THE SUCCESS FOR ALL MODEL OF SCHOOL REFORM

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

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

July 2014

mdrc
BUILDING KNOWLEDGE TO IMPROVE SOCIAL POLICY

The Success for All Model of School Reform

Interim 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
Emily Pramik

July 2014

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 Harry and Jeanette Weinberg Foundation, Inc., The Kresge Foundation, Laura and John Arnold Foundation, Sandler Foundation, and The Starr Foundation.

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

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

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

Overview

First implemented in 1987, Success for All (SFA) is a whole-school reform initiative whose goal is to help all elementary school students become competent readers. Its key elements include:

- Reading instruction marked by an emphasis on phonics and on comprehension, a highly structured curriculum, use of cooperative learning strategies, across-grade ability grouping, frequent assessments, and tutoring for students who need extra help
- Components that address students' noninstructional issues
- Strategies to secure teacher buy-in, provide teachers and leaders with initial and ongoing training, and foster shared leadership

SFA was selected to receive a five-year scale-up grant under the U.S. Department of Education's Investing in Innovation (i3) competition. This, the second of three major reports from MDRC's independent evaluation of the scale-up effort, discusses the program's implementation and impacts in 2012-2013, the second year of operations at the study sites. The impact evaluation uses a cluster random assignment design involving 37 schools serving students in kindergarten through grades 5 or 6 (K-5 or K-6) and located in five school districts; 19 schools were randomly selected to receive the SFA program, while the remaining 18 "control group" schools did not receive the intervention. The report considers the experiences of school staff members and compares the reading performance in first grade of a cohort of kindergarten students who remained either in the SFA schools or in the control group schools for two years (and therefore received the maximum program "dosage").

Key Findings

- During the second year, schools strengthened their implementation of SFA. New program practices were implemented, and the proportion of classrooms within a school where SFA-prescribed practices were in evidence increased. Teachers also reported feeling more comfortable with the program.
- By the end of the second year, 16 of the 19 program group schools were judged to meet the Success for All Foundation's standards for adequate implementation fidelity.
- SFA reading classes continued to be distinguished from reading classes in the control group schools by greater use of cooperative learning, more extensive ability grouping of students, and close adherence to the curriculum. Tutoring, although a key program element, was not implemented more widely in SFA schools than in control group schools.
- First-graders who had been enrolled in SFA schools since kindergarten significantly outperformed their counterparts who had been continuously enrolled in control group schools on two measures of phonetic and decoding skills, although not on measures of fluency and comprehension, which are higher-order reading skills. Impact findings for subgroups of students defined by various demographic characteristics are, for the most part, consistent with the main findings.

Contents

List of Exhibits

Table

Figure

Preface

At a time when school districts are facing straitened economic conditions, it is all the more important to adopt policies and programs that have been shown to work. The U.S. Department of Education's Investing in Innovation (i3) program has provided funding to support the expansion of interventions that have previously been shown to be effective. At the same time, it has required independent evaluations of such initiatives, in order to measure their impacts at scale and in new settings.

MDRC's evaluation of the Success for All (SFA) program examines anew an approach to early reading instruction that, over its nearly 30-year history, has built a strong record of boosting students' reading achievement. The program, which includes curriculum materials and professional development for teachers, emphasizes both phonics and comprehension; its structured lesson plans call for extensive use of cooperative learning methods.

Much has changed since SFA was first put into practice, both in the program itself and in the larger world of reading instruction. This makes a reevaluation of SFA all the more important. Does the SFA model remain substantially different from other reading approaches? Can it be replicated with adequate fidelity? And, most important, does it still produce positive impacts on students' reading skills?

This report, the second of three, updates an earlier report and suggests encouraging, affirmative answers to all these questions. The final report, to be issued in 2015, will further explore these issues, as well as examine SFA's cost-effectiveness and the relationship between implementation and impacts.

Gordon L. Berlin
President

Acknowledgments

This report could not have been completed without the contributions of many people. First and foremost, we want to acknowledge the principals, research liaisons, and teachers at the schools that participated in this study. Their assistance and cooperation were vital for providing the 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, Sharon Fox, and Jill Hanson responded to our many requests for information. Nancy Madden and Robert Slavin provided helpful critiques of earlier drafts.

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, Nicole Morris, 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. Tola Sean and Douglas Bruce assisted with the printing of the instructional logs, and Cammie Brown helped with analysis of the qualitative data.

Fred Doolittle ensured that the team consistently received material and moral support. 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 handled a variety of administrative and other tasks. Robert Weber edited the report sensitively and sensibly, and Stephanie Cowell and Carolyn Thomas prepared the report for publication.

The Authors

Introduction

This is the second of three reports from MDRC's evaluation of the Success for All (SFA) scale-up demonstration, funded under the U.S. Department of Education's Investing in Innovation (i3) competition. The report presents updated findings on SFA's implementation and impacts in the scale-up sites participating in the evaluation.

SFA is one of the best-known and most thoroughly studied 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 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.

Table 1 shows the program's key features.

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 two groups of schools were similar on all school-level characteristics at baseline, although they were not fully representative of all schools participating in SFA's i3 scale-up. The evaluation schools tended to be larger than other schools participating in the 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.[1]

The 19 program group schools received SFA. The 18 control group schools did not get the intervention and, instead, either continued with the same reading program that they had used

[1]Please see Quint et al. (2013) for a more complete description of the program, the evaluation, and the study sites.

The Success for All Evaluation

Table 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 many students leaving their homeroom to receive reading instruction from another teacher ("walk 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

(continued)

Table 1 (continued)

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

previously or, in the case of some schools, adopted a new one.[2] The study compares the experiences of adults and the performance of students in the two groups of schools.

The first report from the i3 evaluation examined 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. Its 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. The report's key findings are that:[3]

- While the majority of teachers agreed that SFA benefited their schools, they acknowledged struggling to implement its structured curriculum. They felt that they had received inadequate preparation for teaching in an SFA classroom; they worried about whether classes were moving too quickly for struggling students; and they found the program's data system complicated and demanding. As the year drew to a close, however, many teachers reported feeling more comfortable with the program.

- By the end of the first year, all but one of the study schools were deemed to have met the standards for adequate first-year implementation established by the Success for All Foundation (SFAF), the organization that provides materials, training, and support to schools operating the program. At the same

[2]The control group schools (like the program group schools before SFA was put in place) generally teach reading with commonly used basal programs available from leading educational publishers. In broad terms, the programs used by the control group schools are quite similar to SFA in striking a balance between decoding and comprehension skills.

[3]Quint et al. (2013).

time, there was considerable room for improving both the breadth and the depth of that implementation.

- Key factors that differentiated reading instruction in SFA and control group schools included more extensive use of cooperative learning and cross-grade ability grouping and regrouping, along with a greater emphasis on comprehension, in the SFA schools. Along other dimensions — including the length of the reading block, the principal's leadership in reading instruction, and the use of data to monitor students' reading progress — there were no statistically significant differences between the two sets of schools.

- At the end of the first year, SFA produced a positive and statistically significant impact on one of the two measures of phonetic skills for the main sample of kindergarten students. The program impact on this measure was robust across a range of demographic subgroups as well as across students with different levels of literacy skills at baseline.

This second report tracks the literacy growth of the initial group of kindergartners as they advanced through first grade, and it also measures the reading skills of students in grades 3 through 5.

Like the first report, this report uses quantitative and qualitative data from a variety of sources.[4] Through teacher and principal surveys, implementation summaries completed by SFAF staff, logs completed by teachers to describe the instruction that they provided to individual students, interviews and focus groups with school personnel conducted in the course of site visits, school district databases, and individual and group assessments of students' reading skills, it addresses three main questions:

1. To what extent were SFA's features implemented during the program's second year?

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

3. Did SFA continue to produce impacts on students' reading skills as the students progressed through first grade?

In brief, the report finds that, during the second year, schools strengthened their implementation of SFA, and teachers were more at ease with it. Reading instruction in SFA schools continued to differ from instruction in control group schools in a number of respects, although

[4]Appendix A describes these data sources and their purposes and presents response rates.

in other ways the two groups of schools were similar. Finally, first-graders who had been enrolled in SFA schools since kindergarten significantly outperformed their counterparts who had been continuously enrolled in control group schools on two measures of phonetic and decoding skills, although not on measures of higher-order reading skills. At this point, the impact findings about the students' academic trajectories are consistent with those reported in the major previous experimental study of SFA.[5]

SFA Implementation During the Second Year

- **A quantitative analysis indicates that the program group schools improved their implementation of SFA during the second year: They put in place new practices that they had not previously implemented, and they increased the proportion of classrooms within a school where SFA-prescribed practices were in evidence.**

The extent of implementation is measured quantitatively using an instrument known as the "School Achievement Snapshot" (the "Snapshot," for short) — a form created by SFAF to guide schools in a continuous improvement process. SFAF coaches work with schools and rate the extent to which each school has put in place 99 program practices.[6] Two of these practices relate to SFAF's provision of training and materials. The remaining 97 practices fall into three categories: Schoolwide Structures, Instructional Processes, and Student Engagement.[7] MDRC

[5]Borman et al. (2007).

[6]When they visit the schools, SFAF coaches meet with school personnel, visit classrooms, and examine program documents; the coaches then use this information to complete the Snapshot, once per quarter if possible but at least at the end of each school year.

[7]Examples of the 41 items measuring Schoolwide Structures practices include "A ninety-minute uninterrupted reading block exists"; An accurate Grade Summary Form is maintained for every grading period"; and "Cross-grade regrouping is used each grading period in all grades except pre-K and kindergarten." Coaches rate Schoolwide Structures items as "in place" or "not in place."

Examples of the 30 items measuring Instructional Processes practices include "Teachers 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" and "Teachers use team scores to help students set goals for improvement, and students receive points for meeting goals."

