Empirical Education

RESEARCH REPORT

Comparative Effectiveness of *Scott Foresman Science*:

A Report of a Randomized Experiment in St. Petersburg Catholic Schools

Gloria I. Miller
Andrew Jaciw
Minh-Thien Vu
Empirical Education Inc.

June 18, 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 St. Petersburg Catholic Schools for their assistance and cooperation in conducting this research and for providing access to their data under an agreement with Empirical Education Inc. The research was sponsored by Pearson Education, which provided Empirical Education Inc. with independence in reporting the results.

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 a Randomized Experiment in St. Petersburg Catholic Schools

Table of Contents

INTRODUCTION 1
METHODS 1
RESEARCH DESIGN 1
INTERVENTION 1
Scott Foresman Science Materials 2
Table 1. Scott Foresman Supplied Materials 2
District Science Materials 3
SITE DESCRIPTIONS 3
St. Petersburg Catholic Schools, FL 3
Table 2. St. Petersburg Racial Makeup 3
SAMPLE AND RANDOMIZATION 3
Recruiting 3
Randomization 4
Table 3. Participating Teachers at St. Petersburg Site 5
Sample Size 5
DATA SOURCES AND COLLECTION 5
Observational and Interview Data 5
Survey Data 5
Table 4. Survey Response Rates 6
Achievement Measures 6
Testing Schedule 7
STATISTICAL ANALYSIS AND REPORTING 7
RESULTS 8
FORMATION OF THE EXPERIMENTAL GROUPS 8
Groups as Initially Randomized 8
Table 5. Distribution of the SFScience and Control Groups by Schools, Teachers, Grades, and Counts of Students 9
Years of Teaching Experience 9
Table 6. Years Teaching Experience 9
Table 7. Distribution of Years Teaching Experience 9
Table 8. Years Teaching in Grade Level (not necessarily consecutive) 10
Table 9. Years Teaching Science 10
Table 10. Science Coursework in College 10
Table 11. Recent Professional Development (PD) for Science Instruction 11
Post Randomization Composition of the Experimental Groups 11
Table 12. Test Data for Students in SFScience and Control Groups 11
Student Variables 11
Table 13. Ethnicity for SFScience and Control Groups 12
Table 14. Gender for SFScience and Control Groups 12
Characteristics of the Experimental Groups Defined by Pretest 12
Table 15. Difference in Pretest Scores Between Students in the SFScience and Control Groups 13
ATTRITION AFTER THE PRETEST 13

IMPLEMENTATION RESULTS 13
Comparison of *SFScience* and Control Groups 13
Classroom Settings for Instruction 13
Opportunities for Learning 14
Control Materials 15
Table 16. Primary Sources for Science Instruction 15
Table 17. Percentage of Time Devoted to Hands-on Science Activities 16
Planning Time 16
Density of Science Inquiry Reflected in the Classroom 16
Implementation of *SFScience* 17
Training and Support 17
Availability and Use of Materials 18
Table 18. Percent of Teachers Covering Each Chapter in Unit A-Life Science 18
Table 19. Chapters in Unit B-Earth Science Covered 18
Table 20. Chapters in Unit C-Physical Science Covered 18
Table 21. Chapters in Unit D-Space & Technology Covered 19
Table 22. For Unit A, How Well Was the Content Aligned to State Standards? 19
Table 23. For Unit B, How Well Was the Content Aligned to State Standards? 19
Table 24. For Unit C, How Well Was the Content Aligned to State Standards? 20
Table 25. For Unit D, How Well Was the Content Aligned to State Standards? 20
Rating the Level of Implementation 21
Summary of Implementation 21
QUANTITATIVE IMPACT RESULTS 21
Science Outcomes 22
Analysis Including Pretest 22
Table 26. Overview of Sample and Impact of SFScience on Science Achievement 22
Figure 1. Impact on Science Achievement: Unadjusted Pre- and Posttest Means for Control and SFScience (Left); Adjusted Means for Control and SFScience (Right) 23
Analysis Including Pretest as a Moderator 23
Table 27. The Impact of SFScience on Science Achievement 24
Figure 2. Comparison of Estimated and Actual Outcomes for SFScience and Control Group Students 25
Analysis Including Gender as a Moderator 25
Table 28. Moderating Effect of Gender on Science Achievement 26
Reading Outcomes 26
Analysis Including Pretest 26
Table 29. Overview of Sample and Impact of SFScience on Reading Achievement 27
Figure 3. Impact on Reading Achievement: Unadjusted Pre- and Posttest Means for Control and SFScience (Left); Adjusted Means for Control and SFScience (Right) 28
Analysis Including Pretest as a Moderator 28
Table 30. The Impact of SFScience on Reading Achievement 29
Figure 4. Comparison of Estimated and Actual Outcomes for SFScience and Control Group Students 30
Classroom Process and Science Achievement 30
Figure 5. Relationships for Exploratory Analysis of Implementation Variables 31
Instructional Time 31

Table 31. The Impact of SFScience on Hours of Science Instruction Time 31
Table 32. Relationship of Instructional Time to Student Outcome 32
DISCUSSION 33

Introduction

Pearson Education contracted with Empirical Education Inc. to conduct five randomized experiments to determine the effectiveness of its *Scott Foresman Science* products curriculum and associated materials. This research project consists of a randomized experiment in St. Petersburg Catholic Schools.

The question being addressed by the research is whether *Scott Foresman Science* is as effective as the current curricula being used by the participating campuses of the Roman Catholic Diocese of St. Petersburg . The research focuses on 3rd, 4th, and 5th grade students. The outcomes were measured by student achievement on standardized tests administered at the beginning and end of the project. Two test areas were selected as the outcome measures: the Northwest Association's Science Concepts and Processes, and Reading Achievement.

The overall comparison between Scott Foresman Science (*SFScience*) and the programs used in the control classrooms was just 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 may help to explain the results.

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 a program — in this case, *Scott Foresman Science* — or to continue using their current teaching materials and methods.

Random assignment to experimental conditions does not, however, 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. This report provides 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 materials used in the district (control group). Teachers volunteered for participation and from the pool of volunteers, the researchers 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

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 by the teachers and 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.

The publisher provided a one-half day workshop to familiarize the treatment teachers with the curriculum and discuss the implementation expectations. All *SFScience* teachers agreed to:

- Complete two units of instruction with at least one Full Inquiry module (student designed investigation)
- Complete one unit assessment
- Use the Leveled Readers
- Use the Science Kit materials for hands-on inquiry

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.

Scott Foresman Science Materials

The *SFScience* teachers were supplied with the following materials specific to their grade level:

Table 1. Scott Foresman Supplied Materials

Teacher Materials (one each unless otherwise specified)	Student Materials (one for every 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	
Lesson Modeling Videos (Activity DVD)	

District Science Materials

All teachers had textbooks for students. Several different texts were in use: older versions of Scott Foresman, Discovery Works and Horizons both by Silver Burdett Ginn, 1999 (now Houghton Mifflin), SRA Science Laboratory by McGraw Hill, Harcourt Brace Science 2002, and Destinations in Science by Addison Wesley. Teachers had a variety of laboratory materials that included materials designed to accompany the texts and those created by the teachers.

Site Descriptions

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.

Table 2. St. Petersburg Racial Makeup

Race/Ethnicity	% of Population
White	71.36
African American	22.36
American Indian/Native Alaskan	0.31
Asian	2.67
Native Hawaiian or Pacific Islander	0.05
Other Race	1.07
Two or more Races	2.17
Hispanic Origin (of any race)	4.23

Source: All population data including racial/ethnic categories and breakdown are excerpted from the 2000 U.S. Census and 2003/04 projections

St. Petersburg Catholic Schools incorporates the five counties surrounding the St. Petersburg area: Pinellas, Hillsborough, Pasco, Hernando, and Citrus. They operate a total of 29 elementary, 2 special purpose schools, 13 early childhood centers, and 6 high schools; six elementary schools participated in this study. Since these schools are private, additional demographic information is not available for the schools.

