Empirical Education

RESEARCH REPORT

Comparative Effectiveness of *Scott Foresman Science:*

A Report of Randomized Experiments in Five School Districts

Gloria I. Miller
Empirical Education Inc.
Stanford University

Andrew Jaciw
Boya Ma
Empirical Education Inc.

March 28, 2007

Empirical Education Inc.
www.empiricaleducation.com
425 Sherman Avenue, Suite 210
Palo Alto, CA 94306
(650) 328-1734

Acknowledgements

We are grateful to the people in the following school districts for their assistance and cooperation in this research and for providing access to their data under an agreement between Empirical Education Inc. and the respective district: Federal Way Public Schools, Ogden City School District, Reynoldsburg City Schools, St. Petersburg Catholic Schools, and Visalia Unified School District. The research was sponsored by Pearson Education through a contract with Empirical Education.

This report was presented as a paper discussion at the Annual American Education Research Association conference in March 2008 (Division H-School Evaluation and Program Development / Section 1: Applied Research in the Schools).

About Empirical Education Inc.

Empirical Education Inc. was founded to help school districts, publishers, and the R&D community assess new or proposed instructional and professional development programs through scientifically based pilot implementations. The company draws on the expertise of world-class researchers and methodologists assuring that the research is objective and takes advantage of current best practice in rigorous experimental design and statistical analysis. The company's findings let educators quantify the value of programs and help them partner with providers to implement those most effective for their students.

Comparative Effectiveness of *Scott Foresman Science*:

A Report of Randomized Experiments in Five School Districts

Executive Summary

We investigated whether *Scott Foresman Science* is more effective than current science programs in five diverse sites. Although we found no evidence that it improved science achievement beyond the regular programs, boys and girls performed equally well, whereas the control group boys outperformed girls. Our results also show that under some conditions the program can enhance reading achievement.

Introduction. Pearson Education contracted with Empirical Education Inc. to conduct randomized experiments to determine the effectiveness of its *Scott Foresman Science* products *(SFScience)* compared to the elementary science programs already in place in five geographically and demographically diverse sites. We compared science and in reading outcomes for classes using the *SFScience* curricular materials and control classes using each district's current materials.

Scott Foresman Science, a year-long curriculum intended for daily use, provides a sequence of structured and supportive inquiry activities and text materials to develop students' independent investigative skills. Science kits contain materials for hands-on activities, while Leveled Readers help the teacher differentiate instruction and provide reading support at, below, and above grade level. During the half-day training, teachers learned how the materials were to be used and how much was to be covered. Control teachers typically used state, district, and teacher developed materials, magazines, videos, online resources, and older science texts for science instruction.

Findings for Science. Overall, we found that students in the *SFScience* classrooms improved in science achievement at the same rate as the students in the established program. The following graph shows the comparison combining the results from all five districts. The set of bars on the left indicate the pre and post results for the control and *SFScience* groups. The bar graph on the right shows the results for control and *SFScience* as predicted by our statistical model that took pretest and other factors into account. The overlapping confidence intervals at the top of the bars indicate there is no statistical difference between the two groups.

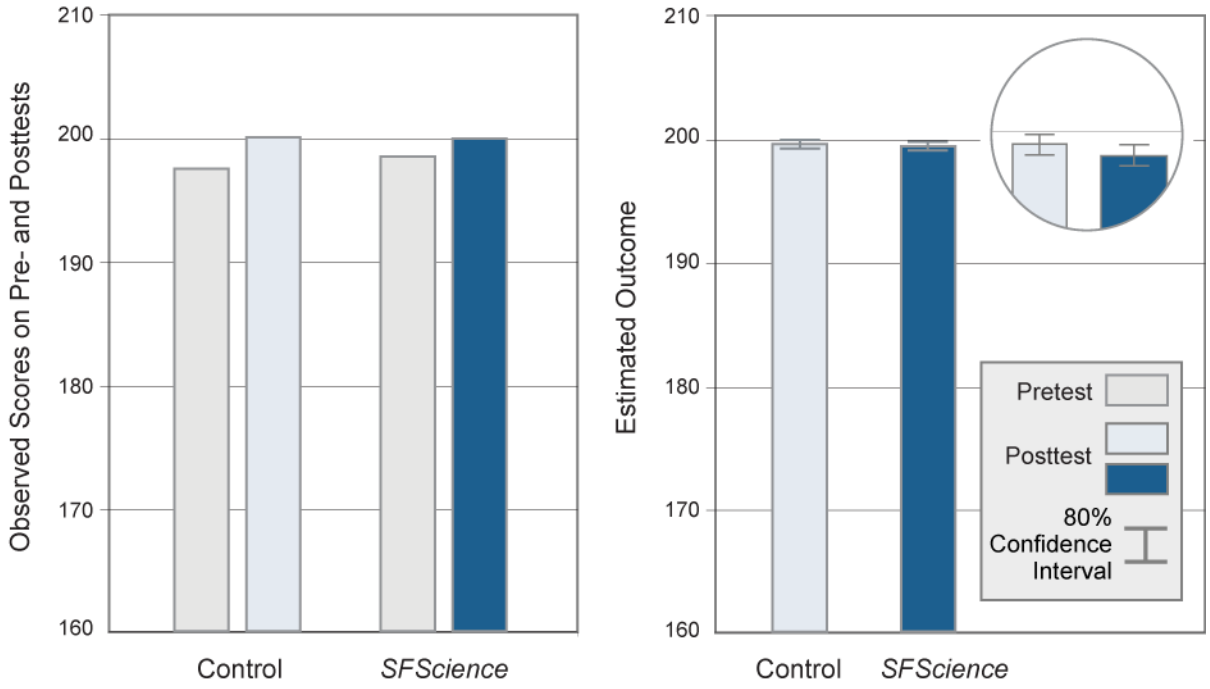

Figure 1. Impact on Science Achievement: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)

One interesting finding is that boys in the control group outperformed girls in science, whereas in *SFScience* boys and girls performed equally well.

Using data from observations, interviews, and surveys, we monitored the overall level of implementation at each site and we considered classroom process measures such as the amount of instructional time teachers devoted to science and the extent of inquiry teaching. These variables appeared not to impact science achievement. Nor did we find differences across grade or prior achievement levels or teacher experience.

Findings for Reading. Because *SFScience* provides a significant reading component, we also determined the amount of reading improvement that can be accounted for by the science program. Figure 2 compares the overall results for reading across the five sites and in combination. The combined results are positive, and two sites show positive results within the 80% confidence interval. The point-and-whiskers shows our estimates (the center points) within an interval representing 80% confidence; that is, if we consider each site separately, we can be 80% sure that the true value of the impact lies within the interval. In two sites, *SFScience* caused a small increase beyond expected gains for the schools' reading program by itself. When all sites are combined, however, this positive difference is insufficient to give us confidence that the difference was not due to chance.

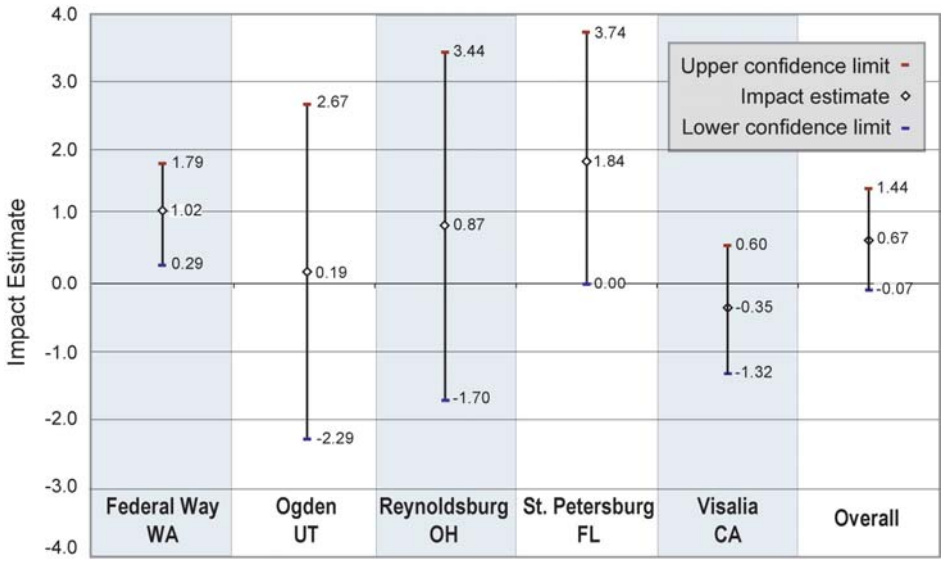

Figure 2. Estimated Reading Impacts Across Districts

Overall, it appeared that sites were more successful in teaching reading than science, reflecting relative emphasis on the two subjects. It is also relevant that this was the first year of use of *SFScience* and the teachers' initial unfamiliarity may have affected implementation, which differed at each individual site.

Our conclusion is that *SFScience* stands up to other science programs in schools. Educators may find the program attractive in the equal help it gave to boys and girls compared to other programs in place. The reading component's capacity for improving reading achievement under some conditions points to a potentially important strength of the program.

Design and Analysis. This study was a multi-site group randomized trial in which volunteer teachers within each district were assigned by coin toss to use the new program or continue with their current program for approximately one school year. Statistical analyses were based on 92 teachers/classes (46 *SFScience* and 46 control) and 2,638 students in grades 3–5. The primary outcomes, as well as pretest measures, are student-level test scores on the Northwest Evaluation Association (NWEA) test in two areas: Science Concepts and Processes and Reading. The mean impact is estimated using multi-level models. The impacts were estimated using multi-level models run in SAS PROC MIXED.

Table of Contents

Introduction

Pearson Education contracted with Empirical Education Inc. to conduct five randomized experiments to determine the effectiveness of its *Scott Foresman Science* program compared to the elementary science programs already in place in those districts. The five districts were geographically and demographically diverse:

- Federal Way Public Schools (WA)
- Ogden City School District (UT)
- Reynoldsburg City Schools (OH)
- St. Petersburg Catholic Schools (FL)
- Visalia Unified School District (CA)

Each experiment is reported separately, with detail on the implementation in that location and the specific results for those teachers and students. This report pools the data from all the sites into a single analysis that allows us to look at commonalities and differences as well as to address questions that can only be answered from a larger sample. These overall analyses were possible because the same pre- and posttests of science and reading were given in each location.

The question being addressed by the research is whether *Scott Foresman Science* is as effective as or more effective than the curriculum being used at each site. Since *Scott Foresman Science* provides a significant reading component, we also determined the amount of reading improvement that can be accounted for by the science program. The outcomes were measured by student achievement on standardized tests administered at the beginning and end of the study. Two test areas were selected as the outcome measures: Science Concepts and Processes and Reading Achievement. The research focuses on third-, fourth-, and fifth-grade students.

The overall comparison between *Scott Foresman Science* (*SFScience*) and the programs used in the control classrooms was the first step in our investigation. We also wanted to understand how the product was implemented and other ways that science instruction differed between the two groups. In addition, we sought to understand how characteristics of the students and of the teachers may have moderated the impact, that is, whether *SFScience* was more effective with students or teachers with differing abilities or experience. Finally, we explored the extent to which the groups differed in amount of time devoted to science or specifically to inquiry and whether those differences might help to explain the results. The reports for the individual districts in some cases addressed student characteristics important in that location; these are not reported here, but are left to the individual site reports.

The design of our experiment reflects the requirements of the No Child Left Behind Act, which directs schools to consult reports of rigorous research in making adoption decisions about instructional programs. A randomized experiment such as we have conducted provides a rigorous test of the program because it removes sources of bias. In particular, we reduce selection bias by tossing a coin to assign teachers to use *SFScience* or to continue using their current teaching materials and methods.

Random assignment to experimental conditions does not assure that we can generalize the results beyond the districts where they were conducted. We designed our study to provide useful information to support local decisions that take into account the specifics of district characteristics and their implementation of the program. The results should not be considered to apply to school districts with practices and populations different from those in this experiment. The individual reports provide a rich description of the conditions of implementation in order to assist the district in strengthening its program and to provide the reader with an understanding of the context for our findings.

Methods

Research Design

Our study is a comparison of outcomes for classes taught using the *Scott Foresman Science* curricular materials (*SFScience* group) and classes taught with the current science materials used in the district (control group). Teachers volunteered for participation, and we randomly assigned approximately equal numbers to the *SFScience* and control groups. The outcome measures are student-level test scores in science and in reading. In a group randomized trial such as this, analyses of covariance are used to increase the precision of estimates. Covariates at the class and student levels are also used to test for interactions with the experimental conditions.

Intervention

The intervention consists of core science curricular materials and one half-day training for the teachers. At each of the five sites, materials were deployed and training was provided as summarized in the following table.

Table 1. Research Milestones for the Five Experiments

Research Milestone	Federal Way, WA	Ogden, UT	Reynoldsburg, OH	St. Petersburg, FL	Visalia, CA
Randomization meeting	14 June 05	25 May 05	22 Sep 05	28 June 05	26 May 05
Product training	15 Sep 05	15 Sep 05	26 Oct 05	14 Sep 05	10 Sep 05
Intervention begins	26 Sep 05*	19 Sep 05[a]	04 Dec 05	19 Sep 05[a]	12 Sep 05
Pretesting completed[b]	22 Nov 05	18 Nov 05	29 Nov 05	15 Dec 05	29 Oct 05
First survey	07 Dec 05	07 Dec 05	07 Dec 05	07 Dec 05	07 Dec 05
Observations & interviews	21 Feb 06	19 April 06	20 Mar 06	28 Mar 05	13 Mar 06
Post-testing complete	19 May 06	19 May 06	19 May 06	19 May 06	19 May 06
Debrief meeting	22 June 06	30 May 06	30 May 06	19 May 06	08 June 06

[a] These are implementation dates for third and fourth grade only. Fifth-grade materials were on backorder. Refer to individual site report for more details concerning intervention start dates.

[b] These are the dates when the greater majority (>70%) of the testing was completed.

Pearson Education's *Scott Foresman Science* is a year-long science curriculum intended to be used as daily instruction. Based on inquiry-rich content with a sequence of structured and supportive inquiry activities, the science curriculum provides materials for both students and teachers in print, video, and online. This method of developing scientific knowledge is called scaffolded inquiry and is aimed at developing the independent investigative skills of the students through hands-on activities and through the use of text materials. Science kits containing materials for hands-on activities designed to minimize set-up time for the teachers and to maximize the students' time on exploration and data gathering provide the substance of the inquiry-driven investigations. A main feature of the curriculum is the

Leveled Reader. These are student readers designed to provide the teacher with an easy way to differentiate instruction and provide reading support at, below, and above grade level.

Training

Each site was provided one-half day of training with the materials (see Table 1 for individual site dates of the training) by a Scott Foresman representative. Details of the individual sessions are provided in the individual site reports.

All *SFScience* group teachers agreed to carry out four tasks for the study:

- Complete two units of instruction with at least one Full Inquiry module (student-designed investigation)
    - The text is segmented into four units: A-Life Science, B-Earth Science, C-Physical Science, and D-Space & Technology
- Complete one unit assessment (assessments developed to accompany the text)
- Use the Leveled Readers as needed to support differing reading abilities
- Use the Science Kit materials for hands-on inquiry

Scott Foresman Science Materials

SFScience teachers were supplied with the following materials specific to their grade level and, when possible, with state-aligned teacher and student texts (Ohio only).

