Empirical Education

RESEARCH REPORT

Comparative Effectiveness of Scott Foresman Science:

A Report of a Randomized Experiment in Visalia Unified School District

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 Visalia Unified School District 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, who 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 Visalia Unified School District

Table of Contents

Introduction

Pearson Education contracted with Empirical Education Inc. to conduct five randomized experiments to determine the effectiveness of its *Scott Foresman Science* products compared to the elementary science programs already in place in those districts. This research project consists of a randomized experiment in Visalia Unified School District.

The question being addressed by the research is whether the *Scott Foresman Science* is as effective as or more effective than the curriculum being used at each site. Since Scott Foresman Science provides a significant reading component, we also determined the amount of reading improvement that can be accounted for by the science program. The outcomes were measured by student achievement on standardized tests administered at the beginning and end of the project. Two test areas were selected as the outcome measures: Science Concepts and Processes, and Reading Achievement. The research focuses on 3rd, 4th, and 5th grade students.

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 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. The individual reports provide a rich description of the conditions of implementation in order to assist the district in strengthening its program and to provide the reader with an understanding of the context for our findings.

Methods

Research Design

Our study is a comparison of outcomes for classes taught using the *Scott Foresman Science* curricular materials (*SFScience* group) and classes taught with the current materials used in the district (control group). Teachers volunteered for participation and, from a pool of volunteers, the researchers randomly assigned approximately equal numbers to *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 for the teachers and to maximize the students' time on exploration and data gathering provide the substance of the inquiry-driven investigations. A main feature of the curriculum is the Leveled Reader. These are student readers designed to provide the teacher with an easy way to differentiate instruction and provide reading support at, below, and above grade-level.

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 carry out four tasks for the study:

- 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 teachers regarding the frequency of the instruction. 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	

Site Descriptions

Visalia, CA

The city of Visalia is located in the San Joaquin Valley situated almost equidistant between San Francisco and Los Angeles. The city proper encompasses 29 square miles with a population of just over 100,000 according to a 2003 population estimate.

Table 2. Visalia Racial Makeup

Race/Ethnicity	% of Population
White	69.5
African American	1.9
American Indian/Native Alaskan	1.4
Asian	5.1
Native Hawaiian or Pacific Islander	0.1
Other Race	17.8
Two or more Races	4.2
Hispanic Origin (of any race)	35.6

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

Visalia Unified School District, CA

The Visalia Unified School District comprises 23 elementary schools, a newcomer language center, four middle schools, four comprehensive high schools, a continuation high school, a charter alternative academy, a charter independent study school, a K-8 charter home school and a school that serves orthopedic handicapped students; two elementary schools participated in this study. One of these schools is designated Title I.

The unit of randomization at this site is the teacher. Matched pairs were formed and a coin was tossed to determine assignment, either Treatment (using *SFScience*) or control (using current district identified materials). This site had some deviation from the original assignment pairs. The recruitment of additional teachers after the original pairs were formed caused new pairs to be formed. At least two teachers were reassigned and one teacher was allowed to be a *SFScience* teacher based on perceived need to address a large ESL population.

Table 3. Background of the Visalia Unified School District

Visalia Unified School District	
Total schools	37
Total teachers	1399
Student to teacher ratio	22.1
Grades	PK -12
Student population	25,794
Migrant students	n.a.
ELL students	20.1%

Source: CDE (DataQuest) Public School District Data for 2005-2006

Sample and Randomization

Recruiting

Pearson Education, the parent company of Scott Foresman, worked with a separate marketing company to identify districts interested in participating in research involving science curriculum. The Visalia Unified School District was identified and contact information was forwarded to us. After contacting the district and identifying the specific schools, we met with district staff members and principals to explain the details and procedures of the study. Principals identified eligible teachers, who were then invited to an after-school meeting. The initial meeting for the research experiment in the Visalia Unified School District occurred on May 26, 2005 with 13 teachers, two principals, and two district-level curriculum specialists. Researchers presented an overview of the study and methodology. We provided samples of the SF Science materials for teachers' review. A question-and-answer period followed the presentation, ending with a call for volunteers. Two teachers were represented by their principal. Of the 13 teachers present, all filled out consent forms, and the principal filled-out contact information for the absent teachers, who could not be present for the meeting because of previous engagements. These two teachers were contacted first by email and then with a phone call to explain the particulars of the study

Randomization

The unit of randomization at this site is the teacher. A total of twenty-one teachers were assigned using a coin toss to either *SFScience* (the treatment condition) or to control (classes that would continue using current district identified materials).