Sample and Randomization

Recruiting

We held a phone conference on June 28, 2005 with district administrators and Principals to explain the details and procedures of the study. The district leadership and the Principals provided us with background information for each the six schools so that researchers could identify and prioritize the criteria for matching grade levels before randomization. A second phone conference was held with the district leadership only to conduct the randomization. Principals were informed of the outcomes of the randomization via email. Principals then identified eligible teachers, who were contacted by the researchers via email and asked to participate in the study. Twenty-two teachers volunteered to participate by the first week of September, 2005. Approximately three weeks later one teacher moved away from district for personal reasons and no longer participated in the study.

Randomization

The unit of randomization at this site was the grade-level. Matched grade-level pairs were formed among the participating schools and a coin was tossed to determine assignment, either treatment (using *SFScience*) or control (using district identified materials). The entire grade-level at any one school participates as either *SFScience* or control.

The purpose of randomization is to establish statistically equivalent groups i.e., groups between which there exist only chance differences and that can be compared fairly. There are various ways to randomize participants to conditions. Usually it is not feasible to randomize students so we randomize classes or teachers. In this district teachers tend to work together in grade-level teams, therefore, in order to not disrupt the processes of teaming within grade levels, we randomized whole grade-level teams. That is, all the students at a particular grade level in each school were assigned either to treatment or control.

A variant of the straight randomization process is a matched pairs design whereby we first pair similar units and then randomize one to treatment and the other to control. A pairing strategy will often result in a more precise measurement of the treatment impact. Our randomizing strategy reflected our belief that school differences matter, therefore we randomized within schools whenever it was possible. That is we paired adjacent grade level within schools and randomized one to each condition (this was done for six pairs.) Where this was not possible, we randomized pairs of classes from the same grade level across schools (this was the case for three pairs.) As a result of the randomization, there were two more classes in the control condition than in the treatment condition in third grade, four more in the treatment condition than in the control condition in fourth grade, and two more in the control condition than in the treatment condition in fifth grade.

This is a potential source of bias for some of the analyses. For example, the estimate of the average impact of the intervention may reflect the fact that the treatment group has more of a certain grade level than the control group. Importantly, however, the extra number of treatment classes in fourth grade is offset by the two extra control classes in each of third and fifth grades. Assuming that growth in achievement is linear from third to fifth grade (the growth from third to fourth grade has the same magnitude as the growth from fourth to fifth), which is a reasonable assumption, bias due to imbalance in grades will be minimal. Further, we model the pretest score on a scale that is vertically aligned, which adjusts for the imbalance on grade levels. In analyses involving the interaction of the prior score with the treatment condition; that is, where we look to see if there is a differential effect of treatment across grades, this imbalance is not an issue because we are looking at the overall shapes of the scatterplots of post- versus pre-test scores. The imbalance changes the densities of the scatterplot, but not their shape, which preserves the lines of interest that we draw.

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 years of teaching experience – a factor that presumably affects the outcome.

The total numbers of participating teachers are displayed in the table below. In some of the 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.

Table 3. Participating Teachers at St. Petersburg Site

Teacher Assignment Status	Number Participating
SFScience	9
Control	13
Total	**22**

Note: One control teacher left the district after the third week of the study and was considered a non-participant. Subsequent tables will indicate the adjusted control teacher count.

Because specialization causes some teachers to have more than one group of students for instruction, the number of classes involved in this study exceeds the total number of teachers participating. There were a total of 17 classrooms assigned to the control condition and 14 classrooms assigned to the *SFScience* condition. No individual teacher taught more than three classes of science.

Sample Size

Sample size is one of the things that determine how precisely we can measure an effect of a given size. With smaller samples we are usually only able to detect larger effects. We often 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.

Based on the available sample size, and certain assumptions about other parameters that affect the size of the effect that we can detect, we calculated that we can detect an effect size as small as 0.47. This is computed assuming false-positive and false-negative error rates of .05 and .20, respectively. Raising the false positive rate to .20 reduces the size of the effect that we can detect to 0.34. We emphasize that the matching design that we used likely further lowers this value. From this we see that the experiment is not powered to detect a very small effect which may be real but not discernable given the number of teachers in the study.

Data Sources and Collection

In addition to the quantitative data we also collected qualitative data. The data are collected over the entire period of the experiment beginning with the initial phone conference 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.

Observational and Interview Data

In general, observational data are used to inform the description of the learning environment, instructional strategies employed by the teachers, and student engagement. These data are minimally coded. Our observation of the initial training in the use of *Scott Foresman Science* materials was conducted on September 14th, 2005. Classroom observations were conducted during the week of March 28th. In all, 18 teachers were observed, all nine in the *SFScience* and nine control group teachers.

Interview data is used to elaborate survey responses, characterize the teacher's schedule, and to provide descriptions of the overall experience teaching with the *Scott Foresman Science* curriculum. Short interviews of both groups were conducted throughout the timeframe of the study.

Survey Data

Surveys were deployed to both *SFScience* and control group teachers beginning on December 5, 2005 and continuing on a bi-weekly basis until late May 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 9 teachers in the *SFScience* group and 12 teachers in the control group. All response rates were calculated based on these expectations. Table 4 summarizes the topics and response rate by survey number. A total of nine surveys were deployed with an overall response rate of 87.83% for both groups, an 93.83% response rate for the *SFScience* teachers, and a 83.33% response rate for the control teachers.

Survey data are used to quantify the extent of exposure to the materials (opportunities to learn with the curriculum). In an effort to collect data equally from both groups, we sent the same survey to all of the teachers on all but one occasion. In Survey 9. the final survey, the topics were modified to allow for the differences between the learning environments across the two groups. Survey 9 focused on the content covered and teachers' overall experience with the various materials.

The quantitative survey data are analyzed using descriptive statistics; these are summarized by individual teacher and by assignment group (*SFScience* and control), and are compared by group. The free-response portions of the surveys are minimally coded.

Table 4. Survey Response Rates

Survey number	Date	Topic	*SFScience* response rate	Control response rate	Overall response rate
Survey 1	Dec. 5 - 9	Science Schedule & Instructional Time	66.67%	66.67%	66.67%
Survey 2	Jan. 16 - 20	Resources	100%	100%	100%
Survey 3	Jan. 23 - 27	Interactions with materials/Students	100%	83.33%	90.48%
Survey 4	Feb. 6 - 10	More Interactions	100%	100%	100%
Survey 5	Feb. 20 - 24	Time & Preparation	100%	91.67%	95.24%
Survey 6	Mar. 6 - 10	Materials & Resources	100%	91.67%	95.24%
Survey 7	Mar. 20 - 24	Assessments	100%	83.33%	90.48%
Survey 8	May 1 - 5	More Interactions	88.89%	100%	95.24%
Survey 9T[a]	May 26	Final Survey	88.89%	N/A	88.89%
Survey 9C[b]	May 26	Final Survey	N/A	33.33%	33.33%

[a] Asked only of *SFScience* teachers.
[b] Asked only of control teachers.

Achievement Measures

The primary outcome measures are student-level test scores on the Northwest Evaluation Association (NWEA) test in two areas: Science Concepts and Processes and Reading. We refer to these tests when reporting Science Achievement and Reading Achievement 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 placement tests, referred to as locator tests. During the second and subsequent administrations, the student is automatically assigned to a level based on previous results. Teachers were provided a one hour review of the testing procedures and given a Proctor manual. Researchers provided additional support by pre-packaging all testing materials on an individual teacher basis.

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 range from about 150 to 300 and indicate a student's current achievement level along a curriculum scale for a particular subject. 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 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.

Testing Schedule

The pretests were given between October and December, 2005 and posttest was given May, 2006 using the same tests with placements provided by the NWEA for all of those students having pretest results. Any newly enrolled student was administered the locator test followed by the appropriate leveled test if they were enrolled within the pretesting period. Students that came into either the *SFScience* or control condition after the pretesting period were not considered subjects in the study because they lacked pretest scores.

There were no anomalies reported in the administration of either the pretest or posttest. Teachers did report that 3rd grade students had some difficulty in completing the tests and some students took 2 or more hours finishing each test. Other teachers reported that some of their higher achieving 5th grade students took long periods of time with each test. All teachers perceived that the tests were rather difficult and that students were not accustomed to being tested in this way (two test administrations each with a locator test component.)

