation fidelity: (1) a relatively simple measure of the proportion of program elements that SFAF coaches judged to be in place and (2) a more refined measure that takes into account the depth of implementation as well as other factors. The third question above pairs the Snapshot with survey and interview data to examine the perspectives of school personnel regarding the challenges associated with putting a new intervention into place and the program-specific and contextual factors that made SFA implementation more difficult.

Key Findings

- SFA schools achieved a moderate level of implementation of high-priority program items. The schools had room to improve the breadth and depth of their implementation but, on average, appeared to be on track.
- There was significant variation in program implementation among the SFA schools. Not all of them were able to implement some key objectives.
- Principals, facilitators, and teachers noted that staffing proved to be a major obstacle to implementation of key objectives. Staff interviews revealed frustration associated with putting some core SFA practices in place.

- SFA schools with teachers who had more years of overall teaching experience tended to have higher scores on implementation of priority objectives as well as higher overall scores.

- Despite the challenges in launching SFA, principals and teachers agreed that the program benefited their schools.

SFA's Implementation After One Year

SFAF coaches complete the School Achievement Snapshot, and, in so doing, they identify each school's areas of implementation strength and the areas where improvement is expected. The coaches then provide feedback and targeted assistance to schools in the requisite areas.[1] SFAF notes in its materials that program implementation evolves and usually improves over the initial three years. Both feedback and the school's own experience with the SFA model are expected to improve the level of achieved implementation over time.

To this end, the program identifies practices or systems that should have greater priority. This chapter focuses on the practices that SFAF set as having the highest priority for implementation in the first year. The final report of this study will assess progress in implementation across the three years for which implementation data are being collected.

Snapshot Organization

The coaches use the Snapshot to rate SFA schools in three domains that correspond roughly to the program operations section of the theory of change: Schoolwide Structures, Instructional Processes, and Student Engagement. *Schoolwide Structures* include systems to assess student progress in academics and behavior and to organize staff to monitor instructional and noninstructional goals. Examples of Schoolwide Structures include confirmation that students are regrouped frequently, that teams meet regularly, and that a full-time facilitator position exists at the school. *Instructional Processes* concern how teachers manage instruction and student behavior in the reading classroom. Examples include pacing of active instruction during the reading block and specific kinds of teacher interactions with students, such as using certain questioning techniques. *Student Engagement* items address student behaviors and cooperative learning. Examples of Student Engagement items include whether students respond to questions in complete sentences (an aspect of instruction that the program emphasizes) and whether they exhibit positive behaviors and self-regulation (reflective of SFA's whole-school

[1] Coaches use the Snapshot (Appendix Figure E.1) to rate schools up to four times during a year; this study relies on the rating at the end of the year. The same coach in a district may rate multiple schools in that district.

approach). In each of these content areas, the Snapshot differentiates between items that are of high priority for implementation in the first year and those that may be implemented later.

The Snapshot measures how many program structures were implemented at the school level and what proportion of the school's reading classrooms demonstrated clear use of the instructional and student engagement practices. The Snapshot does not measure dosage or quantity at the student level. Instead, the summative measures derived from the Snapshot can be interpreted as school-level measures of the extent of program implementation and, to some degree, of its quality as well.

First-Year Success on Snapshot Items

Sixty-six items from the Snapshot are included in this analysis of early program implementation fidelity; of these, 23 are considered to be high-priority items whose implementation is expected during the first year of program operations.[2] While SFA schools could and did implement some of the remaining 43 items during the first year, these items have lower priority for immediate implementation. On average, schools implemented 54 items, or 82 percent of the items rated in 2011-2012, as shown in Table 4.1.[3]

Considering first the 23 high-priority items, Figure 4.1 shows that the program group schools implemented, on average, 19.6, or 85 percent, of these items. SFAF's "stretch" goal is for all the program group schools to have achieved all high-priority items by the end of the first year. Only one school achieved that goal. The SFA schools made inroads on implementation of lower-priority items as well, as the figure demonstrates.

These results suggest that program group schools have achieved moderate SFA implementation for Year 1 and are making progress toward the final goals but that there is still considerable room for improvement.

[2] The most recent version of the Snapshot includes 99 items, of which 33 were excluded from the first-year fidelity analysis. For most of the 33 items, the information across the program group schools was not rated in 2011-2012 because the component or practice had not yet begun or was not measured. For other items, the rating was inconsistent among schools, in part because the Snapshot changed during the 2011-2012 school year and because some coaches used different forms for different schools. The analysis includes only items that were rated consistently for all schools in 2011-2012. As the evaluation moves forward, all 99 Snapshot items will be included in the fidelity analysis.

[3] To contextualize outcomes for kindergarten students — the cohort of interest for the impact analysis — the study team assessed the subset of Snapshot items that are specific to the kindergarten component. Schools implemented about 61 percent of 14 items specific to the KinderCorner component. Of the 14 items, two were targeted for Year 1 implementation. All schools had these two practices in place.

The Success for All Evaluation

Table 4.1

Snapshot Items in Place (Implementation Year 2011-2012)

	Schoolwide Structures	Instructional Processes	Student Engagement	Total
Number of possible items in 2011-2012	24	22	20	66
Mean number of items in place at any level	19.00	19.05	16.16	54.21
Mean proportion of items in place at any level	0.79	0.87	0.81	0.82

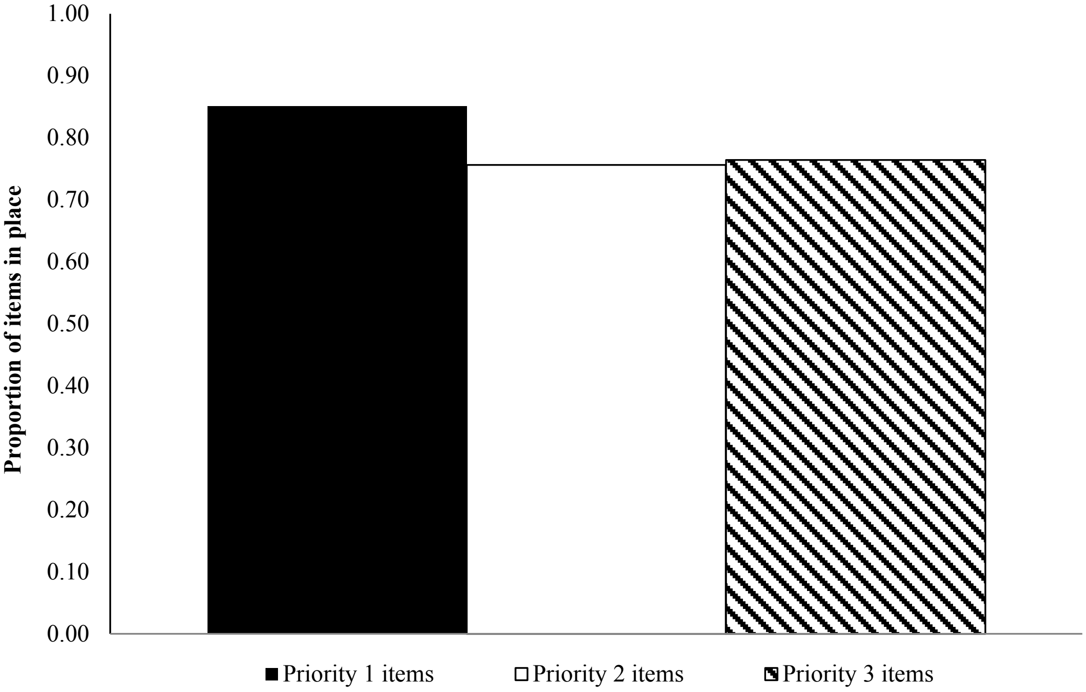
SOURCE: Success for All Snapshot (spring 2012).

NOTE: The table shows averages across 19 SFA schools.

The Success for All Evaluation

Figure 4.1

Snapshot Items in Place During the First Year, by Priority



SOURCE: Success for All Snapshot (spring 2012).

Variation by Content Area and Among Program Group Schools

Although SFA schools are on track, school staff did not describe program implementation as easy or without issues. The challenges can be seen in the variation in implementation fidelity across content areas and among program sites, especially when a more comprehensive measure of implementation is used.

Scoring the Snapshot to Reflect Depth of Implementation

The Snapshot is a school-level rating that also needs to reflect classroom-level implementation in order to indicate the extent of program implementation within an SFA school. For example, some schools could have implemented a practice — such as students' working in pairs — in just a handful of reading classrooms, while other schools could have implemented that practice in the majority of classrooms.[4] To account for these differences in implementation coverage among schools, the evaluation team worked in partnership with SFAF to create a system for scoring the Snapshots. (Appendix E discusses the details of score construction.)

The fidelity score captures the extent of implementation, by weighting key practices and taking into account the proportion of classrooms demonstrating those practices. Such a score may allow for a fairer comparison of fidelity among schools than simply recording the total number of items implemented. The score does not necessarily reflect implementation quality beyond what the Snapshot items themselves capture. For example, a management team meeting at the school to review student reading performance may lead to more successful efforts to address students' problems in one school than another, but the score cannot account for such differences.

Variation in the Snapshot Score

On average, SFA schools earned just over half the maximum possible score for the Snapshot items rated in 2011-2012 (55.8 of a maximum possible score of 102).[5] As shown in Figure 4.2, the lowest-scoring school achieved just 42 percent of the maximum possible score, while the highest-scoring school achieved 78 percent of the maximum possible. The higher-scoring schools not only pursued more practices but also implemented those practices in more classrooms.

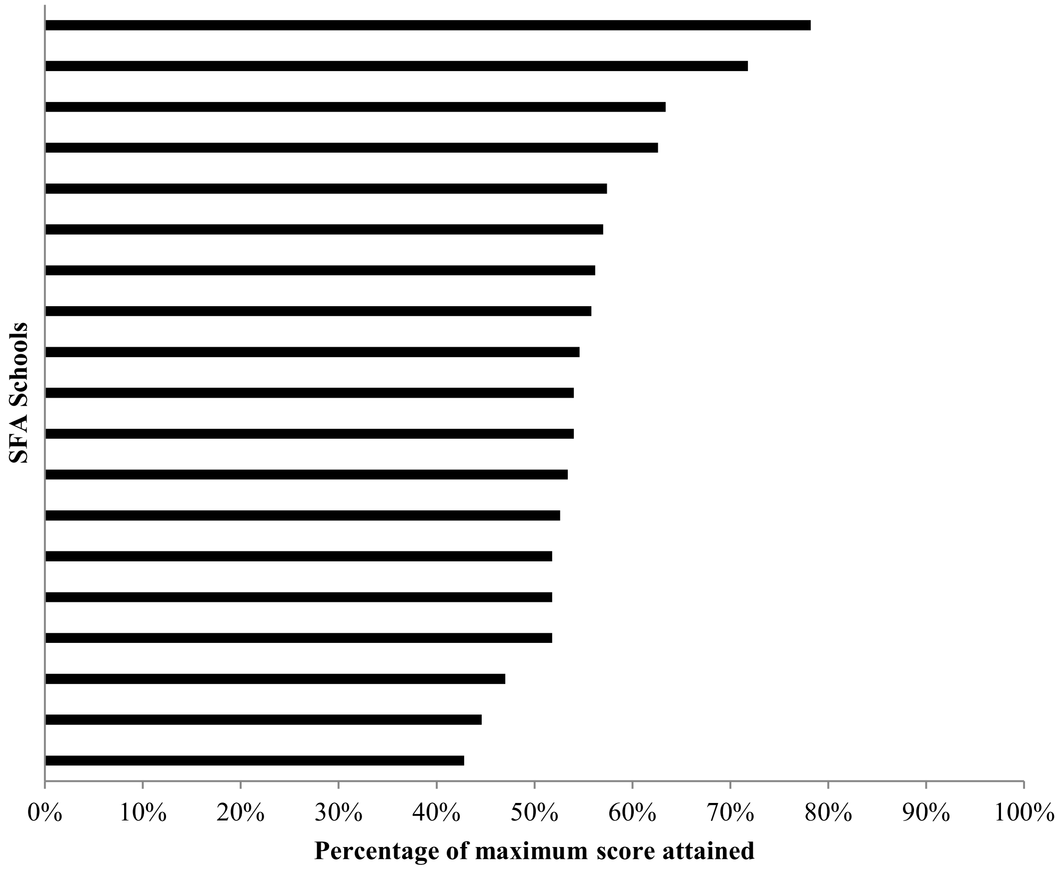
Across all schools, the extent of implementation is lowest in the content area of Student Engagement. Table 4.2 shows that while schools achieved nearly four-fifths of the possible

[4] The chapter refers to classrooms rather than teachers because multiple adults can be serving students in a given classroom. The coach records the proportion of classrooms that have a practice in place.

[5] Not surprisingly, the number of items in place is highly correlated with the overall score. (The correlation coefficient is 0.81, a p-value of 0.0001.)

Figure 4.2

SFA Schools Ranked by Overall Score



SOURCE: Success for All Snapshot (spring 2012).

NOTE: The score attained in Year 1 is shown as a percentage of the maximum possible score in the first year of program implementation.

score for Schoolwide Structures, on average schools attained just 47 percent of the possible score for Instructional Processes and just 42 percent of the possible score for Student Engagement. This result partly reflects that only one Student Engagement practice item is deemed high priority and is targeted for implementation in Year 1; the other practices are targeted mainly for implementation at a later point. In addition, many of the Instructional Processes and Student Engagement practices involve student exposure to enough of the SFA program so that they can

The Success for All Evaluation

Table 4.2

Snapshot Rating Scores, by Content Area (Implementation Year 2011-2012)

	Schoolwide Structures	Instructional Processes	Student Engagement	Total
Maximum possible score in 2011-2012	28	41	33	102
Mean unweighted score	19.00	10.63	8.86	38.49
Mean weighted score[a]	22.74	19.36	13.74	55.83
Mean proportion of maximum total score, by content area[b]	0.81	0.47	0.42	0.55

SOURCE: Success for All Snapshot (spring 2012).

NOTES: The table shows averages across 19 SFA schools.

[a]Per the Success for All Foundation, a weight of 2 is given to "essential" items and to items for Reading Wings.

[b]In Year 1, the denominator is the maximum possible weighted score of 102.

learn and apply their own strategies of self-regulation and cooperative learning. Given that the program was in its first year, the number of schools and proportion of classrooms displaying these practices is low.

The variation across content areas and among SFA school scores reflects the fact that while the program group schools have begun implementation of many items, the degree of implementation (or the proportion of classrooms demonstrating those practices) is still limited in some sites. Thus, the SFA schools have to keep developing some practices already in place while moving on to address the new practices on which they will be rated in subsequent years.

Factors Impeding and Promoting Program Implementation

This section draws on Snapshot ratings, survey results, and interview responses to illuminate some of the key issues that schools faced in implementing the SFA program during the first year. It first lays out some of the barriers to implementation that schools faced: lack of staff and difficulty in adapting to some aspects of the program model. Even so, greater experience on the part of teachers was associated with higher implementation fidelity.

Factors Impeding Program Implementation: Staffing and Funding

Insufficient Staff

The program group schools' experiences in implementing four high-priority Schoolwide Structure elements that serve as cornerstones for SFA — tutoring, having a full-time facilitator, convening management teams, and regrouping students across grades — are highlighted because these practices are critical to launching SFA and because they illustrate a major challenge that schools commonly faced during the first year: lack of staff. Because of insufficient staff complements, teachers and facilitators felt that they did not have enough time to complete what was expected of them to fully implement SFA.

Tutoring

Tutoring was an area where both the Snapshot rating and the interviews indicate some challenges. According to the Snapshot, more than half the SFA schools did not have in place the capacity to tutor the required proportion of students in grades 1 through 3. Half the schools with data available did not provide daily tutoring to targeted students.[6] The interviews confirm these results. At about half the program group schools, teachers, facilitators, and principals indicated that they lacked sufficient time or staff to serve all intended students daily. In the districts where Spanish-dominant students received SFA instruction in Spanish, facilitators and teachers expressed concern that these students needed tutoring the most but that the SFA tutoring program was not computerized for those students. This meant that students being taught in Spanish needed one-on-one tutoring. But as one facilitator noted, "We just don't have enough manpower, enough people who can actually be pulled to dedicate that time every single day."

Regrouping

Periodic regrouping of students by reading performance across grade levels is another fundamental feature of the SFA instructional program, and one whose success is itself contingent on the successful implementation of other program features. It requires the administration of quarterly assessments that reliably measure students' progress and that identify students who might benefit from being moved to a higher- or lower-performing group.

In contrast to tutoring, the Snapshot rating indicates that *all* the program group schools implemented all the objectives associated with this key feature of SFA. But the interviews reveal underlying concerns about the process and resources required to achieve those objectives in some schools. Recalling the difficulty of testing each student individually, one principal said,

[6] The surveys show that most SFA schools were providing tutoring, but the survey item did not ask about daily tutoring.

"Sometimes the testing and the regrouping has been kind of a challenge for us . . . just because we have so many kids." Facilitators and teachers alike complained about the lack of staff who could get to know students' needs and could assess all students on schedule.

Once students have been tested, they need to be placed in an appropriate reading level. But insufficient staff proved to be an obstacle, resulting in a mismatch between the number of students identified for instruction at a certain level and the number of teachers prepared to teach at that level. Sometimes class sizes exceeded the SFAF target of 20 students in Reading Roots classes (the lower reading level). Schools sometimes had to combine classes with students from different reading levels because there were too few teachers trained in a particular reading level or too few students assigned to distinct reading levels. The teacher survey responses reflect these concerns about class size. On the survey, only 49 percent of teachers agreed with the statement "Reading groups are small enough in your reading class for individual students to receive adequate attention."[7]

Solutions Teams

SFAF requires that schools establish "Solutions Teams" — five management teams made up of teachers and other school staff that tackle both instructional and noninstructional aspects of SFA, such as attendance, parent involvement, and intervention (providing instructional and behavioral supports to struggling students). The first-year SFA objective is to identify the five teams and ensure that they meet.