Table 2. Scott Foresman Supplied Materials

Teacher Materials (one each unless otherwise specified)	Student Materials (one for every treatment student in the study)
Teacher Edition	Student Edition
Activity Flip Chart	Activity Book
Vocabulary Cards (set)	Workbook
Teacher's Edition Package	Science Kits (one for each of the four units, sufficient supplies for a class of 32, eight groups of four)
Teacher's Resource Package	Leveled Readers Super Kit: includes six copies of each of 12 Below-Level, On-Level, and Advanced Leveled Readers).
Assessment Book	
Ever Student Learns (Guide to Differentiated Instruction)	
Teacher Guides: Activity Book, Workbook, Leveled Readers, Activities for each of four units	
ExamView Test Generator and Activity (both on DVD)	
Graphic Organizer and Test Talk Transparencies	
Content Transparencies	
Audio Text CD-ROM (audio of textbook materials)	
Teacher Online Access Pack	

District Science Materials

Each district had a variety of materials for science instruction. For the most part students did not have individual science textbooks and teachers used state, district, and teacher developed materials, magazines, videos, online materials, and older science texts to carry the weight of instruction. When students did have textbooks, they were from the following publishers: Harcourt Brace, Foss, and older versions of Scott Foresman Science (2002, 2003). Only the Florida site had classroom sets of the textbooks, the other four sites had insufficient number of textbooks or no textbooks for their students. At one site they used Activities Integrating Mathematics and Science (AIMS) together with United Streaming Videos for science instruction.

Site Descriptions

Four of the five sites (WA, UT, OH, and CA) are public school districts. The Florida site is composed of six private schools. More detailed information regarding demographics of the area and the school district is provided in the individual site reports. Table 3 summarizes the number of participants at each site.

Table 3. Participants Overall

	Schools	Randomized Teachers	Classes	Students
WA	3	21	21	530
UT	4	19	24	547
OH	3	11	17	359
FL	6	21	31	762
CA	2	20	20	616
Totals	**19**	**92**	**113**	**2814**

Federal Way Public Schools, WA

The city of Federal Way is located 25 miles south of Seattle and eight miles north of Tacoma. It is the sixth largest city in Washington State with a population of nearly 86,000 people in a 22-mile-square area. The majority of students in this district is identified as White at 54.7% of the total student population of 22,594.The next largest ethnic group attending Federal Way Public Schools is Asian/Pacific Islander, which constitutes 16.7%. The English Language Learner population in this district is small at 10% of the total (the state average is 7.1%). The economically disadvantaged population is relatively high as compared to the rest of the state, 39.6% and 28.5% respectively. The students in this district achieve higher reading and math proficiencies as compared to state averages and have made steady gains in both reading and math state assessments since 2002.

Federal Way Public Schools serve a larger area than the city of Federal Way, including two other cities located in King County. The district operates 23 elementary schools; three participated in this study.

The unit of randomization at this site is the teacher. Eleven matched pairs were formed and a coin was tossed to determine assignment either to the treatment group (*SFScience)* or control (classes that would continue using current district identified materials). After randomization one teacher left the district for personal reasons unrelated to the study and was noted for study purposes as "inactive" in September 2005. In total 21 teachers with 21 classrooms and 530 students participated in the study.

Table 4. Participating Teachers at WA Site

Teacher assignment status	Number participating
SFScience	11
Control	10
Inactive	1
Totals	**22**

Ogden City School District, UT

The city of Ogden is located approximately 35 miles north of Salt Lake City and is Utah's sixth largest city, encompassing 27 square miles. It has an estimated population of 77,000 according to the 2000 census. The majority of the students in this district are identified as White at 53.9% of the total student population of 12,963.The next largest ethnic group attending Ogden City Schools is Hispanic, which constitutes 39.7%. The English Language Learner population in this district is

comparatively large at 24% of the total (the state average is 10%). The economically disadvantaged population is twice the state average, 64.9% and 32.1% respectively.

Ogden City School District considers itself an inner-city district enriched by multi-cultural diversity. The district operates a total of 15 elementary schools; four Title I schools participated in this study.

The unit of randomization at this site is the teacher. Ten matched pairs were formed and a coin was tossed to determine assignment either to *SFScience* or control groups. Well into the study (March 2006), one teacher reported that she had not taught any science and because of scheduling conflicts was unlikely to teach any for the rest of the year. This teacher was marked "inactive" for the rest of the study and her students' tests scores were not used in the analysis.

In some of these schools, science is considered a "specialty" subject. Teachers can specialize in science instruction and teach other students not assigned to their self-contained classroom. In these cases, all students under the teacher's science instruction are considered part of the study; the teachers of record for these classes are marked "teachers of registration." In total 19 teachers with 24 classrooms and 547 students participated in the study.

Table 5. Participating Teachers at UT Site

Teacher assignment status	Number participating
SFScience	10
Control	10
Teachers of registration[a]	7
Inactive	1
Totals	**28**

[a]Teachers of registration are those teachers whose students are participating in the study, but are not the actual teacher of instruction. *SFScience* or control teachers are the teachers of instruction for these classes.

Reynoldsburg City Schools, OH

The city of Reynoldsburg is located 12 miles east of Columbus and is generally considered part of the larger metropolitan area. With a population of 32,000 in 11 square miles, it is a small residential community. The majority of students in this district are identified as White at 63.9% of the total student population of 6,064.The next largest ethnic group attending Reynoldsburg City Schools is African American, which constitutes 25.4%. The English Language Learner population in this district is very small, 2.2% of the total. The economically disadvantaged population is smaller than the state average, 25.2% and 34.1% respectively.

Reynoldsburg City Schools operate six elementary schools, three middle schools, and one high school; two elementary (K-4) schools and one middle school (grades 5 and 6) participated. The unit of randomization at this site is the teacher. Five matched pairs and one singleton were formed and a coin was tossed to determine assignment to *SFScience* or control. All teachers who were randomized continued for the duration of the study.

As with Ogden City Schools, in some Reynoldsburg schools, science is considered a "specialty" subject, with teachers specializing in science instruction instructing students not assigned to their self-contained classroom. All students under the teacher's science instruction are considered part of the study: the teachers of record for these classes are marked "teachers of registration." In total 11 teachers with 17 classrooms and 359 students participated in the study.

Table 6. Participating Teachers at OH Site

Teacher assignment status	Number participating
SFScience	5
Control	6
Teachers of registration[a]	6
Inactive	0
Totals	**17**

[a]Teachers of registration are those teachers whose students are participating in the study, but are not the actual teacher of instruction. *SFScience* or control teachers are the teachers of instruction for these classes.

St. Petersburg Catholic Schools, FL

The city of St. Petersburg is located on a peninsula between Tampa Bay and the Gulf of Mexico and is Florida's fourth largest city, encompassing 59 square miles. It has an estimated population of 248,000 according to the 2000 census.

The St. Petersburg Catholic Schools serve the five counties surrounding the St. Petersburg area: Pinellas, Hillsborough, Pasco, Hernando, and Citrus. They operate a total of 31 elementary schools, several middle and high schools; six elementary schools participated in this study.

The unit of randomization at this site was the grade level. Nine matched grade-level pairs were formed among the participating schools and a coin was tossed to determine assignment to *SFScience* or control groups. Thus the entire grade level at any one school participates as either *SFScience* or control. One teacher left the area for personal reasons unrelated to the study. There is no public information regarding the overall demographics for this site, but in general the majority population is white, with few students considered economically disadvantaged.

Here, as for Ogden and Reynoldsburg, science is considered a "specialty" subject where teachers specializing in science instruction instruct students not assigned to their self-contained classroom. All students under the teacher's science instruction are considered part of the study: the teachers of record for these classes are marked "teachers of registration." In total 21 teachers with 31 classrooms and 762 students participated in the study.

Table 7. Participating Teachers at FL Site

Teacher assignment status	Number participating
SFScience	11
Control	12
Teachers of registration[a]	10
Inactive	1
Totals	**34**

[a]Teachers of registration are those teachers whose students are participating in the study, but are not the actual teacher of instruction. *SFScience* or control teachers are the teachers of instruction for these classes.

Visalia Unified School District, CA

The city of Visalia is located in the San Joaquin Valley situated almost equidistant between San Francisco and Los Angeles. The city proper encompasses 29 square miles with a population of just over 100,000 according to a 2003 population estimate. The majority of the students in this district are identified as Hispanic at 53.2% of the total student population of 25,794.The next largest ethnic group attending Visalia Unified is white which constitute 36.5%. The English Language Learner population in this district is slightly lower at 20.3% than the overall state average of 25.2%. The economically disadvantaged population is slightly higher than the average for the state, 51.7% and 49.7% respectively. The students in this district on average achieve five points lower in reading and math proficiencies as compared to state averages but have made steady gains in both reading and math state assessments since 2003.

The Visalia Unified School District comprises 22 elementary schools, 4 middle schools, 5 high schools, four charter schools, and one adult school; two of the newest schools (both opened in 2004) participated. One school was designated as Title I.

The unit of randomization at this site is the teacher. The initial randomization was conducted with 15 teachers, seven matched pairs and a singleton. A coin was tossed to determine assignment to *SFScience* or control groups. This site had some deviation from a true random assignment. Six new teachers were introduced at one school at the start of the school year (the original randomization was conducted in May 2005). Two of the original pairs were broken and the teachers matched with two new teachers and re-randomized. At least two teachers were reassigned and one teacher was allowed to be in the treatment group based on perceived need to address a large ESL population. One teacher left the district early in the school year and was marked “inactive”. In total 20 teachers with 20 classrooms and 616 students participated in the study.

Table 8. Participating Teachers at CA Site

Teacher assignment status	Number participating
SFScience	10
Control	10
Inactive	1
Totals	**21**

Sample and Randomization

Recruiting

The Pearson Education Scott Foresman group hired a consulting agency to identify districts as potential research sites. Districts were asked to complete a questionnaire with contact information. After Scott Foresman received the contact information, they forwarded it to us. We met with district staff members and principals to explain the details and procedures of the study. Principals identified eligible teachers, who were then invited to after-school meetings. The initial meetings for the research experiment in these district occurred on the dates noted in Table 1.

Randomization

Ninety-two teacher volunteers were assigned using a coin toss to either the *SFScience* condition or to control. Randomization ensures that, on average, characteristics other than the intervention that affect the outcome are evenly distributed between treatment and control groups. This prevents us from confusing the intervention’s effects with some other factors, technically called “confounders”, that are not evenly distributed between groups and that affect the outcome. For example, through

randomization we try to achieve balance between treatment and control conditions on the average years of teaching experience – a factor that presumably affects the outcome.

There are various ways to randomize teachers to conditions. We used a matched-pairs design whereby we first identified pairs of similar teachers (or in the case of the Florida site, matched grade level pairs) and then, within each pair, we randomized one teacher (in Florida, grade level pairs) to treatment and the other to control. Similarity was based on whether teachers were in the same grade level and whether they shared common meeting times. A pairing strategy will often result in a more precise measurement of an intervention's impact.

Sample Size

Sample size is one of the factors that determine how precisely we can measure the magnitude of an effect. With smaller samples we are usually able only to detect larger effects. We usually measure the size of an effect in terms of standard deviation units, which tells us how big the effect is, controlling for the spread in observed scores.

Our research design assumed that we would report the results for the five districts independently as well as with the five districts combined. With the combined data we estimated that 92 teachers would be a sufficient sample to detect an effect size as small as 0.21. An effect size is derived at by dividing the effect by a measure of how dispersed the data points are (called the standard deviation). An effect size of .2 is two tenths of a standard deviation.

The determination of the minimum detectable effect size involves making educated assumptions about design parameters. We assumed that we would be working with a fairly substantial correlation between the pre- and posttests (.64). We also had to be concerned with how much of the variability in student outcomes was due to average differences at the teacher level. This intra class correlation (ICC) is important in designs that involve more than one level. In this case, randomization was done mostly with teachers (St. Petersburg site was randomized at the grade level), but the outcome measures came from the students. Intuitively, the ICC is the proportion of the variability in student scores that can be accounted for by differences in teacher-level averages of the student scores. When the ICC is very large, for instance, much of the variation in student scores is accounted for by differences among teachers in their students' scores. If the differences among teachers are large and/or the differences within classes are small, then the sample size that matters the most for the experiment is the number of teachers. If the differences among teachers are small so that most of the variation is attributable to differences within classes, then the sample size that matters most is the number of students. In general we need larger samples to detect smaller effects, and the ICC allows us to calculate how small an effect we can detect given available numbers of students and teachers. In this experiment we assumed a fairly conservative intra class correlation of .22.

In our calculation of a .21minimum detectable effect size we also assumed conventional levels of tolerance for false-positive and false-negative outcomes, setting them at .05 and .20, respectively.

We also believed that the pairing of teachers prior to randomization would give our experiment additional power to detect effects in the 0.2 range.

Our experiment spanned five districts, and can be regarded as a multi-site trial. However, because randomization was done at each site, we can consider results at each location separately. In other words, a separate experiment was performed at each site. The sample size at each location was smaller than the combined sample size. The minimum detectable effect size at each location therefore will be larger than .21. However, we believe that with the use of a matched-pairs design and with the willingness to set tolerance for false-positive outcomes above conventional levels, we are able to draw valid inferences about the impact of the intervention within each site.

We did not design our experiment specifically to detect results for subgroups of teachers at the same 0.2 level. *We caution that in the case where we are looking at results within-sites for subgroups of teachers, the minimum detectable effect size may be quite large, and failure to find an effect may be the result of not having adequate statistical power.* With a small number of

teachers there can easily be chance imbalances of teacher characteristics that affect the outcome and thereby compromise our conclusions.

Examination of subgroups of students such as males and females is possible because each teacher will have some members of each subgroup among their students. The power of an experiment where the intervention and randomization is conducted at the teacher level is largely dependent on the number of teachers rather than on the total number of students.

Data Sources and Collection

District Supplied Information

This data requested from the school district included the student records for the students who were taught by participating teachers as well as other background data including demographic information relevant to NCLB categories of disaggregation. Specifically, the districts were asked to provide the following data:

- Student name or unique ID
- Gender
- Free/reduced lunch status (socio-economic level)
- Ethnicity
- Home language
- ELD status
- Disability status (special education)
- Age
- Classroom teacher
- School they attend

All student and teacher data having any individually identifying characteristics were removed and were stored using security procedures consistent with the provisions of the Family Educational Rights and Privacy Act (FERPA)

Class Rosters and Student Demographics

Typically, besides reporting gender and English language learner population, we would report on the other NCLB demographics of interest: socio-economic level, and special education. Here, because several districts do not report these statistics for individual students, we did not provide these for any of the sites.

Achievement Measures

The primary outcome measures are student-level scores on the Northwest Evaluation Association (NWEA) test in two areas: Science Concepts and Processes and Reading Achievement. We refer to these tests as Science achievement and Reading achievement when referring to these specific assessments throughout the report. In the fall of 2005, the NWEA Science and Reading tests were administered to the students at the various schools as a pretest measure. As a posttest measure, the Science and Reading tests were administered in the spring of 2006. The paper-and-pencil versions of these tests are referred to as ALT tests and all sites were provided these materials. Both of these tests are adaptive and comprehensive, and are designed to measure growth over time. The sets of tests consist of multiple levels, with overlapping degrees of difficulty. Several different levels are given within the same classroom. To ensure a good match of student to test, there are five test levels for Science and eight test levels for Reading. The first time a student is tested, the appropriate test level is determined by use of a placement test, referred to as a locator test. The locator test is a 20 item test whose sole purpose is to identify which of the leveled test a

student is best aligned with the student's anticipated achievement level. Once the level is determined, the student is then provided with that leveled test which is then officially scored by the NWEA. It is this score that is used in the subsequent analyses. During the second and subsequent administrations of the ALT, the student is automatically assigned to a level based on previous results.