Statistical Analysis and Reporting

The basic question for the statistical analysis was whether, following the intervention, students in *SFScience* classrooms had higher NWEA 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 those 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 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 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 $p \leq .05$. (This is the level of confidence conventionally referred to as "statistical significance.")
2. We have some confidence when $.05 < p \leq .15$.
3. We have limited confidence when $.15 < p \leq .20$.
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]. The following tables address the nature of the groups. Table 5 displays 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 September 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 5. Distribution of the *SFScience* and Control Groups by Schools, Teachers, Grades, and Counts of Students

	No. of schools	No. of teachers	No. of classes	Students in Grade 3	Students in Grade 4	Students in Grade 5	Total students
SFScience	6	9	14	49	215	90	**354**
Control	6	12	17	192	35	139	**366**
Totals	**6**[a]	**21**	**31**	**241**	**250**	**229**	**720**

[a] Each of the 6 schools participated in both conditions.

Years of Teaching Experience

As part of our data collection we asked the teachers to provide us with information regarding their backgrounds. In this study, since we randomized at the grade level and not at the teacher level, we were particularly interested in the years of teaching experience each teacher had to check to see if the groups had a balanced distribution of experience. Several correlational studies indicate that this may in part determine student achievement outcomes. We stratified according to this variable, to check for discrepancies between conditions in years of teaching experience. But as seen in the following two tables, no such imbalance occurred.

Table 6. Years Teaching Experience

		Early career (0-3 years)	Emerging professional (4-6 years)	Mid-career professional (7-15 years)	Highly experienced professional (15+ years)
	Number of teachers	%	%	%	%
SFScience	8	11%	11%	44%	22 %
Control	11	8%	8%	25%	50%

Note: One *SFScience* (11%) and one control teacher (8%) did not provide this information.

Table 7. Distribution of Years Teaching Experience

	Number of Teachers		
Condition	0 to 3 years	4 or more years	Totals
SFScience	1	7	**8**
Control	1	10	**11**
Totals	**2**	**17**	**19**

Note: One *SFScience* (11%) and one control teacher (8%) did not provide this information.

The following tables further describe the background characteristics of the teachers in the study. In general, most teachers in the study are well established in their careers and hold college degrees with some coursework in science.

One difference noted is that five control teachers have attended recent professional development pertaining to science and that only two of the *SFScience* teachers had. One teacher in the *SFScience* group noted that she had not taught science in eight years and was especially concerned over the laboratory activities.

Table 8. Years Teaching in Grade Level (not necessarily consecutive)

	Number of teachers	0-3 years %	4-6 years %	7-15 years %	15+ years %
SFScience	9	33%	22%	33%	11%
Control	11	25%	17%	42%	8%

Note: One control teacher (8%) did not provide this information.

Table 9. Years Teaching Science

	Number of teachers	0-3 years %	4-6 years %	7-15 years %	15+ years %
SFScience	7	33%	0%	33%	11%
Control	10	25%	0%	33%	25%

Note. Two *SFScience* (22%) and two control teachers (17%) did not provide this information.

Table 10. Science Coursework in College

	Number of teachers	None %	Some %	Minor %	Major %
SFScience	9	0%	78%	0%	22%
Control	10	0%	75%	8%	0%

Note. Two control teachers (17%) did not provide this information.

Table 11. Recent Professional Development (PD) for Science Instruction

	Number of teachers	Attended PD in last two years	No PD in the last two years
		%	%
SFScience	9	22%	78%
Control	12	42%	58%

Post Randomization Composition of the Experimental Groups

In this section we analyze the nature of any attrition and the distribution of various student level characteristics. Table 12 shows that between the initial grade-level randomization and class assignment and the later gathering of pretest data, there was about 10% attrition (approximately 12% in control and 8% in *SFScience*. We do not believe that there were any *SFScience*-related reasons for the differential attrition and therefore it does not represent an obvious source of bias.

Table 12. Test Data for Students in *SFScience* and Control Groups

	Students enrolled in the study	Students having pretests	Students having posttests	Students having both pre- and posttests
SFScience	354	326	328	326
Control	366	323	330	323
Totals	**720**	**649**	**658**	**649**

In checking for balance in the composition of the experimental groups, we examine, student characteristics such as English proficiency, ethnicity, and gender, and student pretest outcomes.

From the previous tables, we see that a total of 720 students were enrolled in the study. The data for this analysis was provided by the individual schools because no district aggregated records are kept. Individual record keeping varies and so you may note discrepancies between student counts in the various categories. For analysis of student outcomes, we would normally remove any students identified as requiring special education services, but since no students were reported, we can only assume that none were enrolled. Additionally, no individual student socio-economic data was provided. Hence, the following analyses are based on a sample size of 649 students, those students having both pretest and posttest scores. Not all categories were reported by every school.

Student Variables

English Proficiency

There was only one student designated as an English learner. Since this was the only data point for this category, no analyses were conducted using English Proficiency as covariate.

Ethnicity

Table 13 summarizes the distribution of student ethnicity. The majority of students are White, which coincides with the general ethnicity of the greater St. Petersburg area. As a result of

random assignment, the ethnicity of the students is evenly distributed across the *SFScience* and control groups. The result of the statistical test is consistent with this assertion.

Table 13. Ethnicity for *SFScience* and Control Groups

	Ethnicity					
Condition	**Asian**	**Hispanic**	**Black**	**White**	**Multi-racial**	**Totals**
SFScience	16	63	14	257	1	**351**
Control	11	47	17	284	1	**360**
Totals	**27**	**110**	**31**	**541**	**2**	**711**

Statistics	*p* value
Fisher's Exact Test	.26

Gender

Table 14 summarizes the distribution of gender. There is a small chance imbalance between conditions on gender.

Table 14. Gender for *SFScience* and Control Groups

Condition	Gender		
	Male	**Female**	**Totals**
SFScience	194	160	**354**
Control	182	184	**366**
Total	**376**	**344**	**720**

Statistics	DF	Value	*p* value
Chi-square Test	1	1.86	.17

Characteristics of the Experimental Groups Defined by Pretest

We also checked whether randomization resulted in balance on pretest scores, a variable that we include in most of our analyses to increase the precision of our impact estimate. Table 15 shows the results of students in grades 3 to 5 for whom pretests were available. The *SFScience* and control groups had slightly different average pretest scores on NWEA Science and Reading. However, the large p-value .94 indicates that there is balance between the *SFScience* and control group.

Table 15. Difference in 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	202.78	9.19	326	0.51	0.27
Control	200.32	9.32	323	0.52	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
Condition (*SFScience* – control)	2.46		647	-3.39	.94

[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

We have no cases of students who took the pretest but did not take the posttest.

Implementation Results

In this section we describe more fully the aspects of the implementation that characterize this intervention. We used the following 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 possible variables related to the outcome measures. Our perspective takes into account three levels of resources needed to implement science instruction: those resources provided by either the district or by Scott Foresman, those provided by the individual schools, and those provided by the teacher.

Implementing a new curriculum can be challenging. There are a number of factors that play into how well a program is incorporated into an already established routine. The curriculum, the school, and the teacher all play a role in the ability to implement and the quality of the implementation. For example, did Scott Foresman supply appropriate amounts of materials and in a timely manner? Was the training for the program adequate and sufficient? On a school level, did the school have the resources necessary to implement the program effectively? Did the school have adequate staffing and space for instruction? These variables are all involved in providing ideal implementation before the teacher even has a chance to use the curriculum. On a teacher level, have all the components of the program been appropriately modeled and demonstrated? Does the teacher have sufficient subject-matter knowledge and pedagogical knowledge to teach science?

Although we do not rate the level of implementation in each individual classroom, we provide a sufficient level of detail to draw overall conclusions as how much science instruction took place, how it was conducted and which materials were covered in the *SFScience* condition.

Comparison of *SFScience* and Control Groups

Six schools participated in the study; five were considered PK-8 and one considered K-8 only. All schools had science laboratories available for use at least once a week.

Classroom Settings for Instruction

The classroom setting was observed during the week of March 28th, 2006. The classroom observations were conducted once during the length of the intervention. Most teachers were observed for approximately 40 to 60 minutes, the length of the science instruction time period.