Nine of the program group schools did not have Solutions Teams that had been identified or that met regularly, according to the Snapshot. Some schools cited the challenge of finding time and staff. Several facilitators mentioned that their school had been doing some version of these practices already — such as setting attendance goals, providing tutoring, or monitoring behavior and academic progress — under the auspices of another reading model. The challenge, they said, was to shift from their previous approach to an SFA approach. In the lagging schools, the shift took time and did not quite meet the SFA goals or modes of administration.

The SFA Facilitator

One of the key school staff to launch and sustain the instructional and noninstructional components is the SFA facilitator. Yet several program group schools struggled to assign and maintain a full-time facilitator. The Snapshot indicates that only four schools were missing a full-time facilitator. However, in interviews, most SFA schools said that they did not have a true full-time facilitator position.

[7] The mean response was 2.39 on a 4-point agreement scale where 1 = Strongly Disagree and 4 = Strongly Agree.

Several principals noted that their facilitator had to fulfill multiple responsibilities at their schools besides those of SFA, including coordinating state testing, coordinating other reading programs, and handling compliance associated with federal programs. According to several facilitators, the time required for these "mandates" often disrupted progress on SFA implementation.

Facilitators and principals at several schools noted that the district had not fully funded the SFA facilitator position or recognized that it is a full-time job. One facilitator said: "The biggest thing with this district is . . . that they don't realize that a facilitator is a job in itself, and so . . . they put you on the spot; they also expect you to do other tasks." A principal in another district observed that "in the grant, [the district] should have written the SFA facilitator position in, because in this district there is no money and there is no one to do that . . . full-time job." Issues related to recognition and funding of the facilitator's job came up repeatedly across schools.

Regardless of whether the position was full time in practice or just on paper, at 14 of 19 program group schools, principals and facilitators expressed a need for more time to complete the SFA tasks required of facilitators. Several facilitators described recruiting additional leaders in the school to manage the workload associated with SFA: "As a facilitator, it's a very, very vast job. Without the help of everyone else, [it is] very difficult."

Factors Impeding Program Implementation: Elements of the SFA Model

In addition to the impeding factors cited above concerning too few staff and too little time, teacher and principal survey respondents cited some aspects of the program model that also made it challenging to implement SFA.

Program Scriptedness

Many teachers did not appreciate beforehand the extent to which SFA is scripted. Teachers were inclined to complain that it stifled their creativity and, on the teacher survey, were likely to agree that their reading program was too rigid or too scripted. One facilitator acknowledged that the program is scripted but disagreed that it undermined teachers' creative impulses:

> SFA makes it so easy on the teachers, because it is a scripted program. You might hear . . . that it limits their creativity. I beg to differ. If a teacher wants to add the bells and whistles to this, it's so easy to do it. Because you don't have to spend the amount of time pre-planning the lesson. You can spend that time adding [to the lesson].

Other teachers understood that SFA is scripted but were under the mistaken impression that this would mean that they did not have to plan their lessons.

Principals noted that while experienced teachers bristled at being asked to change established practices, new teachers seemed to benefit from the structure of SFA. One principal noted that her school has "some younger teachers" and said: "We know that sometimes, in their methods courses, they don't get everything that they should. But they're . . . excited about their new profession, and [SFA] really has some very good professional staff development for them."

Pace of the Program

For both novice and veteran teachers, the rapid pacing of the SFA lessons was a primary teaching challenge. In order to implement the SFA reading curriculum with fidelity, teachers must go from one activity or topic to the next within a specified, and short, period of time. In other words, topics are revisited with the expectation that students may not understand or master concepts during the first encounter (an approach sometimes called "spiraling"). However, many teachers were more accustomed to staying on a topic until their students reached some level of comprehension, and it proved difficult for them to move on to something else when students did not appear to understand the current topic.

In 12 program group schools, the pacing of SFA was cited as a concern. One teacher said that she tried to explain to the SFA coaches "that sometimes [students] just needed more time and more practice opportunities." Another teacher expressed frustration that her expertise as a teacher was being disregarded: "I know how to teach and . . . it requires more time" than 20-minute installments during a reading block. One facilitator referred to SFA's "spiraling" approach as a major obstacle:

> The hardest thing may be for some of us to understand that spiraling. . . . Kids will get it, and you need to move forward. . . . But it's hard to keep going. You feel like your kids haven't mastered it so . . . that's when we try to tweak it and . . . take a little longer . . . on certain things.

Data Use

Data-driven instruction is a foundation of many current reading programs and district reforms. On a scale that combined several items about the frequency and type of data use, teachers reported modest use of data.[8] Despite having some data experience, school staff remarked about how challenging SFA's data system and standards were.

In large part because of limited facilitator time, many schools said that they struggled to launch and use Member Center, the SFA electronic system for entering data. Facilitators and

[8] Mean response on the scale was 2.78, where 1 indicated strong disagreement and 4 indicated high agreement.

teachers are, in principle, supposed to enter data on student progress after each reading unit (every six to nine days) as well as after periodic testing, to provide evidence for students' regrouping assignments. Although the systematic use of Member Center is a practice targeted for full implementation in Year 3 of the program, SFA schools received training and began working with the system during the first year. In many cases, the facilitator served as the point person for training, data entry, and data analysis.

Facilitators at nine SFA schools mentioned that Member Center was overwhelming to learn. They cited technical problems, challenges with understanding the system, and concerns about the amount of time required to enter student data. Several teachers echoed concerns about the time required to enter data, said that they relied on their facilitator, and said that they needed more training. Many principals cited technical problems as a barrier to full implementation. One principal said: "To some extent, [teachers] were not able to start off with some of the right habits on data entry because of [initial problems]. Now that the logistics of the data center are sorted out, [we are] moving to better data entry."

Making use of the data that were entered in Member Center proved to be yet another challenge. One principal noted that the system called for increased use and analysis of data, compared with earlier reading programs: "That was very difficult because [the teachers were] not used to that." Despite these perceived challenges, the Snapshot rating shows that only three SFA schools were not using Member Center regularly by the end of Year 1. This reflects that, as with other aspects of program implementation, SFA schools perceived many challenges associated with ramp-up but eventually achieved even aspirational targets.

Factors Promoting Program Implementation

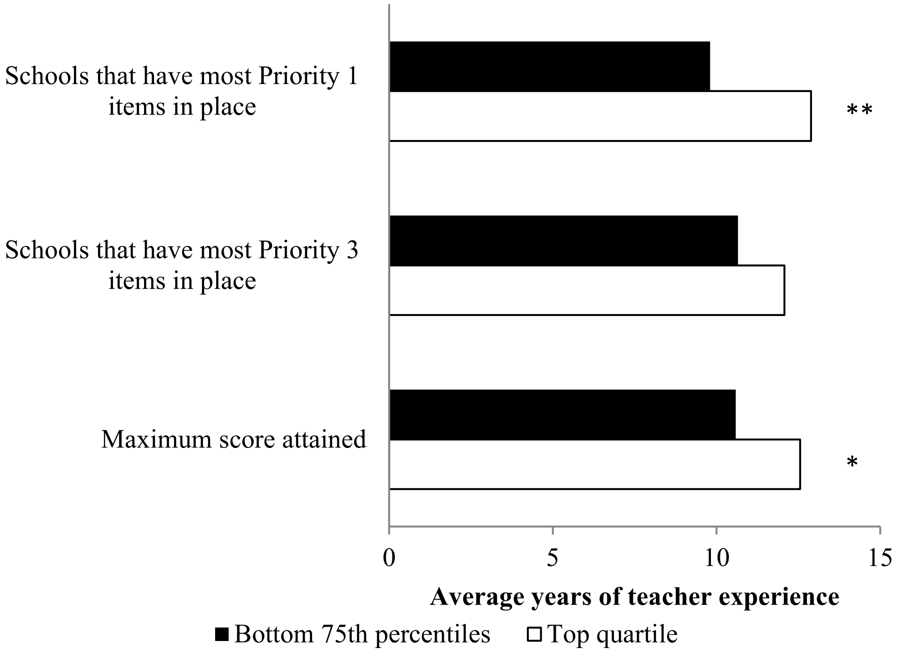
In contrast to the factors impeding SFA implementation that are cited above, it appears that teacher experience is positively associated with more complete program implementation. As shown in Figure 4.3, the SFA schools with the highest scores on the scale measuring fidelity of implementation (those in the top quartile) had teachers with two years' more experience, on average, than schools in the bottom three quartiles. When examining performance just on the high-priority practices, schools with the most high-priority items in place had teachers with nearly three years' more experience, on average (p-value = less than 5 percent).[9] Teachers' experience seems to be associated with slightly greater SFA implementation, though, as the interviews indicate, not necessarily with greater perceptions of success.

[9] The p-values cited should be considered with caution, since the sample size is just 19 SFA schools.

The Success for All Evaluation

Figure 4.3

Average Teacher Experience, by Snapshot Rating Quartiles



SOURCE: Success for All Snapshot (spring 2012).

NOTE: The means reported are means of school-level averages rather than teacher-level averages, so that schools are not weighted based on the number of teachers at the school. The analysis uses 19 school averages and 480 teacher observations.

Mixed models were used to calculate the p-values associated with the mean difference between teacher experience in the top quartile and bottom 75th percentiles. This approach calculates p-values on teacher-level data and adjusts for the correlations among teachers within a given school.

Cross-tabulations present Snapshot ratings and teacher experience, as reported on surveys. The top quartile for each Snapshot variable is constructed based on values, not number of schools; then means of school-level means of teacher experience are calculated by top quartile designation (1 = SFA school was in the top quartile for a given variable; 0 = SFA school was not in the top quartile for a given variable).

Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

Summary

As with the adoption of any new program, Success for All's introduction into the 19 program group schools experienced its share of bumps and challenges along the way. Teachers and principals alike perceived that SFA represented a substantial or extreme change in the way their school taught reading (Table 4.3). A scale that combines the five items about change in practices shows that teachers agreed that they had changed their practices as a result of SFA.[10] By and large, the SFA schools succeeded in getting the program into place without large deviations from the program model.

While staff frequently expressed frustration with the changes that SFA brought — especially at the start of the year — by the spring, there also were hopeful signs. In response to a survey item, all principals and the majority of teachers agreed that SFA had benefited their school (Table 4.4). The interviews suggest that teachers had a steep learning curve to enact SFA but that, by the end of the year, implementation started to seem easier. As one teacher noted:

> Now that we've learned the routines, . . . it's easier for the kids, [and] it's easier for us. . . . It's definitely so much easier than it was in October. You know, 100 times better [now].

The Success for All Evaluation

Table 4.3

Principals' and Teachers' Perceptions of SFA Compared with Prior Reading Approach

Survey: "Compared with your school's reading approach in the 2010-2011 school year, how different does SFA seem?"

	Percentage Reporting That SFA Represents a Substantial Change
Principals	100.0
Teachers	86.0

SOURCES: Teacher survey (spring 2012) and principal survey (spring 2012).

NOTE: These values represent the mean of school means for the given item.

[10] The mean response was 2.76 on a 4-point scale of agreement.

The Success for All Evaluation

Table 4.4

SFA Teachers' and Principals' Survey Responses Related to the Changes and Benefits from SFA

	Teacher Survey Scale Measuring Changes in Practice Resulting from SFA Implementation[a]
Teachers (mean)	2.76
	Agree or Strongly Agree with: "Overall, your school has benefited from the SFA program."[b]
Teachers	71.4
Principals	100.0

SOURCES: Teacher survey (spring 2012) and principal survey (spring 2012).

NOTES: [a]The scale ranges from 1 (Strongly Disagree) to 4 (Strongly Agree).

[b]These values represent the mean of school means on the given item.

Chapter 5

What Makes SFA Schools Distinctive?

Chapter 3 describes a number of key processes involved in putting the Success for All (SFA) reform program into place, and Chapter 4 assesses the extent to which central features of the program were implemented during the first year of operations at the 19 program group schools participating in the SFA evaluation. 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.

But impacts 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 of their instructional and whole-school features, there is little reason to expect program-control differences in measures associated with the intermediate pathways that — according to the SFA theory of change (Chapter 1, Figure 1.1) — lead to changes in student outcomes or in the student outcomes themselves. Before turning in the next chapter to SFA's impacts on student outcomes, it is important, therefore, to broaden the focus, turning from a consideration of implementation in the program group schools to an examination of the ways in which SFA schools and control group schools resemble each other and the ways in which they differ.

The chapter begins with a brief overview of the reading programs used in the control group schools. Then the discussion examines program-control differences on variables associated with different stages of the theory of change. Because that theory treats professional development as a precondition of effective program implementation, the analysis first compares the professional development received by teachers in the SFA schools and by teachers in the control group schools. Attention next turns to the schools in operation, comparing the two groups of schools with respect to instructional attributes — the aspects of reading that teachers emphasize in day-to-day practice and other features of instruction — and to the whole-school initiatives that are not directly tied to instruction. Finally, the analysis compares SFA schools and control group schools along dimensions related to the intermediate pathways that are hypothesized to produce changes in student achievement and other outcomes. The chapter relies on data from principal and teacher surveys administered at the program group and control group schools and on instructional logs that teachers at both sets of schools completed.

Key Findings

- With respect to content coverage, the curricula of the SFA schools and the reading programs used in the control group schools appear to be quite similar: All emphasize both decoding and comprehension skills, and all programs provide materials that include a teacher's manual, reading selections for students, assessments, and strategies for assisting struggling readers.

- Teachers in the SFA schools were more likely to report having received professional development in reading instruction and to have received professional development on a greater number of reading-related topics than their counterparts in the control group schools. These patterns hold whether SFA teachers are compared with teachers in all control group schools or only with teachers at those control group schools that also adopted new reading programs. In general, SFA teachers also rated the professional development that they had received as being more helpful than did control group teachers.

- Four key factors appear to have differentiated instruction in the SFA schools and control group schools: Teachers in SFA schools placed more emphasis on comprehension than did control group teachers; teachers in SFA schools made much more extensive use of cooperative learning strategies; the SFA schools much more consistently employed cross-grade grouping of students by skill level for reading, with regrouping occurring at regular intervals through the school year; and SFA teachers were much more likely to say that they did not deviate from their reading program's lesson plans.

- There are no statistically significant differences between program group and control group schools along other dimensions of reading instruction that SFA deems important: the length of the reading period, the principal's role as an instructional leader, the use of data to monitor students' progress, and the provision of tutoring to students who need extra help.

- There are no statistically significant differences between SFA schools and control group schools in the extent to which the schools had staff members charged with undertaking whole-school improvement measures.

- On the survey measures that are the closest analogues to the intermediate pathway outcomes posited by the theory of change, there are no statistically significant differences between SFA schools and control group schools that favor the SFA schools.

Reading Programs in the Control Group Schools

During the 2011-2012 school year, before adopting SFA, the majority of both the program group schools and the control group schools taught reading/language arts with commonly used basal programs available from leading educational publishers (including Macmillan/McGraw-Hill, Houghton Mifflin Harcourt, and Scott Foresman). In broad terms, these reading programs are quite similar to SFA. Like SFA, they cover 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.[1] Like SFA, too, the programs used by the control group schools strike a balance between decoding and comprehension skills. The programs 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. Finally, the newer of these programs, like SFA, integrate the Common Core standards.[2] The discussion below in this chapter suggests that it is not the curriculum per se but the specific ways that it is enacted that differentiate SFA schools from control group schools.

According to information from surveys and from interviews conducted with principals in the control group schools, the majority of them (11 of 18) continued to use the same programs during the 2011-2012 school year as they had previously been using. Of the remainder, four switched their basal programs, with three of them changing from one Houghton Mifflin Harcourt series to another; for the other three control group schools, the information was insufficient to ascertain whether the reading program had changed. In general, schools within the same district used the same reading program.

Implementing the Reading Programs: Professional Development in SFA Schools and Control Group Schools

Chapter 3 examines several measures involved in SFA's initial implementation in the program group schools: the adoption decision, the delivery of program materials, and the provision of initial training and ongoing professional development to school personnel. Teacher surveys administered in both the SFA and the control group schools illuminate the last of these topics.

[1] National Institute of Child Health and Human Development (2000).

[2] 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 (K) 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 46 states and the District of Columbia (Web site: http://www.corestandards.org).

Table 5.1 displays the results.[3] It shows that while the large majority of teachers in both sets of schools received professional development in reading during the 2011-2012 school year, the mean percentage was significantly higher in the SFA schools than in the control group schools considered together. Furthermore, when asked about six specific areas in which they could have received professional development, SFA teachers reported receiving training in more of these areas than their control group counterparts (an average of 5.6 areas for SFA teachers, compared with 4.9 areas for control group teachers).[4]

As Chapter 3 makes clear, SFA teachers expressed reservations to interviewers about some of the professional development that they had received, but their survey responses concerning that training were more favorable than not. The bottom row of Table 5.1 indicates that SFA teachers valued their professional development more than control group teachers did, as shown by their higher scores on a scale created by the research team to measure the usefulness of professional development.[5] Teachers were asked to rate the helpfulness of each kind of professional development that they had received (on a scale of 1 to 4, where 1 indicates strong disagreement and 4 indicates strong agreement that the professional development had been helpful). An average was then calculated for each teacher and across all teachers in the two groups of schools.[6] SFA teachers gave a higher rating (2.73) to the professional development that they had received than did control group teachers (2.58); both ratings are above the neutral midpoint of 2.5. Data not shown in the table provide further insights into program-control differences on this dimension. For three of the six possible areas, SFA teachers who received professional development in that area were significantly more likely to agree or strongly agree that the professional development had been helpful. The difference is largest with respect to professional development on how to implement cooperative learning techniques: Among the

[3] In all the survey-based tables in this chapter, the statistics shown represent the averages for teachers or principals at SFA or control group schools within each district, summed across districts and weighted by the proportion of all program group or control group schools that the district represents.

[4] A seventh survey item asked teachers about professional development that helped them implement their school's reading program properly. This item was excluded from the data presented in Table 5.1, since SFA teachers — all of whom are implementing a new program — would be more likely to receive professional development in this area than control group teachers.

[5] See Appendix D for information about the items included in the scales that appear in this report, along with the statistical properties of these scales.

[6] Scores for teachers who reported receiving professional development in three or more of the seven possible areas covered on the survey were included in the calculations.