Examples of the 26 items measuring Student Engagement practices include "Student talk equals or exceeds teacher talk" and "Teams are engaged in highly challenging discussions, in which students explain and offer evidence from the text to support their answers."

Coaches rate Instructional Processes and Student Engagement items according to the proportion of teachers for whom the item has been verified, with four possible ratings: P (Power Schoolwide), representing 95 percent to 100 percent of teachers; M (Mastery), representing 80 percent to 94 percent of teachers; S (Significant Use), representing 40 percent to 79 percent of teachers); and L (Learning), representing less than 40 percent of teachers.

(continued)

worked closely with SFAF to convert the Snapshot ratings into numerical scores. Item scores are then summed to yield, for each school, an overall implementation score and a score for each of the three categories.

Because SFAF did not expect schools to implement all the program elements at an early stage, coaches rated only 63 of the 97 practices during the program's first year. These 63 practices were rated during the second year as well, so that, for these 63 practices, it is possible to assess the extent to which schools put in place new practices during the second year or, conversely, dropped practices that they had previously implemented. (While some schools may well have implemented some of the remaining 34 practices early on, there is no way to know from the available data whether these practices were initially put in place in the first or the second year of program operations.) Table 2 shows the results: Among practices rated in both years, 15 of the 19 schools (79 percent) experienced an increase in the number of practices put in place in the second year (that is, they added more practices than they dropped).

Schools also improved implementation by increasing the percentage of classrooms within the school in which program practices were in place. This was especially the case for the domain of Instructional Processes. On average, more teachers adopted such practices as using the basic lesson structure, discussing students' answers with them, and providing time for cooperative learning activities. Some two-thirds of SFA schools increased the proportion of classrooms implementing these three practices between the first and second years.

- **While in general the number of program elements in place grew over time, there was also some implementation slippage from the first year to the second, and, in both years, tutoring proved difficult to put in place.**

Table 2 also shows that three program schools (16 percent of the total) dropped more practices than they added. (One school did not experience a change one way or the other.) Schools were especially likely to drop Schoolwide Structures practices. For example, five schools did not implement the SFA data collection system, known as "Member Center," in the second year but had done so in the first year. Similarly, six schools had implemented quarterly leadership meetings in the first year but not in the second year.

The Snapshot data also indicate that schools had a particularly difficult time implementing SFA's tutoring component (known as "Team Alphie") for students who were lagging behind their peers: Six schools did not have the capacity to provide tutoring to the proportions

(Note 7, continued) It should be noted that while Schoolwide Structures and Instructional Processes are inputs into program implementation, the Student Engagement items on the Snapshot reflect the *results,* or *outputs,* of that implementation.

Table 2

Percentage of Schools That Show a Net Increase, Net Decrease, or No Change in the Number of Snapshot Items in Place at Any Level Across Items Rated in Both Years, by Content Area

	Schoolwide Structures	Instructional Processes	Student Engagement	Total
Percentage of schools with a net increase	52.63	68.42	57.89	78.95
Percentage of schools with a net decrease	42.11	5.26	15.79	15.79
Percentage of schools with no change	5.26	26.32	26.32	5.26

SOURCE: Success for All Snapshot (spring 2013).

NOTES: The number of items a school was rated on in both years differs slightly across schools. On average, in both years, schools were rated on 21 items in the Schoolwide Structures content area, 22 items in the Instructional Processes content area, and 19 items in the Student Engagement content area.

For schools that increased the number of Schoolwide Structures items put in place, the average increase was 1.8 items. Schools that decreased the number of Schoolwide Structures items in place did so by an average of 2.1 items. For schools that increased the number of Instructional Processes items put in place, the average increase was 2.9 items. Schools that decreased the number of Instructional Processes items in place did so by an average of 5 items. For schools that increased the number of Student Engagement items put in place, the average increase was 3.3 items. Schools that decreased the number of Student Engagement items in place did so by an average of 2.3 items.

of students specified by SFA guidelines (30 percent of first-graders, 20 percent of second-graders, and 10 percent of third-graders), and five schools did not provide daily tutoring in either the first or the second year of the program. Resource constraints limited the schools' ability to support this component at the scale intended.

- **Over time, SFAF expects schools to implement more sophisticated practices that require greater teacher skill; the findings indicate that the study schools were able to put these practices into place.**

The schools' improved implementation of the program is also revealed through an analysis that takes into account the complexity of the items that they put into place.

The SFAF designates each Snapshot item as Level 1, Level 2, or Level 3. The 23 Level 1 items are basic elements that are critical for the program to function successfully. In contrast,

the 74 Level 2 and Level 3 items entail more complex program elements and call for more sophisticated instructional practices.[8]

The height of the bars depicted in Figure 1 represents the total number of Level 1 items and of Levels 2 and 3 items that SFA coaches rated during each implementation year. The figure shows that almost all the Level 1 items (20 of the 23) were rated during the first year of program implementation; as basic and critical elements, the coaches expected to see them put in place early on. In contrast, the coaches began rating 33 of the 74 items designated as Levels 2 or 3 only in the second year, as schools gained greater experience with the program.

The figure also shows that the schools did more in order to meet the greater expectations that were placed on them. The shaded portion of the bars represents the average number of items at that level that schools were deemed to have put in place to at least a moderate extent.[9] Thus, on average, during the second implementation year, schools implemented 62 of the 74 Level 2 or 3 items.

- **By the end of the second year, 16 of the 19 program schools were judged to meet SFAF's standards for adequate implementation fidelity, although there was considerable variation within this group.**

As shown in Table 3, under the scoring system that MDRC established in conjunction with SFAF, schools can achieve a maximum of 142 points on the full set of 97 Snapshot items rated in 2012-2013. As noted above, the Snapshot is intended to pinpoint areas where schools need to improve implementation, and achieving a very high score on the instrument was not expected by SFAF. It is also difficult, as a hypothetical example makes clear: A strongly performing school in which all Level 1 items were in place, 75 percent of the teachers adopted all Level 2 and Level 3 Instructional Processes practices, and 75 percent of classrooms

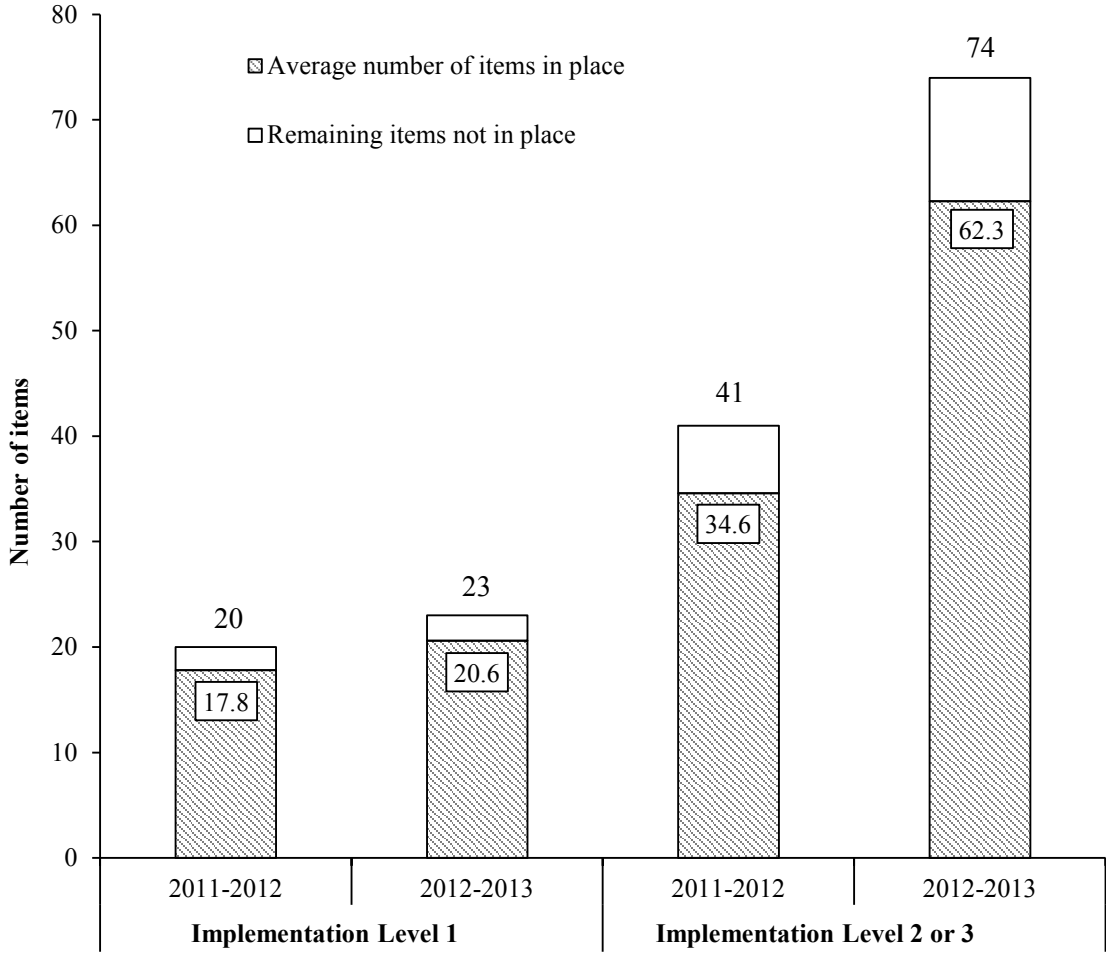
[8]Examples of Level 1 items are "Cross-grade regrouping is used each grading period in all grades except pre-K and kindergarten" and "Teachers use the basic lesson structure and objectives. Teachers use available media regularly and effectively." Examples of Level 2 items are "Parent involvement essentials are in place" and "Teachers provide time for partner and team talk to allow mastery of learning objectives by all students." Examples of Level 3 items are "A positive schoolwide behavior plan is in place and used consistently" and "During class discussion, teachers ask students to share both successful and unsuccessful use of strategies, such as clarifying, questioning, predicting, summarizing, and graphic organizers."