Teachers were not asked to prepare specific lessons for observation, but we made an effort to coordinate the observation schedule with the teacher prior to observation.

Most teachers in both groups had traditional classroom layouts consisting of individual student desks arranged in rows and facing towards a white/blackboard, the designated "front" of the classroom.

The laboratories seemed to be well supplied and orderly, with plenty of room to conduct experiments. Participating teachers reported that although they were able to use the facilities, longer term storage of experiments that were to be used for ongoing observations were accommodated in the classroom for easy access. Teachers had some storage cupboards in their classrooms as well, but these were filled with other materials. In general, all teachers commented that storage of materials is always a challenge.

There were two schools that supplied laptops for their students and consequently all students in both groups had access to this technology. At these two schools there was a strong emphasis on the integration of technology and one classroom was observed to have an interactive whiteboard used for instruction. At another school, teachers report having "Smartboards" in use. In the other four schools, most teachers had some computer stations in the classroom, but not enough for every student. Groups of students worked at the computers for 10 to 20 minutes at a time then returned to their desks. The activities at the computer were Web quests and scavenger hunts or looking up science facts to supplement instruction. The computer activities were practiced by both *SFScience* and control groups. Televisions and video playback/recorder systems were in evidence or accessible by both teacher groups. Some teachers liked to supplement instruction using video, other teachers reported that they rarely used videos but instead used the Internet. Every teacher had an overhead projector that they used periodically.

Overall, most teachers had the materials they needed to teach science, but working space was at a premium for the hands-on activities. All teachers supplemented instruction with some sort of Internet activities. Outside activities such as a communal garden and field trips were less common for the control teachers (5.6%) than for the *SFScience* group (17.5%).

Opportunities for Learning

This site was identified before the beginning of the school year and most of the 3rd and 4th grade materials arrived within the first two weeks of the start of instruction. The 5th grade materials and in some cases other materials were on backorder until late November. The 5th grade teachers used other materials and methods to teach science until December. Full implementation of the *SFScience* materials began in early December.

At four of the six schools science is taught as a specialty subject. The remaining two schools teachers provide science instruction as self-contained subject. When science is taught as a specialty, one teacher is responsible for teaching several classrooms and students are typically rotated in exchange for other subjects, such as reading and mathematics. This system of rotation is more typical of middle school and high school scheduling, but it is becoming common practice in elementary school as an informal way of organizing instruction and taking advantage of teachers' expertise and inclination. As a specialty subject the teacher of instruction may teach the same lesson more than once in a short period of time making adjustments to the lesson similar to what happens in high schools, where the teacher makes adjustments to the lessons according to students' responses often creating a better aligned lesson by the end of the day.

For the self-contained classroom teacher, science is taught as part of all subjects taught to the students. Teachers typically alternate science instruction with social studies. An alternating schedule allows the teacher to plan and gather resources to provide instruction for two to three weeks at a time. Not all teachers followed this scheduling pattern. Some teachers scheduled science instruction for approximately 30 to 50 minutes daily for three or four days every week, allocating the longer periods to days scheduled with hands-on activities.

We surveyed the teachers regarding how much time they spent with their students in science learning as a standalone subject, meaning as a subject unto itself, not used as part of reading

or another program. We also asked if they taught science integrated with other subjects such as reading, mathematics, or social studies and if so, how much time they spent teaching it in this manner. Two control teachers did not report instructional times on a consistent basis, and so we averaged times for all other teachers and did not include data from teachers missing more than 2 data points out of the five times they were asked to report. *SFScience* teachers reported an average of 37.4 total hours and control teachers reported an average of 42.8 total hours of instruction for the length of the implementation. As we observe later in Table 31, we have limited confidence that the actual difference is different from zero.

One important issue that distinguishes this site from the other four sites that were part of the larger research agenda, because these schools are private, they do not need to strictly adhere to the Florida State Curriculum standards (Sunshine State standards). This enabled the teachers to follow the lesson flow as designed in the *SFScience* materials and they were able to take advantage of all of the components of the program.

Control Materials

As noted before, there were several types of textbooks in evidence, older versions of Scott Foresman, Discovery Works and Horizons both by Silver Burdett Ginn, 1999 (now Houghton Mifflin), SRA Science Laboratory by McGraw Hill, Harcourt Brace Science 2002, and Destinations in Science by Addison Wesley. When asked about materials usage through surveys some control teachers responded as shown in Table 16. When asked during informal interviews about science materials in general, teachers responded that they had textbooks for all of their students.

Table 16. Primary Sources for Science Instruction

Which materials constitute the primary resources that you use to teach Science? Check all that apply.							
		District Developed Materials	Textbook	Periodicals	Magazines	Internet	Video
Number of respondents	4	0%	100%	25%	25%	100%	75%

For conducting laboratory activities control teachers indicated that they have no set pattern of usage. It depends on the topic and the availability of materials. For at least one teacher every lesson included an inquiry activity. Teachers felt that they needed an average of 60 minutes of instruction to include hands-on inquiry activities into their lessons. Additionally they indicated that needed better designed classrooms to accommodate science inquiry because they felt that the physical aspects of their rooms were limiting their choices of activities. Teachers are providing their own materials brought from home approximately 5% of the time.

Table 17. Percentage of Time Devoted to Hands-on Science Activities

How much time was spent on hands-on science activities (where students practiced science inquiry steps: investigation, hypothesis, observation and data collection, presentation of results)?						
		90-100%	50-89%	30-49%	10-29%	Less than 10%
Number of respondents	4	0%	0%	50%	50%	0%

Planning Time

Planning time for science instruction is also an important factor for implementing curriculum. Twelve of the possible twelve control teachers responded that they spent approximately 27.5% while *SFScience* teachers reported 37.5% (30 to 60 minutes per week) of their total available planning time on science instruction. The difference between the two groups is due in part to the lack of familiarity with the new *SFScience* materials. The curricular materials used by the teachers in the control group were very familiar because they had been in place for a number of years.

Just three teachers in the control group report not having enough planning time and only two teachers in the *SFScience* group say they could use more time. All teachers coordinate their schedules with the rest of their grade level group.

Density of Science Inquiry Reflected in the Classroom

Sections of the surveys were constructed to collect data on the aspect of science inquiry as a method for teaching/learning science since Scott Foresman specifically designed the curriculum using inquiry as theme and pedagogy.

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

- 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 sequenced activities of the SF science 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)
- Prediction or hypotheses (DI: students are given a prediction for conducting a descriptive investigation; FI: students are guided to make a prediction for a guided investigation; FI: students develop logical/reasonable predictions)
- Investigate (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).

When we asked the teachers on the surveys, we asked about time spent doing these different activities. Both *SFScience* and control group teachers were asked these questions. The variable "science inquiry" is a composite of the time spent in six different aspects of the inquiry process as a percentage. Hence, it is on a scale of 0 to 100 and 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. The average percentage density for the *SFScience* group was 22.7 and for the control group it was 27.4. While a greater amount of density is noticed for *SFScience* condition, statistically we have no confidence that this difference between the groups is different from zero.

Implementation of *SFScience*

Training and Support

The one-half day training took place on September 14, 2005 at one of the participating school's library. During the training, the two Scott Foresman representatives gave a demonstration of the science kits and the pedagogical method of hands-on inquiry. A video was used to initially model the science kits usage and features. A set of videos is packaged with the materials to provide teachers with additional lesson support.

Teachers participated in three different investigations, discussed the various methods that students could use to share their findings, and how the activities integrated into the reading materials. 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. Emphasis was placed on the using the development of inquiry skills by using the materials as sequenced from Directed Inquiry (DI) to Guided Inquiry (GI) and finally to Full Inquiry (FI). The trainers highlighted the different ways that teachers could use to plan the lessons, when time was short, when teaching a lesson without labs, and when a lesson could be delivered fully.

Overall, the teachers were enthusiastic about the materials and the training session provided a good introduction. For a complete list of the materials supplied by Scott Foresman refer to Table 1. Teachers also received an online log-in so that they could reference additional materials. Teachers also indicated that there was a lot of material to cover and it was difficult to digest all of the ideas in such a short period.

One teacher noted that it had been several years since she had last taught science and she had never used any hands-on activities. Another teacher commented that she would like more time to research current topics of interest to make the curriculum relevant to her students. Teachers in general, remarked that they were eager to explore the materials.