These tests are scored on a Rasch unIT (RIT) scale, a measurement scale developed to simplify the interpretation of test scores. This scale is used to measure student achievement and student growth on an equal-interval scale so that a change of one unit indicates the same change in growth, regardless of the actual numerical values. RIT scores typically range from about 150 to 300 and indicate a student's current achievement level along a curriculum scale for a particular subject. Since this is a continuous scale, third grade student scores are usually found lower on scale whereas fifth grade scores are found higher along the scale. The Science Concepts and Processes ALT was specifically selected because we wanted to ensure that differences in state content standards would not be an issue when comparing results across the different grades and across districts. By using a test that emphasizes the concepts and processes of science over specific content, we minimize the impact of the differences in content coverage.

Pretesting

The pretests were given between October and December, 2005. Only one site was familiar with the administration of these assessments. As noted, both tests are adaptive and comprehensive assessments that measure growth over time. The set of tests consist of multiple levels, with overlapping levels of difficulty. Several different levels are given within the same classroom. To ensure a good match of student to test, there are five test levels for science and eight levels for reading. The first time a student is tested, the appropriate level is determined by use of placement tests, referred to as locator tests. During the second and subsequent administrations, the scoring program automatically assigns the level to the student based on previous results.

Most teachers observed some of the third-grade students struggling with the assessment because of the bubble-in answer sheets. This was the first time using this method of recording their answers.

End of Year Posttesting

Posttesting took place in May 2005. April, May, and June are typically high assessment administration months in many school districts. Four districts in the study reported that state assessments were administered during this time period. Teachers noted that some students suffered from testing fatigue. The Florida site did not administer state assessments.

In addition to quantitative data, we also collected qualitative data. Qualitative data were collected over the entire period of the experiment, beginning with the randomization meeting and ending with the academic calendar of the district in June 2006. Training observations, classroom observations, informal and formal interviews, multiple teacher surveys, email exchanges, and phone conversations are used to provide both descriptive and quantitative evidence of the implementation. Refer to Table 1 for the timing of these observations and interviews at each site.

Observational and Interview Data

In general, observational data is used to inform further the description of the learning environment, instructional strategies employed by the teachers, and student engagement. These data are minimally coded.

Interview data is used to elaborate survey responses, characterize the teacher's schedule, and to provide descriptions of the overall experience teaching with the *SFScience* curriculum.

Survey Data

The quantitative survey data are analyzed using descriptive statistics; these are summarized by individual teacher and by assignment group (treatment and control), and are compared by group.

The free-response portions of the surveys are minimally coded. The results of these findings are provided in the individual site reports.

Survey data are used to quantify the extent of exposure to the materials (opportunities to learn with the curriculum). Teacher self-report data on a variety of topics is also reported in the individual site reports. Only a subset of the data is reported here.

Surveys were deployed to both *SFScience* and control group teachers beginning on December 12, 2005 and continued on a bi-weekly basis until April of 2006. Response rates were calculated using a simple percentage calculation based on the ratio of actual received responses to the number of expected responses. There were 46 teachers in the *SFScience* group and 46 teachers in the control group. A total of nine surveys were deployed with a response rate of 88% for the *SFScience* teachers, and an 80% response rate for the control teachers. The overall response rate for both *SFScience* and control group teachers across all of the surveys is 85%, giving reasonable confidence in interpreting the results provided.

The survey topics were developed to account for the various aspects of teacher and student actions associated with instruction and learning. In order to characterize the average time teachers and students spent in science instruction, we used a repeated question strategy. These questions, together with questions regarding the types of activities, allow us to draw inferences about how time was devoted to *SFScience* instruction in both the *SFScience* and control groups. Survey 9 focused on the content covered and teachers' overall experience with the materials.

The following two tables provide a summary of the topics and that were surveyed with both *SFScience* and control group teachers and the average response rate per survey by both groups.

Table 9. Topics for Surveys Deployed from December to May 2006

	Survey topics
S1	Science schedule and instructional time
S2	Resources
S3	Interactions with materials/students
S4	More interactions
S5	Time and preparation
S6	Materials and resources
S7	Assessments
S8	More interactions
S9	Final survey – Usage of materials

Table 10. Survey Response Rates for All Teachers by Site

	Survey Response Rates								
	S1	S2	S3	S4	S5	S6	S7	S8	S9
WA	52%	91%	81%	91%	95%	95%	67%	48%	95%
UT	68%	90%	95%	100%	95%	95%	95%	84%	85%
OH	64%	73%	82%	82%	82%	82%	64%	73%	73%
FL	67%	100%	91%	100%	95%	95%	91%	95%	61%
CA	70%	85%	100%	90%	90%	90%	85%	85%	91%

Instructional and Classroom Descriptions

Data collected for descriptions of the classroom and scheduling come from three sources: classroom observations, interviews and surveys. Teachers self-reported on their planning practices as well as other details of the classroom including students' response to the materials. Researchers conducted classroom observations after confirming schedules with teachers.

Statistical Analysis and Reporting

The basic question for the statistical analyses was whether, following the intervention, students in *SFScience* classrooms had higher science and reading scores than those in control classrooms. The mean impact is estimated using multi-level models that account for the clustering of students in classes, which provides a more accurate, and often more conservative, assessment of the confidence we should have in the findings. We use SAS PROC MIXED (from SAS Institute Inc.) as the primary tool for this work. To increase the precision of our estimate, we include students' pretest scores in the analysis. In our experience, these are good predictors of achievement; including them as covariates in the impact analysis reduces the error variance, which makes it easier to discern the treatment impact.

In addition to the basic analysis of the mean impact, the plan for the study identifies the teacher- and student-level covariates that we expect (through theory or prior research) to make a difference in the effectiveness of the program being tested. The analysis tests for the interactions between the identified covariates and the experimental condition.

In addition to examining impacts and interactions where we anticipate effects, to better understand unexpected results, we use other demographics, teacher characteristics, and supplementary observational data in exploratory analyses to generate additional hypotheses about which factors might potentially moderate or mediate the treatment impact.

Our analyses produce several results: among them are the estimates for fixed effects, effect sizes, and *p* values. These are found in all the tables where we report the results of the statistical models.

Estimates. The estimate can be thought of as a prediction of the size of an effect. Specifically, it is how much we would predict the outcome to change for a one-unit increase in the corresponding variable. We are often most interested in the estimate associated with the experimental conditions, which is the expected change in outcome in going from control to treatment, holding other variables constant.

Effect sizes. We also translate the difference between treatment and control into a standardized effect size by dividing the difference by the amount of variability in the outcome (also called the standard deviation). This allows us to compare the results we find with results from other studies that use different measurement scales. In studies involving student achievement, effect sizes as small as 0.1 (one tenth of a standard deviation) are sometimes found to be important educationally. When possible we also report the effect size of the difference after adjusting for pretest, since that

provides a more precise estimate of the effect (i.e. in theory, with many replications, we would expect the adjusted effect size on average to be closer to the true value).

***p* values.** The *p* value is very important because it gives us a gauge of how confident we can be that the result we are seeing is not due simply to chance. Specifically, it tells us what the probability is that we would get a result with a value as large as – or larger than – the absolute value of the one observed when in fact there is no effect. Roughly speaking, it tells us the risk of concluding that the treatment has had an effect when in fact it hasn't. Thus a *p* value of .1 gives us a 10% probability of that happening. We can also think of it as the level of confidence, or the level of belief we have that the outcome we observe is not simply due to chance. While ultimately depending on the risk tolerance of the user of the research, we suggest the following guidelines for interpreting *p* values:

1. We have a high level of confidence when $p <= .05$. (This is the level of confidence conventionally referred to as "statistical significance.")
2. We have some confidence when $.05 < p <= .15$.
3. We have limited confidence when $.15 < p <= .20$.
4. We have no confidence when $p > .20$.

Results

Formation of the Experimental Groups

Groups as Initially Randomized

The randomization process guarantees that there is no intentional or unintentional bias in the selection of teachers and students into the treatment or the control condition. It does not, however, guarantee that the groups will be perfectly matched. It is important to inspect the two groups to determine whether, in spite of randomization, there are any significant differences on factors that affect the outcome[1]. Table 11 addresses the nature of the experimental groups; it shows the distribution of teachers, classes, grades, and students between *SFScience* and control conditions. This is the complete number of students in the experiment at the time that the experiment began in the fall of 2005.

[1] In technical terms, randomization ensures lack of bias, but we are interested in knowing whether the particular estimate resulting from this randomization may be far from the true value as a result of chance imbalances on factors that affect the outcome

Table 11. Distribution of Participants by Schools, Teachers, Grades, and Counts of Students

Condition	Schools	Teachers	Classrooms	# of Students			Total # of students
				Grade 3	Grade 4	Grade 5	
SFScience	18	46	56	376	644	415	**1435**
Control	18	46	57	557	235	587	**1379**
Totals	**36**	**92**	**113**	**933**	**879**	**1002**	**2814**

Note: The WA and OH sites had identified some classrooms as multi-age (multi-grade) targeted at the Gifted and Talented students in both the *SFScience* and control conditions.

Post Randomization Composition of the Experimental Groups

In checking for balance in the composition of the experimental groups, we examine teacher experience first, followed by student level variables such as gender, grade level, and student pretest outcomes.

Teaching Experience

During the randomization process teachers identified themselves according to years of teaching experience. The initial teacher pair was formed correspondingly so that the bias due to teaching experience would be distributed among the groups evenly. Table 12 summarizes this information.

Table 12. Years of Teaching Experience

Condition	Number of teachers		
	0 to 3 years	4 or more years	Totals
SFScience	4	40	**44**
Control	4	37	**41**
Totals	**8**	**77**	**85**
Statistics		**Value**	***p* value**
Fisher's exact test		0.29	1.00

Notes. Fisher's exact test is reported because 50% of the cells have expected counts less than 10. We are missing information about years of experience from seven teachers.

Randomization resulted in years of teaching experience being evenly balanced between *SFScience* and control teachers. The large *p* value of 1.0 is consistent with this assertion.

Student Variables

From Table 11, we see that 2814 students were enrolled in the study. Of these, 176 students were identified as needing special education support; we will not include those students in the analysis. There maybe other students that were also in this group, but two school districts did

not provide this information. Hence, the following analyses are based on a possible sample size of 2638 students.

Gender Distribution

Table 13 shows the distribution of gender of the students in each group.

Table 13. Gender Distribution for *SFScience* and Control Groups

Condition	Gender		
	Female	Male	Totals
SFScience	655	692	**1347**
Control	632	659	**1291**
Total	**1287**	**1351**	**2638**

Statistics	DF	Value	*p* value
Chi-square test	1	0.03	.87

Randomization resulted in gender being evenly balanced between *SFScience* and control groups. The large *p* value of .87 is consistent with this assertion.

Grade Distribution

Table 14 summarizes the distribution of students across grade-levels.

Table 14. Grades Involved in *SFScience* and Control Groups

	Grade			
Condition	Grade 3	Grade 4	Grade 5	Totals
SFScience	360	594	393	**1347**
Control	519	221	551	**1291**
Totals	**879**	**815**	**944**	**2638**

Statistics	DF	Value	*p* value
Chi-square test	2	224.83	<.01

We see that the number of students per grade was not distributed evenly between the conditions in spite of randomization. There are proportionally more students in grade 5 in the control group than in the *SFScience* group and proportionally fewer students in grade 4 in the control group than in the *SFScience* group. Chi-square tests confirm that this characteristic was not balanced between conditions. The imbalance may lead the estimate of the impact to depart from its true value.

Characteristics of the Experimental Groups as Defined by Pretest

We also checked whether randomization resulted in a balance on pretest scores, a variable that we include in most of our analyses to increase the precision of our estimates. Every student in the study was to take a Science and Reading achievement test for the study. All tests were paper and pencil with the standard bubble sheets. We did not receive test scores for all

students identified by class rosters. Teachers reported that 3rd and 4th grade students were not familiar with this type of testing and as a consequence had difficulty with making the marks dark enough for the testing agency to score. We also noted absentees during the testing periods with insufficient time in the schedule to conduct retests. As a result, the total number of students with pretest scores was much less than the anticipated 2638. The following tables summarize the pretest score attrition. Table 15 and Table 16 summarize pretest attrition according to grade.

Table 15. Students Missing Science Pretests by Grade

Condition	Grade 3	Grade 4	Grade 5	Totals
SFScience	70	43	17	**130**
Control	60	14	35	**109**
Totals	**130**	**57**	**52**	**239**

Statistics	DF	Value	*p* value
Chi-square test	2	20.06	<.01

Table 16. Students Missing Reading Pretests by Grade

Condition	Grade 3	Grade 4	Grade 5	Totals
SFScience	49	93	68	**210**
Control	44	41	63	**148**
Totals	**93**	**134**	**131**	**358**

Statistics	DF	Value	*p* value
Chi-square test	2	10.21	<.01

We can see that the number of missing pretest scores is not distributed evenly between conditions among the grades. For science achievement, there are more 3rd and 4th grade students without pretest scores in the *SFScience* group than in the control group. Additionally, there are fewer students who are missing pretest scores in the control group.

We now look at this same pretest attrition organized by district.

Table 17. Students Missing Science Pretests by District

Condition	WA	UT	OH	FL	CA	Totals
SFScience	54	39	11	8	18	**130**
Control	35	24	15	12	23	**109**
Totals	**89 (37.2%)**	**63 (26.4%)**	**26 (10.9%)**	**20 (8.4%)**	**41 (17.2%)**	**239 (100%)**

Statistics	DF	Value	*p* value
Chi-square test	**4**	**7.87**	**.10**

Table 18. Students Missing Reading Pretests by District

Condition	WA	UT	OH	FL	CA	Totals
SFScience	31	37	20	6	116	**210**
Control	21	21	18	16	72	**148**
Totals	**52 (14.5%)**	**58 (16.2%)**	**38 (10.6%)**	**22 (6.2%)**	**188 (52.5%)**	**358 (100%)**

Statistics	DF	Value	*p* value
Chi-square test	**4**	**10.87**	**.03**

The attrition rate for the science achievement pretest is 9.1% (number missing, 239 divided by 2638, the number expected) of the total identified population. Two sites, Washington and Utah, contributed a disproportionate amount of attrition.

For reading achievement, the attrition rate is 13.6% (358/2638) overall. We can see that the attrition rate for the Florida site was fairly small, whereas the California site had the largest attrition rate for the reading pretest. We believe that attrition in California was unusually high because some students were tested in late Spring of 2005 and test scores were not available. The other three sites contributed similar attrition rates.

We account for the attrition in pretest scores as reported by NWEA (scoring agency) in the following five categories for test sheets that could not be scored: answer sheet did not indicate the level of the test, test was too easy for the student, test was too difficult for the student, too many omitted answers, and student was absent.

Given this information, we are left with a sample size of 2399 students with science pretest scores and 2280 students with reading pretest scores. The following analysis tests for balance in pretest scores and is based on these sample sizes.

NWEA Science Test

The *SFScience* and control groups had slightly different average pretest scores on Science, as shown in Table 19. However, when we accounted for the fact that outcomes for students of the same teacher tend to be related by factoring these dependencies in the model, the *p* value increased to 0.97, indicating that the difference we are seeing is very likely due to chance.

Table 19. Difference in Science Pretest Scores Between Students in the *SFScience* and Control Groups

Descriptive statistics: Pretest outcomes	Raw group means	Standard deviation	Number of students	Standard error	Effect size[a]
SFScience	197.45	10.62	1217	0.30	0.06
Control	196.86	9.58	1182	0.28	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
Condition (*SFScience* – control)	0.59		2397	-1.42	.16

[a] The difference we are measuring is not an effect of treatment (the usual sense of effect size) but a result of chance differences in the randomization.

NWEA Reading Test

As with NWEA Science, the *SFScience* and control groups had slightly different average pretest scores on Reading. Again, when we accounted for the fact that outcomes for students of the same teacher tend to be related by modeling these dependencies, the *p* value increased to 0.91, indicating that this difference is likely due to chance. In the analyses that follow, we add the pretest covariate in order to increase the precision of the impact estimate. (Still, we recognize that, with or without this covariate, the impact estimate is unbiased as a result of the randomization.)