Because the randomization meeting was conducted in May, teaching assignment was not confirmed until August. Thirteen teachers signed consent forms and agreed to participate at the initial meeting, 15 were randomized (7 pairs and 1 solo teacher).

When school reconvened in August, six additional teachers were identified. The new teachers were randomized by school staff. Two of the original pairs were broken and reassigned to two new pairs. The other four new teachers were paired with each other. A coin was tossed resulting in a randomization of 10 pairs and one solo teacher. One teacher was not randomized but assigned to the *SFScience* condition. Additionally, in August, one teacher excused herself from the study because she moved away from the district. This teacher was identified as a non-participant due to reasons unrelated to assignment.

There are various ways to randomize teachers to conditions. We used a matched-pairs design whereby we first identified pairs of similar teachers and then, within each pair, we randomized one teacher to treatment and the other to control. Matched pairs were based on school assignment and grade level taught. We employed a pairing strategy because it will often result in a more precise measurement of the treatment impact.

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 number of participating teachers is displayed in the table below.

Table 4. Participating Teachers

Teacher Assignment Status	Number Participating
SFScience	11
Control	10
Total	**21**

Note: One control teacher that is counted in this table left the district before the school year officially began.

Sample Size

Sample size is one of the factors 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 usually measure the size of an effect in terms of standard deviation units – which tells us how big the effect is, controlling for the spread in observed scores. 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 .46. 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 .34. We emphasize that the matching design that we used will usually lower the size of the effect. From this we see that the experiment is not designed 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. Qualitative data are collected over the entire period of the experiment beginning with the randomization meeting held in May 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 and the context of the study.

Observational and Interview Data

In general, observational data are used to inform the description of the learning environments, 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 10th, 2005. Classroom observations were conducted during the week of March 13th, 2006. Six of the possible 11 teachers in the *SFScience* group were observed, and six teachers from the control group. Five *SFScience* and three control teachers were interviewed individually and in small groups.

Interview data are 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 phone 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. All response rates were calculated based on these expectations. Table 5 summarizes the topics and response rate by survey number. A total of nine surveys were

deployed with an overall response rate of 87.22% for both groups, an 90.9% response rate for the *SFScience* teachers, and a 82.72% 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. For the final survey, survey 9, the topics were modified to allow for the differences between the materials and 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 assignment. The free-response portions of the surveys are minimally coded.

Table 5. 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	72.73%	66.67%	70.00%
Survey 2	Jan. 16 – 20	Resources	90.91%	77.78%	85.00%
Survey 3	Jan. 23 – 27	Interactions with materials/Students	81.81%	80.00%	80.95%
Survey 4	Feb. 6 – 10	More Interactions	100%	100%	100%
Survey 5	Feb. 20 – 24	Time & Preparation	100%	77.78%	90.00%
Survey 6	Mar. 6 – 10	Materials & Resources	100%	77.78%	90.00%
Survey 7	Mar. 20 – 24	Assessments	81.82%	88.89%	85.00%
Survey 8	May 1 – 5	More Interactions	90.91%	77.78%	85.00%
Survey 9T*	May 26	Final Survey	81.82%	N/A	81.82%
Survey 9C**	May 26	Final Survey	N/A	100%	100%

*Asked only of *SFScience* teachers.

**Asked only of control teachers.

The survey topics were developed to account for the various aspects of teacher and student actions associated with instruction and learning. In order to characterize the average time teachers and students spent doing science activities, we used a repeated question strategy. On surveys 2 through 6 we asked two questions: a) Last week, how many days did students spend some amount of time doing science activities (either as a standalone activity or integrated in other subjects)? b) On the average day last week that your students spent time on science activities (either as a stand alone activity or integrated in other subjects) how many minutes did they spend? These questions, together with questions regarding the types of activities, allow us to draw inferences about how time was devoted to science instruction in both the *SFScience* and control groups.