Practices within the Schoolwide Structures category fall under all three levels. In contrast, almost all the practices within the Instructional Processes and Student Engagement categories are classified as Level 2 or Level 3 items.

[9]The Schoolwide Structures included in Figure 1 were counted as implemented if they were rated as "in place." The Instructional Processes and Student Engagement items included in the figure were counted as implemented if they were rated as verified for 40 percent or more of the teachers, a level of adoption that SFAF describes as "significant." (Please see footnote 7.)

Figure 1

Average Number of Snapshot Items Rated and in Place, by Implementation Level and Year



(continued)

manifested all Level 2 and Level 3 Student Engagement practices would nonetheless attain only 59 percent of the maximum score.[10]

[10] As noted above, Student Engagement practices may better be seen as outputs rather than inputs. This fact also makes it difficult for schools to achieve a high score at a relatively early point in the program's implementation, especially because the Student Engagement items account for 45 of the 142 possible points.

Figure 1 (continued)

SOURCE: Success for All Snapshot (spring 2013).

NOTES: The average total number of items rated at each level, by year, is displayed above each bar. Level 1 items are Snapshot items the SFA program considers to be critical to successful functioning implementation, while Level 2 or 3 items reflect more "complex" implementation tasks. The shaded region of each bar represents the number of items that SFA coaches considered to be implemented, averaged across all 19 SFA schools. In the case of Schoolwide Structures items, "implemented" means that the structure or practice asked about in the item was present in the school, whereas, for Instructional Process and Student Engagement items, it means that the SFA coach determined that at least 40 percent of classrooms were implementing the practice. The white segment of each bar represents the items not implemented.

The total number of items rated differed slightly across the 19 schools and was rounded to the nearest whole number. Because of rounding error, the average number of Year 1 items is 61 in this table. In fact, the exact average of Year 1 rated items is 61.2.

The number of possible items on which a school was rated increased in 2012-2013. For some items that began to be rated in 2012-2013, the Snapshot did not indicate whether those items were in place in 2011-2012 as well. The figure cannot be read as displaying growth in the number of items in place, because the newly rated items in Year 2 may or may not have been in place in Year 1. The proportion of items in place does not change materially between years.

Table 3

Mean and Range of Scores Achieved in 2012-2013, by Content Area

	Schoolwide Structures	Instructional Processes	Student Engagement	Total
Maximum possible score at end of 2013	46	51	45	142
Mean	36.42	31.55	25.38	93.35
Minimum	22.00	17.00	10.40	58.20
Maximum	46.00	42.40	37.40	117.80

SOURCE: Success for All Snapshot (spring 2013).

NOTES: SFA coaches completed the Snapshot at the end of the 2012-2013 school year for all 19 program schools. All schools were rated on 41 items in the Schoolwide Structures content area, 30 items in the Instructional Processes content area, and 26 items in the Student Engagement content area in Year 2. The average total score of 93.35 represents a weighted average of the content area scores, weighted by the proportion of the total score that each content area contributes to the total. The minimum and maximum total scores were calculated separately from the component-level minimum and maximum scores.

For accountability purposes, SFAF determined that schools that achieve a total score of 50 percent or more of the maximum possible score should be deemed to have implemented the program with adequate, although not necessarily high, fidelity. As Figure 2 shows, 16 of the 19 schools met or exceeded this threshold. The overall scores for these 16 schools ranged from 51 percent to 83 percent of the maximum score. It is worth noting that the school with the lowest "passing" score of 51 percent of the maximum was able to implement 70 percent of the Level 1 items; in addition, 78 percent of the Level 2 and Level 3 items were each implemented by at least 40 percent of the school's teachers.

Schools met the threshold in different ways. One school that scored 59 percent of the maximum score, for instance, had implemented 96 percent of the Level 1 items, while 34 percent of Level 2 and Level 3 items were put into practice by 80 percent or more of the school's teachers. A second school that achieved 61 percent of the maximum score implemented only 65 percent of the Level 1 items, but 50 percent of the Level 2 and Level 3 items were evident in most classrooms.

- **Teachers in the program group schools reported feeling much more at ease with the SFA initiative in the second year than in the first year, although they continued to express some concerns about the program.**

Teachers participating in focus groups reported that they had greater mastery of the SFA curriculum and that they were better able to teach it as designed. Teachers also agreed that the technical support that they received had improved. For one thing, teachers noted that the school-based facilitators were better able to provide assistance because they were far more conversant with the curriculum. Teachers also told interviewers that they better understood — and felt more receptive toward — input provided by SFA coaches.

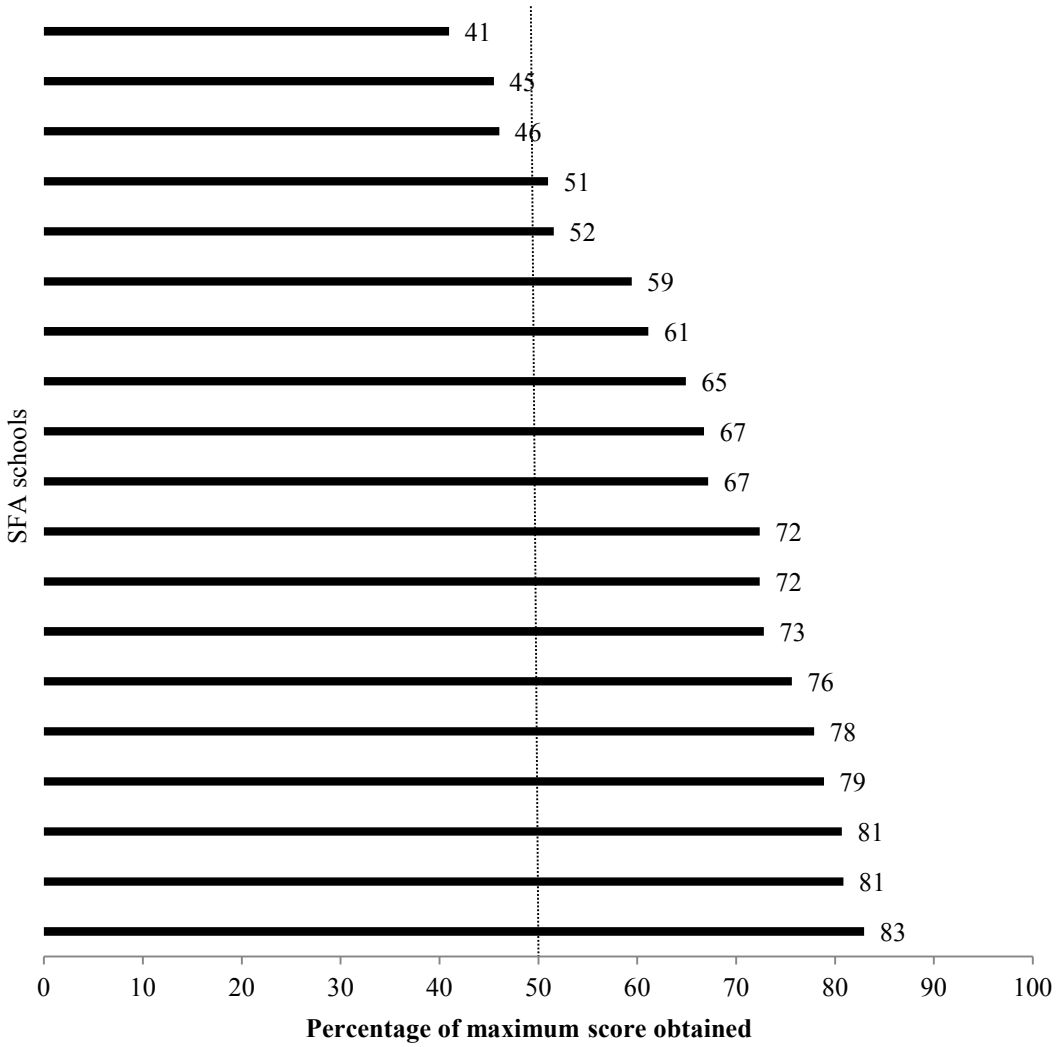
Survey data confirm the qualitative findings: 60 percent of teachers agreed that their school's SFA facilitator was "extremely knowledgeable" about the program, and 76 percent reported that the facilitator had provided them with useful feedback. And 83 percent of the SFA teachers thought that the feedback from the SFA coach was "somewhat" or "extremely" helpful.

Some 66 percent of SFA teachers reported on the survey that, overall, their school had benefited from the program, and 67 percent agreed that the reading program adequately serves most of their students. In focus groups, many teachers also commented that students who had started SFA in kindergarten seemed to be performing especially well.

While teachers reported that they were better able to keep up with SFA's pacing schedule, they continued to critique that pacing, and they questioned some aspects of the program's

Figure 2

Percentage of Maximum Possible Snapshot Score Attained in 2012-2013, by School



SOURCE: Success for All Snapshot (spring 2013).

NOTES: For program implementation to be considered adequate across all schools in Year 2, 80 percent of all schools had to achieve a score of at least 50 percent of the maximum-possible Snapshot score of 142. The mean score was 93.3. The standard deviation was 18.9.

grouping practices. On the teacher survey, only 42 percent agreed that they could get through almost all of the material that they were supposed to cover during each reading class session.[11] And only 45 percent agreed that the pacing allowed most students to learn the critical concepts being taught.[12] (In control group schools, the corresponding percentages were 53 percent and 55 percent, respectively.) In focus groups, teachers also expressed concern that SFA grouping practices do not respond well to the needs of high-functioning and struggling students alike; that grouping sometimes results in mixes of older and younger students, posing an added challenge to classroom management; that SFA's "aggressive placement" policy may move students to a level at which they lack the skills to succeed; and that students unable to advance become bored repeating the same material.[13] Finally, teachers in the SFA schools also noted that special education students did not appear to be achieving satisfactory growth in the program — a point that receives further consideration below.