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.

Availability and Use of Materials

Every teacher assigned to the *SFScience* group received sufficient materials to use with the number of students that they taught whether they taught in a self-contained classroom or in a specialty subject classroom. Several teachers reported missing some materials, the Activity Flip Charts and Audio Text CD's for 3rd, 4th, and 5th grades, while some 5th grade teachers did not have enough student editions. Additionally the science kits were backordered until November for all of the 5th grade.

SFScience group teachers were asked to complete any two of the four units provided in the SF science curriculum. The text materials were segmented into four units: A-Life Science, B-Earth Science, C-Physical Science, D-Space and Technology. At the teacher's discretion she could select the units and chapters she covered with her students.

Eight of the nine *SFScience* teachers responded to the survey questions regarding the content covered in their classrooms. Teachers could select as many chapters within a unit that they covered. 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 18. Percent of Teachers Covering Each Chapter in Unit A-Life Science

		Chapter					
		1	2	3	4	5	6
Number of respondents	8	50%	50%	62.5%	62.5%	75%	75%

Note: One *SFScience* teacher did not respond to this question.

Table 19. Chapters in Unit B-Earth Science Covered

		Chapter			
		7	8	9	10
Number of respondents	8	50%	50%	62.5%	37.5%

Note: One *SFScience* teacher did not respond to this question.

Table 20. Chapters in Unit C-Physical Science Covered

		Chapter				
		11	12	13	14	15
Number of respondents	8	37.5%	25%	12.5%	0%	37.5%

Note: One *SFScience* teacher did not respond to this question. Teachers did not teach any other chapters in this unit and not all teachers taught chapters in this unit.

Table 21. Chapters in Unit D-Space & Technology Covered

		Chapter		
		16	17	18
Number of respondents	4	37.5%	25%	25%

Note: One *SFScience* teacher did not respond to this question.
Not all teachers taught chapters in this unit.

Although the schools in St. Petersburg do not need to follow the state science content standards (Sunshine State Standards), teachers and administrators are aware of them and take note. They use them as loose guidelines to ensure their students are prepared similarly to the public schools. Typically, they make sure they cover the content, but don't worry if they teach beyond the recommended standard. Consequently several teachers were able to complete the desired two units of instruction for the study. Several teachers remarked that some of the hands-on activities were not well aligned to the concepts. Towards the end of the study, teachers had begun substituting the laboratory activities in the lessons with ones they had used in the past because of the stronger alignment and the perceived "robust" nature of the substitute activity. The younger students were not used to having to pay the amount of attention required by the activities. They understood what to do, but did not yet have the skills to understand the connection between "how to do" and "why/when to do".

For each unit we asked teachers to tell us how well they thought the chapters were aligned to the Sunshine Standards. The following tables summarize how teachers viewed the alignment to standards by unit.

Table 22. For Unit A, How Well Was the Content Aligned to State Standards?

		Did not teach any of this unit	Taught too few to have an opinion	Aligned to standards well	Aligned somewhat	Aligned poorly
Number of respondents	8	25%	0%	50%	25%	0%

Note: One *SFScience* teacher did not respond to this question.

Table 23. For Unit B, How Well Was the Content Aligned to State Standards?

		Did not teach any of this unit	Taught too few to have an opinion	Aligned to standards well	Aligned somewhat	Aligned poorly
Number of respondents	8	25%	0%	37.5%	37.5%	0%

Note: One *SFScience* teacher did not respond to this question.

Table 24. For Unit C, How Well Was the Content Aligned to State Standards?

		Did not teach any of this unit	Taught too few to have an opinion	Aligned to standards well	Aligned somewhat	Aligned poorly
Number of respondents	8	37.5%	25%	25%	12.5%	0%

Note: One *SFScience* teacher did not respond to this question.

Table 25. For Unit D, How Well Was the Content Aligned to State Standards?

		Did not teach any of this unit	Taught too few to have an opinion	Aligned to standards well	Aligned somewhat	Aligned poorly
Number of respondents	8	37.5%	12.5%	25%	12.5%	12.5%

Note: One *SFScience* teacher did not respond to this question.

Many teachers had trouble incorporating the Leveled Readers into their science instruction because there were two few to use with the entire class. These classrooms tend to be leveled by reading ability and so most students are at the same reading level. The way the materials are packaged (6 copies each at the three levels) did not allow the teachers to distribute sufficient copies to their students. When they did use them, the students used them to review the content. Still other teachers used them as introductory material for the chapter. Teachers commented that they tried to access the readers online, but had difficulty navigating the online resources. Those teachers that were using the "specialty subject" model of instruction had fewer opportunities to use the Readers, noting that they simply ran out of instructional time.

Whole class reading activities are practiced as part of science instruction in every classroom observed. Some of the classrooms had access to the Audio CDs and frequently incorporated them into the reading routine. Teachers noted that this helped the students with pronunciation and recall because the readings on the CD "were more dramatic"; it also allowed the students to concentrate on the content and associated pictures and figures. It also helped teachers by supporting the inquiry nature of the curriculum by providing pauses and cues ("checkpoints") for asking questions and probing for deeper comprehension.

As for the Science Kits, teachers indicated that not all of the materials were included. At times the instructions did not provide sufficient levels of support because they were incomplete or did not provide fail safe information. Some teachers reported that "things broke and were not sturdy enough for the experiment." The teachers did like the convenience of the kits, specifically having materials ready to hand. They thought it was easy to set-up and clean-up afterwards, but all of the activities took longer than indicated. As noted before, scheduling sufficient time for science instruction with the hands-on materials was a challenge.

All nine *SFScience* teachers used at least one assessment with their students. All of the teachers thought the assessments difficult and poorly aligned to the content and language employed in the chapter/unit. Students performed poorly on initial use of the assessments. One teacher reported that 50% of her students failed on the first exam. Teachers had to find inventive ways to prepare the students for the assessments and later in the study, teachers reported moving away from the assessment booklet and moved towards using the Test-maker software much more. The teachers thought that the CD containing assessment materials were very helpful and much more useful because of the flexibility it allowed to formulate questions. Teachers began using the graphic organizers and preparing study guides that included all new

vocabulary, all of the instrumentation and tools noted, and a chapter outline. Some teachers used the audio CD to prepare the students on the day prior to testing. In all teachers reported spending most of their planning time in organizing new materials for assessments.

Rating the Level of Implementation

We consider the following factors to contribute to a strong implementation:

- Adequate timeframe for instructional patterns to emerge and become routine
- Sufficient training to support teachers' understanding of material usage
- School level resources: storage for materials and teacher professional development
- Sufficient amount of curriculum aligned to standards to keep the pedagogical methodology in tact

We find that for St. Petersburg, implementation was very well aligned with the expectations communicated during the training and so created a strong implementation. All teachers used the materials as intended and had begun exploring new ways to use them. By the March observation period, it was clear that teachers in both groups taught science routinely and that the *SFScience* group had exercised the materials beyond the expectations.

Summary of Implementation

Few barriers to implementation emerged. Leveled Readers presented a logistical challenge because classes tended to be reading-ability grouped and so more copies of each level are required to support each classroom. If anything teachers expectations of the laboratory activities extended beyond what was provided and teachers supplemented instruction with previously used activities (teacher created or from other resources). Teachers were most vocal about the poor alignment of the assessment materials and the need for additional "Study Guide" type materials. Although, the 5th grade science materials did not arrive until late November, because of the daily science instruction schedule used by the majority of teachers the length of the intervention was not problematic.

Quantitative Impact Results

The primary topic of our experiment was the impact of *SFScience* curriculum on student performance on NWEA Science and NWEA Reading. We will first address the impact on Science achievement and then the impact on Reading achievement. Within each content area we provide a statistical analysis of the impact of *SFScience* controlling for pretest and examine the interaction of *SFScience* with pretest, that is, we examine whether students initially scoring higher or lower on the pretest differentially benefited from *SFScience*. We then examine the influence of gender as a potential moderator of the impact of *SFScience* as well as the influence of years of teaching experience.