The Success for All Evaluation

Table 5.1

SFA-Control Comparisons on Survey Variables Related to Implementation Processes: Professional Development (PD)

	Program Group	Control Group	Estimated Difference	P-Value
Variables related to implementation processes				
Percentage of teachers receiving PD in reading	96.4	85.8	10.6	0.001 ***
Average number of topics on which teachers received PD in reading[a]	5.6	4.9	0.8	0.000 ***
Average score on teacher survey scale measuring helpfulness of PD[b]	2.73	2.58	0.15	0.014 **
Number of schools: 36	19	17[c]		

SOURCE: Spring 2012 teacher survey.

NOTES: The calculations in this table are based on a total of six areas of professional development in reading instruction.

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.

[a]This item is based on six possible domains of professional development in reading instruction.

[b]The survey scale includes 4 as the maximum, 1 as the minimum, and 2.5 as a neutral midpoint.

[c]One control group school did not respond.

teachers who received training on this topic, 84 percent of SFA teachers agreed or strongly agreed that it had been helpful, compared with 61.5 percent of control group teachers.[7]

[7] SFA teachers were also significantly more likely to agree that they had received helpful professional development on new instructional techniques for reading instruction and on making effective use of time during the reading period. Statistically indistinguishable percentages of SFA and control group teachers agreed that the professional development that they had received on how to meet the needs of struggling students, how to use classroom materials, and how to use reading assessment data to guide instruction had been helpful.

Because the analysis of program impacts during the first year focuses on kindergartners, it is interesting to look at the response patterns of SFA and control group teachers who taught kindergarten only. The pattern of responses of kindergarten teachers on the variables shown in Table 5.1 is similar to the pattern for all teachers: kindergarten teachers in SFA schools were more likely to receive professional development, received professional development on more topics, and were more likely to find that professional development useful than their control group counterparts.

These overall differences do not appear to have been driven by the fact that the SFA schools were implementing a new reading program. Data not shown here indicate that when the responses of SFA teachers are compared with those of teachers at six control group schools that also reported implementing new reading programs during the 2011-2012 school year, the SFA teachers remained significantly more likely to receive professional development and to receive it about more topics.[8] (Teachers at the SFA schools and at the control group schools that also adopted new programs gave their professional development similar ratings in terms of helpfulness.) In short, it was implementing SFA, not merely implementing any new reading program, that led to increases in the quantity of professional development that teachers received.

The Schools in Operation: Reading Instruction in the Two Groups of Schools

Findings from Teachers' Instructional Logs

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 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.[9] The logs inquired about the topics that were focused on during the reading/language arts period and about instruction that the students who were selected received in two key dimensions of reading: comprehension and word analysis.[10]

Teachers received detailed instructions about how to complete the logs.[11] To demonstrate that they had mastered this information before they began logging for their students, teachers in both the SFA group and the control group were asked to read the same fictional lesson and then to rate the lesson using the logs. The teachers proved to be highly successful in identifying the topics that were the focus of the literacy lesson; 93 percent of SFA teachers and 92 percent of control group teachers correctly marked the log. They were far less able to

[8] Neither the teachers nor the principal at the seventh control group school that changed its reading program completed the surveys for the evaluation.

[9] 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.

[10] Teachers were also asked to provide details about any instruction that selected students received in writing, but this section of the log was not analyzed.

[11] Each teacher was sent a training manual that included a glossary of terms used in the logs. Teachers could also consult the study's Web site, where they could download all log materials, see answers to frequently asked questions, and post questions themselves. Finally, teachers could call a hotline to discuss any questions that they might have about the logs.

identify specific comprehension and word analysis strategies used in the fictional lesson. While they may have been better able to identify the practices used in their own lessons than in the fictional scenario, the results presented below should be regarded with some caution.[12]

A central question on which the logs shed light is whether SFA teachers and control group teachers focus on different topics during their literacy instruction. Figure 5.1 displays the results of a logistic regression analysis conducted to answer this question. It 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]

The figure shows that the instruction offered by SFA teachers differed significantly from that offered by control group teachers in four of the seven areas that were studied. Most notable is SFA's greater emphasis, even in the early grades, on comprehension — work on the meaning of spoken or written language, which has been described as the "essence of reading."[14] The odds that an average SFA teacher focused on comprehension were 1.78 times the odds that an average control group teacher did so. SFA teachers were also significantly more likely to focus on vocabulary (odds ratio = 2.19), which is closely tied to comprehension, with increased vocabulary being linked to gains in comprehension.[15]

On the other hand, SFA teachers were less likely to focus on spelling than their control group counterparts; the odds ratio for this area is 0.18. Teachers in the two groups of schools did not differ in the emphasis that they placed on word analysis, which includes instruction on the structure of words or the sounds and letters that make up words. Students must have a firm foundation in word analysis in order to comprehend text, and teachers tend to pay close attention to it, particularly in the early grades, so it is not surprising that no significant difference was found between the two groups on this dimension.

[12] Only 48 percent of SFA teachers and 38 percent of control group teachers identified the specific comprehension strategies in the fictional lesson correctly, while 44 percent of the teachers in both groups identified the word analysis strategies correctly.

[13] Logistic regression is the preferred method when the outcome is binary: either the event occurred or it did not. An odds ratio of 1 (illustrated by the vertical line in Figure 5.1) indicates that the average SFA and 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. The asterisks indicate whether the odds ratios are significant at the level of 10 percent, 5 percent, or 1 percent.

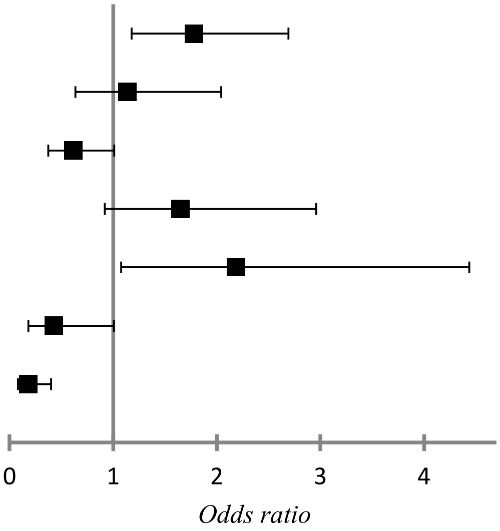
[14] See Durkin (1993).

[15] See National Institute of Child Health and Human Development (2000).

The Success for All Evaluation

Figure 5.1

Instructional Differences Between SFA Schools and Control Group Schools in the Language Arts Topic Focus Across All Lessons



Area of Focus	Odds Ratio	Confidence Interval
Comprehension	1.78 **	(1.18 , 2.69)
Word Analysis	1.14	(0.63 , 2.04)
Writing	0.61	(0.37 , 1.01)
Reading fluency	1.65	(0.92 , 2.96)
Vocabulary	2.19 *	(1.08 , 4.44)
Grammar	0.43	(0.18 , 1.01)
Spelling	0.18 ***	(0.08 , 0.40)

Sample size: 1,383 logs (735 in program group and 648 in control group) from 178 teachers (96 in program group and 82 in control group).

SOURCE: Teacher logs administered in spring 2012.

NOTES: The analysis sample consists of 1,383 teacher logs (735 from program group schools and 648 from control group schools) collected from 178 grade 1 and 2 reading teachers (96 in program group and 82 in control group) in 28 schools (15 program group schools and 13 control group schools).

The figure presents the odds ratio (OR) of an instructional measure occurring in program group schools versus schools in the control group. An OR compares the odds of a certain practice being used in the average SFA school versus 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. By placing a confidence interval around 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 HLM 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 test if the estimated OR is statistically different from 1. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

The upper half of Figure 5.2 indicates the particular strategies for teaching comprehension that were used when comprehension was a focus of instruction.[16] SFA teachers were much more likely than control group teachers to elicit brief answers demonstrating students' understanding of text (odds ratio = 5.77). 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 = 2.40). Through cooperative learning, readers learn to focus on and talk about what they are reading and are exposed to more comprehension strategies; the National Reading Panel concluded that the use of a combination of comprehension strategies produces gains in comprehension as measured by standardized tests.[17]

The lower half of Figure 5.2 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 = 0.28), which are words that students learn to recognize and read by sight, and on structural analysis (odds ratio = 0.38), which entails examining word families, prefixes, suffixes, contractions, and so on.

Finally, while it is noteworthy that SFA teachers were more likely than control group teachers to focus on instruction in comprehension, some areas of comprehension instruction are not very demanding cognitively. For instance, predicting what a text will be about based on the illustrations is far easier than analyzing characters' motivations. Figure 5.3 shows that the comprehension-related instructional strategies used by SFA teachers and control group teachers did not differ in the extent to which they placed cognitive demands on students. That is, while SFA students were more likely to receive instruction focused on comprehension, it was not because their teachers sacrificed any rigor that they might have provided without SFA.

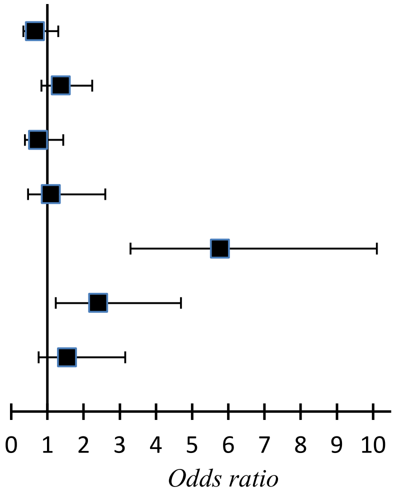
Findings from the Surveys of Principals and Teachers

Table 5.2 presents data from the surveys of principals and teachers about various characteristics of the reading programs at the study schools. The survey items were selected for analysis because they tap the same domains as items in the School Achievement Snapshot discussed in Chapter 4 — that is, aspects of school functioning that SFAF staff deem important for gauging the extent to which the SFA model has been put in place. For example, the Snapshot asks the SFA coach who completes it to indicate whether a 90-minute reading block exists at the school, while the teacher survey asks each teacher to write in the number of minutes per day that she or he teaches reading. Similarly, the coach notes on the Snapshot whether tutoring is provided daily for each tutored student, while the principal survey asks

[16] For this analysis, only those logs in which comprehension was a focus of instruction were examined.
[17] National Institute of Child Health and Human Development (2000).

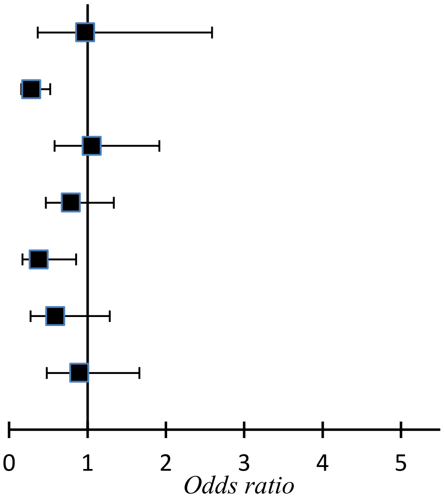
Figure 5.2

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



Comprehension Construct	Odds Ratio	Confidence Interval
Activate knowledge	0.66	(0.33 , 1.30)
Literal comprehension	1.37	(0.84 , 2.24)
Story structure	0.74	(0.38 , 1.43)
Analyze/synthesize	1.10	(0.47 , 2.60)
Brief answers	5.77 ***	(3.29 , 10.11)
Students discuss text	2.40 **	(1.23 , 4.69)
Teacher-directed instruction	1.54	(0.76 , 3.15)

Sample size: 1,083 logs (608 in program group and 475 in control group) from 176 teachers (96 in program group and 80 in control group).



Word Analysis Construct	Odds Ratio	Confidence Interval
Letter-sound relationships	0.97	(0.36 , 2.59)
Sight words	0.28 ***	(0.15 , 0.53)
Use picture/context cues	1.05	(0.58 , 1.92)
Use phonics cues	0.79	(0.47 , 1.34)
Structural analysis	0.38 *	(0.17 , 0.85)
Assess student ability	0.59	(0.27 , 1.28)
Teacher-directed instruction	0.89	(0.48 , 1.66)

Sample size: 748 logs (409 in program group and 339 in control group) from 159 teachers (88 in program group and 71 in control group).

(continued)

Figure 5.2 (continued)

SOURCE: Teacher logs administered in spring 2012.

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

The figure presents the odds ratios (OR) of an instructional measure occurring in program group schools versus schools in the control group. An OR compares the odds of a certain practice being used in the average SFA school versus 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. By placing a confidence interval around 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 HLM 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 1,083 logs (608 logs from program group schools and 475 from control group schools), completed by 176 teachers (96 from program group schools and 80 from control group schools). The analysis sample for word analysis constructs is 748 logs (409 from program group schools and 339 from control group schools), completed by 159 teachers (88 from program group schools and 71 from control group schools).

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

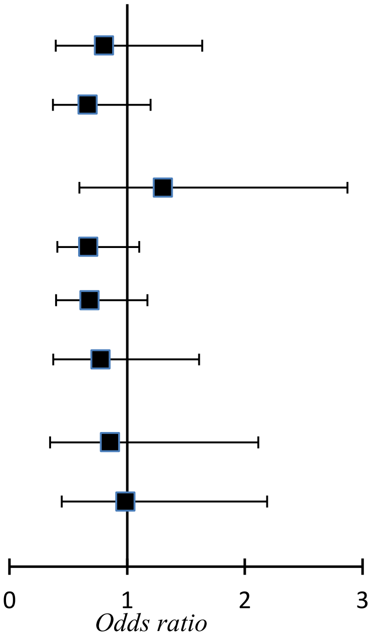
whether tutoring is scheduled to take place more than once a week. With respect to many areas of inquiry, the correspondence between the Snapshot item and relevant items in the surveys is more general. For instance, while the Snapshot asks about the use of several specific forms for recording data, the survey asks principals and teachers broader questions about data reporting and utilization. In any event, the purpose of presenting the survey data in this chapter is not to support or call into question the SFA coaches' Snapshot ratings but to shed light on the extent to which control group schools resemble or differ from SFA schools on these critical program dimensions.[18]

[18] Aside from differences in the wording of items and differences in the experiences and perspectives of SFA coaches and school personnel, the fact that the Snapshot data presented here reflect end-of-year ratings while surveys were generally administered well before the end of the school year complicates direct comparisons of data from the two sources.

Figure 5.3

Impact of SFA on Instruction of Cognitively Demanding Items

Item	Odds Ratio	Confidence Interval
Activate knowledge		
Activating prior knowledge	0.80	(0.39 , 1.64)
Previewing, predicting, surveying text	0.66	(0.37 , 1.20)
Story structure		
Summarizing important details in text	1.30	(0.59 , 2.87)
Sequencing information/events in text	0.67	(0.41 , 1.10)
Using concept maps/frames	0.68	(0.39 , 1.17)
Identifying story structure	0.77	(0.37 , 1.61)
Analyze/synthesize		
Analyzing/evaluating text	0.85	(0.34 , 2.12)
Comparing/contrasting information in text	0.99	(0.44 , 2.19)



Sample size: 1,083 logs (608 in program group and 475 in control group), 176 teachers (96 in program group and 80 in control group).

SOURCE: Teacher logs administered in spring 2012.

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

The figure presents the odds ratios (OR) of an instructional measure occurring in program group schools versus schools in the control group. An OR compares the odds of a certain practice being used in the average SFA school versus 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. By placing a confidence interval around these ORs, instruction in SFA schools can be said to be statistically different from instruction in control 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 HLM 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 test if the estimated OR is statistically different from 1. Statistical significance levels are indicated as follows: *** = 1 percent; ** = 5 percent; * = 10 percent.

As Table 5.2 makes clear, SFA schools and control group schools were similar in many ways. Both sets of schools allocated a substantial amount of time to reading instruction: 103 minutes, on average, in control group schools and 101 minutes in program group schools (a difference that is not statistically significant). Both groups of schools used educational media or technology as part of their reading program. Teachers at both groups of schools gave their principals similar ratings on a scale measuring the principal's instructional leadership in reading (both ratings were above the neutral midpoint of 2.5), and statistically indistinguishable percentages of principals reported that someone at their school — either a group or an individual — was responsible for helping teachers improve their instruction. (Presumably, the principals included themselves in this response.) The widespread emphasis given to using data to improve instruction makes it perhaps not surprising that the items tapping data use did not yield statistically significant differences between the SFA schools and control group schools.[19] Finally, while a higher percentage of SFA principals than control group principals reported that school staff members provided tutoring to students needing extra help with reading, the difference is not statistically significant (94.4 percent, compared with 76.4 percent, respectively).

On the other hand, the survey results point to sizable and statistically significant differences between the two sets of schools along three dimensions that are core principles of SFA instruction: teachers' use of cooperative learning strategies, the grouping and regrouping of students for reading instruction by ability level, and close adherence to the reading program's program lesson plans. Virtually all SFA teachers agreed or strongly agreed that their school's reading program involves students working together in pairs or small groups almost daily. Teachers in the control group schools also used cooperative learning strategies, but to a lesser extent than did SFA teachers.[20] Similarly, teachers in SFA schools were far more likely than their control group counterparts to report that students in their schools were grouped by ability level for reading, that such grouping took place across grades, that students were grouped for reading across grade levels, and that they were regrouped over the course of the year. Again, these practices were not unknown to control group schools, but they occurred much less frequently, according to teacher reports.

All this suggests that the *instructional environment* in SFA schools is quite distinct from the environment in control group schools. That environment is one in which students largely learn through interactions with one another in classrooms where students are all at more or less

[19] This does not mean that SFA and control group schools made identical use of such data; as noted below, using data to group and regroup students by ability level was far more prevalent in SFA schools.

[20] Cooperative learning begins early in SFA schools: SFA kindergarten teachers were also significantly more likely than their control group counterparts to agree or strongly agree that students worked together in pairs or small groups almost daily.