- **Teachers worried about whether SFA was adequately preparing their students for state achievement tests.**

In the focus groups, teachers commented that the SFA curriculum sometimes lacked content covered by the state tests. They also remarked that when older students were grouped with younger ones for reading, they missed out on the grade-specific content included in these tests. Teacher survey data confirm that test preparation was a special concern in the SFA schools: Under half (48 percent) of the SFA teachers agreed with the statement "Your reading program helps prepare students to do well on state achievement tests," compared with 64 percent of teachers at the control group schools, a statistically significant difference.

SFA Schools and Control Group Schools Compared

- **As in the program's first year, SFA reading classes were distinguished from reading classes in the control group schools by greater use of**

[11]In control group schools, 53 percent of the teachers agreed that they could cover the requisite material during the allotted time; the difference between this proportion and the 42 percent reported in the SFA schools is statistically significant.

[12]In control group schools, the corresponding percentage was 55 percent; the difference between this proportion and the 45 percent level of agreement expressed by SFA teachers is not statistically significant.

It is worth noting that SFA's developers do not expect that all students will master key concepts on first exposure to them. Instead, the curriculum calls for these concepts to be reinforced in subsequent lessons; the emphasis, however, is on forward movement. Such rapid pacing is discomfiting to teachers who were trained to continue teaching a concept until all students have mastered it.

[13]In this respect, it is worth recalling the Snapshot data indicating that tutoring — an intrinsic part of the SFA model — was not delivered according to SFA guidelines in a substantial number of program group schools because schools lacked the resources to pay for tutors.

cooperative learning, more extensive ability grouping of students, and closer adherence to the curriculum; SFA teachers also made more intensive use of educational technology.

Table 4 presents teacher and principal survey findings about an array of reading-related instructional practices in SFA and control group schools.

Table 4

SFA-Control Group Comparisons on Survey Variables Related to Reading Instruction

Variable	Program Group	Control Group	Estimated Difference	P-Value	
Characteristics of reading instruction					
Average length (in minutes) of reading instruction period	99.1	107.2	-8.1	0.042	**
Percentage of teachers who agree or strongly agree that they use educational media or technology as part of the reading program	86.6	83.6	3.0	0.390	
Percentage of reading class time during which educational media or technology is used	46.6	30.3	16.4	0.000	***
Percentage of teachers who agree or strongly agree that the reading program involves students working together in pairs or small groups almost daily	99.1	82.5	16.6	0.000	***
Percentage of teachers who agree or strongly agree that the pacing of their school's reading program allows most students to learn critical concepts[a]	44.9	54.5	-9.6	0.112	
Percentage of teachers who agree or strongly agree that the pacing of their school's reading program allows them to get through most of the material they need to cover	41.8	53.1	-11.3	0.032	**
Percentage of principals reporting that a group or individual is responsible for helping teachers to improve their reading instruction	94.1	87.5	6.6	0.524	
Percentage of teachers who agree or strongly agree that the reading program gets students excited about reading	48.5	65.5	-17.0	0.010	**

(continued)

Table 4 (continued)

Variable	Program Group	Control Group	Estimated Difference	P-Value	
Data use					
Percentage of teachers who report that they review reading data at least once per month	78.8	85.0	-6.2	0.088	*
Percentage of principals who report that they review reading data at least once per month	82.4	87.5	-5.1	0.692	
Average score on teacher scale measuring overall data use	2.9	2.9	0.0	0.849	
Percentage of teachers who agree or strongly agree that data are used to identify teachers who need instructional improvement	58.4	37.4	21.0	0.005	***
Percentage of principals who agree or strongly agree that data are used to identify teachers who need instructional improvement[b]	94.1	80.0	14.1	0.242	
Grouping					
Percentage of teachers who report that students are ability-grouped for reading	98.9	59.4	39.5	0.000	***
Percentage of teachers who report that students are ability-grouped for reading across grade levels as a proportion of all teachers who report that students are ability-grouped	97.1	26.8	70.3	0.000	***
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	99.7	91.9	7.8	0.001	***
Tutoring					
Percentage of principals reporting that school staff members provide students with tutoring in reading	88.2	93.7	-5.5	0.596	
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	
Percentage of first-grade students receiving tutoring according to principal report	22.8	22.2	0.6	0.931	
Percentage of third-grade students receiving tutoring according to principal report	16.9	30.5	-13.5	0.069	*

(continued)

Table 4 (continued)

Variable	Program Group	Control Group	Estimated Difference	P-Value	
Prescriptive instruction					
Percentage of teachers who agree or strongly agree that they change parts of the reading program that they do not like or disagree with	44.7	89.6	-44.9	0.000	***
Percentage of teachers who agree or strongly agree that their reading program is too rigid or scripted	59.0	17.2	41.8	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	94.1	87.5	6.6	0.524	

SOURCES: Spring 2013 teacher survey and spring 2013 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 strongly agreed 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.

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.

Response rates for all teacher survey items presented were above 94 percent. The only items in the table that have lower response rates are those that were specifically calculated for subsets of teachers who responded in a certain way to another item. For example, this occurs when the percentage of teachers who group across grades is reported, conditional on their having reported grouping in the prior item.

17 out of 19 princpals at program group schools completed surveys, and 16 out of 18 principals at control group schools completed surveys.

373 teachers at program group schools completed surveys, and 257 teachers at control group schools completed surveys. All schools in the sample had completed teacher surveys.

[a]The difference on the orignal 1-to-4 scale is statistically significant. The program group mean is 2.34, and the control group mean is 2.53. The p-value of the difference is 0.029.

[b]The difference on the orignal 1-to-4 scale is statistically significant. The program group principal mean is 3.24, and the control group principal mean is 2.87. The p-value on the estimated difference is 0.064.

Cooperative learning is a hallmark of SFA, and, in response to a teacher survey question, 99 percent of the SFA teachers reported that their students worked in pairs or small groups daily, compared with 83 percent of control group teachers.[14] Instructional logs data (described

[14]Unless the lack of a statistically significant difference is specifically noted, all differences noted here are statistically significant at the 10 percent level or less.

in greater detail below) also indicate that, in reading lessons focused on comprehension, SFA second-grade teachers were twice as likely as their control group counterparts to have students discuss text with each other. (Interestingly, a number of SFA teachers participating in focus groups noted that they used cooperative learning methods to teach subjects other than reading.)

Ability-grouping practices also differed markedly between the two groups of schools. Almost all teachers at SFA schools noted on the survey that their students were grouped by reading level, compared with 59 percent of control group school teachers. When control group schools employed ability grouping, it appears that students were generally placed in different groups within the same classroom, rather than placed in different classrooms with other students at their same level. Thus, all SFA principals, but only 25 percent of the control group principals, said that their school used a "walk-to-read" model that entails students leaving their homerooms for reading instruction with other readers at a similar skill level. Finally, SFA teachers were far more likely than their control group counterparts — 96 percent and 16 percent, respectively — to report that ability grouping occurred across grades.

SFA teachers hewed far more closely to the SFA curriculum than their control group counterparts did to the curricula that they used. Thus, only 45 percent of the SFA teachers, compared with 90 percent of control group teachers, said that they changed parts of the curriculum that they did not like or with which they disagreed.

Finally, similar proportions of teachers in the two groups of schools reported employing educational media or technology in reading class. However, SFA teachers used technology during a much larger proportion of reading class time — 47 percent of the time, compared with 30 percent for control group teachers, on average.

- **No differences were found between SFA schools and control group schools with respect to other instructional practice elements that SFA developers consider to be important: an extended class period for reading instruction, use of data, and tutoring for students who are not keeping pace with their peers.**

SFA calls for a 90-minute long "reading block."[15] As Table 4 indicates, in SFA schools, the average length of the reading block was 99 minutes, but it was even longer in control group schools: 107 minutes, on average — a difference that is statistically significant.

A number of items on the teacher and principal surveys asked about the purposes for which data were used in their schools, and a scale was created from answers to the teacher

[15]The term "reading block" refers to a period in which reading instruction takes place; it excludes time spent on instruction in spelling and grammar.

survey items.[16] No significant differences were found between the average scale scores in the two groups of schools.[17] Responses to specific scale items did differ, however. For example, both teachers and principals in SFA schools were more likely to report that data were analyzed to identify teachers who needed to improve their instruction; the difference is statistically significant for teachers but not for principals.[18]

The majority of both program and control group schools considered themselves to be implementing Response to Intervention (RtI) practices. Response to Intervention is an approach to identifying and serving students who struggle with reading that has become widely used in schools across the country over the past decade. Under RtI, all students are initially screened, and those found to be lagging behind their peers receive additional assistance, either in small groups or, if they are still not making adequate progress, one-on-one. The large majority of principals in both sets of schools (88 percent of SFA principals and 94 percent of control group principals) reported that staff members at their school provided students with tutoring, a form of intervention for struggling students. Both because students in control group schools received additional academic supports and because many students in SFA schools did not receive the tutoring prescribed by the program, it is perhaps not surprising that identical proportions of first-graders — about 22 percent — received tutoring in the two sets of schools, according to principal surveys. Third-graders in control group schools were actually significantly more likely to receive tutoring than their counterparts in SFA schools; 31 percent of students, compared with 17 percent, respectively, received such assistance.[19]

- **Early reading instruction in SFA classrooms and in control group classrooms differed with respect to content, with SFA reading classes placing more emphasis on fluency and on the use of specific strategies to enhance comprehension.**

[16]The surveys asked respondents how much they agreed with such items as "Since the start of the 2012-2013 school year, your school has used data to identify students struggling with reading" and "Since the start of the 2012-2013 school year, your school has used data to develop strategies to move students from the below basic and basic categories into the proficient category on standardized tests of reading skills."