Achievement Measures

The primary outcome measures are student-level scores on the Northwest Evaluation Association (NWEA) test in two areas: Science Concepts and Processes and Reading Achievement. We refer to these tests as Science achievement and Reading achievement when referring to these specific assessments throughout the report. In the fall of 2005, the NWEA Science and Reading tests were administered to the students at the various schools as a pretest measure. As a posttest measure, the Science and Reading tests were administered in the spring of 2006. The paper-and-pencil versions of these tests are referred to as ALT tests and all sites were provided these materials. Both of these tests are adaptive and comprehensive, and are designed to measure growth over time. The sets of tests consist of multiple levels, with overlapping degrees of difficulty. Several different levels are given within the same classroom. To ensure a good match of student to test, there are five test levels for Science and eight test levels for Reading. The first time a student is tested, the appropriate test level is determined by use of a placement test, referred to as a locator test. The locator test is a 20 item test whose sole purpose is to identify which of the leveled test a student is best aligned with the student's anticipated achievement level. Once the level is determined, the student is then provided with that leveled test which is then officially scored by the NWEA. It is this score that is used in the subsequent analyses. During the second and subsequent administrations of the ALT, the student is automatically assigned to a level based on previous results. Teachers were reasonably familiar with the NWEA testing procedures since they had been using the Language Arts test in the districts prior to this study. Researchers provided teachers with 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 typically range from about 150 to 300 and indicate a student's current achievement level along a curriculum scale for a particular subject. Since this is a continuous scale, third grade student scores are usually found lower on scale whereas fifth grade scores are found higher along the scale. The Science Concepts and Processes ALT was specifically selected because we wanted to ensure that differences in state content standards would not be an issue when comparing results across the different grades and across districts. By using a test that emphasizes the concepts and processes of science over specific content, we minimize the impact of the differences in content coverage.

Testing Schedule and Administration

The pretests were given in October and all posttesting was conducted between the last week of April and May 19th 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 the assessments during the pretest period. Visalia Unified had used the NWEA in previous years and some students had reading test scores from the spring 2005 test administration.

Statistical Analysis and Reporting

The basic question for the statistical analyses 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 the identified covariates and the experimental condition.

In order 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.

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. We report two types of effect size: the unadjusted and adjusted effect size. The unadjusted effect size is the difference between treatment and control, controlling for dependencies of observations within randomized units. (This has implications for p-values, but it also affects the estimate of the difference: it weights some cluster averages more than others – therefore we can expect inconsistency between the estimated difference and the raw difference.) The adjusted effect size adjusts for the pretest as well as other fixed and random effects used in the models with interactions that follow.

***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 has not. Thus a *p* value of .1 gives us a 10% probability that the treatment has had the estimated effect when in fact, it did not happen. 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 table addresses the nature of the groups, displaying 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.

Table 6. Distribution of the *SFScience* Group 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	2	11	11	90	132	125	**347**
Control	2	9	9	133	50	86	**269**
Totals	**2**[a]	**20**	**20**	**223**	**172**	**211**	**616**

[a] Both schools participated in both conditions.

Post Randomization Composition of the Experimental Groups

In checking for balance in the composition of the experimental groups, we examine teacher experience, student characteristics such as English proficiency and gender, and student pretest outcomes. From the previous tables, we see that 616 students were enrolled in the study. Of these, 52 students have been designated as needing special education support; we will not include those students in the analysis. Hence, the following analyses are based on a sample size of 564 students (320 in the *SFScience* group and 244 in the control group).

Teacher Variables

Years of Teaching Experience

During the randomization process, we paired teachers according to additional factors such as the grade level they taught and years of teaching experience. We stratified according to years of teaching experience, which we believed affected student scores, to avoid a potential imbalance in outcomes due to chance discrepancies between conditions in this factor.

[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 7. Distribution of Years Teaching Experience

	Number of Teachers		
Condition	0 to 3 years	4 or more years	Totals
SFScience	0	11	**11**
Control	0	9	**9**
Totals	**0**	**20**	**20**

We ran a *t* test to confirm that balance was achieved. Table 8 shows the results of a statistical test of balance between conditions on teachers' experience. We see that there is no difference in years of teaching experience between treatment and control teachers.

Table 8. Difference in Years of Teaching Experience between Students in the *SFScience* and Control Groups

Descriptive statistics: Years of Teaching Experience	Raw group means	Standard deviation	Number of teachers	Standard error
SFScience	9.82	6.19	11	1.87
Control	8.67	3.61	9	1.20
***t* test for difference between independent means**	**Difference**	**DF**	***t* value**	***p* value**
Condition (*SFScience* – control)	1.15	18	-0.49	.63

The following tables further describe the background characteristics of the teachers in the study.