The Success for All Evaluation

Table 5.2

SFA-Control Comparisons on Survey Variables Related to Reading Instruction

	Program Group	Control Group	Estimated Difference	P-Value
Variables related to reading instruction				
Characteristics of reading instruction				
Average length (in minutes) of reading instruction period	100.6	103.0	-2.4	0.559
Percentage of teachers who agree or strongly agree that they use educational media or technology as part of the reading program[a]	84.5	77.2	7.3	0.127
Percentage of teachers who agree or strongly agree that the reading program involves students working together in pairs or small groups almost daily[b]	99.5	73.1	26.4	0.000 ***
Presence of staff focused on instructional improvement				
Average score on teacher survey scale measuring principal's instructional leadership in reading	2.76	2.75	0.02	0.840
Percentage of principals reporting that a group or individual is responsible for helping teachers to improve their reading instruction	94.4	76.4	18.0	0.137
Assessment and data use				
Percentage of principals who agree or strongly agree that their school uses data to evaluate reading progress of students over time[c]	100.0	100.0	0.0	NA
Percentage of principals reporting that they reviewed reading data 1 or 2 times a month or more	94.4	76.4	18.0	0.137
Percentage of teachers reporting that they reviewed reading data 1 or 2 times a month or more	63.6	70.1	-6.5	0.214
Average score on teacher survey scale measuring data use	2.78	2.80	-0.02	0.691
Student grouping and regrouping				
Percentage of teachers reporting that students are ability-grouped for reading	96.9	54.8	42.1	0.000 ***
Percentage of teachers who say that students are ability-grouped for reading across grades as a proportion of all teachers who report that students are ability-grouped	89.3	40.9	48.4	0.000 ***

(continued)

Table 5.2 (continued)

	Program Group	Control Group	Estimated Difference	P-Value
Percentage of teachers who say that students are periodically regrouped for reading as a proportion of all teachers who report that students are ability-grouped.	90.8	83.5	7.3	0.035 **
Tutoring				
Percentage of principals reporting that school staff members provide students with tutoring in reading	94.4	75.0	19.4	0.117
Percentage of principals who say that tutoring is scheduled to take place at least once a week as a proportion of all principals who reported that school staff members provide students with tutoring in reading	100.0	100.0	0.0	NA
Prescriptive instruction				
Percentage of teachers who agree or strongly agree that they change parts of the reading program that they don't like or disagree with[d]	26.4	66.3	-39.9	0.000 ***
Percentage of teachers who agree or strongly agree that their reading program is too rigid or scripted[e]	54.8	19.8	35.0	0.000 ***
Percentage of principals reporting that they looked for classes following a prescribed or recommended sequence of activities "all" or "most" of the time when observing reading instruction[f]	100.0	86.7	13.3	0.128
Number of schools: 36	19	17[g]		

SOURCE: Spring 2012 teacher survey.

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. Percentages of teachers or principals who agreed or disagreed with an item were obtained by taking the proportion who responded 3 or 4, expressed as a percentage of those who responded to the 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, so as not to give more weight to schools with more teachers or vice versa.

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.

[a] The SFA and control group means and standard deviations for this item are SFA mean = 3.05, control group mean = 2.91, SFA standard deviation = 0.749, control group standard deviation = 0.699.

[b] The SFA and control group means and standard deviations for this item are SFA mean = 3.59, control group mean = 2.85, SFA standard deviation = 0.528, control group standard deviation = 0.660.

[c] The SFA and control group means and standard deviations for this item are SFA mean = 3.56, control group mean = 3.59, SFA standard deviation = 0.511, control group standard deviation = 0.507.

(continued)

Table 5.2 (continued)

[d]The SFA and control group means and standard deviations for this item are SFA mean = 2.14, control group mean = 2.76, SFA standard deviation = 0.751, control group standard deviation = 0.688.

[e]The SFA and control group means and standard deviations for this item are SFA mean = 2.67, control group mean = 2.12, SFA standard deviation = 0.835, control group standard deviation = 0.608.

[f]The SFA and control group means and standard deviations for this item are SFA mean = 3.59, control group mean= 3.07, SFA standard deviation = 0.507, control group standard deviation = 0.799.

[g]One control group school did not respond.

the same level in terms of reading skills (so that teachers do not have to teach and reach students at all points along the achievement distribution in the same class).

Finally, SFA teachers were much more apt to adhere closely to the lesson plans laid out for them in the teacher's guide. As noted in Chapter 4, SFA teachers taking part in focus groups frequently complained that the program was overly rigid, and the survey results in Table 5.2 confirm that SFA teachers were much more likely than their control group counterparts to agree that their reading programs were too scripted. Nonetheless, SFA teachers were much less likely than teachers in the control group schools to report that they changed parts of the reading program that they did not like or with which they disagreed.[21]

Again, it was not the fact that SFA was a new program that led to these differences. When SFA teachers' survey responses were compared with those of teachers at the control group schools that also implemented new reading approaches, the same response patterns were evident: Teachers in the SFA schools were much more likely to find their reading program overly rigid and scripted, but were also much less likely to say that they changed the programs.

It appears that, whatever their opinions of SFA, personnel in the SFA schools saw it as a program that demanded their compliance — and that, for the most part (if the survey results accurately reflect behavior), comply they did.[22]

[21]Unlike teachers in other grades, SFA kindergarten teachers were not significantly more likely than control school kindergarten teachers to agree that their reading program was overly scripted. But like other teachers in SFA schools, they were much less likely than control group kindergarten teachers to say that they changed parts of the reading program that they disagreed with.

For their part, a higher percentage of principals in the SFA schools than in control group schools reported that, when conducting classroom observations, they checked that the teacher was following the prescribed sequence of activities "all" or "most" of the time; the difference narrowly missed being significant at the 10 percent level. Principals of SFA schools were also more likely to report looking to see whether students were working in pairs or teams.

[22]Given the complexities involved in implementing a new and very different reading program, the perceived limits on their autonomy that SFA imposed, and the demands on their time that it created, it is not surprising that SFA teachers were not altogether satisfied with their schools' reading program. If anything, it is

(continued)

The Schools in Operation: Whole-School Features in the Two Groups of Schools

SFA includes a number of noninstructional components that seek to improve the school as a whole. These take the form of committees composed of teachers and other school personnel who are charged with various functions. While not intended to improve academic outcomes directly, the activities of these committees may contribute to better academic performance by promoting a more orderly environment in which learning can occur more readily, by securing better attendance, by increasing parental involvement and support for their children's achievement, and by enlisting community organizations and businesses to help address student needs.

Table 5.3 examines the extent to which these kinds of undertakings are unique to SFA schools. Two survey questions asked principals about the measures their schools took to improve attendance; there are no statistically significant differences between program group and control group schools on these items. Another series of survey questions asked principals whether someone at their schools — an individual or a group — was responsible for carrying out various activities associated with noninstructional whole-school reforms. As the table indicates, while SFA schools were generally more likely than control group schools to have people charged with these responsibilities, the differences are not statistically significant. To be sure, the survey questions provide only limited data; no information is available about what the individuals or groups of people at either set of schools actually did over the course of the year to fulfill their goals. Nonetheless, the data suggest that the SFA schools and control group schools were taking similar steps to improve their environments and meet their students' varied needs.

Intermediate Outcomes in the Theory of Change: How SFA Schools and Control Group Schools Compare

The final section of this chapter turns to the third column in the theory of change discussed in Chapter 1 (Figure 1.1): the intermediate pathways by means of which the SFA program, when implemented, is expected to result in improved student outcomes. As Figure 1.1 shows, some of these intermediate pathways are created in response to the instructional changes brought about in SFA schools; others emerge from the implementation of the program model's whole-school components.

surprising that they expressed the same degree of satisfaction as their control group counterparts. On a scale of 1 to 4, with 4 indicating the highest possible rating and 1 the lowest possible rating, SFA teachers rated their satisfaction with their schools' reading program as 2.71, and control group teachers rated their schools' programs as 2.70. This difference is not statistically significant. (See Appendix Table D.1.)

Table 5.3

SFA-Control Comparisons on Survey Variables Related to Whole-School Aspects of SFA

	Program Group	Control Group	Estimated Difference	P-Value
Variables related to schoolwide structures				
Percentage of principals reporting that a group or individual is responsible for a schoolwide program emphasizing social skills development and conflict resolution	77.8	64.7	13.1	0.407
Percentage of principals reporting that a group or individual is responsible for developing school-wide solutions for students with behavioral challenges	83.3	94.1	-10.8	0.331
Percentage of principals reporting that a group or individual is responsible for fostering relationships with students' families	88.9	76.5	12.4	0.344
Percentage of principals reporting that a group or individual is responsible for building relationships with local businesses and institutions to increase community involvement	50.0	35.3	14.7	0.394
Percentage of principals reporting that staff ask guardians of frequently tardy or absent students to meet with school administrators to discuss student progress and behavior	61.1	71.4	-10.3	0.557
Percentage of principals reporting that staff provide positive recognition to parents whose children attend school regularly	50.0	53.3	-3.3	0.854
Number of schools: 36	19	17[a]		

SOURCE: Spring 2012 principal survey.

NOTES: Items on the principal survey that asked about the principal's levels of agreement were on a 4-point scale: 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree. Percentages of principals who agreed or disagreed with an item were obtained by taking the proportion who responded 3 or 4, expressed as a percentage of those who responded to the 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.

[a] One control group school did not respond.

A limited set of items is available from the principal and teacher surveys to tap the constructs that are included as intermediate pathways. Some of the constructs cannot be assessed at all using the surveys. (For example, there is no measure of perceived instructional quality.) There are no direct measures of other constructs; parental engagement, for example, is gauged through the responses of principals and teachers, not through surveys administered to parents.[23] Despite these limitations, the surveys provide unique data about the extent to which SFA and control group schools are similar or different on the items that best represent the pathway constructs.

Table 5.4 shows the results. Differences that are statistically significant at the 10 percent level or less emerged on three variables; all of these differences favored the control group schools. First, two-thirds of the control group teachers agreed that their reading groups were sufficiently small for students to get the attention that they needed, whereas SFA teachers were almost evenly split between those who agreed and those who disagreed.[24] Control group teachers were also significantly more likely to report that their students were engaged during reading class and that the reading program at their school promoted teacher collaboration — although SFA teachers' and control group teachers' scores on a scale that actually measures teacher collaboration are virtually identical.

Otherwise, the table does not point to strong differences between SFA schools and control group schools on indicators of parental engagement, problem behavior, attention to vision and hearing issues, and teachers' sense of self-efficacy.

*   *   *

This chapter explores areas of similarity and difference between SFA schools and control group schools as a precursor to examining SFA's impacts on students in Chapter 6. The final chapter of the report, Chapter 7, considers the extent to which variables that measure the constructs presented in the SFA theory of change may be associated with any impacts that emerge.

[23] Collecting systematic data from parents would have been prohibitively expensive.

[24] Chapter 4 notes that, in focus groups held with SFA teachers, the teachers frequently voiced their view that their reading classes were too large.

The Success for All Evaluation

Table 5.4

SFA-Control Comparisons on Survey Variables Related to Intermediate Outcomes

	Program Group	Control Group	Estimated Difference	P-Value	
Variables related to logic model pathways					
Greater individualization of instruction					
Percentage of teachers who agree or strongly agree that reading groups are small enough for individual students to receive adequate attention[a]	49.3	67.8	-18.5	0.001	***
More attention to health/vision issues					
Percentage of principals reporting that their school screens all or some students in grades 1, 2, or 3 for vision problems	100.0	100.0	0.0	NA	
Percentage of principals reporting that their school screens all or some students in grades 1, 2, or 3 for hearing problems	100.0	100.0	0.0	NA	
Percentage of principals reporting that their school helps students with hearing or vision problems obtain solutions	93.8	100.0	-6.3	0.377	
Greater parental engagement to support students					
Average score on principal survey measure of parental engagement	2.36	2.46	-0.10	0.558	
More orderly environment					
Percentage of principals who agree or strongly agree that student behavior is a problem at their school[b]	35.3	47.1	-11.8	0.501	
Percentage of teachers who agree or strongly agree that their students are well behaved during reading class[c]	71.3	77.7	-6.4	0.135	
Teachers' greater confidence in ability to reach difficult students and greater belief that such students can succeed					
Percentage of teachers who agree or strongly agree that students come to school ready to learn[d]	53.6	59.7	-6.1	0.276	
Percentage of teachers who agree or strongly agree that they can help most students attain grade-level reading skills by the end of the year, regardless of their family or economic circumstances[e]	73.2	76.0	-2.8	0.511	

(continued)

Table 5.4 (continued)

	Program Group	Control Group	Estimated Difference	P-Value
Percentage of teachers who agree or strongly agree that they can help most students improve reading but not necessarily attain grade-level reading skills by the end of the year[f]	85.6	82.4	3.2	0.348
Greater cooperation among teachers				
Average score on teacher collaboration scale	3.17	3.17	0.00	0.984
Percentage of teachers who agree or strongly agree that the reading program at their school promotes teacher collaboration[g]	47.2	57.1	-9.9	0.097 *
Percentage of teachers who agree or strongly agree that their students are engaged during reading class[h]	78.6	87.9	-9.3	0.012 **
Number of schools: 36	19	17[i]		

SOURCES: Spring 2012 teacher survey and spring 2012 principal survey.

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. Percentages of teachers or principals who agreed or disagreed with an item were obtained by taking the proportion who responded 3 or 4, expressed as a percentage of those who responded to the 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, so as not to give more weight to schools with more teachers or vice versa.

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 SFA and control group means and standard deviations for this item are SFA mean = 2.39, control group mean = 2.72, SFA standard deviation = 0.850, control group standard deviation = 0.816.

[b] The SFA and control group means and standard deviations for this item are SFA mean = 2.29, control group mean = 2.41, SFA standard deviation = 0.772, control group standard deviation = 0.939.

[c] The SFA and control group means and standard deviations for this item are SFA mean = 2.76, control group mean = 2.92, SFA standard deviation = 0.667, control group standard deviation = 0.680.

[d] The SFA and control group means and standard deviations for this item are SFA mean = 2.49, control group mean = 2.59, SFA standard deviation = 0.767, control group standard deviation = 0.713.

[e] The SFA and control group means and standard deviations for this item are SFA mean = 2.84, control group mean = 2.90, SFA standard deviation = 0.692, control group standard deviation = 0.654.

[f] The SFA and control group means and standard deviations for this item are SFA mean = 3.00, control group mean = 2.97, SFA standard deviation = 0.605, control group standard deviation = 0.686.

[g] The SFA and control group means and standard deviations for this item are SFA mean = 2.44, control group mean = 2.58, SFA standard deviation = 0.777, control group standard deviation = 0.689.

[h] The SFA and control group means and standard deviations for this item are SFA mean = 2.90, control group mean = 3.06, SFA standard deviation = 0.620, control group standard deviation = 0.549.

[i] One control group school did not respond.

Chapter 6

Early Impacts of the Success for All Program

Previous chapters in the report describe the launch and implementation of the Success for All (SFA) program in its first year and the differences between SFA schools and control group schools. This chapter examines the last step in the theory of change (Chapter 1, Figure 1.1) and assesses whether the SFA program has produced early impacts on kindergarten students, whose reading performance was measured after one year of program implementation.

Chapter 6 begins with an overview of the analytic approach and outcome measures used for impact analysis in the study. This is followed by a discussion of the SFA program's impact on academic outcomes during the first year of implementation for the full analysis sample. The chapter next presents impacts for key subgroups of sample members. Then the chapter examines whether the program impact varies with students' initial reading levels. The chapter concludes with reflections on the findings.

All analyses reported in this chapter focus on outcomes measured in the spring of the first year that the SFA program was implemented and are intended to examine the immediate impact of the program in that year.

Key Findings

- By the end of the first implementation year, SFA had produced a positive and statistically significant impact on one of the two reading outcomes measured for the main sample of kindergarten students who remained in their study schools for the whole school year and, as a group, had maximum possible exposure to the program. The program impact on the Woodcock-Johnson Word Attack score is 0.55 raw score point, or 0.18 standard deviation in effect size.

- A similar impact on Word Attack test scores was found for the spring analysis sample, which includes all kindergarten students with at least one valid spring test score, including those who moved into a study school over the course of the year.

- The program impact on Word Attack score seems to be robust across a range of demographic and socioeconomic subgroups. Positive and statistically significant impacts were found for male students, female students, black students, and Hispanic students, students in poverty (as defined by each district), and students who were not English language learners.

- The program impact on Word Attack does not vary by students' baseline reading level.
- The SFA program as implemented in the first year did not produce any meaningful impact (positive or negative) on Woodcock-Johnson Letter-Word Identification test scores.

Analytic Approach for the Estimation of Impacts

As explained in Chapter 1, the evaluation randomly assigned 37 study schools to a program group (19 schools) or a control group (18 schools). Given this design, the basic analytic strategy for assessing the impacts of the SFA program is to compare outcomes for schools that were randomly assigned to the program group and received the SFA program with outcomes for schools that were randomly assigned to the control group and used other reading programs. The average outcome in the control group schools represents an estimate of the achievement level that would have been observed in the program group schools if they had not been assigned to the program group. The difference in outcomes between the program group and the control group provides an unbiased estimate of the impact of the SFA program.[1]

Analytically, the primary impact estimation model is a two-level hierarchical model with students nested within schools. The model uses data from all five study districts in a single analysis, treating districts as fixed effects in the model. Separate program impact estimates are obtained for each district and then are averaged across the five districts, weighting each district's estimate in proportion to the number of SFA schools from each district in the sample. Findings in this report therefore represent the impact on student performance in the average SFA school within the five 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.

The impact tables in this chapter report the estimated program impacts on targeted outcome measures as well as the 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 indicates the chance of

[1] Note that all impact estimates are based on an *intent-to-treat* analysis that includes all students in the sample schools at the time that outcome data were collected. Therefore, the impact estimates reported here reflect the impact of assignment to the SFA group or to "business as usual" conditions. Appendix F discusses the impact estimation model in detail.