When performing analyses of experimental impacts, there are alternative statistical models that we can use to try to increase how precise our estimate of the effect of the intervention is. For instance, we routinely add students' pretest scores into the analysis. It is important to decide which of these covariates will be added to the model at the time that the experiment is being planned so that the covariates are not added to the model opportunistically after the fact, which could lead to capitalization on chance and misleading levels of statistical certainty. In this study, we decided to not model the fixed effect of grade-level since the randomization resulted in grades being distributed between conditions in such a way that the observed difference between conditions is unlikely to be due to chance imbalances in grade levels (we described this situation more fully in the section that outlines our randomization strategy.)

In the following sections, our analysis of the quantitative results takes the same form. We present the results of statistical models 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. That is, we test for the interaction of treatment with the prior score. The fixed factor part of the table provides estimates of the factors of

interest, in particular, whether being in a *SFScience* or a control class makes a difference for the average student. 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 will often 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 Science achievement using the NWEA Science Concepts and Processes assessment. Table 26 provides a summary of the sample we used in the analysis and the results for the comparison of *SFScience* and control. The "Unadjusted" row gives information about all the students in the original sample for whom we have a posttest. 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, that is, the size of the difference between the means for *SFScience* and control in standard deviation units, and the *p* value, indicating the probability of arriving at a difference as large or larger than the absolute value of the one observed when there truly is no difference. The "Adjusted" row is based on the students who have both pretests and posttests. This is the sample that we use in the analyses on which we base our results reported in Table 27 and Table 28. The adjusted effect size is based on a model that includes fixed effects for schools as well as pairs within which we randomized.

Table 26. Overview of Sample and Impact of *SFScience* on Science Achievement

	Condition	Means	Standard[a] deviations	No. of students	No. of classes	No. of teachers	Effect size	*p* value[b]
Un-adjusted	***SFScience***	204.14	9.20	326	14	9	0.18	.85
	Control	202.56	8.72	330	17	12		
Adjusted	***SFScience***	202.38	9.04	326	14	9	-0.01	.89
	Control	202.48[c]	8.75	322	17	12		

[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 teachers 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 1 provides a visual representation of specific information in Table 26. The bar graphs represent average performance using Science Achievement as the metric.

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 26.) We can see that the two groups were essentially indistinguishable. The high *p* value for the treatment effect (.89) indicates we should have no confidence that the actual difference is different from zero.

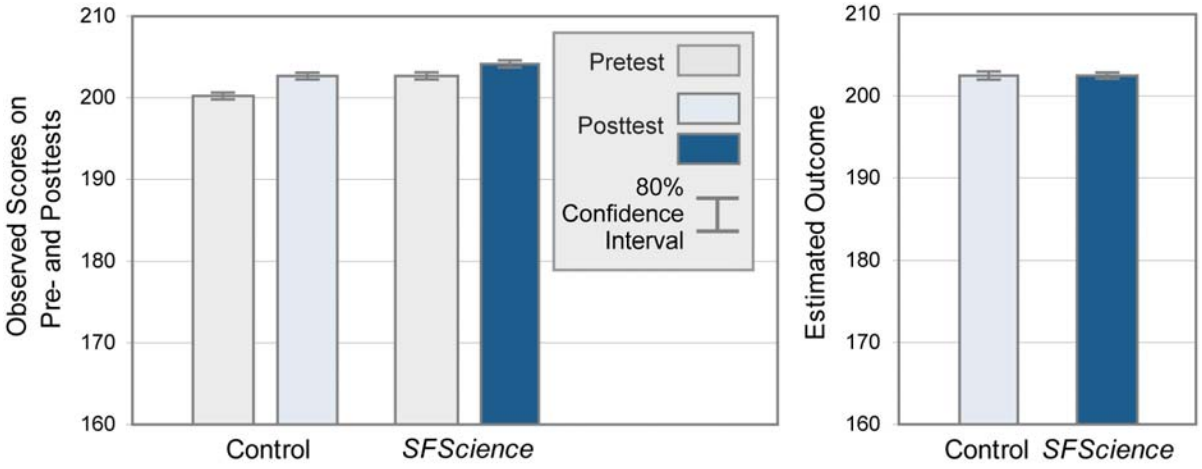

Figure 1. Impact on Science Achievement: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)

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[2]. We begin by examining the moderating effect of the prior score. Since the NWEA tests are on a continuous scale and the experiment involved three grades, we do not interpret low NWEA scores as indicating "low achieving" students within each grade. It is likely that third graders are more heavily represented in the lower range of the scores and fifth graders in the higher end of the scores. Table 27 shows the estimated impact of *SFScience* on the performance in science of a student who has an average score on the pretest as measured by Science achievement, as well as the moderating effect of the prior score.

[2] Before analyzing the results, we select the moderators of interest. In this case we decided that the moderators of interest are prior score and gender. With exception of the prior score, we graph results only for moderators for which the *p* value for the interaction effect is less than or equal to .20 i.e., where we have at least limited confidence that the moderating effect is different from zero.

Table 27. The Impact of *SFScience* on Science Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Estimated value for a control student with an average pretest	205.48	1.47	10	140.16	<.01
Impact of *SFScience* for a student with an average pretest	-0.15	0.72	10	-0.21	.83
Estimated change in control outcome for each unit increase on the pretest	0.71	0.04	623	18.43	<.01
Interaction of pretest and *SFScience*	-0.03	0.05	623	-0.50	.62
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	0.57	0.79		0.73	.23
Within-teacher variation	29.27	1.65		17.70	<.01

[a] Schools, and pairs of grades used for random assignment are also modeled as a fixed factor but are not included in this table. Because of the use of fixed effects, the estimated value for a control student with an average pretest is for a particular school and pair.

[b] Teachers were modeled as a random factor.

The row in the table 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 Science for a student who has an average score on the pretest. The estimate associated with *SFScience* is -0.15. This shows a small negative effect associated with *SFScience*. However, the *p* value of .83 gives us no confidence that the underlying effect is different from zero.

We also estimated the moderating effect of the pretest score on the impact of *SFScience* to see whether the intervention was differentially effective for students at different points along the pretest scale. The *p* value for this effect is .62. We have no confidence that the actual effect is different from zero.

As a visual representation of the results described in Table 27, we present a scatterplot in Figure 2, which shows student performance at the end of the year in science, as measured by Science achievement, against their performance on Science achievement 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 plots 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).

We see that there was growth on average for both the *SFScience* and control groups. However, there is no difference in the result between the two groups.

The tilt in the prediction lines simply shows the fact that scores tend to regress to the mean: on average, those who scored at the higher range of the pretest tend to score lower on the posttest and those who score at the lower range of the pretest tend to score higher on the posttest. This is a normal characteristic of test-retest data and is not an indication of a systematic effect.

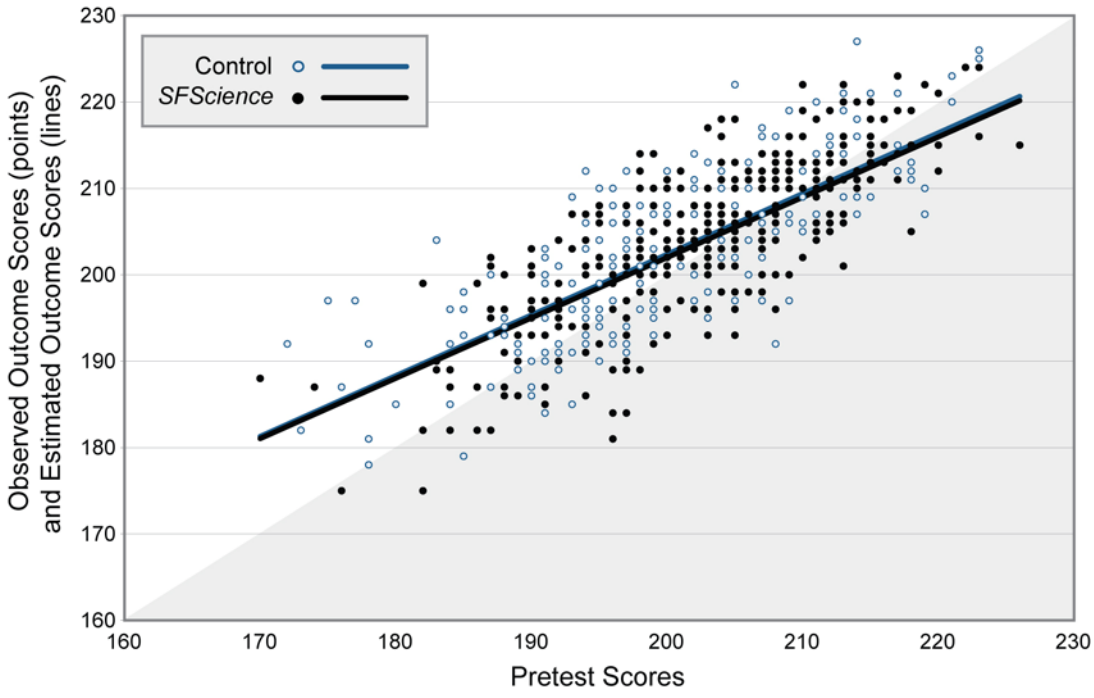

Figure 2. Comparison of Estimated and Actual Outcomes for *SFScience* and Control[3] Group Students