Table 9. Years Teaching Experience

	Number of teachers	Early career (0-3 years)	Emerging professional (4-6 years)	Mid-career professional (7-15 years)	Highly experienced professional (15+ years)
SFScience	11	0%	45. 5%	36.3%	18.2%
Control	9	0%	33.3%	66.6%	0%

Table 10. Years Teaching in Grade Level

	Number of teachers	0-3 years	4-6 years	7-15 years	15+ years
SFScience	11	27.3%	63.6%	9.1%	0%
Control	9	66.7%	33.3%	0%	0%

Table 11. Years Teaching Science

	Number of teachers	0-3 years	4-6 years	7-15 years	15+ years
SFScience	11	9.1%	36.4%	36.4%	18.2%
Control	9	22.2%	22.2%	55.6%	0%

Table 12. Science Coursework in College

	Number of teachers	None	Some	Minor	Major
SFScience	11	27.3%	72.7%	0%	0%
Control	9	0%	100%	0%	0%

Table 13. Recent Professional Development (PD) for Science Instruction

	Number of teachers	Attended PD in last two years	No PD in the last two years
SFScience	11	45.5%	54.5%
Control	9	66.7%	33.3%

Student Variables

English Proficiency

We observe in Table **14** that English proficiency was not distributed evenly between the conditions in spite of randomization. There are proportionally more proficient students in the *SFScience* group than in the control group. Chi-square tests confirm that this characteristic was not balanced between conditions. The imbalance may lead the estimate of the impact to depart from its true value.

Table 14. English Proficiency for *SFScience* and Control Groups

Condition	English Proficiency		
	Not proficient	Proficient	Totals
SFScience	59	261	**320**
Control	60	184	**244**
Totals	**119**	**445**	**564**

Statistics	DF	Value	*p* value
Chi-square	1	3.15	.08

Gender

Table 15 summarizes the distribution of gender. As a result of random assignment, the balance of males and females are evenly distributed across the *SFScience* and control groups. The result of the statistical test is consistent with this assertion.

Table 15. Gender for *SFScience* and Control Groups

Condition	Gender		
	Male	Female	Totals
SFScience	153	167	**320**
Control	121	123	**244**
Totals	**274**	**290**	**564**

Statistics	DF	Value	*p* value
Chi-square	1	0.18	.68

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 estimates. Table 16 shows the results of students without disabilities in grades 3 to 5 for whom pretests were available.

NWEA Science

Table 16. 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	196.46	9.84	254	0.62	0.16
Control	194.87	9.80	196	0.70	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
Condition (*SFScience* – control)	1.60		448	-1.71	.09

[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.

The *SFScience* and control groups had slightly different average pretest scores on the NWEA test, as shown in Table 16. However, when we accounted for the fact that outcomes for students of the same teacher tend to be related by factoring these dependencies in the model, the *p* value increased to 0.54, indicating that the difference we are seeing is very likely due to chance.

Attrition

Based on the cases of non-disability students from grades 3, 4, and 5, a high percentage of students did not take the NWEA science pretests. Out of a total number of 564 students, 114 students or (20%) did not have pretest scores. Of these remaining 450 students, no one is missing posttest scores. Twenty students have posttest scores but are missing pretest scores.

Based on the cases of non-disability students from grades 3, 4, and 5, a high percentage of students did not take the NWEA reading pretests. Out of a total number of 564 students, 243 students or (43%) did not have the pretest scores. Of these remaining 321 students, no one is missing NWEA reading posttest scores. One hundred fifty-four students have posttest scores but are missing pretest scores.

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 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 to how much science instruction took place, how it was conducted and which materials were covered in the *SFScience* group.

Comparison of *SFScience* and Control Groups

Two elementary schools participated in the study; one school had a PreKindergarten, and both covered Kindergarten through 6th grade.

Classroom Settings for Instruction

The classroom setting was observed once during the length of the intervention, during the week of March 13th, 2006. Teachers were observed during classroom instruction for about 50 minutes to an hour, the typical length of a science lesson. Teachers who were not observed were interviewed in small groups during lunch or after school. Teachers were not asked to prepare specific lessons for observation, but we made an effort to coordinate the observation with the teacher prior to observation. Six *SFScience* teachers and six control teachers were observed, eight teachers were interviewed (5 *SFScience*, 3 control), two teachers could not be scheduled for either an observation or an interview. 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.

Some teachers had a few older computer stations in the classroom, but not enough for every student. Televisions and video playback/recorder systems were in evidence or accessible by both teacher groups. Most all of the teachers in both groups supplemented instruction using videos accessed through United Streaming. Two other supplemental materials used by both groups of teachers were "BrainPop" (animated Internet web content) and "Bill Nye the Science Guy" (DVD video). Other teachers reported that they also used the Internet for general research. Every teacher had an overhead projector that he/she used periodically.