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

In addition, to provide context for interpreting the estimated impacts, the impact tables also show regression-adjusted mean outcome levels for the program and control groups. The mean outcome levels were calculated for the two groups using the same regression models.[2]

Outcome Measures

Two developmentally appropriate and individually administered measures of reading achievement were used as outcome measures for the impact analysis after one year of SFA program implementation. These two tests are the Woodcock-Johnson Letter-Word Identification and Word Attack scales, which together form the "Basic Reading" achievement cluster of the Woodcock-Johnson III Tests of Achievement.[3] While both tests are considered tests of students' basic reading skills focusing on knowledge of alphabetics and phonics, each test has its own emphasis.[4]

The **Letter-Word Identification** test measures a student's word identification skills and tests reading decoding. Initial items in the test require a student to identify individual letters in bold type. The majority of items in the test require a student to read words of increasing difficulty in isolation. (Words are presented in list form rather than in context.) The test measures students' cognitive ability in feature detection and analysis (for letters) and recognition of visual word forms and/or phonological access to pronunciations associated with visual word forms (that is, the words may or may not be familiar).

The **Word Attack** test measures a student's ability to apply phonic/decoding skills to unfamiliar words. The initial items require a student to produce sounds for a small set of single letters. The majority of items require students to pronounce nonsense words of increasing complexity. The test measures the grapheme-to-phoneme translation of pseudo-words that are not contained in the lexicon and a student's skill in applying phonic and structural analysis skills to the pronunciation of unfamiliar printed words.

[2] Black et al. (2008); Garet et al. (2008).

[3] See Jaffe (2009).

[4] See Wendling, Schrank, and Schmitt (2007).

In addition, as noted in Chapter 2, for a subgroup of students who were instructed primarily in Spanish during their kindergarten year, the Spanish versions of the Letter-Word Identification and Word Attack tests were administered in addition to the English versions, and test scores based on the Spanish tests are used as outcome measures for this subgroup as well.

Early Impacts for the Main Analysis Sample

Table 6.1 displays the impacts of the SFA program on the main analysis sample, comprising students who remained in a study school throughout the first year of program implementation and who had a valid test score in the spring. The SFA program produced a statistically significant impact of 0.18 standard deviation in effect size for the Word Attack raw score. The estimated program impact on the Letter-Word Identification measure, on the other hand, is slightly negative in direction (effect size = 0.0 standard deviation) and is not statistically significant (p-value = 0.991).

Given that two statistical tests were conducted, measures need to be taken to deal with the multiple hypotheses testing issue. For example, if the significance level is set at 5 percent, then for each individual impact estimate, there is a 5 percent chance of falsely obtaining a statistically significant result if there was no true impact on the outcome. Similarly, if one sets the significance level at 10 percent, then there is a 10 percent chance of obtaining a statistically significant result when the impact is actually zero. The more impact estimates one tests, the greater the likelihood of falsely rejecting a “zero impact” hypothesis simply by chance. Following the What Works Clearinghouse guideline, the Benjamini-Hochberg procedure was applied to the individual outcome findings reported in Table 6.1. With this adjustment, the impact on the Word Attack score cannot be considered statistically significant at the 5 percent level, but it is considered significant if the significance threshold is set at 10 percent.[5]

Impacts on Spanish Test Scores for a Subsample of Students

A subset of the students in the main sample primarily spoke Spanish when they entered kindergarten and, therefore, might have had difficulty receiving instruction in English during that year. Different districts have different policies for such students. In one of the districts in the sample,

[5] What Works Clearing House (2013). The Benjamini-Hochberg procedure (Version 3.0, 2013) deals with the multiple hypotheses testing issue by controlling for the False Discovery Rate (FDR). In this procedure, the p-value for an individual estimate is compared with an “adjusted” p-value threshold. The adjustment depends on total number of significance tests conducted and the rank order of the given test among all tests (ranked by p-values in ascending order). In this case, the rank order for the Word Attack test is 1, and the total number of tests is 2, so the adjustment is 1/2. Appendix F provides a more detailed explanation.

The Success for All Evaluation

Table 6.1

Early Impact of SFA on Kindergarten Student Reading Achievement for the Main Analysis Sample (Implementation Year 2011-2012)

Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value
Stable student subgroups					
Woodcock-Johnson Letter-Word Identification	20.02	20.02	-0.01	0.00	0.991
Woodcock-Johnson Word Attack	5.87	5.32	0.55	0.18	0.032 **
Number of schools: 37	19	18			

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

NOTES: The "main 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 and who was enrolled in a study school during the fall of the same year.

The student sample size for the Woodcock-Johnson Letter-Word Identification test is 2,564 students (1,331 in the program group and 1,233 in the control group).

The student sample size for the Woodcock-Johnson Word Attack test is 2,564 students (1,334 in the program group and 1,230 in the control group).

The impact analyses for student 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.

Effect sizes were computed using the full control group's standard deviations for the respective measures. The control group standard deviations are as follows:

WJLWI: 6.98

WJWA: 3.07

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.

these kindergarten students were provided reading instruction primarily in Spanish. For these students, the study team tested their reading achievement in both English and Spanish.

Table 6.2 provides impact estimates on both the English and the Spanish versions of the two tests for this group of students.[6] SFA did not produce any statistically significant impacts for this group on any of the measures. It is worth noting that, for this subsample, the average

[6] Given that all students in this subsample are in one district, there is no district fixed effect in the estimation model used, and the program effect estimates are not weighted across districts.

The Success for All Evaluation

Table 6.2

Early Impact of SFA on Kindergarten Student Reading Achievement for the Spanish Analysis Sample (Implementation Year 2011-2012)

Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value
Subsample of students with Spanish test scores					
Woodcock-Johnson Letter-Word Identification	11.67	12.63	-0.96	-0.14	0.257
Woodcock-Johnson Word Attack	3.71	3.73	-0.02	-0.01	0.952
BATLWI (Spanish version of WJLWI)	22.68	22.84	-0.16	-0.01	0.941
BATWA (Spanish version of WJWA)	12.62	11.67	0.95	0.13	0.495

SOURCES: Woodcock-Johnson Letter-Word Identification test (spring 2012), Woodcock-Johnson Word Attack test (spring 2012), Spanish versions of the same tests (spring 2012), and student records data collected from one district in the study sample.

NOTES: The “Spanish analysis sample” 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 258 students (171 in the program group and 87 in the control group).

The student sample size for the WJWA test is 259 students (172 in the program group and 87 in the control group).

The student sample size for the BATLWI test is 259 students (172 in the program group and 87 in the control group).

The student sample size for the BATWA test is 259 students (172 in the program group and 87 in the control group).

The impact analyses for student 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.

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: 6.98
WJWA: 3.07
BATLWI: 12.68
BATWA: 7.31

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.

scores for the English versions of both tests are much lower for both the program and control group than for students in the main analysis sample, reflecting the fact that English is not the primary language used by this group of students.

Impacts for Other Demographic and Socioeconomic Subgroups of Students

The study team also explored potentially heterogeneous impacts across different subgroups of students defined by demographic and socioeconomic characteristics. These subgroups include those defined by the student's gender, race/ethnicity, English language learner status, poverty status, and special education status.[7] Overall, the findings indicate that the SFA program, as implemented in the first year, produced positive and significant impacts on the Word Attack test scores for male and female students, black students, Hispanic students, students in poverty (as defined by each district), non-English language learners, and students not in special education.[8] This pattern of findings suggests that the program's significant impact on the Word Attack score is robust across different subgroups, with the exception of English language learners and special education students**.** On the other hand, the program did not produce any significant impact on the Letter-Word Identification test scores across a range of subgroups.

Appendix Table F.2 presents detailed findings for each of the subgroups of the main analysis sample.

Impacts for the Spring Sample

During the first implementation year, some students who were in the study schools at the beginning of the school year 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") were not tracked for outcome data collection and, therefore, are not in the spring analysis sample; the second group (the "in-movers") were tested at follow-up. Thus, while the majority of students tested in the spring received the full "dosage" of the SFA intervention, some students in the spring sample received a lesser dosage. Impact results for the spring sample reflect the effects of SFA when taking student mobility into account.[9]

Table 6.3 presents the results of this analysis. Parallel to the impact findings for the main analysis sample, there is a positive and significant impact of SFA on the Word Attack test scores of students in the spring sample (effect size = 0.18 standard deviation), while the impact

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

[8] Although the findings for these groups are statistically significant, they are not statistically different from the findings for their respective counterpart groups. Therefore, the results need to be interpreted with caution.

[9] Appendix B discusses the relationship of in-movers and out-movers to the student samples considered in this report.

Table 6.3

Early Impact of SFA on Kindergarten Student Reading Achievement for the Spring Analysis Sample (Implementation Year 2011-2012)

				Estimated	
	Program	Control	Estimated	Impact	
Outcome	Group	Group	Impact	Effect Size	P-Value
Full student impact analysis sample					
Woodcock-Johnson Letter-Word Identification	19.67	19.74	-0.07	-0.01	0.903
Woodcock-Johnson Word Attack	5.74	5.21	0.54	0.18	0.032 **

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

NOTES: The "spring analysis sample" is defined as the sample of students who had a valid score on the spring 2012 Woodcock-Johnson Letter Word Identification or Woodock-Johnson Word Attack test. The sample for both outcomes consists of students from 37 schools (19 program group schools and 18 control group schools).

There were a total of 8 students who had only one valid score on a test: 4 had valid WJLWI scores only, and 4 others had valid WJWA scores only. For this reason, the total sample size, by test, is 4 less than the total sample size for the spring analysis sample, which is 2897.

The student sample size for the WJLWI test is 2,893 students (1,518 in the program group and 1,375 in the control group).

The student sample size for the WJWA test is 2,893 students (1,521 in the program group and 1,372 in the control group).

The impact analyses for student 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.

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: 6.98

WJWA: 3.07

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 su ms and differences.

estimate on the Letter-Word-Identification test score is small in magnitude and is not statistically significant. The fact that results for the spring sample are so similar to results for the main analysis sample is not surprising, given that a large majority (about 89 percent) of the students in the spring sample remained in their schools during the first implementation year. More student mobility may occur during the second and third follow-up years, and the team will explore the implications of the mobility issue in future reports of the study.

Differential Effects Based on Student Baseline Achievement

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, the study team re-estimated the student impact model, including the main effect for 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 program group and control group reading achievement level 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.[10]

Only students with valid test scores from both fall and spring were used in this analysis. Because this is not a randomly selected subsample, this analysis should be considered nonexperimental, and results reported here cannot be considered causal.

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 Letter-Word Identification 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.

Table 6.4 reports the estimated coefficients for the interaction terms for all three models. None of them detects any significant association between student's baseline reading achievement level and the program effect, regardless of whether baseline reading level is defined by the baseline score on the PPVT, the baseline score on the Letter-Word Identification test, or both baseline scores. These results suggest that the program's impact does not vary by the student's initial reading level.

The Findings in Context

Overall, the early impact findings presented in this chapter indicate that the SFA program produced a positive impact of about 0.18 standard deviation in effect size on one of the two basic reading achievement outcomes that focuses more on student's phonemic awareness and decoding skills — the Word Attack test. This impact seems to be robust and consistent for the

[10] The model implicitly assumes a linear relationship between baseline student test scores and the program effect.

The Success for All Evaluation

Table 6.4

Interaction of Student Baseline Test Scores and the Effects of SFA (Implementation Year 2011-2012)

	Baseline Test Score Interaction Effect		
Standardized Outcome	Estimate	Standard Error	P-Value
Model 1: Interact fall baseline PPVT[a] score			
WJLWI[b]	-0.01	0.02	0.690
WJWA[c]	0.00	0.01	0.644
Model 2: Interact fall baseline WJLWI score			
WJLWI	0.00	0.02	0.839
WJWA	0.00	0.01	0.557
Model 3: Interact baseline WJLWI and PPVT score			
Spring WJLWI			
Interact with PPVT	-0.01	0.01	0.573
Interact with WJLWI	0.01	0.02	0.540
Spring WJWA			
Interact with PPVT	0.00	0.01	0.906
Interact with WJLWI	0.00	0.01	0.726

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

NOTES: This analysis uses the sample of students who were present in the sample schools in the fall and spring of the 2011-2012 school year and who have valid fall and spring test scores. The sample for all models and outcomes consists of students from 37 schools (19 in the program group and 18 in the control group).

The student sample size for the Woodcock-Johnson Letter-Word Identification test (for each of Models 1, 2, and 3) is 2,522 students (1,307 in the program group and 1,215 in the control group).

The student sample size for the Woodcock-Johnson Word Attack test (for Models 1 and 2) is 2,522 students (1,310 in the program group and 1,212 in the control group).

The impact analyses for student 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 treatment indicator. Only the estimated coefficient for the interaction term in each model is reported here.

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).

full analysis sample, as it persists across various large subgroups and does not seem to vary with students' baseline reading achievement levels. Yet the program does not seem to have affected students' test scores on the Letter-Word Identification measure.

Not only are these findings consistent within the study sample, but they are also consistent with previous findings about the effects of the SFA program on early-grade reading. For example, Borman et al. find a significant SFA effect of 0.32 standard deviation in effect size on Word Attack for kindergarten students after one year of implementing the SFA program.[11]

To put the estimated effect on the Word Attack test into context: Calculations based on national norming samples for seven major standardized tests show that, during the kindergarten year, an average student's reading achievement test score grows 1.52 standard deviations in effect size.[12] This indicates that the impact on Word Attack experienced by the program group students in this study represents about 12 percent of the annual growth for an average kindergarten student.

In addition, this result is comparable to the impacts of other similar school reform programs. For example, Borman et al. used a meta-analysis to show that the overall effect across 29 of the most widely deployed comprehensive school reforms ranged between 0.09 and 0.15 standard deviation in effect size.[13] Similarly, a synthesis of observational studies on the effectiveness of the federal Title I program put its effect at around 0.11 standard deviation.[14] Furthermore, the Tennessee Student-Teacher Ratio (STAR) study found that reducing early-grade classes from their standard size of 22 to 26 students to only 13 to 17 students significantly increases average reading performance in elementary schools by 0.11 to 0.22 standard deviation in effect size.[15]

All these comparisons indicate that the magnitude of the SFA impact on Word Attack reported here is meaningful and on a par with what is to be expected from this kind of intervention. This is especially promising, given that the SFA program had been on the ground for less than a full year in the program group schools at the time that students' achievement was measured.

[11] The study by Borman, Hewes, Overman, and Brown (2003), like this one, used a school-level randomization design that included 35 schools and about 2,409 students in the analysis sample. Also like this one, it did not find a significant impact of SFA on Letter Identification or Word Identification scores for kindergarten students. There were, however, positive and significant impacts for SFA first-graders on Word Attack and for SFA second-graders on Passage Comprehension.

[12] Hill, Bloom, Black, and Lipsey (2007).

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

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

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

Chapter 7

Reflections and Conclusions

Success for All (SFA) is a complex and far-reaching intervention that includes both instructional and whole-school reforms. It requires that schools put in place new structures and processes that are time-consuming and labor-intensive. It requires even experienced elementary school reading teachers to adopt new practices that may appear unfamiliar to them and whose effectiveness may seem questionable (at least until they have seen the outcomes of these practices). It is not surprising that the survey and interview responses of teachers in the SFA schools that are presented in this early report reflect the schools' initial difficulties in implementing a new and complicated program. Although teachers in the SFA schools rated the professional development in reading that they received as both more extensive and more helpful than did their counterparts in the control group schools, many teachers believed that it did not prepare them adequately for teaching in SFA classrooms.

Nonetheless, by the end of the first year, almost all the program schools had put in place three-quarters or more of the elements of SFA that the Success for All Foundation (SFAF) considers to have the highest priority, along with many lower-priority elements. And many teachers were beginning to feel more comfortable with the program and were looking forward to a smoother second year of operations.

Instructional logs and principal and teacher surveys point to key ways in which reading instruction in SFA schools and control group schools differed. Reading lessons in SFA schools were more likely than those in control group schools to emphasize comprehension. Teachers in SFA schools were also more likely to emphasize aspects of instruction that are hallmarks of the SFA program: grouping and regrouping of students for reading, from first grade on, by ability level and across grades, and, from kindergarten on, regular use of cooperative learning techniques and close adherence to a highly structured curriculum. Even teachers who were not sold on SFA followed its lesson plans closely.

The first-year impact analysis of the study centers on students who entered kindergarten in the program group and control group schools in fall 2011; it assesses their reading skills in spring 2012. (Subsequent reports will also measure effects on students in higher grades, but the kindergarten cohort is the focus of attention, since students will have received SFA reading instruction since the beginning of their educational careers.) These findings are encouraging: They show that kindergartners in SFA schools scored significantly higher than their control group counterparts on one of two standardized measures of early literacy. The impact is robust: It holds up across several subgroups and remains significant at the 10 percent level when a procedure that corrects for multiple hypothesis testing is applied. The effect size (p-value = 0.18) is on a par with that achieved by a number of other prominent school reform initiatives.

These results are preliminary, for a number of reasons. First, students were tested after only one year of exposure to the program. Second, the measures used for kindergartners test phonetic skills; what ultimately matters for reading is comprehension, and this will not be assessed until students are slightly older. Third, it is difficult to measure kindergartners' reading skills accurately. Fourth, teachers are likely to be able to implement SFA in their classrooms more easily and more smoothly in subsequent years than in this first year. And, finally, it is anticipated that a number of program elements not now in place in the SFA schools will be added over time.

The implementation analysis points to two components that are in need of strengthening: tutoring and instruction for English language learners. Half the program schools were unable to provide tutoring for the proportion of students specified by SFAF or were unable to provide tutoring daily. Since tutoring is a key intervention for struggling students, its absence is an important gap. Second, teachers pointed out the limitations of the SFA Spanish-language program used in schools where bilingual instruction is available for students whose dominant language is Spanish. Teachers complained that the Spanish-language materials arrived later than the materials in English, that they contained errors (subsequently corrected), and that they did not include all the resources that the English program offered. Moreover, a computerized version of tutoring for the Spanish language program was not available. SFAF has signaled its awareness of this issue and is working to improve its program for Spanish-speaking students.

The early implementation and impact findings presented in this report will be updated in two additional reports, to be published in 2014 and 2015. These reports will also explore additional topics, including:

- Students' comprehension skills
- The relationship of Snapshot scores to impacts
- Aspects of implementation that are especially related to the i3 scale-up process
- A comparison of impacts found in this evaluation with the impacts found in earlier evaluations of SFA.