Items included in the principal and teacher surveys were identical, except that the principal survey included one additional question about the use of data for grouping students by ability level.

[17]On average, SFA teachers scored 2.89 on the data-use scale, and control group teachers scored 2.90. The average score on the data-use scale was 3.55 for SFA principals and 3.20 for control group principals.

[18]Fifty-eight percent of SFA teachers, compared with 37 percent of control group teachers, responded that data were used for this purpose. Among principals, 94 percent of SFA principals, compared with 80 percent of control group principals, agreed that they used data to identify teachers needing instructional improvement.

[19]For the most part, tutoring operated similarly at the two sets of schools. Typically, it was scheduled to take place more than once a week, and it was more likely to involve tutors working with small pullout groups of students than with individual children. However, all the control group principals, but only 71 percent of the principals in SFA schools, reported that most tutors were certified teachers.

Teachers of early reading were asked to maintain instructional logs that detailed the nature of the instruction that they provided over eight days to eight different students (one day per student).[20] These logs supply the best evidence about the contents of reading instruction.

Table 5 summarizes key findings from the logs.[21] Most notably, SFA teachers placed greater emphasis on reading fluency — the ability to read sentences, stories, and other connected text easily and quickly. The odds that an SFA teacher focused on reading fluency were 1.62 the odds that an average control group teacher did so. On the other hand, SFA teachers were less likely to focus on grammar and spelling and on writing than their control group counterparts; the last is not surprising, given that the SFA reading curriculum spends only limited time on writing.

In its landmark review of effective reading practices, the National Reading Panel concluded that the use of a combination of comprehension strategies can produce gains in comprehension as measured by standardized tests.[22] The data presented in Table 5 suggest that teachers in both SFA and control group schools used a variety of strategies to teach comprehension but that somewhat different strategies were emphasized in the two groups of schools. For example, first-grade SFA teachers were more likely than their control group counterparts to employ strategies focused on literal comprehension; that is, they more frequently asked questions whose answers were directly stated in the text, and they explained to their students how to find these answers. SFA teachers in both grades were also much more likely to ask questions eliciting brief answers that demonstrated students' understanding of text. As noted above, in the lessons in which comprehension was a focus, SFA second-grade teachers were more likely to have students discuss text with each other than were second-grade control group teachers.

Finally, when teaching students how to read words, SFA teachers were more likely to use picture and context cues, a strategy that emphasizes the meaning of these words. Control group teachers, in contrast, were more likely to examine words isolated from their contexts — having students learn words by sight, for example, and paying attention to such aspects of word structure as the presence of prefixes, suffixes, and contractions.

---

[20]In the SFA schools, teachers of Reading Roots (beginning reading) classes and of lower-level Reading Wings (more advanced reading) classes completed the logs; in control group schools, the logs were completed by first- and second-grade teachers.

[21]The numbers in the table are *odds ratios* produced by logistic regression, the preferred analytic method when an outcome is binary (that is, either occurred or did not occur). An odds ratio of 1 indicates that the average SFA teacher and the average control group teacher were equally likely to have focused on a particular topic or to have used a specific instructional method; an odds ratio greater than 1 means that the average SFA teacher was more likely to adopt this practice than the average control group teacher; and an odds ratio of less than 1 means that the average SFA teacher was less likely to adopt this practice than the average control group teacher.

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

Table 5

Instructional Differences Between SFA Schools and Control Group Schools (Implementation Year 2012-2013)

Construct	All Grades Odds Ratio	Grade 1 Odds Ratio	Grade 2 Odds Ratio
Language arts focus[a]			
Comprehension	1.50	1.01	1.71
Word Analysis	0.72	0.70	0.72
Writing	0.54 **	0.28 ***	0.86
Reading fluency	1.62 *	1.82	1.33
Vocabulary	0.85	1.00	0.72
Grammar	0.30 ***	0.38 ***	0.25 ***
Spelling	0.15 ***	0.19 ***	0.12 ***
Comprehension[b]			
Activate knowledge	0.99	0.56	1.33
Literal comprehension	1.85 **	2.16 **	1.70
Story structure	0.77	0.57	0.88
Analyze/synthesize	0.72	0.59	0.82
Brief answers	4.93 ***	5.64 ***	4.32 ***
Students discuss text	1.37	0.76	2.21 *
Teacher-directed instruction	0.81	0.57	1.11
Word analysis[c]			
Letter-sound relationships	1.28	1.06	1.43
Sight words	0.36 **	0.34 **	0.40 *
Use picture/context cues	1.36	2.32 *	0.85
Use phonics cues	1.76	2.49 **	1.10
Structural analysis	0.32 ***	0.40 *	0.33 **
Assess student ability	1.09	1.11	1.14
Teacher-directed instruction	0.81	0.69	0.77
Number of schools	36	35	34

(continued)

- **SFA teachers and control group teachers expressed similar levels of satisfaction with their reading programs — and similar doubts about the ability of these programs to meet the needs of struggling students.**

In response to a survey item, 60 percent of the teachers in SFA schools reported that they were satisfied with the overall quality of their reading program, as did 66 percent of teachers in control group schools; this difference is not statistically significant. At the same time, less than half the teachers in either group (47 percent of SFA teachers and 46 percent of control group teachers) agreed that their reading program adequately serves the most struggling students. (These data are not shown in tables.)

Table 5 (continued)

Construct	All Grades Odds Ratio	Grade 1 Odds Ratio	Grade 2 Odds Ratio
Cognitively demanding items[d]			
Activate knowledge			
Activating prior knowledge	0.81	0.47 *	1.09
Previewing, predicting, surveying text	0.96	0.54	1.43
Story structure			
Summarizing important details in text	1.17	0.92	1.35
Sequencing information or events in text	0.51 **	0.54 *	0.51 **
Using concept maps/frames	1.06	0.67	1.48
Identifying story structure	0.69	0.69	0.64
Analyze/synthesize			
Analyzing/evaluating text	0.94	0.86	0.96
Comparing/contrasting information	0.71	0.78	0.75
Number of schools	36	35	34

SOURCE: Teacher logs administered in spring 2013.

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

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.

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

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

[a]The analysis sample for language arts focus items consists of 2,183 teacher logs (1,281 from program group schools and 902 from control group schools) collected from 282 grade 1 and 2 reading teachers (168 in the program group and 114 in the control group) in 36 schools (18 program group schools and 18 control group schools). The grade 1 subset consists of 1,033 teacher logs (647 from program group schools and 386 from control group schools) collected from 161 teachers (111 in the program group and 50 in the control group) in 35 schools (18 program group schools and 17 control group schools). The grade 2 subset consists of 1,149 teacher logs (634 from program group

(continued)

Table 5 (continued)

schools and 515 from control group schools) collected from 197 teachers (132 in the program group and 65 in the control group) in 34 schools (18 program group schools and 16 control group schools).

[b]The analysis sample for comprehension constructs was restricted to include only those logs where teachers indicated comprehension as "a focus of instruction." The sample consists of 1,470 teacher logs (898 from program group schools and 572 from control group schools) collected from 267 grade 1 and 2 reading teachers (155 in the program group and 112 in the control group) in 36 schools (18 program group schools and 18 control group schools). The grade 1 subset consists of 684 teacher logs (430 from program group schools and 254 from control group schools) collected from 147 teachers (98 in the program group and 49 in the control group) in 35 schools (18 program group schools and 17 control group schools). The grade 2 subset consists of 786 teacher logs (468 from program group schools and 318 from control group schools) collected from 176 teachers (112 in the program group and 64 in the control group) in 34 schools (18 program group schools and 16 control group schools).

[c]The analysis sample for word analysis constructs was restricted to include only those logs where teachers indicated word analysis as "a focus of instruction." The sample consists of 1,053 teacher logs (589 from the program group schools and 464 from the control group schools) collected from 234 grade 1 and 2 reading teachers (129 in the program group and 105 in the control group) in 36 schools (18 program group schools and 18 control group schools). The grade 1 subset consists of 635 teacher logs (388 from the program group schools and 247 from the control group schools) collected from 137 teachers (88 in the program group and 49 in the control group) in 35 schools (18 program group schools and 17 control group schools). The grade 2 subset consists of 417 teacher logs (201 from program group schools and 216 from control group schools) collected from 140 teachers (83 in the program group and 57 in the control group) in 33 schools (17 program group schools and 16 control group schools).

[d]The analysis sample for cognitively demanding items was restricted to include only those logs where teachers indicated comprehension as "a focus of instruction." The items are subcategories of the construct "comprehension," as discussed in Rowan, Camburn, and Correnti (2004). The sample consists of 1,470 teacher logs (898 from program group schools and 572 from control group schools) collected from 267 grade 1 and 2 reading teachers (155 in the program group and 112 in the control group) in 36 schools (18 program group schools and 18 control group schools). The grade 1 subset consists of 684 teacher logs (430 from program group schools and 254 from control group schools) collected from 147 teachers (98 in the program group and 49 in the control group) in 35 schools (18 program group schools and 17 control group schools). The grade 2 subset consists of 786 teacher logs (468 from program group schools and 318 from control group schools) collected from 176 teachers (112 in the program group and 64 in the control group) in 34 schools (18 program group schools and 16 control group schools).