Analysis Including Gender as a Moderator

We were also interested in whether *SFScience* was differentially effective for males and females. Table 28 shows that there is no difference between boys and girls in science and no differential effect of *SFScience* depending on gender. In other words, boys and girls performed equally as well on NWEA Science when using the *SFScience* curriculum.

[3] Displaying estimated values can be confusing when we model separate intercepts for upper-level units. The estimated values are shifted vertically for each unique intercept value. For ease of displaying the estimated interaction effect we graph the results of a simpler model. We exclude the upper-level fixed effects and graph the result only if the estimate of the interaction is consistent with the original more complex model in the following two ways: 1) the direction of the interaction is the same as it was for the model that included fixed effects (i.e. the estimate does not change signs); and 2) the *p* value does not go from ≥ .20 to <.20 (or from <.20 to ≥.20).

Table 28. Moderating Effect of Gender on Science Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Outcome for a girl with an average pretest in the control group	204.87	1.41	10	144.80	<.01
Average *SFScience* effect for girls	0.66	0.80	10	0.83	.43
Estimated change in outcome for each unit increase on the pretest[c]	0.69	0.03	620	26.44	<.01
Difference (boys minus girls) in average performance in the control condition	0.41	0.61	620	0.67	.51
Difference (boys minus girls) in the average *SFScience* effect	-0.91	0.86	620	-1.05	.29
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	0.48	0.72		0.67	.25
Within-teacher variation	28.36	1.61		17.66	<.01

[a] Schools and pairs of grades used for random assignment were modeled as a fixed factor but the estimated effects are not included in this table; because we estimated fixed effects for schools and assignment pair, the estimated value for a female control with an average pretest applies to a particular school and assignment pair.

[b]Teachers were modeled as a random factor.

[C]The prior score was centered at the mean, therefore, the estimate for the average outcome for a girl in the control group applies to student who had an average score on the pretest.

Reading Outcomes

Analysis Including Pretest

Our next set of analyses address Reading outcomes as measured by Reading achievement. Table 29 provides a summary of the sample we used in the analyses and the results for the comparison of *SFScience* and control. The “Unadjusted” row gives information about all the students in the original sample for whom we have a posttest. 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, that is, the size of the difference between the means for *SFScience* and control in standard deviation units, and the *p* value, indicating the probability of arriving at a difference as large 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 pretests and posttests. This is the sample that we use in the analyses on which we base our results reported in Table 30.The adjusted result is the effect estimate in standard deviation units from a model that includes the pretest as a covariate. On average, including the pretest should increase the precision of the effect estimate.

Table 29. Overview of Sample and Impact of *SFScience* on Reading Achievement

	Condition	Means	Standard deviations[a]	No. of students	No. of classes	No. of teachers	Effect size	*P* value[b]
Un-adjusted	***SFScience***	212.45	12.65	316	14	9	.34	.92
	Control	209.50	11.70	330	17	14		
Adjusted	***SFScience***	211.05	12.41	312	14	9	.16	.23
	Control	209.18[c]	11.53	318	17`	14		

[a] The standard deviations used to calculate the adjusted and unadjusted effect sizes are calculated 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 figures in clustering of students in teachers but does not adjust for any other covariates. The *p* value for the adjusted effect size is computed using a model that controls for clustering and that includes as covariates both the pretest and, where relevant, indicators for upper-level units within which the units of randomization are nested.

[c] Modeling separate intercepts for upper-level units leads to estimates of performance, in the absence of treatment, that are specific to those units. For purposes of display, to set the performance estimate for the control group, we compute the average performance for the sample of control cases used to calculate the adjusted effect size. The estimated treatment effect, which is constrained to be constant across upper-level units, is added to this estimate to show the relative advantage or disadvantage to being in the treatment group.

Figure 3 provides a visual representation of specific information in Table 29. The bar graphs represent average performance in 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 30.) We can see that the two groups were essentially indistinguishable. The *p* value for the treatment effect (.23) indicates we should have no confidence that the actual difference is 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 could very well be due to chance.

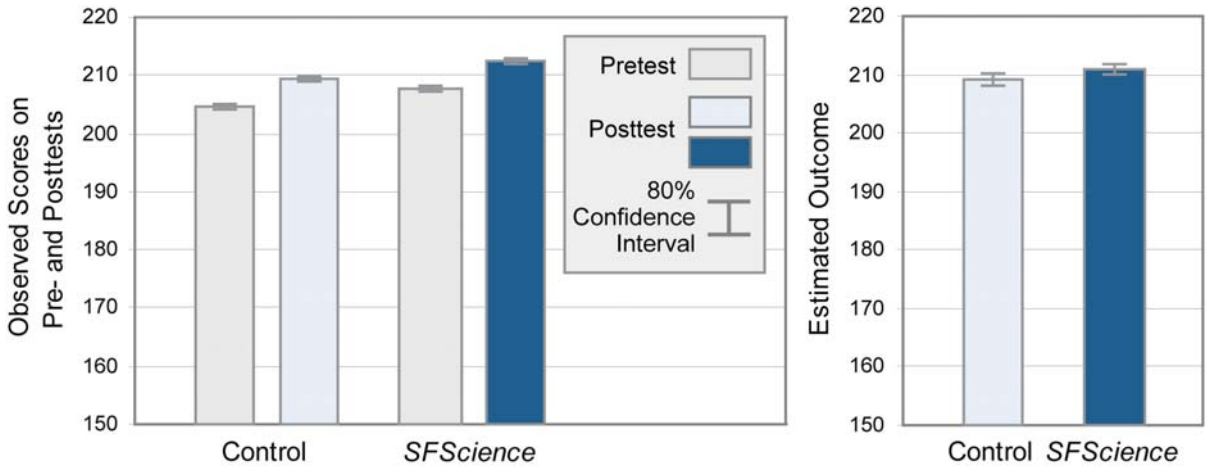

Figure 3. Impact on Reading Achievement: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)

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. Since the NWEA tests are on a continuous scale and the experiment involved three grades, we do not interpret low NWEA scores as indicating "low achieving" students within each grade. It is likely that third graders are more heavily represented in the lower range of the scores and fifth graders in the higher end of the scores. Table 30 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 30. The Impact of *SFScience* on Reading Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Estimated value for a control student with an average pretest	209.29	2.97	10	70.51	<.01
Impact of *SFScience* for a student with an average pretest	1.79	1.47	10	1.22	.25
Estimated change in control outcome for each unit increase on the pretest	0.77	0.03	605	26.22	<.01
Interaction of pretest and *SFScience*	-0.05	0.04	605	-1.21	.23
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	5.30	3.45		1.54	.06
Within-teacher variation	31.85	1.83		17.42	<.01

[a] Schools and pairs of grades 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 the table labeled “Impact of *SFScience* for a Student with an Average Pretest” tells us whether *SFScience* made a difference in Reading achievement for a student who has an average score on the pretest. The estimate associated with *SFScience* is 1.79. This shows a positive effect of *SFScience*. However, the *p* value of .25 gives us no confidence that the effect being estimated 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 .23. We have no confidence that the actual effect being estimated is different from zero.