Appendix A

Data Sources and Response Rates for the Success for All Evaluation

The Success for All Evaluation

Appendix Table A.1a

Data Sources and Overall Response Rates

Instrument and Purpose	Number Targeted	Number of Respondents	Response Rate (%)
Principal survey[a] 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	36	97.3
Teacher survey[b] 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.	995	896	90.1
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	36	97.3
Facilitator interviews: Interviews with the SFA facilitator at program group schools to learn about his or her duties and the SFA implementation story.	19	17	89.5

(continued)

Appendix Table A.1a (continued)

Instrument and Purpose	Number Targeted	Number of Respondents	Response Rate (%)
Teacher focus groups: Focus group with teachers at program group schools to learn about implementation of SFA in the classrooms.	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[c] 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 classroom practices.	2,264	1,383	61.1
Baseline tests			
Woodcock-Johnson Letter-Word Identification test was administered to all sample students in fall 2011. Spanish versions were administered to Spanish speakers without English mastery. Test scores were used as covariates in the impact estimation model.	2,956	2,849	96.4
Peabody Picture Vocabulary Test was administered to all sample students in fall 2011. Spanish versions were administered to Spanish speakers without English mastery. Test scores were used as covariates in the impact estimation model.	2,956	2,835	95.9
Follow-up tests			
Woodcock-Johnson Letter-Word Identification test was administered to all sample students in spring 2012. Spanish versions of the tests were administered to students without English mastery. Test scores serve as an outcome variable in the impact estimation model.	3,003	2,893	96.3

(continued)

Appendix Table A.1a (continued)

Instrument and Purpose	Number Targeted	Number of Respondents	Response Rate (%)
Woodcock-Johnson Word Attack test was administered to all sample students in spring 2012. Spanish versions of the tests were administered to students without English mastery. Test scores serve as an outcome variable in the impact estimation model.	3,003	2,893	96.3
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.	5	5	100.0

NOTES: [a]36 of 37 school principals returned the survey. At one of the 19 program group schools, the principal did not receive the SFA portion of the survey. Thus, the responses for the items pertaining to the SFA program represent a maximum of 18 principals at SFA schools.

[b]36 of 37 schools returned surveys from their teachers. The school that did not return its surveys was a control group school with 18 teachers. 910 teachers of 995 returned surveys. Of these, 14 teachers were dropped because they did not teach reading. This left 896 teachers in the analytic sample: 404 in control group schools and 492 in program group schools. Roughly half of teacher respondents come from just one district.

[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.

The Success for All Evaluation

Appendix Table A.1b

Data Sources and Response Rates, by Program or Control Status

	Program Group			Control Group			P-Value of
	Number	Number of	Response	Number	Number of	Response	Response Rate
Instrument and Purpose	Targeted	Respondents	Rate (%)	Targeted	Respondents	Rate (%)	Difference[a]
Principal survey[b] 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.	19	19	100.0	18	17	94.4	0.311
Teacher survey[c] 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.	523	492	94.1	472	404	85.6	0.141

(continued)

Appendix Table A.1b (continued)

	Program Group			Control Group			P-Value of
	Number	Number of	Response	Number	Number of	Response	Response Rate
Instrument and Purpose	Targeted	Respondents	Rate (%)	Targeted	Respondents	Rate (%)	Difference[a]
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.	19	19	100.0	18	17	94.4	0.311
Facilitator interviews: Interviews with the SFA facilitator at program group schools to learn about his or her duties and the SFA implementation story.	19	17	89.5	–	–	–	–
Teacher focus groups: Focus group with teachers at program group schools to learn about implementation of SFA in the classrooms.	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	–	–	–	–

(continued)

Appendix Table A.1b (continued)

	Program Group			Control Group			P-Value of
	Number	Number of	Response	Number	Number of	Response	Response Rate
Instrument and Purpose	Targeted	Respondents	Rate (%)	Targeted	Respondents	Rate (%)	Difference[a]
Teacher logs[d] 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 classroom practices.	1,102	735	66.7	1,162	648	55.7	0.690
Baseline tests							
Woodcock-Johnson Letter-Word Identification test was administered to all sample students in fall 2011. Spanish versions were administered to Spanish speakers without English mastery. Test scores were used as covariates in the impact estimation model.	1,542	1,480	96.0	1,414	1,369	96.8	0.338
Peabody Picture Vocabulary Test was administered to all sample students in fall 2011. Spanish versions were administered to Spanish speakers without English mastery. Test scores were used as covariates in the impact estimation model.	1,542	1,468	95.2	1,414	1,367	96.7	0.136

(continued)

Appendix Table A.1b (continued)

	Program Group			Control Group			P-Value of
	Number	Number of	Response	Number	Number of	Response	Response Rate
Instrument and Purpose	Targeted	Respondents	Rate (%)	Targeted	Respondents	Rate (%)	Difference[a]
Follow-up tests							
Woodcock-Johnson Letter-Word Identification test was administered to all sample students in spring 2012. Spanish versions of the tests were administered to students without English mastery. Test scores serve as an outcome variable in the impact estimation model.	1,571	1,513	96.3	1,432	1,380	96.4	0.978
Woodcock-Johnson Word Attack test was administered to all sample students in spring 2012. Spanish versions of the tests were administered to students without English mastery. Test scores serve as an outcome variable in the impact estimation model.	1,571	1,516	96.5	1,432	1,377	96.2	0.683
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.	–	–	–	–	–	–	–

(continued)

Appendix Table A.1b (continued)

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

[a] Some measures were intended only for the program group; therefore, it was not possible to test the difference in response rates between the program and control groups.

[b] 36 of 37 school principals returned the survey. At one of the 19 program group schools, the principal did not receive the SFA portion of the survey. Thus, the responses for the items pertaining to the SFA program represent a maximum of 18 principals at SFA schools.

[c] 36 of 37 schools returned surveys from their teachers. The school that did not return its surveys was a control group school with 18 teachers. 910 teachers of 995 returned surveys. Of these, 14 teachers were dropped because they did not teach reading. This left 896 teachers in the analytic sample: 404 in control group schools and 492 in program group schools. Roughly half of teacher respondents come from just one district.

[d] 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

The Student Samples and Student Mobility

Formation of Student Samples

As shown in Appendix Table B.1, the Success for All (SFA) evaluation makes use of three main samples of kindergarten students who were enrolled in the 37 study schools and who were not in self-contained special education classes: (1) the student baseline sample, (2) the main analysis sample, and (3) the spring analysis sample. Within the main samples but not shown in the table, an important subgroup consists of (4) Spanish-speaking students who received bilingual instruction and who were tested on both the English and the Spanish versions of the tests.

The Student Baseline Sample

Out of a total of 2,962 kindergarten students who were enrolled in fall 2011, three program group students were unable to be tested due to limited English capabilities. The remaining 2,959 students (1,545 program group and 1,414 control group students) make up the baseline sample.

The Main Analysis Sample

The main analysis sample includes any kindergarten student who was in the baseline sample and who had a valid test score on one of the two tests administered in spring 2012: the Woodcock-Johnson Letter-Word Identification test and the Woodcock-Johnson Word Attack test. A test score is considered valid if the test was administered correctly and the child was not designated as having special needs or having limited English abilities.

The Spring Analysis Sample

The spring analysis sample includes any kindergarten student who had a valid test score on one of the two tests administered in spring 2012, whether or not he or she was in the baseline sample.

Terminology Related to Student Mobility

- **Not eligible for testing:** any student identified as having special needs that could not be accommodated at testing, including students who were unable to take the English-language version of the test that was administered
- **Out-mover:** any student who was enrolled in a study school in fall 2011 but who was not enrolled in a study school in spring 2012
- **In-mover:** any student who was enrolled in a study school in spring 2012 but who was not enrolled in a study school in fall 2011

The Success for All Evaluation

Appendix Table B.1

Derivation of Study Samples

Student Status	SFA Schools (Number of Students)	Control Group Schools (Number of Students)
Deriving the baseline sample		
All students enrolled in fall 2011	1,545	1,414
Ineligible for testing	(3)	–
Total: baseline sample	1,542	1,414
Eligible for testing but does not have a valid test score[a]	(62)	(41)
Total: students with at least one valid test score at baseline[b]	1,480	1,373
Deriving the main analysis sample from the baseline sample		
All students enrolled in fall 2011	1,545	1,414
Students moving to a nonstudy school between fall and spring (out-movers)	(161)	(139)
Eligible for testing in both fall and spring but has no valid spring test scores	(42)	(36)
Eligible for testing in fall but ineligible in spring	(7)	(5)
Ineligible for testing in fall and remained ineligible in spring	(1)	–
Total: main analysis sample[c]	1,334[d]	1,234
Deriving the spring analysis sample by combining the main analysis sample and in-movers with valid spring tests		
Main analysis sample	1,334	1,234
Students not enrolled in fall but enrolled in spring (in-movers)	206	164
Ineligible for testing in spring	(5)	(8)
Eligible for testing in spring but has no valid spring test scores	(14)	(14)
Total: spring analysis sample[e]	1,521	1,376

NOTES: [a]This could happen because of student withdrawal, persistent absence on testing dates, parental refusal for the child to be tested, or test administration errors that invalidated scores. The vast majority of such cases were due to student withdrawal.

[b]This is not a sample of primary focus in this report and is presented here for documentation purposes.

[c]This includes any student enrolled in a study school in fall 2011 and who had at least one valid test score in spring 2012.

[d]One of these students was ineligible for testing at baseline.

[e]This includes any student with at least one valid test score in spring 2012, regardless of whether the student was enrolled at a study school in fall 2011.

- **Nonresponse:** includes any student who was eligible for testing but who did not have a valid test score[1]

Mobility Between Study Schools

Appendix Table B.2 shows the percentage of students who transferred from a study school into another study school after fall testing but before spring testing, expressed as a percentage of students eligible for fall testing. There is no statistically significant relationship between a student's research group (program group or control group) and whether he or she transferred schools between fall and spring testing.

The Success for All Evaluation

Appendix Table B.2

Percentage of Students Transferring Between Study Schools, by Program or Control Status

	Spring 2012	
	Program group	Control group
Fall 2011		
Program group	0.5% (N = 7)	0.9% (N = 14)
Control group	0.6% (N = 9)	0.8% (N = 11)

NOTE: Percentages are calculated using the number of students enrolled at baseline: 1,545 program group students and 1,414 control group students. The values of 0.5 percent and 0.9 percent in the top row were calculated using a denominator of 1,545, while percentages in the bottom row were calculated using a denominator of 1,414.

There is not a statistically significant difference between the proportion of out-movers in the program group and control group. Additionally, there is not a statistically significant difference between the proportion of in-movers in the two groups.

Appendix Table B.3 summarizes, by research group, the distribution of in-movers and out-movers. The first column shows the percentage of out-movers expressed as a percentage of all students enrolled in fall 2011, by research group. For example, 10.4 percent of all students

[1] Reasons for nonresponse include persistent student absence on testing dates, student withdrawal, parents' refusal that their child be tested, or test-administration errors that invalidate scores.

The Success for All Evaluation

Appendix Table B.3

Percentage of Students Moving into and out of Study Schools, by Program or Control Status

	Out-Movers	In-Movers
Program group	10.4% (N = 161)	13.0% (N = 206)
Control group	9.8% (N = 139)	11.4% (N = 164)

NOTES: The percentages of out-movers by experimental status were calculated using the number of students by experimental status at baseline: 1,545 program group students and 1,414 control group students. The percentages of in-movers by experimental status were calculated using the number of students by experimental status in the spring: 1,585 program group and 1,444 control group students. These numbers are slightly different from those used in analysis, in which a student's baseline experimental status was used in the event that the student switched experimental statuses between fall and spring, and each student was required to have at least one valid spring test score.

enrolled in program group schools in fall 2011 transferred into a nonstudy school by spring 2012. Likewise, 9.8 percent of all students enrolled in control group schools in fall 2011 transferred to a nonstudy school by spring 2012. The rightmost column of the table shows the analogous information for in-movers in the two research groups.

Appendix Tables B.2 and B.3 show different aspects of student mobility. The former summarizes the distribution of students who transferred from one study school to another study school, whereas the latter summarizes the distribution of students who transferred between study and nonstudy schools. Statistical tests were carried out separately on each kind of mobility, as noted above. However, the research team also tested whether a student's research status is related to whether he or she transferred at all (within study schools or between nonstudy and study schools). In this way, both kinds of mobility — as summarized in the two tables — were tested simultaneously. No statistically significant relationship was found between a student's research status and whether or not he or she transferred schools.

Appendix C

Baseline Equivalence Tests for Additional Samples

The Success for All Evaluation

Appendix Table C.1

Selected Characteristics of the Main Analysis Sample, by Program or Control Status (Fall of School Year 2011-2012)

	Program Group	Control Group	Estimated Difference	P-Value for Estimated Difference
Age (years)[a]	5.5	5.5	0.0	0.538
Students eligible for free and reduced-price lunch (%)	88.0	88.5	-0.4	0.841
Race/ethnicity (%)				
White	12.7	13.5	-0.8	0.594
Black	20.0	19.5	0.5	0.918
Hispanic	63.9	64.9	-1.0	0.841
Asian	1.5	1.0	0.5	0.466
Other	1.9	1.0	0.9	0.111
Male (%)	50.2	49.0	1.1	0.583
English language learners (%)	24.5	17.9	6.6	0.073 *
Special education status (%)	7.5	7.6	-0.1	0.947
Fall 2011 Peabody Picture Vocabulary Test				
Scaled test score	91.5	92.4	-0.9	0.450
Percentile equivalent[b]	30	30	–	–
Fall 2011 Woodcock-Johnson Letter-Word Identification				
Scaled test score	93.5	95.4	-1.8	0.081 *
Percentile equivalent	34	40	–	–

(continued)

Appendix Table C.1 (continued)

SOURCE: Student-level test files on the Peabody Picture Vocabulary Test (PPVT) and the Woodcock-Johnson Letter-Word Identification (WJLWI) test taken in fall 2011.

NOTES: The "main analysis sample" is defined as any student who had a valid test score on at least one spring test and who was present in the fall baseline sample. The main analysis sample consists of 2,568 students: 1,334 are in the program group, and 1,234 are in the control group.

Sample size for the main analysis sample by test

2,530 students had valid PPVT test scores in fall 2011: 1,310 are in the program group, and 1,220 are in the control group. 2,540 students had valid WJLWI test scores in fall 2011: 1,321 are in the program group, and 1,219 are in the control group.

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

The values 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 values for the control group 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.

[a]Age is calculated as the age (in years) of a student as of September 1, 2011.

[b]The percentile equivalent corresponds to the mean standard score for a given study group.

Appendix Table C.2

Selected Characteristics of the Subsample of Students with Spanish Test Scores, by Program or Control Status (Fall of School Year 2011-2012)

Characteristics	Program Group	Control Group	Estimated Difference	P-Value for Estimated Difference
Age (years)[a]	5.5	5.5	0.0	0.903
Students eligible for free and reduced-price lunch (%)	98.8	100.0	-1.2	0.315
Male (%)	55.2	50.1	5.2	0.459
Special education status (%)	4.1	6.7	-2.6	0.524
Fall 2011 PPVT				
Scaled score	72.5	73.4	-0.9	0.629
Percentile equivalent	37	39		
Fall 2011 WJLWI test				
Scaled score	85.0	84.7	0.3	0.902
Percentile equivalent	4	4		
Fall 2011 TVIP				
Test score	33.8	32.0	1.8	0.499
Fall 2011 BATLWI test				
Scaled score	101.8	98.9	2.9	0.302
Percentile equivalent	55	47		

(continued)

Appendix Table C.2 (continued)

SOURCES: Student-level test files on the Peabody Picture Vocabulary Test (PPVT), Woodcock-Johnson Letter-Word Identification (WJLWI) test, and the Spanish-language versions of these tests (the TVIP and the BATLWI test, respectively) taken in fall 2011.

NOTES: The "subsample of students with Spanish test scores" is defined as the set of students who were enrolled in a study school in fall 2011 and had at least one valid score on the BATLWI or BATWA test administered in spring 2012.

Sample sizes for the subsample of students with Spanish test scores

247 students took the PPVT: 163 program group students and 84 control group students.

252 students took the WJLWI test: 167 program group students and 85 control group students.

230 students took the TVIP: 147 program group students and 83 control group students.

235 students took the BATLWI test: 150 program group students and 85 control group students.

Due to missing values, the number of students included varies by characteristic. Sample sizes reported here are for the subsample of students with Spanish test scores.

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

The values for the program group are the weighted average of the observed district means for schools or students randomly assigned to that group (using the number of program group schools in each district as weight). The values for the control group 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.

[a]Age is calculated as the age (in years) of a student as of September 1, 2011.

[b]The percentile equivalent corresponds to the mean standard score for a given study group.

Appendix D

Scale Construction and Statistical Properties of Scales

Overview of the Scale Creation Process

In order to reduce the error in measuring a construct, it is common to combine survey items associated with that construct into a scale. Several constructs were of interest to the Success for All (SFA) research team, and the teacher surveys were designed with these constructs in mind.

The research team hypothesized that certain items would cluster together to form a scale measuring a particular construct. The reliability of each candidate item for a given scale was tested using Cronbach's alpha. This essentially measures the intercorrelation among a set of items and, thus, whether it is statistically feasible that a single construct is being tapped by them. Items were removed from the candidate set when doing so increased the reliability, and this process was continued until the reliability of the set of items had an alpha of at least 0.7. Usually it was not necessary to remove any items from the original pool of items; at most, two items were removed. All but one scale turned out to have a reliability of 0.8 or higher.

Once the team had identified a final set of items pertaining to the proposed scale, the scale score for each respondent was formed by taking the mean value of all items pertaining to the construct, subject to two requirements. First, a scale had to have at least three items. Second, a respondent could receive a scale score only if he or she answered at least half the items in the scale, rounding down for scales containing an odd number of items (except for five-item scales, for which it was required that the respondent answer at least three items). Because all scales turned out to contain between five and seven items, this means that the respondent had to answer at least three items in a given scale to be included in the calculations.