The control group teachers had fewer packaged materials to teach science and used a variety of materials such as a weekly science newspaper called Science Studies Weekly.

Opportunities for Learning

Although this site was identified before the beginning of the 2005-2006 academic year in September, certain materials did not arrive in time for the start of school. The school dealt directly with the publisher to obtain the materials and was able to begin full classroom implementation by October, 2005.

At these schools science is taught as part of all subjects taught to the students (self-contained classrooms). Some teachers taught science daily for about 30 minutes, some others taught two to three times a week depending on the week for about 40 minutes each lesson, and still others used an alternating schedule. There were two types of alternating schedule, two to three weeks at a time or two months out of every trimester. An alternating schedule allows the teacher to plan and gather resources to provide instruction for a certain time, teaching science everyday for those weeks and then switching to teach another subject for the next time period. The alternating pattern fit well with the teachers using teacher developed materials.

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. Five of the surveys asked these questions as pertaining to the week immediately preceding the survey so we were able to obtain a sample of data points that we averaged and

then multiplied by the number of weeks of the implementation. One control teacher provided only two data points that were not included in computing the average. This provided an estimate for each teacher of the total amount of science teaching time. *SFScience* teachers reported an average of 47.3 total hours and control teachers reported an average of 45.16 total hours of instruction for the length of the implementation. As we observe later in Table 34, we have no confidence that the actual difference is different from zero.

Control Group Materials

As noted before, there were some textbooks in evidence, but for the most part few reading materials were available consistently for the control group students. When asked about materials usage some control teachers responded as shown in Table 17. Three teachers practiced whole class science reading for more than half of the time they spent on science, but less than 90%. Three teachers reported using whole class reading activities less than half of the time, and three teachers reported whole class reading activities constituting less than 30% of the time, leading us to conclude that this was a relatively common activity.

Table 17. 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	9	33.3%	0%	11.1%	22.2%	55.5%	22.2%

For conducting laboratory activities four control teachers indicated that they spend more than 50% but less than 90% of their time on laboratory activities, but they rarely have sufficient supplies to do them as whole class activities. They tend to be demonstration activities. Five teachers reported spending between 30% and 45%of the time on laboratory activities and one teacher reported less than 30% of the time spent on laboratory activities. Teachers reported that time to incorporate the activities fully, the space to conduct laboratory and availability of materials were all concerns. Teachers did agree that students found the activities fun, but both teachers and students had to work at making the connection with the concepts.

Table 18. 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	9	0%	44.4%	55.6%	0%	0%

Planning time for science instruction is also an important factor for implementing curriculum. Eight of the possible ten control teachers responded that they spent approximately 31% (20

minutes per week) of their total available planning time on science instruction. Two teachers in the *SFScience* group report spending almost no time planning science instruction, the remaining eight report percentages of planning time from 20% to 40% (approximately 12 to 25 minutes).

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 the inquiry process 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 similar types of questions as used in the curriculum 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 *SFScience* program as a continuum:

- Questions (DI: students use a question provided by the teacher, materials or some other source; GI: students are guided to refine and clarify questions; FI: students investigate their own questions)
- 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 22.27 and for the control group it was 30.80. While a greater amount of density is noticed for the control condition, statistical tests (*p* value of .24) indicate that we have no confidence that the actual difference between the groups is different from zero.

Implementation of *SFScience*

Training and Support

The one-half day training took place on September 10, 2005 at one of the schools. During the training, the Scott Foresman representative gave a demonstration of the science kits and the pedagogical method of hands-on inquiry. The trainer emphasized inquiry instruction and provided the teachers a model lesson with an opportunity to use hands-on activities based on the materials provided in the kits. 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 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 trainer 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. Besides the materials listed in Table 1, teachers also received an online log-in so that they could reference additional materials. However, 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.

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 they taught.

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.

Only nine of the ten *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 19. Percent of Teachers Covering Each Chapter in Unit A-Life Science

		Chapter					
		1	2	3	4	5	6
Number of Respondents	9	44.4%	44.4%	33.3%	44.4%	44.4%	33.3%

Note: Two teachers did not provide information.

Table 20. Chapters in Unit B-Earth Science Covered

		Chapter			
		7	8	9	10
Number of Respondents	9	33.3%	22.2%	55.5%	22.2%
Note: Two teachers did not provide information.					