- **SFA and control group school principals were equally likely to report that their schools had personnel and processes addressing a variety of whole-school improvement efforts.**

The SFA model includes a number of components that seek to improve the school as a whole. These take the form of Solutions Teams — committees composed of teachers and other

school personnel who are charged with various functions: helping implement a schoolwide program emphasizing social skills development and conflict resolution, developing strategies for students with behavioral and academic issues, fostering closer relationships with students' families, engaging the support of local businesses and institutions, and finding solutions for frequent tardiness and absenteeism.

A series of survey questions asked principals whether someone — an individual or a group of people — was responsible for activities associated with each of these whole-school reforms. As Table 6 shows, no statistically significant differences between the responses of SFA and control group principals emerged. Each schoolwide function was addressed by the majority of SFA schools — and by the majority of control group schools as well.[23]

SFA's Impacts on Students' Reading Abilities

- **First-graders who had entered SFA schools as kindergartners — that is, students who had been in SFA classrooms for two years — scored higher on two measures of phonetic skills than their counterparts in the control group.**

The main sample of interest for the impact analysis consists of the cohort of students who enrolled in an SFA or control group school as kindergartners in the fall of 2011 and remained enrolled in a school of the same type in the spring of 2013.[24] In essence, these children had received all their formal reading instruction either through SFA or through a different reading program. Students were individually assessed using well-established instruments that measure several different reading skills: phonics and decoding abilities, fluency, and comprehension.

Table 7 shows the estimated program impacts on these outcome measures, along with the effect size and p-value of each impact estimate.[25] SFA produced a positive and statistically

[23]Only limited information is available about what the individuals or groups of people actually did over the course of the year in pursuit of their objectives.

[24]For further details about the analysis samples described here and about the impact analysis more generally, please see Quint et al. (2013).

[25]The *effect size* indicates the magnitude of the estimated effect; it is calculated as the difference between the average score for program and control group students, divided by the standard deviation of the outcome measure for the control group. The *p-value* indicates the likelihood of obtaining an impact as large as the estimated impact if, in fact, there were no true impact and the difference that was measured occurred simply by chance. If a result is considered statistically significant at the 5 percent level (that is, the p-value of the estimate is less than or equal to 0.05), there would be no more than a 5 percent chance of obtaining the impact if there were no true effect. Because results that are not statistically significant may have occurred by chance, they do not provide strong evidence about the program's impact.

The Success for All Evaluation

Table 6

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

Variable Related to Schoolwide Structures	Program Group	Control Group	Estimated Difference	P-Value
Percentage of principals reporting that a group or individual is responsible for a schoolwide program emphasizing social skills development and conflict resolution	88.2	75.0	13.2	0.340
Percentage of principals reporting that a group or individual is responsible for developing school-wide solutions for students with behavioral challenges	88.2	87.5	0.7	0.950
Percentage of principals reporting that a group or individual is responsible for developing schoolwide solutions for students with learning challenges.	88.2	93.7	-5.5	0.596
Percentage of principals reporting that a group or individual is responsible for fostering relationships with students' families	88.2	100.0	-11.8	0.167
Percentage of principals reporting that a group or individual is responsible for building relationships with local businesses and institutions to increase community involvement	64.7	75.0	-10.3	0.535
Percentage of principals reporting that a group or individual is responsible for improving attendance	94.1	87.5	6.6	0.524
Number of schools: 37	19	18		

SOURCE: Spring 2013 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 strongly agreed 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.

Table 7

Early Impact of SFA on First-Grade Student Reading Achievement for the Main Analysis Sample (Implementation Year 2012-2013)

Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value	
Woodcock-Johnson Letter-Word Identification	31.09	30.27	0.82	0.09	0.084	*
Woodcock-Johnson Word Attack	12.78	10.69	2.09	0.35	0.000	***
Test of Word Reading Efficiency	30.66	29.72	0.95	0.06	0.305	
Woodcock-Johnson Passage Comprehension	15.08	14.92	0.16	0.03	0.565	
Number of schools: 37	19	18				

SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2013), Woodcock-Johnson Word Attack test (Spring 2013), Woodcock-Johnson Passage Comprehension test (Spring 2013), Test of Word Reading Efficiency (Spring 2013), and 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 baseline year.

The student sample size for the Woodcock-Johnson Letter-Word Identification test is 2,243 students (1,182 in the program group and 1,061 in the control group).

The student sample size for the Woodcock-Johnson Word Attack test is 2,251 students (1,184 in the program group and 1,067 in the control group).

The student sample size for the Test of Word Reading Efficiency is 2,147 students (1,129 in the program group and 1,018 in the control group).

The student sample size for the Woodcock-Johnson Passage Comprehension test is 2,248 students (1,185 in the program group and 1,063 in the control group).

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

The impact analyses for 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. Rounding may cause slight discrepancies in calculating sums and differences.

Effect sizes were computed using the full control group's standard deviations for the respective measures. The control group standard deviations are as follows: 8.84 for the Woodcock-Johnson Letter Word Identification Test, 6.05 for the Woodcock-Johnson Word Attack test, 16.00 for the Test of Word Reading Efficiency, and 5.36 for the Woodcock-Johnson Passage Comprehension test.

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

significant impact on two measures of phonic and decoding skills: the Woodcock-Johnson Word Attack test and the Woodcock-Johnson Letter-Word Identification test.[26] On the Word Attack measure, the p-value of the program-control group difference is less than 0.000, and the effect size is 0.35. On average, a first-grade student's reading achievement test score grows by about 0.97 standard deviation in effect size over the course of the school year.[27] Therefore, a 0.35 effect size is equivalent to about 35 percent of the annual reading gain experienced by first-grade students. The impact on the Letter-Word Identification measure is smaller (the effect size is 0.09) but is still significant at the 10 percent level.[28]

- **No significant differences were found between SFA and control group students on measures of two more advanced reading skills: fluency and comprehension.**

Phonics and decoding are core skills that first-graders must master in order to develop into proficient readers. Two measures of more advanced skills were also administered to students in the research sample. The Test of Word Reading Efficiency (TOWRE) assesses fluency: It measures the number of words on a vertical list that students can read accurately and easily within a limited time period. The Woodcock-Johnson Passage Comprehension test asks students to read a short passage and supply a missing word that makes sense in the context of the passage. On neither of these measures did the scores of students in SFA schools differ significantly from those of students in the control group schools.[29]

- **Impact findings for subgroups defined by various demographic characteristics are, for the most part, consistent with the main findings.**

Table 8 shows the effects that were registered for different subgroups of students within the main sample on the various outcome measures. (In the table, "+" indicates that the difference favors the SFA group and is statistically significant; "0" means that there is no statistically significant difference one way or the other; and "-" means that the difference favors the control group and is statistically significant.) Positive and significant impacts on the Woodcock-Johnson Word Attack test were observed for the majority of subgroups examined, including

[26]While the two tests tap essentially similar skills, the majority of items on the Letter-Word Identification test require a student to read real words of increasing difficulty, while the majority of items on the Word Attack test require students to read nonsense words of increasing difficulty.

[27]See Bloom, Hill, Black, and Lipsey (2008), Table 8.

[28]This result remains statistically significant at the 10 percent level after making the Benjamani-Hochberg adjustment to account for testing two outcome measures within the same domain of phonics skills.

[29]It may be difficult to measure comprehension at such an early stage in children's reading development. It may also be that while SFA reading instruction stresses fluency and comprehension, the measures of these skills are not particularly well aligned with that instruction.

The Success for All Evaluation

Table 8

Direction of Early Impacts of SFA on First-Grade Student Reading Achievement for Subgroups of the Main Analysis Sample (Implementation Year 2012-2013)

Subgroup	Woodcock-Johnson Letter-Word Identification	Woodcock-Johnson Word Attack	Test of Word Reading Efficiency	Woodcock-Johnson Passage Comprehension
Black	0	+	0	0
White	0	0	0	+
Hispanic	+	+	0	0
Female	+	+	0	0
Male	0	+	0	0
Special education	-	0	-	-
English language learner	0	0	0	0
Not English language learner	0	+	0	0
Not poverty	0	0	0	0

SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2013), Woodcock-Johnson Word Attack test (Spring 2013), Woodcock-Johnson Passage Comprehension test (Spring 2013), Test of Word Reading Efficiency (Spring 2013), and student records data collected from the five districts in the study sample.

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

Program and control group sample sizes for each of the above subgroups, as well as more detailed information about subgroup effects, can be found in Appendix Table B.1. Due to small sample sizes, estimates could not be computed for race/ethnicity groups other than white, black, and Hispanic.

Because students in the "Poverty" and "Not special education" subgroups make up most of the overall student sample, results for these groups are similar to results for the main sample and are not included in this table.

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

boys, girls, and students who are black and Hispanic; they were also found for female, white, and Hispanic students on the Woodcock-Johnson Letter-Word Identification test. The lack of impacts on the fluency and comprehension measures is reflected in most subgroup findings as well.[30]

- **There is reason to question whether SFA is as effective with special education students as it is with other students.**

Special education students in the SFA schools fared significantly worse on three of the four measures than their control group counterparts. Just what should be made of this finding is unclear, since the impact estimates for this subgroup are based on a very small number of students — 56 or 57 control group students and between 56 and 58 SFA students across the five school districts for most of the tests, and only 39 control group and 38 program group students for one of the tests — and the estimates vary depending on just who is included in the sample.[31] But the fact that, under alternative sample specifications, special education students in SFA schools scored significantly lower than such students in control group schools on the two assessments that measure higher-level reading skills (fluency and comprehension) may be a reason for concern. Success for All has not developed a separate program for special education students. Instead, it seeks as much as possible to serve such students within the regular classroom, supplemented by small-group and individual tutoring and by the Solutions Teams. Because of resource constraints, however, many schools did not offer the prescribed tutoring, and special education students may have been especially disadvantaged by this implementation shortfall.