Table 20. Difference in Reading Pretest Scores Between Students in the *SFScience* and Control Groups

Descriptive statistics: Pretest outcomes	Raw group means	Standard deviation	Number of students	Standard error	Effect size[a]
SFScience	199.84	15.69	1137	0.47	-0.02
Control	200.07	14.39	1143	0.43	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
Condition (*SFScience* – control)	-0.23		2278	0.37	.71

[a] The difference we are measuring is not an effect of treatment (the usual sense of effect size) but a result of chance differences in the randomization.

Attrition after the Pretest

NWEA Science Test

Based on the on the information above there are 2399 who have science pretest scores. However, we did not receive science posttest scores for all of these students. Of these participating students 2079 have both pretest and posttest scores. Another 117 students have posttest scores but are missing pretest scores.

Table 21 shows the attrition of enrolled students that occurred after taking the pretest. Chi-square tests confirm that this attrition of students was balanced between conditions.

Table 21. Missing Science Tests for *SFScience* and Control Groups

Condition	Missing Tests		
	Students having pretest and posttest	Students missing posttest scores[a]	Totals
SFScience	1059	158	1217
Control	1020	162	1182
Totals	**2079**	**320**	**2399**
Statistics	**DF**	**Value**	***p* value**
Chi-square test	1	0.27	**.60**

[a] These students are taken from the population of participants having both pretest and posttest scores. They have pretest scores but are missing posttest scores.

We observe that 320 students (or 13%) are missing posttest scores due to a variety of reasons including being absent during testing or not being able to complete the test. Table 22 shows that students with no score for the posttest (having pretest scores only) scored lower on the pretests. The low *p* value confirms that there is a bias toward including higher scoring students. Thus, we can be less confident of the applicability of findings for lower scoring students.

Table 22. Difference in Pretest Scores for Students Having Pre- and Posttest Scores versus Pretosttest Only

Descriptive statistics: Pretest outcomes	Raw group means	Standard deviation	Number of students	Standard error	Effect size[a]
Missing posttest scores	194.2	10.52	320	0.59	-0.34
Have both pre- and posttest scores	197.62	9.99	2079	0.22	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
(Missing posttest) – (Have posttest)	-3.41		2397	5.65	<.01

[a] The difference we are measuring is not an effect of treatment (the usual sense of effect size) but a result of chance differences in the randomization.

NWEA Reading Test

Based on the information above, there are 2280 who have Reading achievement pretest scores. However, we did not receive Reading posttest scores for all of these students. Of these participating students, 1908 have both pretest and posttest scores. Another 246 students have posttest scores but are missing pretest scores.

Table 23 shows us that there are proportionally more students in the *SFScience* group who were originally enrolled but did not take the reading posttest, as compared to the control group. Chi-square tests confirm that this attrition was not balanced between conditions. The imbalance may lead the estimate of the impact to depart from its true value.

Table 23. Missing Reading Tests for *SFScience* and Control Groups

Condition	Missing Tests		
	Students having pretest and posttest	Students missing posttest scores[a]	Totals
SFScience	934	203	1137
Control	974	169	1143
Totals	**1908**	**372**	**2280**

Statistics	DF	Value	*p* value
Chi-square test	1	3.93	.05

[a] These students are taken from the population of participants having both pretest and posttest scores. They have pretest scores but are missing posttest scores.

We observe that 372 students (or 16%) are missing posttest scores. Table 24 shows that students with no score for the posttest (having pretest scores only) scored lower on the pretests. The low *p* value confirms that there is a bias toward including higher scoring students. Thus, we can be less confident of the applicability of findings for lower scoring students.

Table 24. Difference in Pretest Scores for Students Having Pre- and Posttest Scores versus Pretest Only

Descriptive statistics: Pretest outcomes	Raw group means	Standard deviation	Number of students	Standard error	Effect size[a]
Missing posttest scores	195.99	16.35	372	0.85	-0.32
Have both pre- and posttest scores	200.72	14.66	1908	0.34	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
(Missing posttest) – (Have posttest)	-4.74		2278	5.59	<.01

[a] The difference we are measuring is not an effect of treatment (the usual sense of effect size) but a result of chance differences in the randomization.

Implementation Results

In this section we describe more fully the aspects of the implementation that characterize this intervention. We used two questions to guide our descriptions and analysis: What resources are

needed to manifest the *SFScience* condition? Are there differences in the extent, quality, and type of implementation of the materials? We also studied the features of the implementation to identify variables possibly related to the outcome measures.

Comparison of *SFScience* and Control Groups

Classroom Settings for Instruction

For this combined report, we only report information at a very broad level of detail. For specifics regarding classroom conditions and resources, refer to the individual site reports.

Most of the schools participating in the study are elementary schools, K-5/6. The Florida site schools are all K-8 and one school in Ohio is called a middle school with grades 5 and 6. The elementary school classrooms typically do not have storage for the science equipment, nor do they have space to keep long term observational experiments. Some teachers, because of small classroom spaces, also noted that conducting hands-on science activities was difficult. For some teachers, this lack of space constitutes a barrier for using the science kits regularly and more generally for hands-on science activities. Lack of space was also noted for storing the Leveled Readers, but this was less of a problem because teachers already accommodated libraries.

Because they had older students who required laboratory access for their science instruction, the Florida schools and the one Ohio school had designated laboratory classrooms accessible on a rotating basis to the teachers. For these teachers the science kits and activities did not present storage challenges.

Many of the schools had televisions, computer stations, and video/sound playback available in the classrooms. Several teachers noted that they would use the audio version of the text with their students during whole-class science reading activities. Many teachers in both *SFScience* and control groups supplemented all science instruction with videos. Bill Nye the Science Guy was very popular as was United Streaming and Discovery Channel.

During the *SFScience* classroom observations we noted two basic modes of instruction: 1) whole class reading, where students took turns reading aloud to the rest of the class with corrections supplied by the teacher and 2) hands-on science activities, where students interacted directly with materials, guided by their teacher. In the control classrooms, instruction strategies similar to those used by the *SFScience* teachers were noted. More often we noted whole-class reading activities in the control classrooms.

A large part of *SFScience* is composed of reading materials, specifically the Leveled Readers, designed to provide three different levels of reading difficulty. The schools in our study all provide separate reading instruction to students; that is, similar to science instruction, reading instruction is treated as a “specialty” subject where one teacher in the grade level teaches several class groups throughout the day. Thus teachers exchange students for both science and reading. Consequently, the Leveled Readers were used by teachers other than those teachers assigned to the *SFScience* group. This may be thought of as a type of contamination because the treatment was provided by others not assigned to the treatment condition, but to our knowledge no control group students received instruction using the Leveled Readers. So strictly speaking this is not contamination, although it did introduce teachers and situations where no data was collected.

Some *SFScience* classrooms were noted to include students who were selected on the basis of reading ability. This caused some difficulties using the Leveled Readers since they are supplied in packets of six per level. In classrooms where students of low reading ability were concentrated, there were insufficient copies to supply the whole class. Therefore the students took turns using the Leveled Readers and teachers had to plan alternate activities for those students not engaged with the Readers. We noted that some control classrooms were also organized according to the reading level of the students, but this did not cause teachers to plan alternative activities because they provided sufficient materials for all students.

Leveled Readers were also used during free reading time, when students elected reading materials. More than half of the teachers reported that students often selected the *SFScience* Leveled Readers as their reading of choice. .

At the majority of the schools, teachers reported that students did not have the requisite foundational experiences to use the *SFScience* materials as organized. Teachers reported that students needed pre-learning activities before they could begin using the text or the science kits. Often the teachers would show a video or prepare a short introductory lesson before beginning instruction with the *SFScience* materials. Control teachers did not have these issues since they typically organized the lessons themselves using a variety of materials.

In summary, the *SFScience* classrooms were not markedly different from the control classrooms except for the materials in use and the sequencing of the materials. Teachers reported that the comprehensive materials provided excellent resources for teaching science.

Demographics of the Classrooms

The table below summarizes two of the NCLB categories of interest. The other two categories, socio-economic level and Special Education needs are not reported because some school districts do not provide this information for individual students. These demographics are reported specifically as they apply to the students in this study.

Table 25. Student Demographics Relevant to NCLB

	% English Language Learners	Gender: %F / %M
WA	n.a.	47 / 53
UT	26.9	50 / 50
OH	n.a.	44 / 56
FL	0.1	47 / 53
CA	21.6	51 / 49

Note. n.a. means not applicable because the schools in these districts had low levels of students in this category.

Opportunities for Learning

Another challenge faced by all of the public schools in the study is the scheduling of science instruction. The actual time available for teaching science is relatively low given the stress placed on reading and mathematics instruction in grades 3 through 5. Both *SFScience* and control group teachers reported having an alternating schedule, wherein they teach science daily for two weeks approximately 30 minutes per day and then teach social science for two weeks. Still other teachers indicated that they taught science for two months out of a trimester as time allowed. The only teachers reporting having daily science classes were from the Florida schools. There, all schools were PreK-8 and had a science laboratory. Every teacher was scheduled to use the laboratory once a week (or more often as time allowed) and science was taught on a daily schedule, usually for 30 to 40 minutes per session. Visalia Unified was the only public school site that allotted science instructional time as often as the Florida schools.

A summary of the approximate instructional times per site and by group as calculated from teacher surveys is presented in Table 26.

Table 26. Science Instruction Time

	Total Average Instruction Time in Hours	
	Control	*SFScience*
WA	13.04	19.98
UT	25.20	24.00
OH	21.88	23.80
FL	42.78	37.35
CA	47.30	45.16
Overall	32.29	30.84

It should be noted that in Florida, the control curriculum had been in place for years and the teachers using the materials had well established routines. In Visalia Unified, the schools were new with many new teachers with common planning for grade levels. In Washington, the control group teachers depended on district developed science kits. These kits were in limited supply and so were scheduled to rotate through the district. Consequently, control teachers had to wait for their turn, which caused larger gaps in scheduling of science instruction.

Density of Science Inquiry Reflected in the Classroom

Because Scott Foresman designed the *SFScience* curriculum using inquiry as theme and pedagogy, sections of the teacher surveys were constructed to collect data on the aspect of science inquiry as a method for teaching/learning science.

Specifically, Scott Foresman designed the curriculum to "scaffold" the inquiry process. Here is a brief description of how inquiry is reflected in the structure of the curriculum. First, the publisher conceptualized learning science through the process of inquiry as a series of developmental stages. At the beginning, students might not know the process or have used the process in science, so a chapter in every unit begins with a "Directed Inquiry" (DI). This activity is usually teacher-led and introduces the essential features of the inquiry process. The activity has a step-by-step process attached to it that allows for practice of both the process and methods. The next activity in the chapter is called a "Guided Inquiry" (GI). Now the teacher acts more as a facilitator. The activity is outlined as a series of goals rather than step-by-step process and allows students to practice with guidance. The final activity in the unit (after all of the chapters have been completed) is a "Full Inquiry" (FI) and is aimed at giving students practice at creating their own inquiry activity. Only the inquiry framework is provided as support.

For each site we used the same group of questions to create a composite variable that indicates the degree of inquiry density. We used five essential elements of the framework to measure inquiry density:

- Questions are scientifically oriented.
- Learners use evidence to evaluate explanations.
- Explanations answer the questions.
- Alternative explanations are compared and evaluated.
- Explanations are communicated and justified.

This framework is reflected in the in the sequenced activities of the *SFScience* program as a continuum:

- Questions (DI: students use a question provided by the teacher, materials, or some other source; GI: students are guided to refine and clarify questions; FI: students investigate their own questions)
- Predictions or hypotheses (DI: students are given a prediction for conducting a descriptive investigation; FI: students are guided to make a prediction for an investigation; FI: students develop logical/reasonable predictions)
- Investigations (DI: students are given the procedures and materials to conduct an investigation; GI: students are given suggestions for procedures and materials; FI: students devise a plan for the investigation).

On teacher surveys, we asked *SFScience* and control group teachers about time spent doing these different activities. The variable "science inquiry" is a composite of the time spent in six different aspects of the inquiry process as a percentage. Hence, on a scale of 0 to 100, it can be thought of as a measure of "inquiry process density" with 100 being an indication that the teacher and students were practicing the inquiry process every time science was taught.

A summary of the approximate science inquiry density calculated per site and by group as informed by survey responses are presented in the Table 27.

Table 27. Percentage of Science Inquiry Density by Condition and Site

	% of Average Science Inquiry Teaching	
	Control	*SFScience*
WA	27.84	18.42
UT	28.33	30.33
OH	33.15	43.75
FL	27.42	22.69
CA	30.80	22.27
Overall	29.13	25.11

Implementation of *SFScience*

Training and Support

Each site was provided one-half day of training with the materials by a Scott Foresman representative. (Table 1 shows individual site dates of the training.) During the training, a demonstration of the science kits and the method of hands-inquiry were presented to the teachers. A common vision of how the materials were to be used and how much material was to be covered was shared with the teachers. Each facet of the curriculum was discussed: teacher edition, student edition, workbook, activity book, audio tapes, assessment book, science kits, graphic organizers, and additional materials. Overall, the teachers were enthusiastic about the materials and the training session provided a good introduction. Details of the individual sessions are provided in the individual site reports. (Table 2 contains a complete list of the materials supplied by Scott Foresman.)

No specific instructions were given to the teachers regarding the frequency of the instruction, and teachers understood this to mean that they were to use the materials when they normally schedule science instruction with their students.

Timeline of the Implementation

Although many of the sites had the training near the start of the academic year, several issues arose that made it difficult to implement *SFScience* for the expected eight months. Materials on backorder caused delays getting started. Moreover, the relatively low priority for teaching science (particularly as compared to reading and mathematics) resulted in teachers varying their schedules. November and December were low activity months for science as teachers prepared students for the holidays, and April and May are months that are usually dedicated to preparation for high stakes testing and test administration. Thus the *ideal* implementation of eight months was truncated to a span of three and one-half to five and one-half months. At one school, teachers reported that science classes were suspended for two months in order to address AYP goals. Specific scheduling issues for the five sites are noted below.

Federal Way

Fourth-grade *SFScience* teachers at one school were able to begin full implementation on or about September 26th, 2005. Third- and fifth-grade teachers were missing essential items: science kits and teacher resource packets, and in fifth grade, student edition textbooks. These teachers reported that they began implementing *SFScience* shortly after November 21st.

At the other two schools third-, fourth-, and fifth-grade *SFScience* teachers were missing science kits and teacher resource packets. Additionally, fifth-grade classrooms were missing student edition textbooks. Treatment is reported to have started shortly after November 21st.

Ogden

All treatment group third- and fourth-grade teachers were able to begin using the *SFScience* materials provided by Scott Foresman shortly after the training workshop on September 19, 2005. Additionally, fifth-grade teachers at one school were able to begin using all of the *SFScience* materials on September 19th.

Fifth-grade teachers at the other three schools were able to begin using the textbooks and other reading materials provided in the curriculum, but not any of the inquiry science kit materials because these were on backorder until late November. These teachers report that they began using all of the materials on or about December 4th, 2005.

Reynoldsburg

This site was identified much later in the process than the other four sites. Science kit materials for third, fourth, and fifth grades were backordered. Treatment teachers were able to begin using the textbooks and other reading materials provided in the curriculum, but not any of the inquiry science kit materials. Teachers indicated that they began using the *SFScience* materials in early December.

St. Petersburg

All treatment group third- and fourth-grade teachers were able to begin using the *SFScience* materials shortly after the training workshop on September 19, 2005. Additionally, fifth-grade treatment teachers at three schools began using the all of the science materials on September 19th.

Fifth-grade teachers at one of the schools were able to begin using the textbooks and other reading materials provided in the curriculum, but none of the inquiry science kit materials because these were on backorder. At this school fifth-grade treatment began full use of the materials on November 21st.