Table 21. Chapters in Unit C-Physical Science Covered

		Chapter				
		11	12	13	14	15
Number of Respondents	9	66.7%	66.7%	22.2%	44.4%	44.4%
Note: Two teachers did not provide information.						

Table 22. Chapters in Unit D-Space & Technology Covered

		Chapter		
		16	17	18
Number of Respondents	9	44.4%	55.6%	22.2%
Note: Two teachers did not provide information.				

As can be seen, teachers mostly taught from all sections of the text, selecting units on the basis of best alignment to the California content standards.

Alignment to standards continues to be a big issue and a challenge at all grades levels. No one teacher completed a full unit because not every chapter is part of the state requirement. Teachers were very vocal about needing texts that strictly align with the standards because it takes much more planning time to make changes and supplement instruction. Each chapter had applicable activities but then the DI-GI-FI sequence was greatly compromised. Only a few teachers completed the inquiry sequence so that they could give the students a Full Inquiry experience. Third grade students were not used to having to pay the amount of attention required by the activities. They understand what to do, but did not yet have the skills to understand the connection between “how to do” and “why/when to do”. Teachers thought that the textbook was too difficult for their students because many students were lacking “foundational experiences.” Many teachers commented that they would like to see video sequences begin the chapter and/or end the chapter.

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

Table 23. 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	9	11.1%	11.1%	22.2%	55.6%	0%

Note. Two teachers did not provide information.

Table 24. 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	9	0%	11.1%	33.3%	55.6%	0%

Note. Two teachers did not provide information.

Table 25. 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	9	0%	0%	55.6%	22.2%	22.2%

Note. Two teachers did not provide information.

Table 26. 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	9	33.3%	11.1%	33.3%	22.2%	0%

Note. Two teachers did not provide information.

Many teachers incorporated the Leveled Readers into their science instruction and also used it successfully with their reading instruction. All of the teachers remarked that the Leveled Readers were very successful for their students. They noted two difficulties, the packaging – insufficient numbers to be used with whole class small groups, and the vocabulary was too difficult for their English Language Learners and would have preferred even lower levels readers.

As for the Science Kits, teachers did like the convenience of the kits, specifically having all of the materials ready to hand. They thought it was easy to set-up and clean-up afterwards. Several teachers commented that not all materials were included in the kit. Also, several

teachers reported that getting the experiment to "work" was sometimes very problematic. Additionally, teachers commented that space to conduct the experiment was problematic and on occasion resorted to the demonstration model only.

Some *SFScience* teachers used the assessments with their students. In general teachers thought the questions too difficult for their students. Most teachers created their own assessments. Only two teachers had discovered the Test-maker CD and indicated that they preferred the flexibility offered by the ability to modify and select questions 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 Visalia, implementation was weaker than the desired ideal model, mainly due to standards alignment.

Summary of Implementation

Certain factors emerged as barriers to a smooth implementation. For this site the large ELL population requires teachers to support the text with video and additional prior experiences before students can begin the chapters. Lack of alignment to the standards also contributed to the overall modified implementation. Teachers tended to supplement and modify the chapters as necessary to meet the demands of the student population. Lack of space contributed to the inability to use the *SFScience* materials as often as intended. The length of implementation was approximately 6 months.

Quantitative Impact Results

The primary topic of our experiment was the impact of *SFScience* curriculum on student performance on the NWEA test. Impact is measured in terms of the difference between performance of the *SFScience* students and the control students. We will first address the impact on Science achievement and then the impact on Reading achievement.

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

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

We note that the number of cases used to compute the effect size 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 Achievement

Analysis Including Pretest

Our first analysis addressed Science achievement using the NWEA Science Concepts and Processes scale. Table 27 provides a summary of the sample we used in the analysis and the results for the comparison of *SFScience* and control. 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 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 truly there is no difference. The "Adjusted" row is also 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 28 and Table 29. The means, and therefore the effect size, are adjusted to take into account the student pretest scores. The adjusted effect size is based on a model that includes fixed effects for schools as well as the matched pairs within which we randomized. It also figures in the effect of students being grouped within teachers.

Table 27. Overview of Sample and Impact of *SFScience* on Science Achievement

	Condition	Means	Standard deviations[a]	No. of Students	No. of Classes	No. of Teachers	Effect Size	*p* value[b]
Un-adjusted	***SFScience***	197.24	10.39	263	11	11	<. 01	.96
	Control	197.18	9.63	207	9	9		
Adjusted	***SFScience***	196.18	10.45	254	11	11	-0.13	.25
	Control	197.51[c]	9.72	196	9	9		

[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 teacher but does not adjust for any other covariates. The *p* value for the adjusted effect size is computed using a model that that figures in clustering and includes the pretest as a covariate, as well as other fixed effects, as needed.