- **Positive effects on the Word Attack measure were found for SFA students in a second "full" sample made up of first-grade students who were present in the study schools at the end of 2012-2013 school year, regardless of how long they had been there.**

Student turnover is a fact of life, especially in schools serving low-income students, so whether SFA has positive impacts on students who do not attend program schools from kindergarten on is a policy-relevant question. A separate analysis indicates that, for this secondary sample, SFA registered a positive and statistically significant impact on the Word Attack measure but not on the Letter-Word Identification test. Results for this sample appear in Appendix C.

[30]Appendix B presents more detailed statistical information about the subgroup effects.

[31]To be included in the analysis of subgroup effects, a district had to have 20 students in that subgroup across the two groups of schools. One district had only 12 special education students and was excluded for that reason. If those 12 students were included, SFA's impact on the Word Attack measure for special education students would be positive and statistically significant.

- **Students in third, fourth, and fifth grades constitute the "auxiliary sample" for the impact evaluation; no statistically significant effects were found for this sample, either on tests of vocabulary and reading comprehension administered specifically for the study or on state reading tests used to measure school performance and establish accountability.**

Older students in SFA schools did not first learn to read "the SFA way," but the evaluation sought to determine whether they had benefited from close to two years of SFA instruction. The auxiliary sample comprises third-, fourth-, and fifth-graders in the SFA and control group schools, who were in grades 1 through 3, respectively, when the evaluation began. These students were administered grade-specific Gates-MacGinitie Reading tests in vocabulary and reading comprehension. Students in these grades also take state tests that are used to measure reading performance and establish school accountability. On none of these measures did SFA students fare either better or worse than their control group counterparts. Results for the auxiliary sample appear in Appendix D.

Reflections

This report brings much good news about Success for All's implementation and impacts during the second year of the program. While most program schools met SFAF's standards for adequate implementation at the end of the first year, there was much room for improvement. During the second year, schools *did* improve their implementation of the program, both putting in place new elements and increasing the proportion of classrooms adopting practices that previously had been implemented less systematically. Moreover, teachers expressed greater ease and confidence in their ability to deliver the curriculum.

The fact that previous evaluations had demonstrated the effectiveness of Success for All in improving students' reading was central to the program's selection as one of only four recipients of scale-up grants under the U.S. Department of Education's initial i3 competition. Continued evaluation of SFA is nonetheless important both because the SFA program has evolved over time and because many other school reading programs have also changed their practices since the earlier studies of SFA were conducted. Thus, the current version of SFA places greater emphasis on the use of technology in the classroom and on the deployment of school district personnel trained by SFAF to provide professional development to schools along with the SFA coaches; for their part, control group schools have strengthened the teaching of phonics. And, as noted above, large numbers of elementary schools have incorporated Response to Intervention techniques for struggling readers.

Given these changes, it is striking that the impact results reported here closely replicate those found in the most important prior evaluation of SFA, conducted by Borman et al.[32] During the first year of program implementation, both that study and this one find positive and statistically significant effects for kindergartners on the Woodcock-Johnson Word Attack measure of phonetic skills. During the second year of implementation, both studies find positive and statistically significant effects for first-graders on another measure of phonetic skills, as well as a persistent effect on the Word Attack measure.[33]

While these results are promising, what ultimately matters for reading is comprehension. At the two-year point, neither the Borman evaluation nor this one shows a positive, statistically significant impact on the Woodcock-Johnson Passage Comprehension measure used in both analyses. By the third year, however, Borman and his colleagues did find such an impact. Thus, the key question to be addressed in the final report from the i3 evaluation — to be produced in 2015 — is whether, in the third year of scale-up, SFA will prove more effective than other reading programs in promoting students' understanding of what they read.

One finding may cast a shadow over this otherwise bright picture. Special education students in SFA schools scored significantly lower than their control group counterparts on three of the four impact measures used in the study. While the subgroup of special education students is small, the consistency of the results suggests that this is a population to which SFA's developers may want to pay more attention.

In addition to discussing implementation and impacts during the third year, the next, and final, report from this evaluation will consider the program's scale-up process.

---

[32]Borman et al. (2007).

[33]It is interesting to try to compare the SFA impacts with those reported by other studies of elementary reading programs with a rigorous evidentiary base — Reading Partners, the Experience Corps, and Reading Recovery. There are important differences, however, in the nature of the interventions, the students tested, and the measures used. Thus, while SFA is a classroom-focused intervention for all students, the other three programs all involve upward of 90 minutes a week of one-to-one tutoring for struggling students. The effects of Reading Partners were measured for students in grades 2 through 5; those for the Experience Corps were measured for students in grades 1 through 3; and those for Reading Recovery and SFA were measured for first-graders.

These differences make comparisons difficult. For example, while Reading Partners, unlike SFA, reported small but statistically significant effects on fluency and comprehension, these effects were measured for students in grades 2 through 5; in SFA, the main sample of interest consists of first-graders. Examining effect sizes without regard to the domains analyzed, the effect sizes for the Word Attack and Letter-Word Identification measures reported above for SFA (0.10 and 0.35, respectively) are larger than or similar in magnitude to those reported by Reading Partners and the Experience Corps (with effect sizes ranging from 0.00 to 0.10, depending on the outcome) but are smaller than the effect size on the Iowa Test of Basic Skills measure reported by Reading Recovery (0.68).

Appendix A

Data Sources and Response Rates

The Success for All Evaluation

Appendix Table A.1

Data Sources and Response Rates, by Program or Control Group Status (Implementation Year 2012-2013)

Instrument and Purpose	Program Group: Number Targeted	Program Group: Number of Respondents	Program Group: Response Rate (%)	Control Group: Number Targeted	Control Group: Number of Respondents	Control Group: Response Rate (%)	P-Value of Response Rate Difference[a]
Principal survey Survey administered to all principals at both program and control group schools. Program group surveys also included questions about SFA. The survey provides information about the school's reading program, professional development, and school practices and supports. Additionally, it describes the launch and implementation of SFA in program group schools.	19	17	89.5	18	16	88.9	0.956
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.	410	373	91.0	310	257	82.9	0.054 *

(continued)

Appendix Table A.1 (continued)

Instrument and Purpose	Program Group			Control Group			P-Value of Response Rate Difference[a]
	Number Targeted	Number of Respondents	Response Rate (%)	Number Targeted	Number of Respondents	Response Rate (%)	
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	32	86.5	–	–	–	–
Facilitator interviews: Interviews with the SFA facilitator at program group schools to learn about his or her duties and the SFA implementation story.	19	16	84.2	–	–	–	–
Teacher focus groups: Focus group with teachers at program group schools to learn about implementation of SFA in the classrooms.	19	18	94.7	–	–	–	–
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.1 (continued)

Instrument and Purpose	Program Group			Control Group			P-Value of Response Rate Difference[a]
	Number Targeted	Number of Respondents	Response Rate (%)	Number Targeted	Number of Respondents	Response Rate (%)	
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.	1,492	1,281	85.9	1,110	902	81.3	0.609
Baseline tests							
Woodcock-Johnson Letter-Word Identification test was administered to all sample students in fall 2011. Spanish versions of the tests were administered to students without English mastery. Test scores serve as an outcome variable in the impact estimation model.	1,630	1,089	66.8	1,461	995	68.1	0.692
Peabody Picture Vocabulary test was administered to all sample students in fall 2011. Spanish versions of the tests were administered to students without English mastery. Test scores serve as an outcome variable in the impact estimation model.	1,630	1,081	66.3	1,461	997	68.2	0.560

(continued)

Appendix Table A.1 (continued)

Instrument and Purpose	Program Group: Number Targeted	Program Group: Number of Respondents	Program Group: Response Rate (%)	Control Group: Number Targeted	Control Group: Number of Respondents	Control Group: Response Rate (%)	P-Value of Response Rate Difference[a]
Follow-up tests							
Woodcock-Johnson Letter-Word Identification test was administered to all sample students in spring 2013. Spanish versions of the tests were administered to students without English mastery. Test scores serve as an outcome variable in the impact estimation model.	1,630	1,565	96.0	1,461	1,387	94.9	0.793
Woodcock-Johnson Word Attack test was administered to all sample students in spring 2013. Spanish versions of the tests were administered to students without English mastery. Test scores serve as an outcome variable in the impact estimation model.	1,630	1,565	96.0	1,461	1,397	95.6	0.744
Woodcock-Johnson Passage Comprehension test was administered to all sample students in spring 2013. Spanish versions of the tests were administered to students without English mastery. Test scores serve as an outcome variable in the impact estimation model.	1,630	1,563	95.9	1,461	1,394	95.4	0.806
Test of Word Reading Efficiency was administered to all sample students in spring 2013. Test scores serve as an outcome variable in the impact estimation model.	1,630	1,477	90.6	1,461	1,325	90.7	0.573

(continued)

Appendix Table A.1 (continued)

Instrument and Purpose	Program Group			Control Group			P-Value of Response Rate Difference[a]
	Number Targeted	Number of Respondents	Response Rate (%)	Number Targeted	Number of Respondents	Response Rate (%)	
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.	1,630	1,586	97.3	1,461	1,422	97.3	0.991

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]37of 37 schools returned surveys from at least some of their reading teachers.

[c]Log response rates were calculated based on the number of logs distributed to a given teacher, which was typically eight logs. The statistical test was computed at the level of logs, and it tests whether the experimental status of the school to which a teacher belonged affected the probability that the teacher would return a completed log.