At another school fifth-grade teachers could not implement any portion of the curriculum because student textbooks were on backorder, as were the science kits. These teachers reported that they began using all the materials on or about December 4th, 2005.

Visalia

All treatment group third-, fourth-, fifth-grade teachers were able to begin using the *SFScience* materials shortly after the training workshop on September 12, 2005.

Curriculum Covered

Treatment group teachers were asked to complete any two of the four units provided in the *SFScience* curriculum. The units are as follows: Unit A, Life Science; Unit B, Earth Science; Unit C, Physical Science; and Unit D: Space & Technology. The table below summarizes the *SFScience* teachers' responses to an inquiry regarding which chapters in each of the units they covered. The reader is cautioned to note that these are not all of the chapters covered during the academic year, but only those that were covered by most of the teachers and those chapters that the fewest teachers reported addressing with their students. Note that content presented in chapters vary by grade level. This data is presented as an overall idea of what was used by the teachers and not specific to any one grade level.

Table 28. Most and Least Covered Chapters by Unit

	Chapter(s) Most Covered	Chapter(s) Least Covered
Unit A: Life Science	1,4	3
Unit B: Earth Science	9	10
Unit C: Physical Science	11	15
Unit D: Space & Technology	16,17	18

Teacher Opinions

In this section we report on a subset of the information collected regarding various aspects of the materials, whether teachers would recommend them, and how they thought their students had experienced them. We specifically addressed the areas that were required of the teachers for implementation: Leveled Readers, Science Kits, Inquiry Cycle, and Assessments. The numbers of teachers noted in the tables are the actual number of teachers who responded to the questions. These may be less than the total numbers of teachers assigned to the *SFScience* group.

Leveled Readers

Teachers' responses to the Level Readers were mixed. Although the teachers liked the concept and uniformly agreed that they motivated students' interest, implementation was problematic in some cases. The following figures depict the teachers' responses when asked the extent to which they agreed to questions about the Leveled Readers, whereas the table shows satisfaction at individual sites.

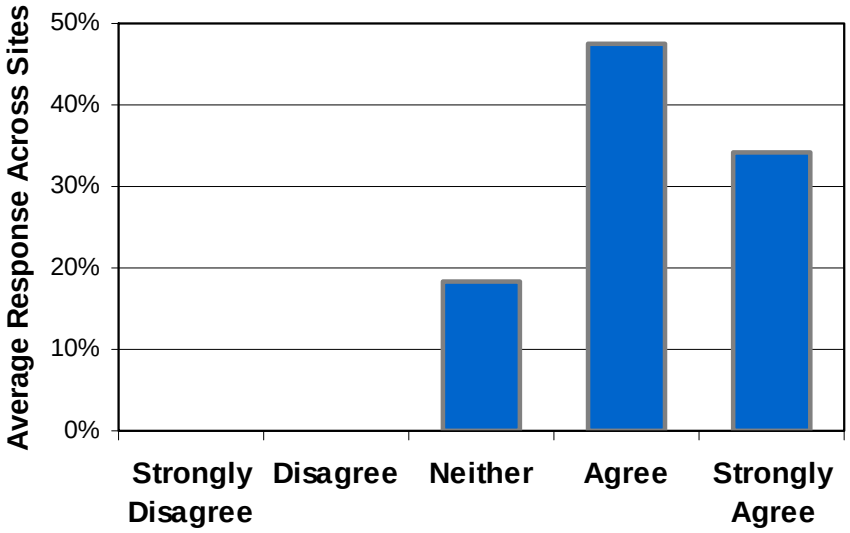

Figure 1. Did the materials help motivate students' non-fiction reading?

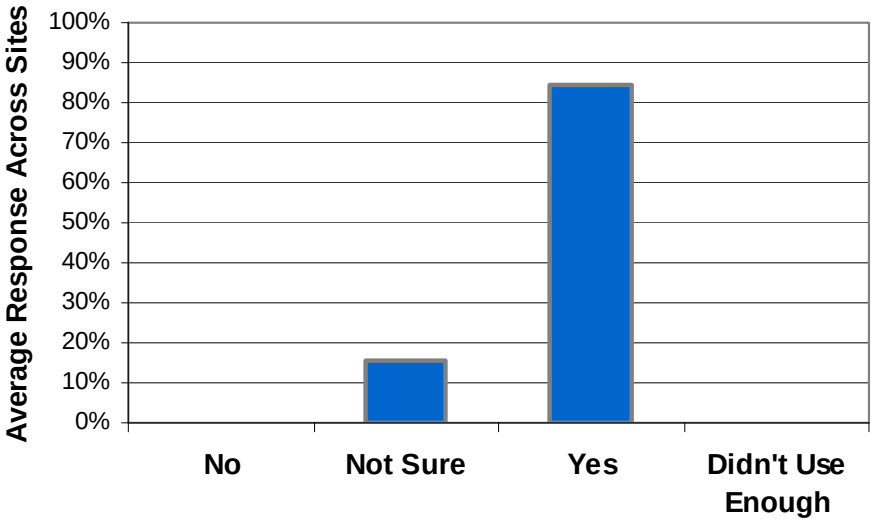

Figure 2. Would you recommend the Leveled Readers to other teachers in your grade?

Table 29. *SFScience* Teachers' Reported Satisfaction with the Leveled Readers

School district	Number[a]	Very dissatisfied %	Somewhat dissatisfied %	Neither %	Somewhat satisfied %	Very satisfied %
WA	11	0.00	0.00	0.00	27.27	72.72
UT	6	0.00	0.00	0.00	16.67	83.33
OH	4	0.00	25.00	0.00	50.00	25.00
FL	8	12.50	12.50	0.00	75.00	0.00
CA	9	11.11	0.00	11.11	44.44	33.33

[a]These numbers reflect the number of teachers responding to the survey, not the total number of teachers assigned to the *SFScience* group.

The figures highlighted in red for Florida reflect the teachers' desire for more readers at one level. At this site, because science instruction is grouped by reading ability, teachers need an entire class set of readers at one level.

The figures highlighted in red for California reflect the teachers' desire for lower level readers. Because of the high population of English language learners at this site, teachers need a wider range of levels beyond the three levels supplied by the readers as provided. Teachers noted that two grade levels below grade would be more appropriate for their students.

Two quotes from teachers' reflected positive opinions on the Leveled Readers:

> *"They were helpful so students at all reading levels were able to understand the text and concepts in a small group setting."*
>
> *"I teach in differentiated groups. I liked that all students had similar content on their level and the same vocabulary."*

Two quotes depicted the challenges teachers confronted in using the Leveled Readers:

> *"I am having great difficulty using the leveled readers because my students are ability grouped. Also, I do not teach reading. I do not have enough books for the entire class. It would be a great help if the books were available online."*
>
> *"Having only 6 copies at each level for 68 students proved to be too few individual copies to used effectively. Having access to the leveled readers on line would be a very effective and useful approach to utilizing them. Also I did not have the teacher's guide for the readers."*

Science Inquiry/Hands-on Learning

Since much of the emphasis of the *SFScience* materials is to develop the student's ability to use the process of scientific inquiry to learn science, we asked teachers to respond regarding the process of inquiry. The following graphs depict the extent to which teachers agree with a question about science as a way of exploring the world. We report the Ogden site separately because all teachers responded uniformly.

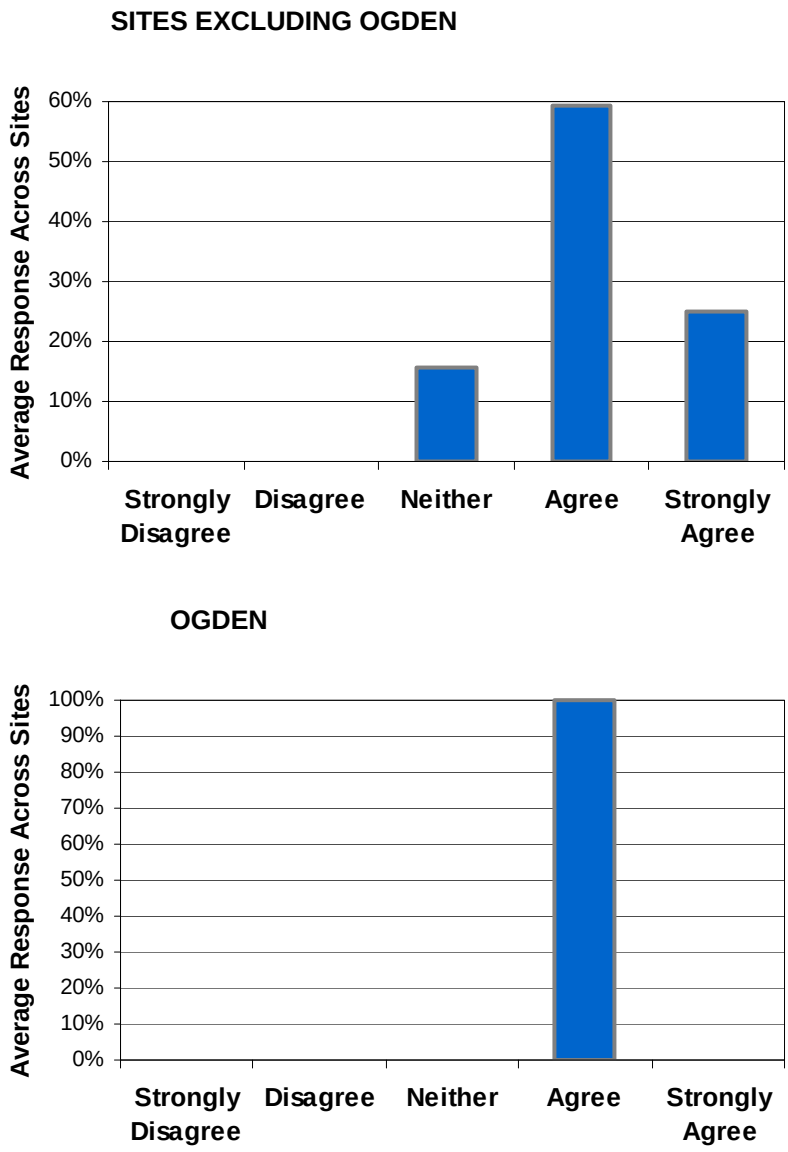

Figure 3 and Figure 4. Did the materials help students learn "science is everywhere"?

Three quotes reflected teachers' positive views about the Science Inquiry/Hands-on Learning aspect of *SFScience*:

> *"This is my preferred way to teach science. When we had the time to spend, I knew they would understand on a deeper level, and that they would remember what they learned better."*
>
> *"Students connect with their learning. Also students retain information better."*
>
> *"The students were able to experience what they were learning. They did the most tremendous job on the last unit where they developed the activities themselves."*

Science Kits

Teachers used the science kits with varying degrees of success, as indicated by their responses to questions about their level of satisfaction.

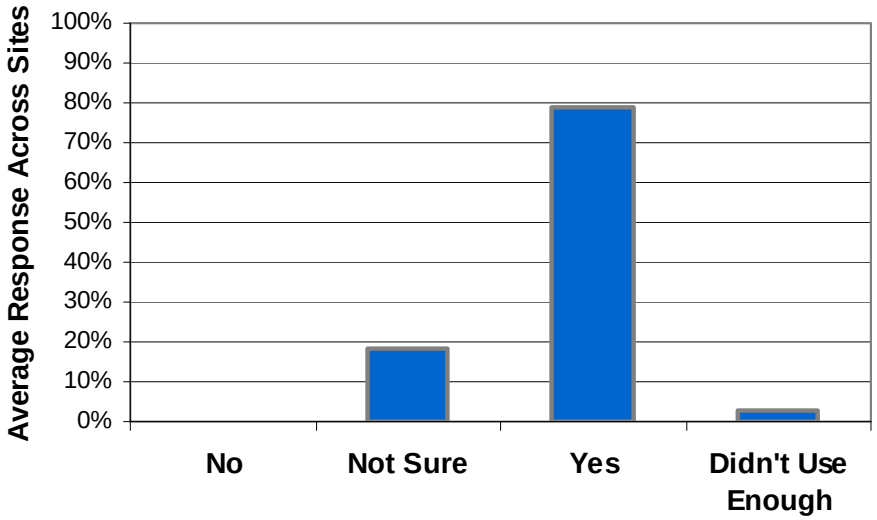

Figure 5. Would you recommend the science kits to other teachers in your grade?

Table 30. *SFScience* Teachers' Reported Satisfaction with the Science Kits

School district	Number[a]	Very dissatisfied %	Somewhat dissatisfied %	Neither %	Somewhat satisfied %	Very satisfied %
WA	11	0.00	9.09	18.18	54.55	18.18
UT	6	0.00	0.00	0.00	50.00	50.00
OH	4	0.00	0.00	50.00	50.00	0.00
FL	8	0.00	0.00	0.00	50.00	50.00
CA	9	0.00	0.00	22.22	33.33	44.44

[a]These numbers reflect the number of teachers responding to the survey, not the total number of teachers assigned to the *SFScience* group.

The figure highlighted in red for Washington reflects the teachers' desire for materials better aligned with their state standards.

These positive quotes reflect three teachers' experiences with the science kits:

"I loved having everything right there ready to go -- It was GREAT! The kids LOVED the experiments we did."

"The students were involved with the concepts. The audio tapes gave back-up to the lower readers. The photos provided a visual that helped to make the concepts concrete. The lesson manual was laid out well for presentation."

"When [the science kits] aligned to our state standards it was great! My students really enjoyed them. I liked how organized they were and easy to use."

Four quotes illustrate the challenges of working with the science kits:

"The kits did not have all the items we needed for the experiments and I found that I needed to prepare for the experiments in advance."

"I felt restricted by them at times. The concepts could have been expanded upon and the experiments could have allowed for more student/teacher creativity."

"They take up so much room!!! It is so hard to store everything."

"Taking the time to set up materials."

Assessment Materials

As shown in the graph below, the assessment materials caused the greatest difficulty and consequently the greatest dissatisfaction among teachers.

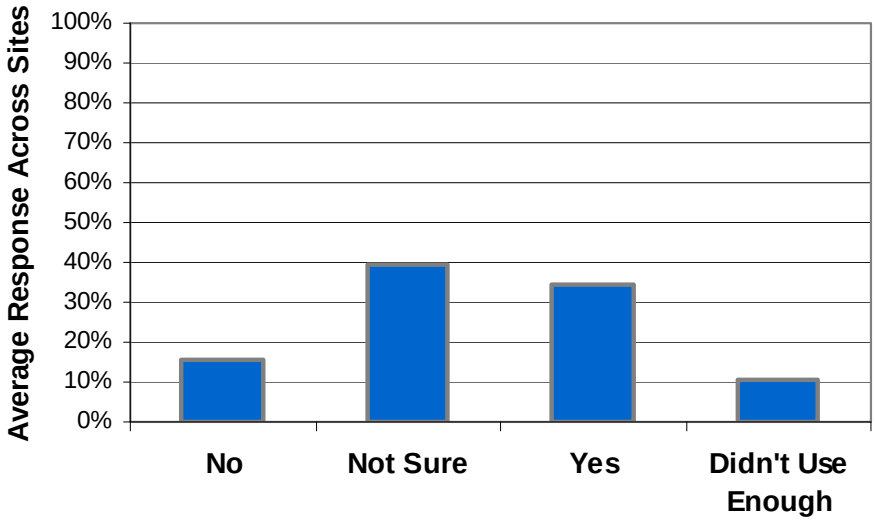

Figure 6. Would you recommend the assessment book materials to other teachers in your grade?