[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 1 provides a visual representation of the information in Table 27. The bar graphs represent average performance in the metric of NWEA Science.

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

The panel on the right shows estimated performance on the posttest for the two groups based on a model that adjusts for students' pretest scores and other fixed effects (i.e., this is a visual display of results from the row labeled 'adjusted' in Table 27.) We can see that the two groups were essentially indistinguishable. The *p* value for the treatment effect (.25) 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 to indicate the range of the possible scores.

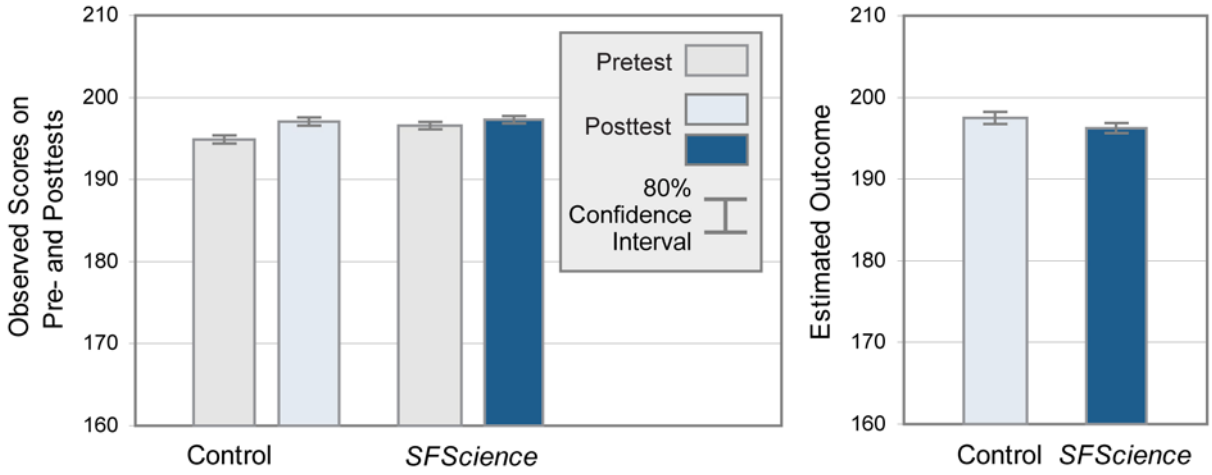

Figure 1. Impact on Science: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)

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] 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 28 shows the estimated impact of *SFScience* on students' performance in science as measured by the NWEA Science test.

[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 English proficiency. 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 28. The Impact of *SFScience* on Student Performance on Science Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Estimated value for a control student with an average pretest	196.44	0.94	17	209.88	<.01
Impact of *SFScience* for a student with an average pretest	-1.31	0.20	17	-1.32	.20
Estimated change in control outcome for each unit increase on the pretest	0.71	0.05	427	14.81	<.01
Interaction of pretest and *SFScience*	0.12	0.06	427	1.92	.06
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	3.37	1.72		1.96	.02
Within-teacher variation	33.52	2.29		14.60	<.01

[a] Schools were modeled as a fixed factor but the estimated effects are not included in this table; because we estimated fixed effects for schools, the estimated value for a control student with an average pretest applies to a particular school.

[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 the Science NWEA for a student who has an average score on the pretest. The estimate associated with *SFScience* is -1.31 with a *p* value of .20. Using the criteria outlined earlier in the report, we conclude that we have limited confidence that the true impact 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 .06. We have some confidence that the true effect is different from zero. In other words, the effect of *SFScience* was different depending on the student's prior score. The result indicates that the treatment has a negative effect for students scoring on the lower range (3rd grade) and no effect for students scoring in the mid to upper ranges.

As a visual representation of the results described in Table 28, we present a scatterplot in Figure 2, which shows student performance at the end of the year in science, as measured by NWEA Science, against their performance on NWEA Sciende in the fall. This graph shows where each student fell in terms of his or her starting point (horizontal, x-axis) and his or her outcome score (vertical, y-axis). Each point represents one student's post-intervention score

against his or her pre-intervention score. The darker points represent *SFScience* students; the lighter points, control students. The shaded area in the lower right of the graph is the area of negative change (i.e., where students lost ground).