Appendix B

Subgroup Impacts

The Success for All Evaluation

Appendix Table B.1

Early Impact of SFA on First-Grade Student Reading Achievement for Subgroups of the Main Analysis Sample (Implementation Year 2012-2013)

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]	31.82	30.65	1.17	0.13	0.404		177	135
WJWA[b]	13.03	9.21	3.82	0.63	0.002	***	178	136
TOWRE[c]	30.99	28.63	2.37	0.15	0.410		174	134
WJPC[d]	15.21	15.22	-0.01	0.00	0.996		177	133
White								
WJLWI	34.32	32.24	2.09	0.24	0.293		148	137
WJWA	15.05	13.61	1.45	0.24	0.568		149	136
TOWRE	38.37	34.55	3.81	0.24	0.274		138	116
WJPC	18.99	16.99	2.00	0.37	0.063	*	139	114
Hispanic								
WJLWI	29.05	28.07	0.98	0.11	0.087	*	790	746
WJWA	11.85	9.56	2.29	0.38	0.000	***	790	752
TOWRE	27.30	26.19	1.11	0.07	0.324		743	706
WJPC	13.87	13.60	0.27	0.05	0.446		794	751
Female								
WJLWI	31.63	30.47	1.16	0.13	0.067	*	607	522
WJWA	13.11	10.63	2.48	0.41	0.000	***	609	528
TOWRE	31.49	29.97	1.52	0.09	0.206		585	512
WJPC	15.49	15.07	0.42	0.08	0.272		608	524
Male								
WJLWI	30.48	29.95	0.52	0.06	0.267		573	538
WJWA	12.40	10.41	1.99	0.33	0.000	***	573	538
TOWRE	29.74	29.69	0.06	0.00	0.955		542	506
WJPC	14.62	14.77	-0.16	-0.03	0.637		575	538
Special education								
WJLWI	25.79	27.94	-2.15	-0.24	0.086	*	70	65
WJWA	9.48	8.08	1.39	0.23	0.146		68	66
TOWRE	24.08	28.85	-4.76	-0.30	0.068	*	57	60
WJPC	11.18	13.53	-2.36	-0.44	0.003	***	69	65

(continued)

Appendix Table B.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	31.41	30.39	1.02	0.12	0.038 **	1105	988
WJWA	12.97	10.81	2.17	0.36	0.000 ***	1109	993
TOWRE	31.05	29.57	1.49	0.09	0.131	1065	954
WJPC	15.33	14.95	0.38	0.07	0.222	1109	990
English language learner							
WJLWI	24.59	23.67	0.93	0.10	0.471	327	194
WJWA	8.93	7.91	1.01	0.17	0.197	326	197
TOWRE	25.12	22.90	2.22	0.14	0.360	286	177
WJPC	11.99	11.32	0.68	0.13	0.318	328	198
Not English language learner							
WJLWI	32.27	31.66	0.61	0.07	0.235	850	837
WJWA	13.54	11.42	2.12	0.35	0.000 ***	853	840
TOWRE	31.34	30.85	0.49	0.03	0.610	838	815
WJPC	15.56	15.69	-0.13	-0.02	0.701	852	835
Poverty status							
WJLWI	30.69	29.96	0.73	0.08	0.149	1055	959
WJWA	12.64	10.68	1.96	0.32	0.000 ***	1056	964
TOWRE	29.97	29.13	0.84	0.05	0.402	1003	917
WJPC	14.84	14.76	0.08	0.01	0.797	1057	961
Not poverty status							
WJLWI	33.52	34.24	-0.72	-0.08	0.634	125	101
WJWA	13.62	12.43	1.19	0.20	0.325	126	102
TOWRE	34.50	35.81	-1.31	-0.08	0.669	124	101
WJPC	16.98	16.91	0.07	0.01	0.933	126	101

(continued)

Appendix Table B.1 (continued)

SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2013), Woodcock-Johnson Word Attack test (Spring 2013), Woodcock-Johnson Passage Comprehension test (Spring 2013), Test of Word Reading Efficiency (Spring 2013), and student records data collected from the five districts in the study sample.

NOTES: 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. Rounding may cause slight discrepancies in calculating sums and differences.

Due to small sample sizes, estimates could not be computed for race/ethnicity groups other than white, black, and Hispanic.

Effect sizes were computed using the full control group's standard deviations for the respective measures. The control group standard deviations are as follows: 8.84 for the Woodcock-Johnson Letter Word Identification Test, 6.05 for the Woodcock-Johnson Word Attack test, 16.00 for the Test of Word Reading Efficiency, and 5.36 for the Woodcock-Johnson Passage Comprehension test.

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

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

[b]Woodcock-Johnson Word Attack test.

[c]Test of Word Reading Efficiency.

[d]Woodcock-Johnson Passage Comprehension test.

Appendix C

Full-Sample Impacts

The Success for All Evaluation

Appendix Table C.1

Early Impact of SFA on First-Grade Student Reading Achievement for the Full Student Sample (Implementation Year 2012-2013)

Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value	
Woodcock-Johnson Letter-Word Identification	30.34	29.80	0.54	0.06	0.255	
Woodcock-Johnson Word Attack	12.36	10.51	1.85	0.31	0.000	***
Test of Word Reading Efficiency	29.50	28.73	0.76	0.05	0.415	
Woodcock-Johnson Passage Comprehension	14.69	14.57	0.11	0.02	0.690	

SOURCES: Woodcock-Johnson Letter-Word Identification test (Spring 2013), Woodcock-Johnson Word Attack test (Spring 2013), the Test of Word Reading Efficiency (Spring 2013), and the Woodcock-Johnson Passage Comprehension test (Spring 2013).

NOTES: The "full student sample" is defined as the sample of students who had at least one valid score on the Spring 2013 Woodcock-Johnson exams. The sample for both outcomes consists of students from 37 schools (19 program group schools and 18 control group schools).

The student sample size for the Woodcock-Johnson Letter-Word Identification test is 2,952 students (1,569 in the program group and 1,383 in the control group).

The student sample size for the Woodcock-Johnson Word Attack test is 2,962 students (1,571 in the program group and 1,391 in the control group).

The student sample size for the Test of Word Reading Efficiency is 2,802 students (1,482 in the program group and 1,320 in the control group).

The student sample size for the Woodcock-Johnson Passage Comprehension test is 2,957 students (1,569 in the program group and 1,388 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: 8.84 for the Woodcock-Johnson Letter-Word Identification test, 6.05 for the Woodcock-Johnson Word Attack test, 16.00 for the Test of Word Reading Efficiency, and 5.36 for the Woodcock-Johnson Passage Comprehension test

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

Rounding may cause slight discrepancies in calculating sums and differences.

Appendix D

Auxiliary-Sample Impacts

The Success for All Evaluation

Appendix Table D.1

Gates-MacGinitie and State Test Achievement for the Auxiliary Analysis Sample (Implementation Year 2012-13)

Outcome	Program Group	Control Group	Estimated Impact	Estimated Impact Effect Size	P-Value
Grade 3					
Gates-MacGinitie Comprehension Scale Score	447.89	450.02	-2.13	-0.05	0.38
Gates-MacGinitie Comprehension Percentile Rank	27	30	–	–	–
Gates-MacGinitie Vocabulary Scale Score	447.37	449.85	-2.48	-0.06	0.34
Gates-MacGinitie Vocabulary Percentile Rank	29	32	–	–	–
State Reading Test Z-Score[a]	-0.10	-0.10		-0.09	0.22
Grade 4					
Gates-MacGinitie Comprehension Scale Score	472.80	472.09	0.72	0.02	0.66
Gates-MacGinitie Comprehension Percentile Rank	31	28	–	–	–
Gates-MacGinitie Vocabulary Scale Score	467.90	468.21	-0.31	-0.01	0.88
Gates-MacGinitie Vocabulary Percentile Rank	29	29	–	–	–
State Reading Test Z-Score	0.02	0.02		0.02	0.77
Grade 5					
Gates-MacGinitie Comprehension Scale Score	487.77	488.08	-0.31	-0.01	0.88
Gates-MacGinitie Comprehension Percentile Rank	30	30	–	–	–
Gates-MacGinitie Vocabulary Scale Score	485.63	486.09	-0.45	-0.01	0.87
Gates-MacGinitie Vocabulary Percentile Rank	28	28	–	–	–
State Reading Test Z-Score	0.01	0.01		0.02	0.76
Number of schools: 37	19	18			

(continued)

Appendix Table D.1 (continued)

SOURCES: Gates-MacGinitie Reading Comprehension and Vocabulary subtests (Spring 2013) and student state testing records collected from the five districts in the study sample.

NOTES: The "auxiliary analysis sample" is defined as the set of students who were present in grades 3, 4, or 5 in the sample schools in the 2012-2013 school year and who have state testing scores or vocabulary or reading comprehension subtest scores from the Gates-MacGinitie Reading test.

The sample of third-grade students consists of 2,959 students (1,498 in the program group and 1,461 in the control group). The sample of fourth-grade students consists of 2,993 students (1,585 in the program group and 1,408 in the control group). The sample of fifth-grade students consists of 2,807 students (1,461 in the program group and 1,346 in the control group).

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

Effect sizes were computed using the full control group's standard deviations for the respective measures by grade level. For the Gates-MacGinitie reading comprehension subtest, the control group standard deviations are 39.97 for grade 3 students, 36.57 for grade 4 students, and 37.87 for grade 5 students. For the Gates-MacGinitie vocabulary subtest, the control group standard deviations are 42.57 for grade 3 students, 37.02 for grade 4 students, and 37.36 for grade 5 students.

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

Rounding may cause slight discrepancies in calculating sums and differences.

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

References

Bloom, Howard S., Carolyn Hill, Alison Rebeck Black, and Mark W. Lipsey. 2008. "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions." *Journal of Research on Educational Effectiveness* 1, 4: 289-328.

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.

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.

Quint, Janet, Rekha Balu, Micah DeLaurentis, Shelley Rappaport, Thomas J. Smith, and Pei Zhu with Emma Alterman, Herbert Collado, and Emily Pramik. 2013. *The Success for All Model of School Reform: Early Findings from the Investing in Innovation (i3) Scale-Up*. New York: MDRC.

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.

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.