Table 31. *SFScience* Teachers' Reported Satisfaction with the Assessment Book

School district	Number[a]	Very dissatisfied %	Somewhat dissatisfied %	Neither %	Somewhat satisfied %	Very satisfied %
WA	11	0.00	18.18	36.36	45.45	0.00
UT	6	0.00	0.00	16.67	83.33	0.00
OH	4	50.00	25.00	25.00	0.00	0.00
FL	8	25.00	62.50	12.50	0.00	0.00
CA	9	0.00	11.11	44.44	22.22	22.22

[a]These numbers reflect the number of teachers responding to the survey, not the total number of teachers assigned to the *SFScience* group.

The figures highlighted in red for the sites reflect the teachers' concern with reading level alignment; that is, the assessment reading level was much higher than the text and Leveled Readers. Additionally, teachers were concerned with the content alignment. Teachers reported that the questions in the assessments often did not reflect the concepts presented in the text and in the activities. Teachers reported that they needed to create study sheets and other materials to prepare the students for taking these assessments.

Teachers' positive quotes reflected their experiences with the assessment materials:

"They were already made for me to use. I did not have to type my own test."

"I enjoyed the short answer questions as they really helped you to see what students understood. I also enjoyed the rubrics and written "typical" answer samples for short answer questions or activities."

*"The test making **software** was WONDERFUL!!!! I used it frequently. I would recommend it without reservation."*

Three additional quotes depict the challenges discovered by some teachers:

"These tests were so hard that I used them as teaching tools, not assessments."

"Emphasis was mostly on vocabulary, facts rather than concepts."

"The language was especially difficult for third graders to read and understand. Many of the assessments used vocabulary that was not used/introduced in the book. This was very frustrating for the students and therefore to me. I did not use it very much for this reason."

Summary of Implementation

Overall, teachers felt very positive about *SFScience*, but there were a variety of challenges. Some of these challenges came from the teaching environment at the schools and others came from the teachers facing a new curriculum.

Outside Challenges:

- Science is given lower priority than other subjects
- Instruction time is rarely allotted for science

Curriculum Challenges:

- Assessments are not leveled, nor are they aligned with content covered.
- Some teachers require more than three levels for the Leveled Readers.
- Some teachers did not have enough copies of particular levels of the Leveled Readers.
- Alignment to state standards was not consistent.
- Science kits required classroom management skills different from those possessed by teachers.
- Science learning through inquiry requires deeper teacher understanding and overall knowledge than teachers may currently have.

Quantitative Results

Overview

The primary topic of our experiment was the impact of curriculum on student performance on the NWEA test. We will first address the impact on science achievement and then the impact on reading achievement.

In the following sections, our analysis of the quantitative results takes the same form. Within each content area, we first estimate the average impact of on student performance. These results are presented in terms of effect sizes.

We then show the results of mixed model analyses where we estimate whether the impact of the intervention depends on the level of certain moderator variables. For instance, we show the results of a model that tests whether there is a differential impact across the prior score scale. We also model the potential moderating effects of gender and years of teaching experience. We provide a separate table of results for each of these moderator analyses. The fixed factor part of each table provides estimates of the factors of interest. For instance, in the case where we look at the moderating effect of a student's prior score, we show whether being in a *SFScience* or a control class makes a difference for the average student. We also show whether the impact of the intervention varies across the prior score scale. At the bottom of the table we give results for technical review – these often consist of random effects estimates which are added to the analysis to account for the fact that the individual results that come from a common upper-level unit (e.g., class or teacher) tend to be similar (i.e., the observations are dependent.) In some cases, to account for these dependencies, we model fixed rather than random effects but do not present the individual fixed effects estimates. Modeling the dependencies results in a more conservative estimate of the treatment impact.

We note that the number of cases used to compute the effect size often will be larger than the number used in the mixed model analysis because to be included in the latter analysis a student has to have both a pretest and a posttest score.

Science Outcomes

Analysis Including Pretest

Our first analysis addressed *SFScience* outcomes using the NWEA Science scale. Table 32 provides a summary of the sample we used in the analysis and the results for the comparison of NWEA scores for students in *SFScience* and control groups. The "Unadjusted" row gives information about all the students in the original sample for whom we have a posttest score.

This shows the means and standard deviations as well as a count of the number of students, classes, and teachers in that group. The last two columns provide the effect size, which is the size of the difference between the means for *SFScience* and control in standard deviation units. Also provided is the *p* value, indicating the probability of arriving at a difference as large as, or larger than, the absolute value of the one observed when there truly is no difference. The "Unadjusted" row is based on all students with a posttest and the estimated effect size takes into consideration the clustering of students in upper-level units (i.e., that students are grouped within teachers.) The "Adjusted" row is based on the students who have both pretest and posttest scores. This is the sample that we use in the analyses on which we base our results reported in Table 35 and Table 36. The means, and therefore the effect size, are adjusted to take into account the student pretest scores; hence, these statistics are adjusted for any chance imbalance in the two randomized groups[2]. They also figure in the effect of students being grouped within teachers.

Table 32. Overview of Sample and Impact of *SFScience* on Science Achievement

	Condition	Means	Standard deviations[a]	No. of students	No. of classes	No. of teachers	Effect size	*p* value[b]
Un-adjusted	***SFScience***	199.63	10.70	1124	56	46	.01	.64
	Control	199.57	9.79	1072	57	46		
Adjusted	***SFScience***	199.35	10.67	1059	55	45	-.02	.62
	Control	199.65[c]	9.73	1020	57	46		

[a] The standard deviations used to compute the adjusted and unadjusted effect sizes are computed from the scores of the students in the sample for that row

b The unadjusted *p* value is computed using a model that includes clustering of students within teachers but no other covariates. The adjusted *p* value is computed using a model that includes clustering and pretest as a covariate, as well as fixed effects when needed.

[c] This value is the raw mean of the control group students used in the statistical model. The model using fixed effects does not provide a single value. The adjusted value for *SFScience* is control value plus the estimate for treatment.

Figure 7 provides a visual representation of specific information in Table 32. The bar graphs represent average performance using the metric of NWEA Science.

The panel on the left shows average pre- and posttest scores for the control and *SFScience* groups. The pre- and posttest bars show that both the *SFScience* and control groups on average grew in their science achievement during the year.

The panel on the right shows estimated performance on the posttest for the two groups based on a model that adjusts for students' pretest scores and other fixed effects (i.e., this is a visual display of results from the row labeled 'adjusted' in Table 32.) The overall impact on science as an effect size (standard deviation units) is -.02 which is equivalent to a loss of one percentile point if the median *SFScience* student were placed in the control group. The high *p* value for the treatment effect (.62) indicates we should have no confidence that the actual difference is

[2] We also include any fixed effects used to estimate differences among upper level units.

different from zero. We added 80% confidence intervals to the tops of the bars. The overlap in these intervals further indicates that any difference we see is easily due to chance.

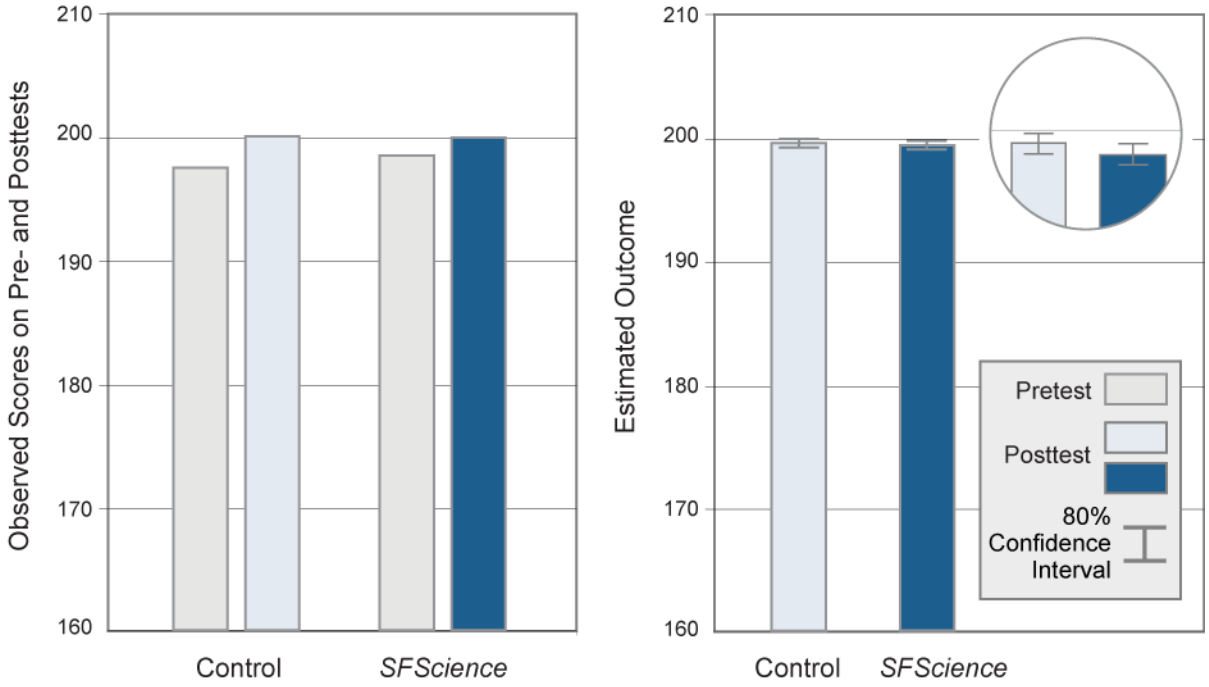

Figure 7. Impact on Science Achievement: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)

Table 33. Difference Made by a School Year of Science Instruction for the Control Group

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Score on the pretest	199.12	0.89	45	223.40	<.01
Gain made over the year	2.45	0.35	45	7.07	<.01

[a] Teachers were modeled as random effects.

Note. In obtaining the average pretest score, we did not model school and assignment pair as fixed effects. However, this model provides similar standard error and gain estimates over the year and with the standard fixed effects model.

Table 34. Difference Made by a School Year of Science Instruction for the *SFScience* Group

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Score on the pretest	198.69	1.05	45	190.01	<.01
Gain made over the year	1.82	0.35	44	5.16	<.01

[a] Teachers were modeled as random effects.

Note. In obtaining the average pretest score, we did not model school and assignment pair as fixed effects. However, this model provides similar standard error and gain estimates over the year and with the standard fixed effects model.

Analysis Including Pretest as a Moderator

We now report on the analyses that examine not just the overall impact of *SFScience* but also the moderating effects of other variables. We begin by examining the moderating effect of the prior score. Table 35 shows the estimated impact of *SFScience* on students' performance in science as measured by NWEA Science as well as the moderating effect of the prior score.

Table 35. Impact of *SFScience* on Science Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Predicted value for a control student with an average pretest	206.62	3.49	42	59.26	<.01
Impact of *SFScience* for a student with an average pretest	0.65	0.6	42	1.1	.28
Predicted change in control outcome for each unit increase on the pretest	0.81	0.02	1802	49.87	<.01
Interaction of pretest and *SFScience*	-0.02	0.02	1802	-1.08	.28
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	3.57	1.42		2.51	.01
Within-teacher variation	36.83	1.22		30.08	<.01

[a] Pairs of teachers used for random assignment are also modeled as a fixed factor but not included in this table.

[b] Teachers were modeled as a random factor.

The row in Table 35 "Impact of *SFScience* for a student with an average pretest" tells us whether *SFScience* made a difference in terms of student performance on NWEA Science for a student who has an average score on the pretest. The estimate associated with *SFScience* is 0.65. This shows a very small positive effect associated with *SFScience*. The *p* value of .28, indicates that we can expect to see a difference, as large or larger than the absolute value of

the estimate, 28% of the time when there truly is no effect. Using the criteria outlined earlier in the report, we conclude that we have no confidence that the true impact is different from zero. We see a discrepancy in the directions of the estimated effects between Table 32 and Table 35; in Table 32 we describe an average effect whereas here we describe an effect for the average student, and these are not equivalent. However, neither can be distinguished statistically from zero.

We also estimated the moderating effect of the pretest score on the impact of *SFScience* to determine whether it was differentially effective for students at various points along the pretest scale. The *p* value for this effect is .28. We have no confidence that the actual effect is different from zero. In other words, the effect of *SFScience* was the same for students, regardless of where a student started on the pretest.

As a visual representation of the results described in Table 35, we present a scatterplot in Figure 8, which shows student performance at the end of the year in science, as measured by NWEA Science, against their performance in the fall. These graphs show where each student fell in terms of his or her starting point (horizontal x-axis) and his or her outcome score (vertical y-axis). Each point represents one student's post-intervention score against his or her pre-intervention score. The darker points represent *SFScience* students; the lighter points, control students. The shaded area in the lower right of the graph is the area of negative change (i.e., where students lost ground).

The two lines are the predicted values on the posttest for students in the *SFScience* and control conditions as determined using the estimated fixed effects in the model. The fact that these lines are very close together represents the finding that the outcomes for students in *SFScience* and control groups were very similar.

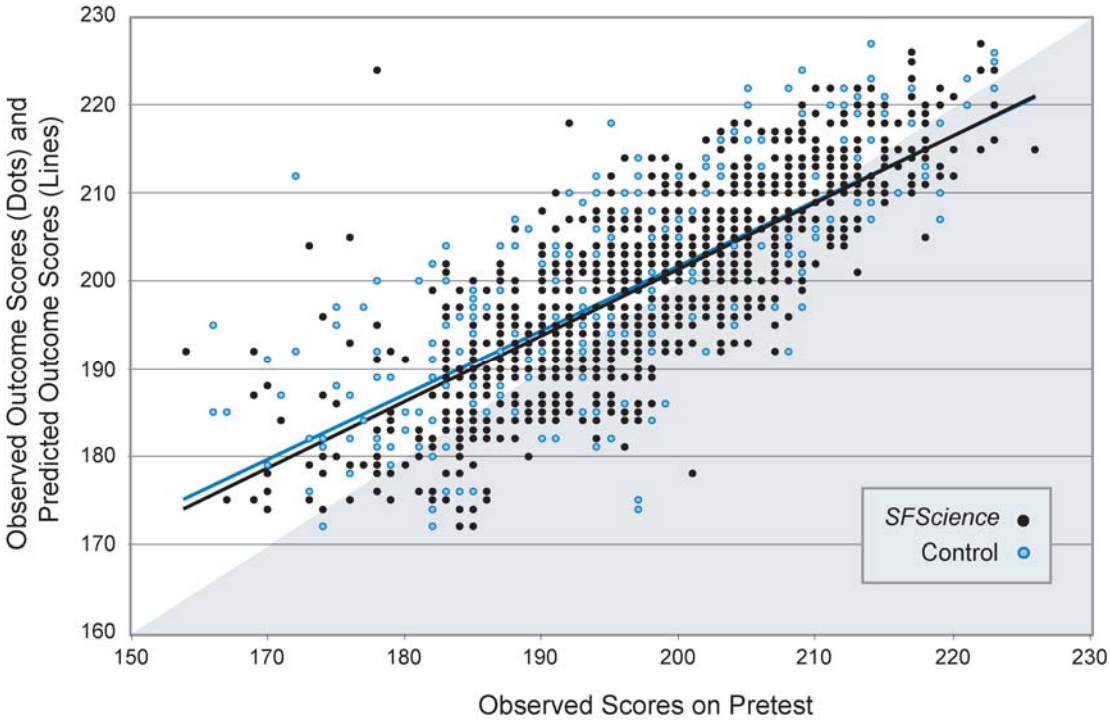

Figure 8. Comparison of Predicted and Actual Outcomes for *SFScience* and Control Group Students (Science Achievement)

Analysis Including Gender as a Moderator

In addition to looking at the main effect of *SFScience,* we estimated the interactions of *SFScience* with the pretest scores and gender of the students. In particular, we were interested in whether the condition's effect was differentially effective for males and females because much of the research literature indicates that gender differences exist in students' performance on science outcomes. Table 36 shows the moderating effect of gender on students' performance on NWEA Science.