It is important to note that not every observation that was eventually assigned a scale score actually contributed to the calculation of the scale's reliability. Reliability tests are computed based only on teacher respondents who had all *nonmissing* items. If a teacher skipped even one item in the proposed scale, his or her answers to the remaining items were excluded in calculating the scale's reliability. If the scale met the reliability threshold of an alpha of 0.7, however, every respondent who answered at least three items in the scale, though not necessarily all of them, was assigned a scale score.

Theoretically, it is possible that teachers who were missing items in the scale are different from the teachers who were not missing any items in the scale, in such a way that it is invalid to assign those who were missing items a scale score. However, such cases are relatively few. The percentage of these cases is highest for control group teachers on the professional development scale; 12 percent of respondents (37 teachers) did not respond to all the items on this scale. The reason for this is readily understood: Control group teachers were more likely to select the option "5: Professional development was not received in this area," which was recoded to missing when testing reliability. (For details, see the section below about the professional development scale.) This suggests that missing values were not, in this case, signs that control group teachers had more trouble understanding the items in the scale than SFA teachers, or that something else was awry, but simply that control group teachers received less professional development then their program group counterparts.

For all other scales, less than 5 percent of program group or control group teachers had missing scale items.

Factor analysis was used to check whether the scales were tapping multiple constructs. None of the scales used in the analysis had more than one factor, providing strong evidence that the scales were indeed tapping a single construct.

Statistical Properties of Scales

Teacher Satisfaction with the School's Reading Program (Appendix Table D.1)

Survey item stem: *To what extent do you agree with the following statements about the reading program at your school during the 2011-2012 school year?* (1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree)

9a: You are satisfied with the overall quality of the reading program at your school.

9b: You have been given the support you need, in terms of additional resources, to implement your school's reading program.

9d: You are satisfied with the overall quality of the reading materials (including technology) that you use.

9g: The reading program at your school adequately serves most of your students.

9k: Your school's reading program gets students excited about reading or learning how to read.

Teacher Evaluation of the Principal's Leadership (Appendix Table D.2)

Survey item stem: *To what extent do you agree with the following statements about your principal's role in the reading program at your school in the 2011-2012 school year?* (1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree)

Your principal has . . .

11a: Served as a knowledgeable source concerning reading standards and curriculum.

11b: Ensured that teachers have time for planning reading instruction.

11c: Provided teachers with adequate classroom materials to improve student reading proficiency.

11d: Ensured that teachers receive adequate professional development in reading.

11e: Reached out to parents to support reading practices at home.

11f: Ensured that teachers receive regular feedback regarding their reading instruction.

Teacher Evaluation of the Helpfulness of Professional Development in Reading Instruction (Appendix Table D.3)

Survey item stem: *Since the start of the 2011-2012 school year, your professional development in reading instruction has . . .* (1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree, 5 = Professional development was not received in this area.)[1]

13a: Helped you learn how to implement your school's reading program properly.

13b: Helped you learn new techniques for reading instruction.

13c: Helped you develop strategies to better meet the needs of the reading students who struggle the most.

13d: Helped you learn how to use classroom materials, including technology, to improve reading instruction.

13e: Helped you learn how to better use the time allocated to reading instruction.

13f: 'Helped you learn how to implement cooperative learning techniques among students.'

13g: 'Helped you learn how to use reading assessment data to guide instruction.'

Extent That Data Are Used in Relation to the Reading Program (Appendix Table D.4)

Survey item stem: *To what extent do you agree with the following statements?* (1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree)

Since the start of the 2011-2012 school year, your school has used data to . . .

16a: Evaluate the reading progress of students over time.

16b: Communicate with and inform parents about student reading performance.

16c: Identify students struggling with reading.

16d: Develop strategies to move students from the below basic and basic categories into the proficient category on standardized tests of reading skills.

16e: Examine school-wide instructional issues related to reading.

[1] It was necessary to recode this response option to "missing." Retaining the value of 5 would imply that the respondent agreed more than "4: Strongly Agree," which is not the case. Recoding to 0 would also be wrong, as it would imply that the respondent disagreed more than "1: Strongly Disagree."

16f: Identify reading teachers who need instructional improvement.

Extent of Teacher Collaboration (Appendix Table D.5)

Survey item stem: *Considering your experiences during the 2011-2012 school year, to what extent do you agree with the following statements?* (1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree)

17d: You approach other teachers when you have a concern or question related to instruction.

17e: You discuss student behavioral challenges with other teachers.

17f: You discuss how to improve instruction with other teachers.

17g: You discuss lesson plans that were not particularly successful with other teachers.

17h: You discuss lesson plans that were successful with other teachers.

17i: You share your students' work with other teachers.

Extent That the Program Represents a Change Toward SFA Practices (Appendix Table D.6)

One of the program group sites mistakenly received the control group version of the survey, which omitted SFA-specific items, and therefore was prevented from obtaining a scale score corresponding to the following items. As a result, 30 observations from this site could not be used; only 18 out of 19 program group schools were used for this scale.

Survey item stem: *To what extent do you agree with the following statements?* (1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree)

22a: As a result of SFA, you received training in reading instruction that you had not received before.

22b: The SFA facilitator provides you with useful feedback.

22d: As a result of SFA, you have changed your process for reviewing student reading data.

22e: As a result of SFA, you have changed your process for grouping students for reading.

22f: Overall, your school has benefited from the SFA program.

The Success for All Evaluation

Appendix Table D.1

Reliability Estimates from Exploratory Factor Analysis: Teacher Satisfaction with the School's Reading Program

Population	Number of Items in Scale	Reliability	Mean[a]	Standard Deviation	Number of Observations Used in Reliability Estimate[b]	Number of Teachers with Scale Scores[c]	Total
Overall	5	0.86	2.70	0.60	865	883	896
Program group		0.87	2.71	0.65	478	486	492
Control group		0.84	2.70	0.52	387	397	404

NOTES: Reliability is measured using Chronbach's alpha. All values for this table were obtained using PROC CORR in SAS Enterprise Guide 4.3.

[a]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.

[b]To be included in the reliability estimate, the respondent had to answer all questions in the scale (that is, no items could have missing data).

[c]To be assigned a scale score, the respondent had to answer at least 3 items in the scale.

The Success for All Evaluation

Appendix Table D.2

Reliability Estimates from Exploratory Factor Analysis: Teacher Evaluation of the Principal's Leadership

Population	Number of Items in Scale	Reliability	Mean[a]	Standard Deviation	Number of Observations Used in Reliability Estimate[b]	Number of Teachers with Scale Scores[c]	Total
Overall	6	0.89	2.75	0.61	859	878	896
Program Group		0.90	2.75	0.64	480	483	492
Control Group		0.88	2.74	0.56	379	395	404

NOTES: Reliability is measured using Chronbach's alpha. All values for this table were obtained using PROC CORR in SAS Enterprise Guide 4.3.

[a]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.

[b]To be included in the reliability estimate, the respondent had to answer all questions in the scale (that is, no items could have missing data).

[c]To be assigned a scale score, the respondent had to answer at least 3 items in the scale.

The Success for All Evaluation

Appendix Table D.3

Reliability Estimates from Exploratory Factor Analysis: Teacher Evaluation of the Helpfulness of Professional Development in Reading Instruction

Population	Number of Items in Scale	Reliability	Mean[a]	Standard Deviation	Number of Observations Used in Reliability Estimate[b]	Number of Teachers with Scale Scores[c]	Total
Overall	7	0.93	2.67	0.63	690	777	896
Program group		0.92	2.74	0.61	413	463	492
Control group		0.94	2.58	0.66	277	314	404

NOTES: Reliability is measured using Chronbach's alpha. All values for this table were obtained using PROC CORR in SAS Enterprise Guide 4.3.

[a] 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.

[b] To be included in the reliability estimate, the respondent had to answer all questions in the scale (that is, no items could have missing data).

[c] To be assigned a scale score, the respondent had to answer at least 3 items in the scale.

The Success for All Evaluation

Appendix Table D.4

Reliability Estimates from Exploratory Factor Analysis: Extent That Data Are Used in Relation to the Reading Program

Population	Number of Items in Scale	Reliability	Mean[a]	Standard Deviation	Number of Observations Used in Reliability Estimate[b]	Number of Teachers Scores[c]	Total
Overall	6	0.84	2.79	0.50	830	875	896
Program group		0.85	2.78	0.53	451	481	492
Control group		0.82	2.81	0.46	379	394	404

NOTES: Reliability is measured using Chronbach's alpha. All values for this table were obtained using PROC CORR in SAS Enterprise Guide 4.3.

[a] 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.

[b] To be included in the reliability estimate, the respondent had to answer all questions in the scale (that is, no items could have missing data).

[c] To be assigned a scale score, the respondent had to answer at least 3 items in the scale.

The Success for All Evaluation

Appendix Table D.5

Reliability Estimates from Exploratory Factor Analysis: Extent of Teacher Collaboration

Population	Number of Items in Scale	Reliability	Mean[a]	Standard Deviation	Number of Observations Used in Reliability Estimate[b]	Number of Teachers with Scale Scores[c]	Total
Overall	6	0.86	3.15	0.45	866	885	896
Program group		0.86	3.15	0.45	476	487	492
Control group		0.87	3.16	0.45	390	398	404

NOTES: Reliability is measured using Chronbach's alpha. All values for this table were obtained using PROC CORR in SAS Enterprise Guide 4.3.

[a] 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.

[b] To be included in the reliability estimate, the respondent had to answer all questions in the scale (that is, no items could have missing data).

[c] To be assigned a scale score, the respondent had to answer at least 3 items in the scale.

The Success for All Evaluation

Appendix Table D.6

Reliability Estimates from Exploratory Factor Analysis: Extent That the Program Represents a Change Toward SFA Practices

Population	Number of Items in Scale	Reliability	Mean[a]	Standard Deviation	Number of Observations Used in Reliability Estimate[b]	Number of Teachers with Scale Scores[c]	Total
Program group	5	0.73	2.76	0.50	420	448	492

NOTES: Reliability is measured using Chronbach's alpha. All values for this table were obtained using PROC CORR in SAS Enterprise Guide 4.3.

[a] 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.

[b] To be included in the reliability estimate, the respondent had to answer all questions in the scale (that is, no items could have missing data).

[c] To be assigned a scale score, the respondent had to answer at least 3 items in the scale.

Appendix E

The Success for All Foundation (SFAF)
School Achievement Snapshot

Appendix E first describes the School Achievement Snapshot, an instrument used by the Success for All Foundation (SFAF) coaches to describe the extent of program implementation in Success for All (SFA) schools. The Snapshot thus measures the breadth of practices among administrators, teachers, and students from the point of view of an SFAF expert. The appendix then discusses how the SFA evaluation team has used the Snapshot to derive two quantitative measures of the extent of implementation fidelity. Finally, the appendix considers the limitations of this analysis.

The Contents of the Snapshot and Its Rating by SFAF Coaches

The Snapshot includes 99 items, 66 of which are used in the analysis of first-year program implementation. These items may be organized along a number of different dimensions:

- **Content area (school, classroom, or student practice).** Items in the rubric fall into three content areas — Schoolwide Structures, Instructional Processes, and Student Engagement — that reflect the SFA program's key instructional and whole-school improvement goals and components.

- **Reading level (kindergarten reading, lower-level primary reading, and higher-level primary reading).** Many items are rated separately for the kindergarten component (KinderCorner), lower-level primary reading classes (Reading Roots), and higher-level primary reading classes (Reading Wings).

- **Implementation priority.** Every item is accompanied by a notation of 1, 2, or 3, with 1 indicating items that are highest priority. While dubbed "priorities for implementation," the numbers also correspond roughly to the year in which a school is likely to implement that item with mastery. For example, SFAF's intention is for all Priority 1 items to be implemented during a school's first year operating the program. In contrast, many of the Priority 2 and 3 items require more time for effective implementation (for example, students' practicing self-control and other positive behaviors) and thus are expected to be attained after the first year, although some schools may implement these items earlier.

- **Whether an item is "essential" to the SFA program.** SFAF considers 16 of the 99 items — spanning the three content areas — to be essential. These items concern such practices as having a 90-minute reading block, conducting cross-grade regrouping, and structuring classroom conversation for cooperative learning and sharing.

SFAF coaches rate the Schoolwide Structures as “in place” or “not in place.” They rate the Instructional Processes and Student Engagement items, which concern classroom practice, in terms of the proportion of classroom teachers using the practice. For an item about teachers using the basic lesson structure, the school would be rated as having reached the Power level if 95 percent to 100 percent of teachers implement the practice, as reaching the Mastery level if 80 percent to 94 percent of teachers use the practice, as at the level of Significant Use if 40 percent to 79 percent of teachers demonstrate it, or as still in Learning mode if less than 40 percent of teachers use the basic lesson structure.

Creating Implementation Measures from the Snapshot

Extent of SFA program implementation is measured in two ways in this report:

1. **Number of items in place.** The first method looks at the number of items rated as in place as a proportion of the 66 items that coaches rated during the 2011-2012 school year. (As noted above, 99 items are eventually expected to be in place for rating by the 2013-2014 school year.)

2. **Numerical score.** The second method creates a numerical score for each school, considering not only the items in place but also the weights that SFAF personnel, in consultation with the research team, assigned to each item. First, the team assigned a point value to each item. Schoolwide Structures items were given a value of 1 if they were in place or 0 if they were not in place. Instructional Processes and Student Engagement items were assigned numerical values of 1, 0.8, 0.4, and 0 to indicate the proportion of the school’s classrooms with the objective in place. Second, SFAF staff and the evaluation team assigned weights to key items. Items across the content areas that SFAF considers “essential” were given a double weight. Items that pertain to implementation of the Reading Wings curriculum were also weighted twice as much as items pertaining to other reading levels. Reading Wings requires mastery over a greater number of teaching and classroom management strategies, and a school is likely to have more Reading Wings classes than Reading Roots classes across all grades.

Chapter 4 presents the results of this analysis, by content area and by implementation priority.[1]

Limitations of the Snapshot Analysis

A few notes of caution about the results of the Snapshot analysis are in order. First, because the Snapshot is intended as a tool for coaches to assess schools qualitatively but quickly, some items evaluate multiple constructs. Second, although SFAF trains coaches and provides detailed guidance on observation forms that create the inputs for the final rating for each item, generally there is one coach per school and district. Thus, any similarities among schools' ratings within a district may be confounded with the rater. Third, within this first year, it is difficult to make any inferences about change or growth, since the study lacks any rating before SFA implementation. Fourth, the Snapshot was used only in SFA schools for this study, so the construct validity of items in the instrument has not been tested in other contexts.

Finally, some items were not rated or were missing for Year 1. The Snapshot changed during the 2011-2012 school year, and some coaches used different forms for different schools. Although it appears on the form that there are more than 66 items that could have been rated in 2011-2012, due to challenges with the standardization of forms and missing data, the study team could use only 66 items across all schools. All items and ratings were standardized before and during data entry for this analysis. Appendix Table E.1 shows the Snapshot items that were used to create the numerical scores for this study.

[1] Although the Snapshot contains an additional set of items to be rated by school year 2013-2014, the relative scores and rank of each school would have remained the same, since all schools were rated only for the 66 items that were available in 2011-2012.

Appendix Table E.1
Success for All Snapshot

Schoolwide Structures

B	1	2	3	4	IP = In place; N = Not in place
Fundamentals					
					❶All leaders and staff have received essential training. (1)
					❶Materials necessary for program implementation are complete. (2)
					❶Schoolwide Solutions coordinator has been identified and given time to fulfill Solutions responsibilities. (3)
					❶Facilitator is a full-time position. (4)
					❷Classes in Reading Roots do not exceed twenty students. (5)
					❶A ninety-minute (elementary) or sixty-minute (secondary) uninterrupted reading block exists. (6)
					❶The principal is fully involved with SFA implementation. (7)
					❶Instructional component teams meet regularly to address professional-development needs and connect teachers to online and print resources for program support. (8)
					❶All Schoolwide Solutions teams have been identified and meet regularly as specified. (9)
					❶Getting Along Together structures are in place in every classroom (Class Council meetings, Peace Paths, Think It Through sheets). (10)
					❷Getting Along Together structures are in place schoolwide (Peace Paths; Think It Through sheets; using conflict stoppers in cafeteria, on playground, in hallways, etc.). (11)
					❶Attendance plans are complete and effectively implemented. At least 95% of children are in school on time every day. (12)
					❶The Intervention team meets weekly and uses the Solutions Sheet process to create individualized achievement plans. (13)
					❷Read and Respond forms are collected each week, and return is celebrated. Return rate is 80% or better. (14)
					❷Parent involvement essentials are in place. (15)
					❷Volunteer listeners are in place. (16)
▨	▨	▨	▨	▨	❸A positive schoolwide behavior plan (e.g., PBIS, Checkpoints for Success, CMCD) is in place and used consistently. Emergency schoolwide disciplinary procedures are clear and functional. School climate is positive, calm, and orderly. (17)
▨	▨	▨	▨	▨	❸A community-supported vision program is in place. (18)
Assessment					
					❶An accurate Grade Summary Form is maintained for every grading period. (19)
					❶Formal reading-level assessments with consistent measures are conducted at the beginning of the year and at the end of each grading period. (20)
					❶Teacher cycle record forms or weekly record forms are used by all teachers to record classroom data throughout the grading period. (21)
					❷A Classroom Assessment Summary is submitted quarterly by each teacher. (22)
					❸Member Center (or equivalent) data-collection and reporting tools are used consistently. (23)