As a visual representation of the results described in Table 30, we present a scatterplot in Figure 2, which shows student performance at the end of the year in reading, as measured by NWEA Reading, against their performance on Reading achievement 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 plots 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).

We see that there was growth on average for both the *SFScience* and control groups. However, there is no difference in the result between the two groups.

The tilt in the prediction lines simply shows the fact that scores tend to regress to the mean: on average, those who score at the higher range of the pretest tend to score lower on the posttest and those who score at the lower range of the pretest tend to score higher on the posttest. This is a normal characteristic of test-retest data and is not an indication of a systematic effect.

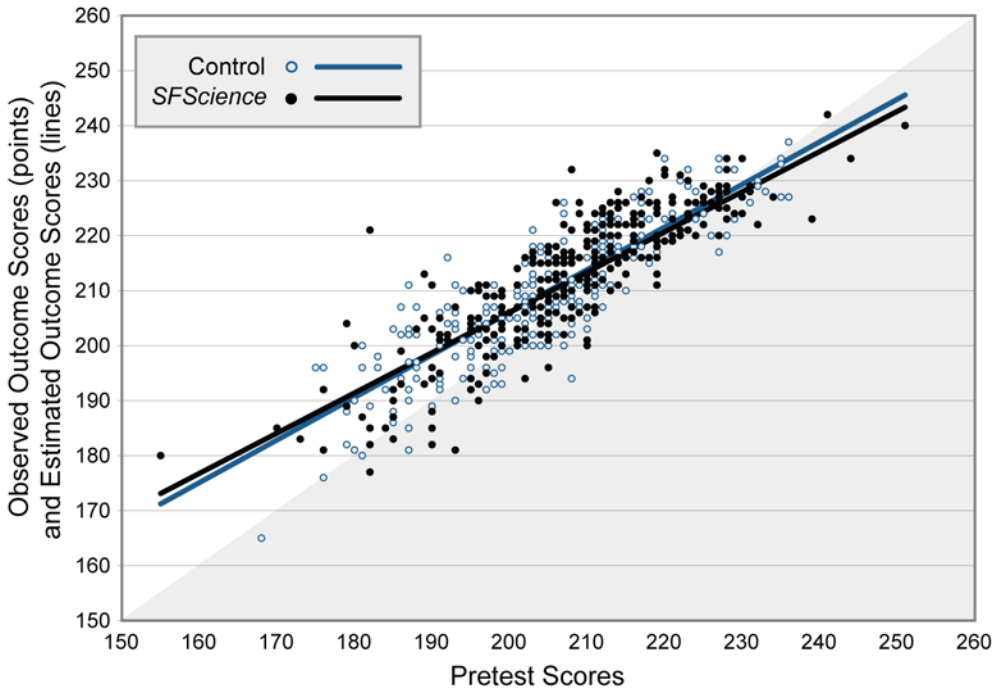

Figure 4. Comparison of Estimated and Actual Outcomes for *SFScience* and Control Group Students[4]

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 under the implementation results, we measured the amount of instructional time the teachers devoted to science.

When dealing with implementation variables, we can understand them as a path for the impact of the treatment as illustrated in Figure 5. That is, *SFScience* can have a direct impact on both the outcomes and on instructional time. The link from instructional time to the outcome is correlational but an important relationship to explore.

[4] Displaying estimated values can be confusing when we model separate intercepts for upper-level units. The estimated values are shifted vertically for each unique intercept value. For ease of displaying the estimated interaction effect we graph the results of a simpler model. We exclude the upper-level fixed effects and graph the result only if the estimate of the interaction is consistent with the original more complex model in the following two ways: 1) the direction of the interaction is the same as it was for the model that included fixed effects (i.e. the estimate does not change signs); and 2) the *p* value does not go from >= .20 to <.20 (or from <.20 to >=.20).

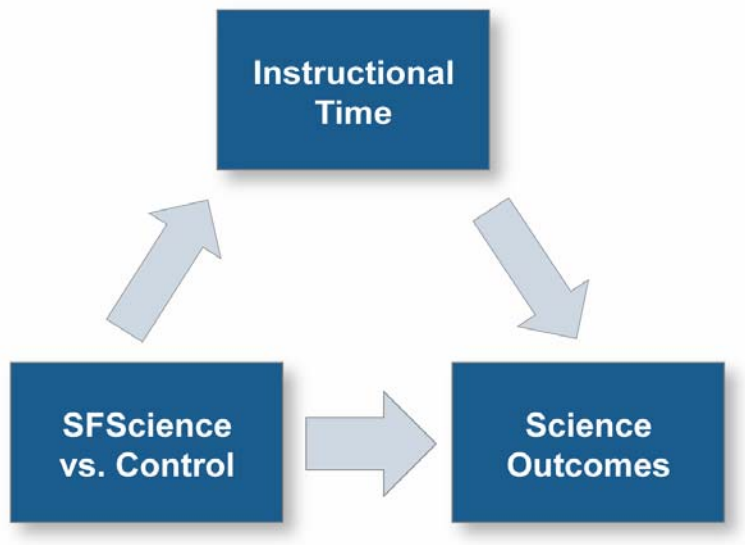

Figure 5. Relationships for Exploratory Analysis of Implementation Variables

Instructional Time

We wanted to explore the relationship between how much time was spent teaching science and science outcomes. 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 s/he spent using *SFScience* per week, from which we calculated hours spent teaching science per year.. 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 on instructional time. Table 31 shows *SFScience* teachers taught approximately 13 fewer hours of science during the year. The *p* value of .17 gives us limited confidence that *SFScience* causes a reduction in the number of minutes used on science instruction.

Table 31. The Impact of *SFScience* on Hours of Science Instruction Time

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Hours of science for a control teacher	57.98	17.36	8	3.34	.01
Impact of *SFScience* on hours of science instruction	-12.98	8.68	8	-1.50	.17
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Residual teacher variance	225.98	112.99		2.00	.02

[a] Schools and pairs of grades that were used for random assignment are modeled as fixed factors. The estimates of these effects are not included in this table.

Given that there are differences in the amount of instructional time across the teachers in the experiment, we next explored whether there was a correlation between time spent and student achievement. The result of this investigation is purely correlational—we cannot be sure whether

it is instructional time or some other variable which is correlated with instructional time (e.g., teacher enthusiasm) that is the true cause of the student outcome. A test of the correlation between instructional time and student performance in science reveals a slight positive relationship between *SFScience* usage and the student outcome. The *p* value for this effect is .17, which gives us limited confidence that the true relationship is in fact different from zero.

Table 32. Relationship of Instructional Time to Student Outcome

Fixed effects	Estimate	Standard error	DF	t value	p value
Intercept	203.74	1.64	10	124.33	<.01
Estimated change in outcome for each unit increase on the pretest	0.68	0.03	625	25.87	<.01
Estimated change in outcome for each additional hour of science time	0.04	0.02	41	1.49	.17
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	0.45	0.80		0.57	.29
Within-teacher variation	30.18	1.70		17.72	<.01

[a] Schools and pairs of grades 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.

Discussion

We began this research in St. Petersburg Catholic Schools with the question of whether *Scott Foresman Science* was as effective as or more effective than their existing programs we were comparing it to. Our 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.

We found no overall difference between the science or reading scores of students taught using *SFScience* as compared to the established program. We also did not find any statistically significant difference in the value for science or reading depending on the student's initial achievement. The very small difference for the average student between *SFScience* and control cannot be distinguished from zero because of the relatively small sample of teachers and students in the experiment. This same difference, when analyzed in the context of the other four experiments did fall within our region of limited confidence. This result is suggestive and may be strengthened with more systematic use of the program's reading materials.

We also looked at the relationship of *SFScience* to gender and found that there is no differential effect of *SFScience* on gender for science achievement.

Our experiment in St. Petersburg was small, involving only 21 teachers. With small numbers we must caution that we have limited ability to detect with any statistical confidence small differences that may be important educationally. This experiment was part of a larger five-district national study but we recognize that the specific resources, demographics, and educational agendas make analyses of specific cases worthwhile, although often not applicable outside of the participating district.

This report is not intended to provide widely generalizable results and the reader should consider the characteristics of this district to evaluate the applicability of the findings.