Table 36. Moderating Effect of Gender on Science Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Average outcome for girl in control group	198.9	2.04	43	97.32	<.01
Average *SFScience* effect for girls	0.21	0.55	43	0.38	.71
Predicted change in outcome for each unit increase on the pretest	0.73	0.02	1960	46.56	<.01
Difference (boys minus girls) in average performance in the control condition	0.93	0.36	1960	2.6	.01
Difference (boys minus girls) in the average *SFScience* effect	-1.04	0.50	1960	-2.08	0.04
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	2.5	1.04		2.4	.01
Within-teacher variation	30.81	0.98		31.34	<.01

[a] All of these values apply to a student with an average score on the pretest

These results show two effects; first, that there is a strong effect of gender on science achievement. That is, not considering *SFScience*, boys in the control group did significantly better than girls. Second, there is an interaction between gender and *SFScience* such that the performance of boys and girls was drawn even under the treatment condition.

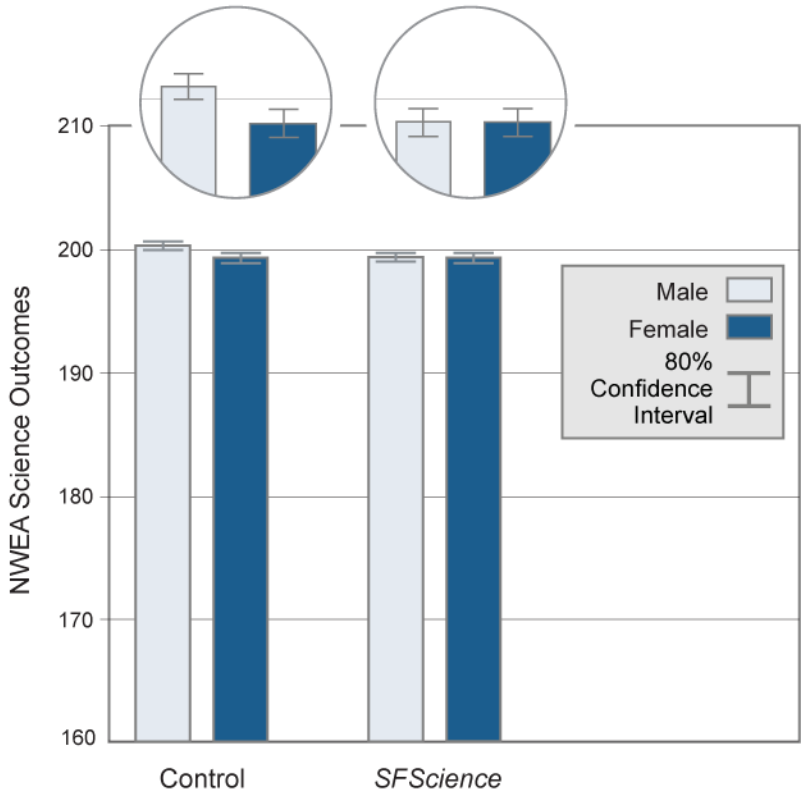

Figure 9. Results for Male and Female Students in the Control and *SFScience* Conditions

Analysis Including Teaching Experience as a Moderator

We also considered whether the treatment impact is differentially effective for students who had relatively inexperienced teachers (3 years or fewer) versus those with more experienced teachers (4 or more years). Table 37 shows the moderating effect of years of teaching experience on students' performance on NWEA Science. There is no difference in the value of the program across levels of experience.

Table 37. Moderating Effect of Teaching Experience on Science Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Average outcome for a student with an experienced teacher in the control condition	199.27	2.17	39	91.76	<.01
Average *SFScience* effect for a student with an experienced teacher	-0.26	0.59	39	-0.44	.66
Predicted change in control outcome for each unit increase on the pretest	0.72	0.02	1814	42.96	<.01
Difference (score for student with an inexperienced teacher minus score for student with an experienced teacher)in performance in the control condition	-0.93	2.08	39	-0.45	.66
Difference (score for student with an inexperienced teacher minus score for student with an experienced teacher) in the average *SFScience* effect	-0.70	1.72	39	-0.41	.69
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	2.87	1.25		2.29	.01
Within-teacher variation	33.01	1.09		30.19	<.01

[a] All of these values apply to a student with an average score on the pretest

Comparison of Results Across Locations

Figure 10 compares the overall results for student science achievement across the five sites. We conducted an analysis to confirm that the sites were similar enough to combine into a single analysis. As a result we were able to produce a point-and-whiskers plot showing estimates of the impact of *SFScience* on science outcomes separately for each site as well as in combination[3]. The estimates are the center points within each interval. Each interval is an 80% confidence interval. That is, if we consider each site separately, we can be 80% sure that the true value of the impact lies within the interval shown. All of the intervals cross zero, including the one for the overall impact. This is consistent with the finding of no impact described in Table 35.

[3] These are expressed in terms of the NWEA Science scale, not as standardized effect sizes.

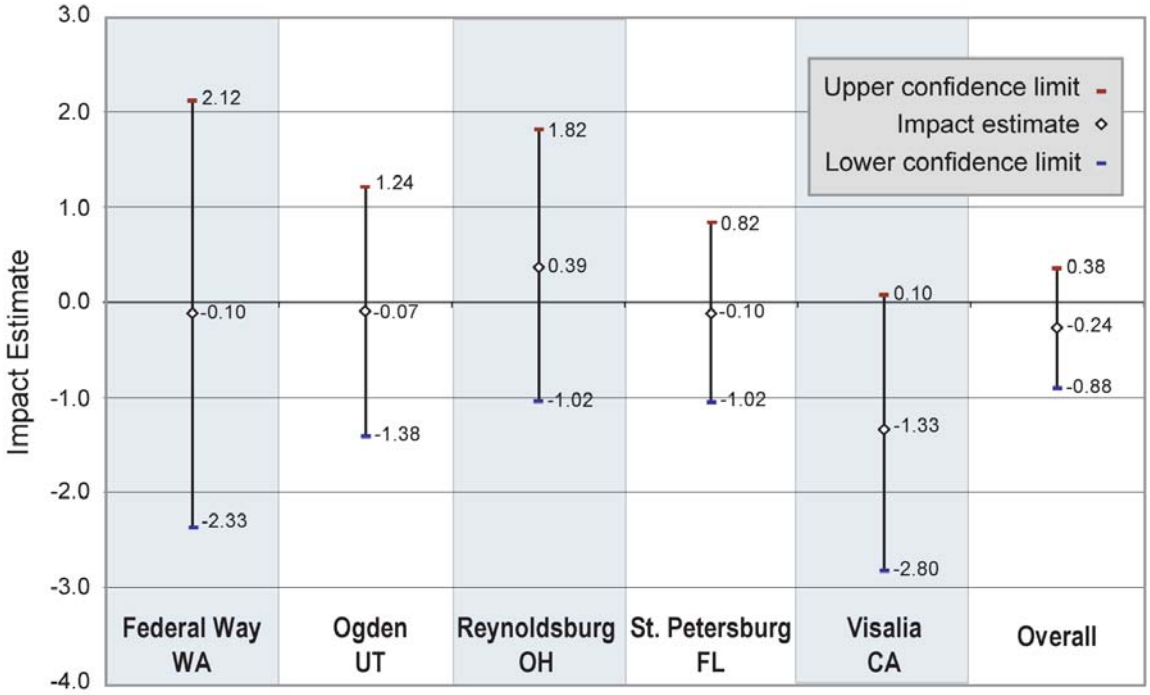

Figure 10. Estimated Science Impacts Across Districts

Reading Outcomes

Analysis Including Pretest

Our next set of analyses addresses reading achievement as measured by NWEA Reading. Table 38 provides a summary of the sample we used in the analyses and the results for the comparison of *SFScience* and control group performance for reading. The interpretation of this table is the same as for Table 32. The information for the adjusted effect size is based on the sample that we use in the analyses on which we base our results reported in Table 41 through Table 43. The means, and therefore the effect size, are adjusted to take into account the student pretest scores.

Table 38. Overview of Sample and Impact of *SFScience* on Reading Achievement

	Condition	Means	Standard[a] deviations	No. of students	No. of classes	No. of teachers	Effect size	*p* value[b]
Un-adjusted	***SFScience***	205.26	15.13	1086	56	46	.02	.61
	Control	204.94	13.79	1068	57	46		
Adjusted	***SFScience***	205.69	14.93	934	55	45	.05	.26
	Control	205.01[c]	13.61	974	57	46		

a The standard deviations used to compute the adjusted and unadjusted effect sizes are computed from the scores of the students in the sample for that row

b The *p* value for the unadjusted effect size is computed using a model that includes clustering of students in schools and pairs but no other covariates. The *p* value for the adjusted effect size is computed using a model that controls for clustering and includes the pretest covariate, as well as fixed effects when needed.

[c] The modeling of fixed effects for upper level units leads to unit-specific estimates of performance in the absence of treatment. For purposes of display, to set the performance estimate for the control group, we compute the average performance for the controls used to calculate the adjusted effect size. The estimated treatment effect is added to this estimate to show the relative advantage or disadvantage to being in the treatment group

Figure 11 provides a visual representation of specific information in Table 38. The bar graphs represent average performance using the metric of NWEA Reading.

The panel on the left shows average pre- and posttest scores for the control and *SFScience* groups. The pre- and posttest bars show that both the *SFScience* and control groups on average grew in their reading achievement during the year.

The panel on the right shows estimated performance on the posttest for the two groups based on a model that adjusts for students' pretest scores and other fixed effects (i.e., this is a visual display of results from the row labeled 'adjusted' in Table 38). The overall impact on reading as an effect size (standard deviation units) is 0.05 which is equivalent to a gain of at least one percentile point if the median control student were placed in the *SFScience* group. The *p* value indicates that there is a .26 probability that the difference is due to chance.

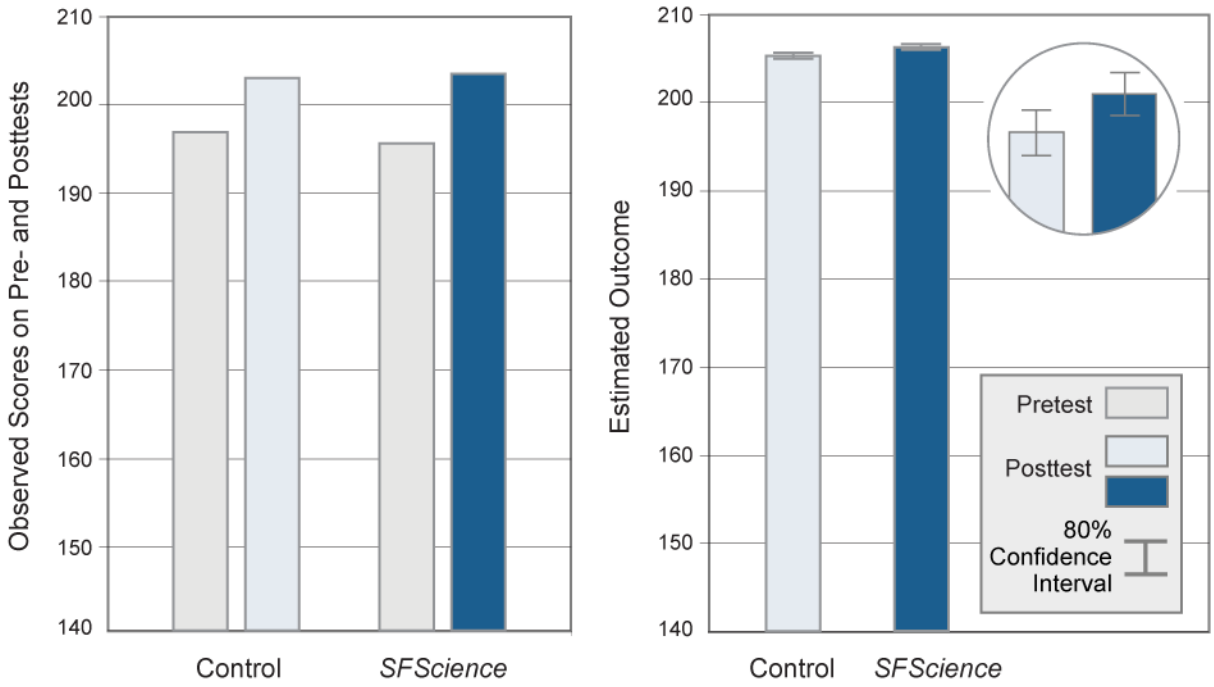

Figure 11. Impact on Reading Achievement: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)

Table 39. Difference Made by a School Year of Science Instruction for the Control Group

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Score on the pretest	204.69	1.22	45	167.61	<.01
Gain made over the year	4.32	0.5241	45	8.24	<.01

[a] Teachers were modeled as random effects.

Note. In obtaining the average pretest score, we did not model school and assignment pair as fixed effects. However, this model provides similar standard error and gain estimates over the year and with the standard fixed effects model.

Table 40. Difference Made by a School Year of Science Instruction for the *SFScience* Group

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Score on the pretest	204.04	1.50	45	136.13	<.01
Gain made over the year	4.23	0.54	44	7.83	<.01

[a] Teachers were modeled as random effects.

Note. In obtaining the average pretest score, we did not model school and assignment pair as fixed effects. However, this model provides similar standard error and gain estimates over the year and with the standard fixed effects model.

Analysis Including Pretest as a Moderator

We now report on the analyses that examine not just the overall impact of *SFScience* on reading but also the moderating effects of other variables. We begin by examining the moderating effect of the prior score. Table 41 shows the estimated impact of *SFScience* on students' performance in reading as measured by NWEA Reading as well as the moderating effect of the prior score.

Table 41. Impact of *SFScience* on Reading Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Predicted value for a control student with an average pretest	206.62	3.49	42	59.26	<.01
Impact of *SFScience* for a student with an average pretest	0.65	0.6	42	1.1	.28
Predicted change in control outcome for each unit increase on the pretest	0.81	0.02	1802	49.87	<.01
Interaction of pretest and *SFScience*	-0.02	0.02	1802	-1.08	.28
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	3.57	1.42		2.51	.01
Within-teacher variation	36.83	1.22		30.08	<.01

[a] Pairs of teachers used for random assignment are also modeled as a fixed factor but not included in this table

[b] Teachers were modeled as a random factor.

The row in Table 41 labeled "Impact of *SFScience* for a student with an average pretest" tells us whether *SFScience* made a difference in terms of student performance on NWEA Reading for a student who has an average score on the pretest. The estimate associated with *SFScience* is

0.65. This shows a small positive effect associated with *SFScience*. However, the *p* value of .28 gives us no confidence that the actual effect is different from zero.

We also estimated the moderating effect of the pretest score on the impact of *SFScience* to see whether it was differentially effective for students at different points along the pretest scale. The *p* value for this effect is .28. We have no confidence that the true effect is different from zero.

As a visual representation of the results described in Table 41, we present a scatterplot in Figure 12, which shows student performance at the end of the year in reading, as measured by the NWEA test, against their performance in the fall. These graphs show where each student fell in terms of his or her starting point (horizontal x-axis) and his or her outcome score (vertical y-axis). Each point represents one student's post-intervention score against his or her pre-intervention score. The darker points represent *SFScience* students; the lighter points, control students.

The two lines are the predicted values on the posttest for students in the *SFScience* and control conditions as determined using the estimated fixed effects in the model. Consistent with the results described above, we see that *SFScience* and the programs used in the control classrooms were equally effective as measured by NWEA Reading.