B	1	2	3	4	IP = In place; N = Not in place	
Aggressive Placement						
					❶Cross-grade regrouping is used each grading period in all grades except pre-K and kindergarten. (24)	
					❶Multiple measures are used to determine placement. (25)	
					❷Placement is aggressive; students are placed at the highest level at which they can be successful. (26)	
Tutoring						
					❶Capacity exists to tutor 30% of first-grade students, 20% of second-grade students, and 10% of third-grade students. (27)	
					❶A certified teacher-tutor coaches other tutors. (28)	
					❶Tutoring is provided daily for each tutored student. (29)	
					❷Team Alphie or Alphie's Alley is used for tutoring. (30)	
Leading for Success						
▨	▨	▨	▨	▨	❸The Leadership team meets monthly to review schoolwide data, monitor Leading for Success teams, and prepare for the quarterly Success Network meetings. (31)	
					❸Members of the school Leadership team know the number and percentage of students achieving at grade level and meeting quarterly proficiency goals. (32)	
					❷Leading for Success quarterly meetings are held at the start of school and quarterly to review schoolwide progress toward achievement goals and Leading for Success team reports. (33)	
					CC/KC	❷Instructional component teams set SMARTS targets based on program data, chart progress, and work collaboratively to meet their targets. (34)
					RR	
					RW	
					❸The facilitator uses the GREATER coaching process to support continuous improvement of student achievement through high-quality implementation. (35)	
					Attendance	❷Schoolwide Solutions teams set SMARTS targets based on program data, chart progress, and work collaboratively to meet their targets. (36)
					Intervention	
					Cooperative Culture	
					Community Connections	
					Parent and Family Involvement	
					❸The Schoolwide Solutions coordinator supports Schoolwide Solutions teams to identify student-achievement targets that guide the teams' efforts. (37)	
▨	▨	▨	▨	▨	❸All Leading for Success teams set targets that are aligned with schoolwide quarterly goals. (38)	

***Please Note:** The shaded areas indicate objectives that may not be rated at your school until the 2013–2014 school year.*

Priorities for implementation: ❶ *mechanical* ❷ *routine* ❸ *refined.*

Instructional Processes*

√	B	1	2	3	4		
						CC/KC	❶Teachers use the basic lesson structure and objectives. Teachers use available media regularly and effectively. (1)
						RR	
						RW	
						CC/KC	❸Active instruction is appropriately paced and includes modeling and guided practice that is responsive to students' understanding of the objective. (2)
						RR	
						RW	
						CC/KC	❷Teachers use Think-Pair-Share, whole-group response, Random Reporter (or similar tools that require every student to prepare to respond) frequently and effectively during teacher presentation. (3)
						RR	
						RW	
						CC/KC	❸Teachers restate and elaborate student responses to promote vocabulary mastery at a high standard of oral expression. (4)
						RR	
						RW	
						CC/KC	❷Teachers provide time for partner and team talk (and lab activities in kindergarten) to allow mastery of learning objectives by all students. (5)
						RR	
						RW	
						CC/KC	❸Teachers facilitate partner and team discussion (and student interaction in labs) by circulating, questioning, redirecting, and challenging students to increase the depth of discussion and ensure individual progress. (6)
						RR	
						RW	
						RW	❷Following Team Talk or other team study discussion, teachers conduct a class discussion in which students are randomly selected to report for their teams; rubrics are used to evaluate responses, and team points are awarded. (7)
						RW	❸During class discussion, teachers effectively summarize, address misconceptions or inaccuracies, and extend thinking through thoughtful questioning. (8)
						RW	❸During class discussion, teachers ask students to share both successful and unsuccessful use of strategies, such as clarifying, questioning, predicting, summarizing, and graphic organizers. (9)
						RR	❷Teachers calculate team scores that include academic achievement points in every instructional cycle and celebrate team success in every cycle. (10)
						RW	
						RR	❸Teachers use team scores to help students set goals for improvement, and students receive points for meeting goals. (11)
						RW	

√	B	1	2	3	4		
						KC	❷Read and Respond forms are collected each week, and return is celebrated. Return rate is 80% or better. (12)
						RR	
						RW	
						GAT	❸Teachers conduct Class Council meetings weekly. The atmosphere is open, and relevant class issues are addressed effectively. (13)
▨	▨	▨	▨	▨	▨	GAT all day	❸Teachers facilitate the use of emotion-control and conflict-resolution strategies throughout the day (including use of the Stop and Stay Cool steps, Think It Through sheets, the Feelings Thermometer, and the Peace Path). (14)

Student Engagement*

√	B	1	2	3	4		
						CC/KC	❶Students are familiar with routines. (1)
						RR	
						RW	
						CC/KC	❸Students speak in full, elaborate sentences when responding to teacher questions. (2)
						RR	
						RW	
						CC/KC	❷Student talk equals or exceeds teacher talk. (Each student should be engaged in partner/team discussion as a speaker or active listener during half of class time.) (3)
						RR	
						RW	
						CC/KC	❷Students are engaged during team/partner practice and labs. If needed, strategies such as talking chips or role cards are in use. (4)
						RR	
						RW	
						CC/KC	❸Partners assist each other effectively with difficult words and use retell every day during partner reading. (5)
						RR	
						RW	
						CC/KC	❸Students use rubrics to meet expectations (e.g., fluency, writing, vocabulary, strategy use, comprehension). (6)
						RR	
						RW	
						RW	❸Teams are engaged in highly challenging discussions, in which students explain and offer evidence from the text to support their answers, or, for writing, students offer thoughtful responses during the revision process. (7)
						RR	❷Students value team scores and work daily to ensure that team members are prepared to successfully report for the team during Random Reporter and to succeed on tests. (8)
						RW	
						RW	❸Students use strategy cards to assist one another during reading and discussion, or students use revision guides to offer helpful feedback during the writing process. (9)
						RR	❸Students know their reading levels and can articulate what they need to do to increase their reading achievement, or, for writing, students know their writing strengths and what they need to do to improve their writing. (10)
						RW	
						GAT all day	❸Students use win-win decision-making skills to solve problems that arise through the use of the Peace Path, conflict stoppers, and Think It Through sheets. (11)
						GAT all day	❸Students can identify the intensity of their feelings and use self-control strategies (Stop and Stay Cool) when needed. (12)

√ = Area of focus

P = Power schoolwide – Objective is verified for 95% of teachers.
M = Mastery – Objective is verified for 80% of teachers.
S = Significant use – Objective is verified for 40% of teachers.
L = Learning – Staff members are working toward verification of this objective.

* Verified by observation or artifacts such as team score sheets, facilitator observation records, videos, audio records, transcripts of instruction, or teacher records of student responses. Leave blank if documentation is not yet available.

Appendix F

Estimation of Program Impacts

The Impact Estimation Model

The primary impact estimation model is a two-level hierarchical model in which students are nested within schools. The model uses data from all five study districts in a single analysis, treating districts as fixed effects in the model. Separate program impact estimates are obtained for each district and then are averaged across the five districts, weighting each district's estimate in proportion to the number of Success for All (SFA) program schools from each district in the sample. Findings in this report therefore represent the impact on student performance in the average SFA 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 the districts participating in the study were selected.

Specifically, the following statistical model is used for all impact estimations reported in Chapter 6 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} = 1 if school k is in district m (m = 1 to 5) and 0 otherwise

T_k = 1 if school k is assigned to receive the SFA program and 0 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_{ijk} = school-level and student-level random error, respectively, assumed to be independently and identically distributed

The error-term structure reflects the "hierarchical" or "nested" structure of the data, which has students nested within schools, 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 the SAS statistical software package.

The weighted average γ_1 (weighted by the number of treatment schools in each district/block) of the estimated γ_{1m} coefficients for the five districts is the estimated program effect on student achievement for the average treatment school in the study sample. A two-tailed t-test is used to assess whether γ_1 differs from zero. Impact results are reported in terms of both scaled scores and effect sizes.

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

Also note that the impact estimates described above provide an "intent-to-treat" analysis of the impact of the program. In other words, the estimates reflect the program impact on all students in the targeted schools, with each student's treatment status being 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

Following are the principles and rules for choosing covariates in the impact model:

- Choose covariates because they are related to outcomes. Do not choose covariates because there are big differences between program group 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 scores on the Peabody Picture Vocabulary Test (PPVT) and the Woodcock-Johnson Letter-Word Identification test are included in the impact model.

- If theory and prior empirical evidence do not provide enough guidance, use the following method to select appropriate covariates: (1) Re-randomize the sample schools to create pseudo program and control groups. (2) Run the 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 student's English language learner status, special education status, age, and gender as covariates for the impact model.

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 Hypotheses Testing

Following the What Works Clearinghouse guideline (Version 3.0, 2013), the Benjamini-Hochberg procedure[2] was applied within each of the outcome domains to adjust for the p-values, in order 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. This procedure works as follows:

- Order the p-values in ascending order
- Let m equal the number of hypotheses to be tested
- Let q equal the desired False Discovery Rate (FDR)
- Reject hypothesis H_i if $p(i) <= i/m * q$

For the report, there are two confirmatory tests, one for the program impact on the Woodcock Johnson Letter-Word Identification score and one for the impact on the Word Attack score. The following unadjusted p-values were obtained from the impact estimation model:

$$\{0.032, 0.991\}$$

If the goal is to limit the FDR to less than 0.05, one would compare the first p-value (0.032) to

$$1/2 * (0.05) = 0.025$$

Since 0.032 is less than 0.025, the first null hypothesis (for the Word Attack score) cannot be rejected at the 5 percent level. Similarly, the next p-value is compared to 2/2 * (0.05) = 0.05.

[1] There is little information on the relative advantages and disadvantages of different imputation methods for covariates in the context of randomized trials. For a detailed discussion of this issue, see Puma, Olsen, Bell, and Price (2010).

[2] Benjamini and Hockberg (1995).

Again, this null hypothesis (for the Letter-Word Identification score) cannot be rejected at the 5 percent level.

If the goal is to limit the FDR to less than 0.10, then one would compare the first p-value (0.032) to

$$1/2 * (0.10) = 0.05$$

Since 0.032 is greater than 0.05, the first null hypothesis (for the Word Attack score) can be rejected at the 1 percent level. On the other hand, the next p-value is compared to 2/2 * (0.1) = 0.1. Again, this null hypothesis (for the Letter-Word Identification score) cannot be rejected at the 1 percent level. This result indicates that the positive impact on the Word Attack score is significant at the 10 percent level after the Benjamini-Hochberg adjustment.[3]

Using 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 comparably aged 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, 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. In the interest of completeness, however, standard scores are presented in Appendix Table F.1.

[3] Benjamini and Hochberg (1995).

The Success for All Evaluation

Appendix Table F.1

Early Impact of SFA on Kindergarten Student Reading Achievement for Subgroups of the Main Analysis Sample, Using Scaled Scores (Implementation Year 2011-2012)

Subgroup and Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value	Number in Program Group	Number in Control Group
Black							
WJLWI[a]	105.42	105.21	0.20	0.02	0.886	197	165
WJWA[b]	112.04	108.91	3.14	0.25	0.169	198	165
White							
WJLWI	109.02	107.19	1.82	0.14	0.631	148	146
WJWA	116.57	111.67	4.90	0.38	0.184	148	146
Hispanic							
WJLWI	100.75	100.53	0.22	0.02	0.853	878	841
WJWA	109.36	107.40	1.97	0.15	0.079 *	879	837
Female							
WJLWI	104.29	104.45	-0.16	-0.01	0.890	661	614
WJWA	111.29	109.92	1.36	0.11	0.189	661	611
Male							
WJLWI	102.94	102.97	-0.03	0.00	0.976	658	604
WJWA	110.82	107.60	3.22	0.25	0.001 ***	660	603
Special education							
WJLWI	94.05	97.80	-3.76	-0.28	0.098 *	91	83
WJWA	102.86	103.12	-0.26	-0.02	0.883	91	81

(continued)

Appendix Table F.1 (continued)

Subgroup and Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value	Number in Program Group	Number in Control Group
Not special education							
WJLWI	104.38	104.07	0.31	0.02	0.757	1222	1128
WJWA	111.73	109.29	2.44	0.19	0.007 ***	1224	1126
English language learner							
WJLWI	91.38	91.78	-0.41	-0.03	0.827	350	215
WJWA	103.19	103.08	0.11	0.01	0.956	351	212
Not English language learner							
WJLWI	106.09	106.40	-0.31	-0.02	0.783	965	973
WJWA	112.81	110.26	2.55	0.20	0.010 ***	966	972
Poverty status							
WJLWI	102.91	103.01	-0.10	-0.01	0.922	1178	1102
WJWA	110.80	108.46	2.34	0.18	0.013 **	1180	1099
Not poverty status							
WJLWI	109.31	110.44	-1.12	-0.08	0.498	141	116
WJWA	114.00	111.01	2.99	0.23	0.055 *	141	115

SOURCES: Woodcock-Johnson Letter-Word Identification test (spring 2012) and Woodcock-Johnson Word Attack test (spring 2012).

NOTES: The norming sample for the Woodcock-Johnson tests were taken between 1996 and 1999. The research team believes that the norming sample is outdated and yields inflated scaled scores. Therefore, the data presented in this table are for documentation purposes only and should not be used to make inferences about the relative standing of the study students to national standards.

The main analysis sample includes any student who had at least one valid spring test score and who was enrolled in a study school during the fall of the same year.

The student sample size for the Woodcock-Johnson Letter-Word Identification test is 2,564 students (1,331 in the program group and 1,233 in the control group).

The student sample size for the Woodcock-Johnson Word Attack test is 2,562 students (1,333 in the program group and 1,229 in the control group).

Effect sizes were computed using the full control group's sample standard deviations for the respective measures. The control group standard deviations for each measure are as follows:

WJLWI = 13.34
WJWA = 12.74

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.

[a]Woodcock-Johnson Letter-Word Identification test.

Impact Findings for Subgroups of the Main Analysis Sample

Appendix Table F.2 presents detailed findings for each of the subgroups of the main analysis sample. It groups these students by race/ethnicity, gender, special education status, English language learner status, and poverty status.

The Success for All Evaluation

Appendix Table F.2

Early Impact of SFA on Kindergarten Student Reading Achievement for Subgroups of the Main Analysis Sample (Implementation Year 2011-2012)

Subgroup and Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value	Number in Program Group	Number in Control Group
Black							
WJLWI[a]	21.04	20.66	0.37	0.05	0.633	197	165
WJWA[b]	5.99	4.97	1.02	0.33	0.036	198	165
White							
WJLWI	21.96	20.75	1.21	0.17	0.582	148	146
WJWA	6.56	6.23	0.33	0.11	0.784	148	146
Hispanic							
WJLWI	18.63	18.33	0.30	0.04	0.659	878	841
WJWA	5.49	4.94	0.55	0.18	0.083 *	880	838
Female							
WJLWI	20.46	20.44	0.02	0.00	0.974	661	614
WJWA	5.96	5.45	0.52	0.17	0.083 *	661	612
Male							
WJLWI	19.64	19.64	0.00	0.00	0.999	658	604
WJWA	5.80	5.05	0.75	0.24	0.010 **	661	603
Special education							
WJLWI	15.84	17.21	-1.37	-0.20	0.178	91	83
WJWA	4.40	4.04	0.36	0.12	0.326	91	81

(continued)

Appendix Table F.2 (continued)

Subgroup and Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value		Number in Program Group	Number in Control Group
Not special education								
WJLWI	20.39	20.19	0.20	0.03	0.732		1222	1128
WJWA	6.01	5.38	0.62	0.20	0.020	**	1225	1127
English language learner								
WJLWI	14.16	14.04	0.12	0.02	0.894		350	215
WJWA	4.32	4.09	0.22	0.07	0.585		352	213
Not English language learner								
WJLWI	21.22	21.38	-0.16	-0.02	0.805		965	973
WJWA	6.24	5.57	0.67	0.22	0.027	**	966	972
Poverty status								
WJLWI	19.62	19.63	-0.01	0.00	0.989		1178	1102
WJWA	5.78	5.13	0.65	0.21	0.015	**	1181	1100
Not poverty status								
WJLWI	23.61	23.70	-0.09	-0.01	0.930		141	116
WJWA	7.02	6.06	0.96	0.31	0.107		141	115

SOURCES: Woodcock-Johnson Letter-Word Identification test (spring 2012) and Woodcock-Johnson Word Attack test (spring 2012).

NOTES: Due to small sample sizes, estimates could not be computed for Asian sample members or for race/ethnicity other than white, black, or Hispanic.

Effect sizes were computed using the full control group's standard deviation for the respective measures. The control group standard deviations for each measure are as follows:

WJLWI = 6.97

WJWA = 3.07

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.

[a]Woodcock-Johnson Letter-Word Identification test.

[b]Woodcock-Johnson Word Attack test.

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. Series B (Methodological),* 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 2008-4022. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

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., Gina M. Hewes, Laura T. Overman, and Shelly Brown. 2003. "Comprehensive School Reform and Achievement: A Meta-Analysis." *Review of Educational Research* 73: 125-230.

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.

Common Core State Standards. 2013. Web site: http://www.corestandards.org.

Durkin, Dolores. 1993. *Teaching Them to Read*, 6th ed. Boston: Allyn and Bacon.

Garet, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Jones, Kazuaki Uekawa, Audrey Falk, Howard Bloom, Fred Doolittle, Pei Zhu, and Laura Sztejnberg. 2008. "The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement." NCEE 2008-4030. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Hill, Carolyn J., Howard S. Bloom, Alison R. Black, and Mark W. Lipsey. 2007. "Empirical Benchmarks for Interpreting Effect Sizes in Research." Working Paper. New York: MDRC.

Jaffe, Lynn E. 2009. "Development, Interpretation, and Application of the W Score and the Relative Proficiency Index." Woodcock-Johnson III Assessment Service Bulletin No. 11. Rolling Meadows, IL: Riverside.

Leading for Success Team. 2012. *Leading for Success Facilitator's Guide*. Baltimore: Success for All Foundation.

National Institute of Child Health and Human Development (NICHD). 2000. Report of the National Reading Panel. *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction.* NIH Publication No. 00-4769. Washington, DC: U.S. Government Printing Office.

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: 127-142.

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: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.
Web site: http://ies.ed.gov/ncee/pubs/20090049/index.asp.

Rowan, Brian, Eric Camburn, and Richard Correnti. 2004. "Using Teacher Logs to Measure the Enacted Curriculum in Large-Scale Surveys: A Study of Literacy Teaching in 3rd Grade Classrooms." *Elementary School Journal* 105: 75-102.

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: University of Pennsylvania, Consortium for Policy Research in Education.

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.

What Works Clearing House. 2013. *Procedures and Standards Handbook,* Version 3.0. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

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.