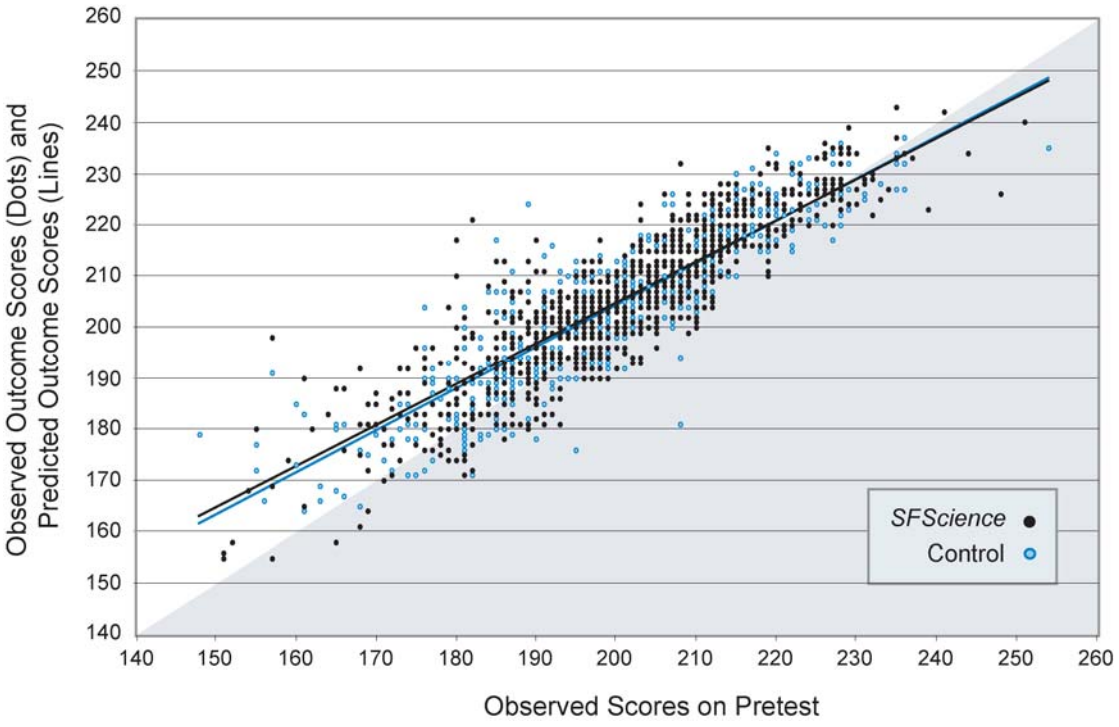

Figure 12. Comparison of Predicted and Actual Outcomes for *SFScience* and Control Group Students

Analysis Including Gender as a Moderator

As with science, we estimated the interactions of *SFScience* with the reading pretest and gender of the students. In particular, we were interested in whether the condition's effect was differentially effective for males and females. Table 42 shows that there is no difference between boys and girls in reading and no differential effect of *SFScience* depending on gender.

.

Table 42. Moderating Effect of Gender on Reading Achievement

Fixed effects	Estimate	Standard error	DF	*t* value	*p* value
Average outcome for girl in control group	206.49	3.48	42	59.26	<.01
Average *SFScience* effect for girls	0.87	0.66	42	1.32	.19
Predicted change in outcome for each unit increase on the pretest	0.80	0.01	1801	67.89	<.01
Difference (boys minus girls) in average performance in the control condition	0.08	0.4	1801	0.21	.84
Difference (boys minus girls) in the average *SFScience* effect	-0.38	0.57	1801	-0.66	.51
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	3.55	1.42		2.50	.01
Within-teacher variation	36.86	1.23		30.07	<.01

[a] All of these values apply to a student with an average score on the pretest

Analysis Including Teaching Experience as a Moderator

We also considered whether *SFScience* is differentially effective for students who had relatively inexperienced teachers (3 years or fewer) versus those with more experienced teachers (4 or more years). Table 43 shows the moderating effect of years of teaching experience on students' performance on NWEA Reading. There is no difference in the value of the program across levels of experience.

Table 43. Moderating Effect of Teaching Experience on Reading Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Average outcome for a student with an experienced teacher in the control condition	206.47	3.58	40	57.6	<.01
Average *SFScience* effect for a student with an experienced teacher	0.72	0.69	40	1.05	.30
Predicted change in control outcome for each unit increase on the pretest	0.8	0.01	1803	68.11	<.01
Difference (score for student with an inexperienced teacher minus score for student with an experienced teacher)in performance in the control condition	-1.01	1.59	40	-0.63	.53
Difference (score for student with an inexperienced teacher minus score for student with an experienced teacher) in the average *SFScience* effect	0.02	1.67	40	0.01	0.99
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	3.86	1.55		2.49	.01
Within-teacher variation	36.83	1.22		30.09	<.01

[a] All of these values apply to a student with an average score on the pretest.

Comparison of Results Across Locations

Figure 13 compares the overall results for student Reading achievement across the five sites. We conducted an analysis to confirm that the sites were similar enough to combine into a single analysis. As a result we were able to produce a point-and-whiskers plot showing estimates of the impact of *SFScience* on reading outcomes separately for each site as well as in combination. The estimates are the center points within each interval. Each interval is an 80% confidence interval. That is, if we consider each site separately, we can be 80% sure that the true value of the impact lies within the interval. Unlike the science outcomes, two of the locations show positive results within the 80% confidence interval shown. Overall, the results are positive which is consistent with the findings from Table 41.

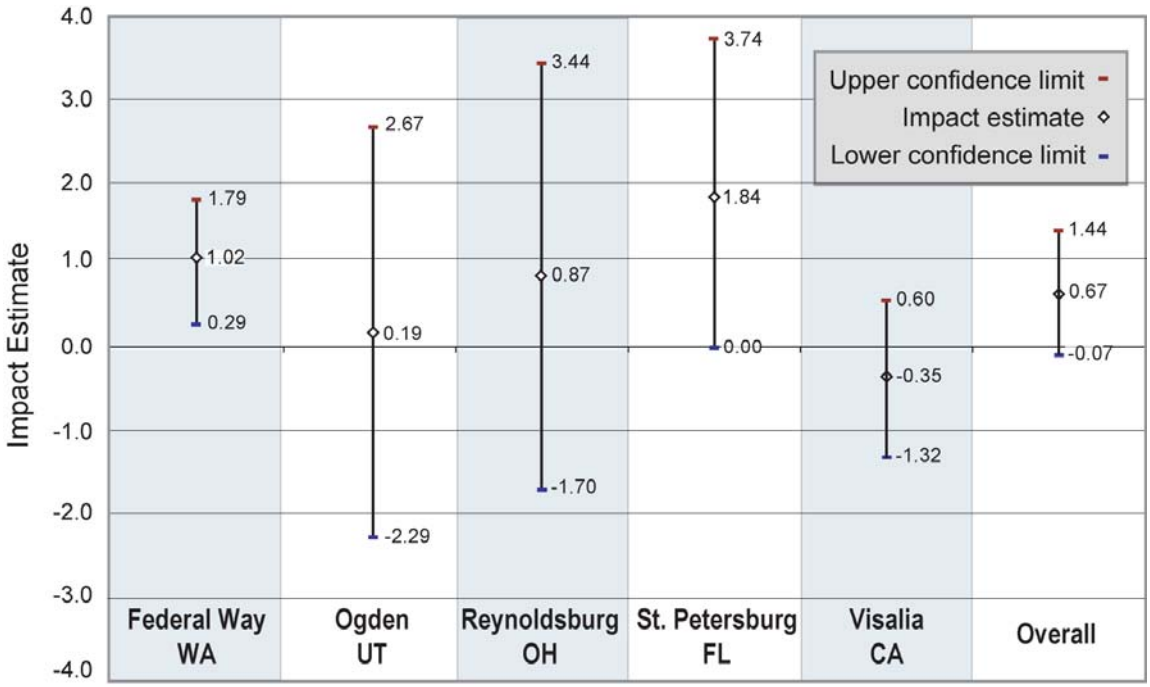

Figure 13. Estimated Reading Impacts Across Districts

Classroom Process and Science Achievement

We also considered a number of measures from the classroom. These processes are potentially outcomes of *SFScience* as well as related to the student achievement outcome. As described in the implementation results section of this report, we measured the amount of instructional time that teachers devoted to science and the extent to which inquiry teaching took place.

When dealing with implementation variables, we can understand them as defining a distinct path or link between the intervention and student-level achievement, as illustrated in Figure 14. Part of the impact of *SFScience* on student outcomes may be mediated by the intermediate variables. *SFScience* can have a direct impact on both student outcomes and on instructional time, a teacher-level outcome. The link from instructional time to the student outcome is correlational but an important relationship to explore.

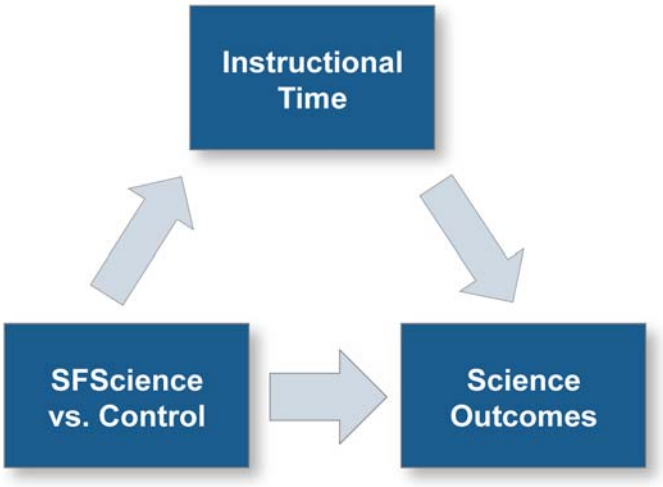

Figure 14. Relationships for Exploratory Analysis of Implementation Variables

Instructional Time

We wanted to explore the relationship between how much time was spent teaching science and science achievement. The surveys provided data on this variable. Our measure is the total hours spent teaching science during the experiment. Instructional time was measured by each teacher's self-report of the number of minutes she or he spent using *SFScience* per week. Results were averaged across eight surveys that were administered every two weeks and adjusted for the number of weeks of implementation at that site.

We look first at the impact of condition on instructional time. Table 44 shows that there was no difference between the two groups of teachers on the amount of instructional time. The high *p* value of .95 gives us no confidence that the actual difference is different from zero.

Table 44. Impact of *SFScience* on Hours of Science Instruction Time

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept	14.70	17.80	31	0.83	.42
Impact of *SFScience*	-0.30	4.67	31	-0.06	.95
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Residual teacher variance	316.67	80.43		3.94	<.01

[a] Pairs of teachers used for random assignment are also modeled as a fixed factor but not included in this table

Even with no difference between the two groups, it is useful to explore whether there is a relationship between amount of science instruction time and student achievement. The result of this analysis is purely correlational; we have not assigned teachers to levels of instructional time with *SFScience*, therefore we cannot be sure whether it is instructional time or some other variable correlated with instructional time (e.g., teacher enthusiasm) that is the true cause of the

student outcome. A test of the correlation, however, reveals no relationship between *SFScience* usage and the student outcome. Table 45 gives us a high *p* value and no confidence that the difference related to instructional time is different from zero.

Table 45. Relationship of Instructional Time to Student Outcome

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Predicted value for student with an average pretest	22.76	15.8	29	1.44	.16
Predicted change in outcome for each pretest point	0.90	0.08	29	11.49	<.01
Predicted change in outcome for hour of science time	0.01	0.02	29	0.34	.74
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Residual	2.36	0.62		3.81	<.01

[a] Districts, schools, and pairs of teachers used for random assignment are also modeled as a fixed factor but not included in this table

Inquiry

Because *SFScience* emphasizes the inquiry processes, we asked similar questions about this process variable. As with instructional time, there is a more complete discussion in the implementation results section. Our question here is whether *SFScience* resulted in a greater amount of inquiry teaching. Table 46 shows the result of this analysis. We have no confidence that *SFScience* teachers were more likely to use inquiry teaching techniques.

Table 46. Mixed Model Estimating the Impact of *SFScience* on Inquiry

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept	7.94	16.38	31	0.48	.63
Impact of *SFScience* on inquiry teaching	-0.44	4.16	31	-0.11	.92
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Residual teacher variance	250.89	63.73		3.94	<.01

[a] Pairs of teachers used for random assignment are also modeled as a fixed factor but not included in this table

The final process question we asked was whether inquiry instruction was related to outcome. As shown in Table 47, we found no relationship between inquiry and student science outcome.

Table 47. Relationship of Inquiry to Student Outcome

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Predicted value for student with an average pretest	20.29	18.59	30	1.09	.28
Predicted change in outcome for each pretest point	0.91	0.09	30	9.88	<.01
Predicted change in outcome for each percent of inquiry teaching	-0.02	0.02	30	0.90	.38
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Residual	3.20	0.83		3.87	<.01

[a] Districts, schools, and pairs of teachers used for random assignment are also modeled as a fixed factor but not included in this table

Discussion

We began this research with the question whether *Scott Foresman Science* was as effective as or more effective than the existing programs it was being compared to. The question applied both to science achievement as well as to whether the science program made a measurable difference in reading achievement beyond the growth resulting from the core reading program.

Overall, we found no difference between the science scores of students taught using *SFScience* as compared to the established program. In reading we found positive effects in two of the five districts involved in the study. In both science and reading, students made progress over the course of the year with progress in reading more robust than in science.

There is no evidence that *SFScience* improved science achievement beyond what the schools could expect from their regular program. This was a consistent finding across the five districts. For reading, the picture is somewhat different. In two districts, there was a discernible difference between classrooms that used *SFScience* and those that did not. This finding means that *SFScience* caused a small increase in reading above and beyond the gains that we would expect to observe for the schools' reading program by itself. When all the sites are combined, this difference, while still positive, is not strong enough to give us confidence that the difference did not occur simply by chance (there is a 26% chance that the difference was not caused by *SFScience)*. These results suggest that under some conditions the science program can help the school with student reading achievement.

We investigated several factors other than condition that may have contributed to the outcome. We found the same result across grade levels and across prior student achievement levels. We also found that teacher experience (dividing teachers between those with fewer than four years of teaching and the more experienced teachers) had no impact on the results. We did find that, while in the control group boys outperformed girls in science, this was not the case in *SFScience* classes, where there was no difference in performance between boys and girls.

We analyzed two aspects of classroom process—the amount of instructional time or opportunity to learn science and the amount of inquiry science teaching. We were interested both in whether *SFScience* classrooms experienced more of either of these and whether classes with a greater opportunity to learn or with more inquiry science activity would also show higher science achievement. We found no differences in achievement related to these classroom processes.

A finding of no difference does not mean that the product is ineffective. It is important to interpret these results in relation to what teachers were using in the control condition classrooms where science and reading were also being taught. Overall, it appeared that the schools were more successful in teaching reading than in teaching science, reflecting the relative emphasis schools place on the two subjects. In one district, for example, science instruction was suspended for a period to give teachers more time to teach reading. It is also relevant that this was the first year of use of *SFScience* and the teachers' initial unfamiliarity may have had an effect on implementation. *SFScience* appears to be equally robust across the grades tested and across levels of teacher experience, and for male and female students.

We designed the set of five experiments to provide useful information to the participating districts. In each district, the implementation and results were somewhat different, as we observe from the graphs comparing science and reading scores across the districts. But even with five diverse districts, this report is not intended to provide widely generalizable results and the reader should consult the individual report of the district most similar to his or her own to consider the applicability of the findings. Overall our conclusion is that *SFScience* stands up to other science programs in schools and science learning in general will benefit from greater emphasis in the elementary grades. Educators may find the program attractive in the equal help it gave to boys and girls compared to the programs in place. The reading component was shown capable of improving reading achievement under some conditions. The result for reading points to a potentially important strength of the program. Additional research focused on that question and with greater attention to implementation of the reading components will increase our confidence in these findings.