The two lines are the estimated values on the posttest for students in the *SFScience* and control conditions as determined using a simple model with no fixed effects.[3] We see that the slopes of the two lines are different, an indication of the interaction effect.[4]

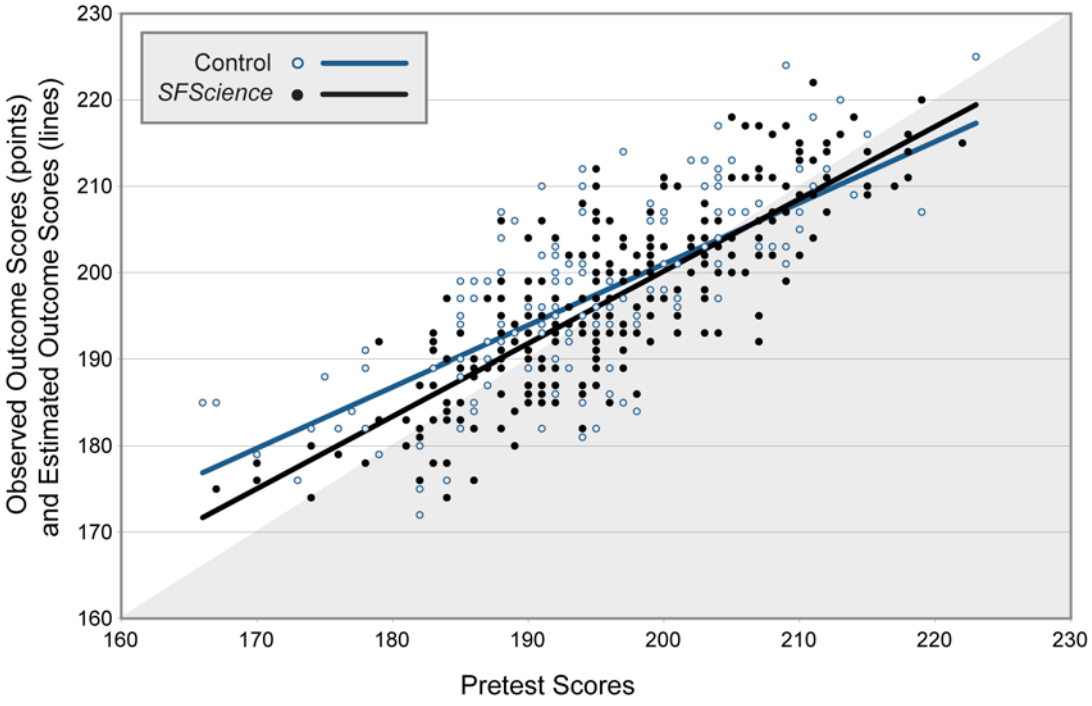

Figure 2. Comparison of Estimated and Actual Science Achievement for *SFScience* and Control Group Students

Figure 3 illustrates the interaction in terms of the estimated difference between the *SFScience* and control groups for different points along the prior score scale. This display of the results

[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).

[4] The lines representing the estimated values are centered on the no-growth line – this reflects that there was very little growth from pre to post. As a result of this, and the fact that extreme scores tend to regress to the mean we see that students with low pretest scores rise above the area of negative gain whereas those with high pretest scores dip into the area of negative gain. This phenomenon is due to regression to the mean. The critical point concerning the interaction is the fact that the lines representing estimated values cross.

allows us to see where *SFScience* had its greatest impact.[5] In this graph the estimated difference between *SFScience* and control groups is expressed as the straight line in the middle of the shaded bands – it is the estimated outcome for a *SFScience* student minus the estimated outcome for a control student. Around the difference line, we provide gradated bands representing confidence intervals. These confidence intervals are an alternative way of expressing uncertainty in the result. The band with the darkest shading surrounding the dark line is the "50-50" area, where the difference is considered equally likely to lie within the band as not. The region within the outermost shaded boundary is the 95% confidence interval—we are 95% sure that the true difference lies within these extremes. Between the 50% and 95% confidence intervals we also show the 80% and 90% confidence intervals. We also add points along the middle line to mark what the estimated treatment effect is for the median student for each quartile of the pretest. Consistent with the results in Table 28, there is evidence of a differential impact of the intervention across the prior score scale as measured by the NWEA Science Test considering the points representing the median student in the bottom and top quartiles, it appears that *SFScience* has no benefit for students scoring in the higher ranges and has a negative effect for students scoring in the lower range.

[5] As with the scatterplot, 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).

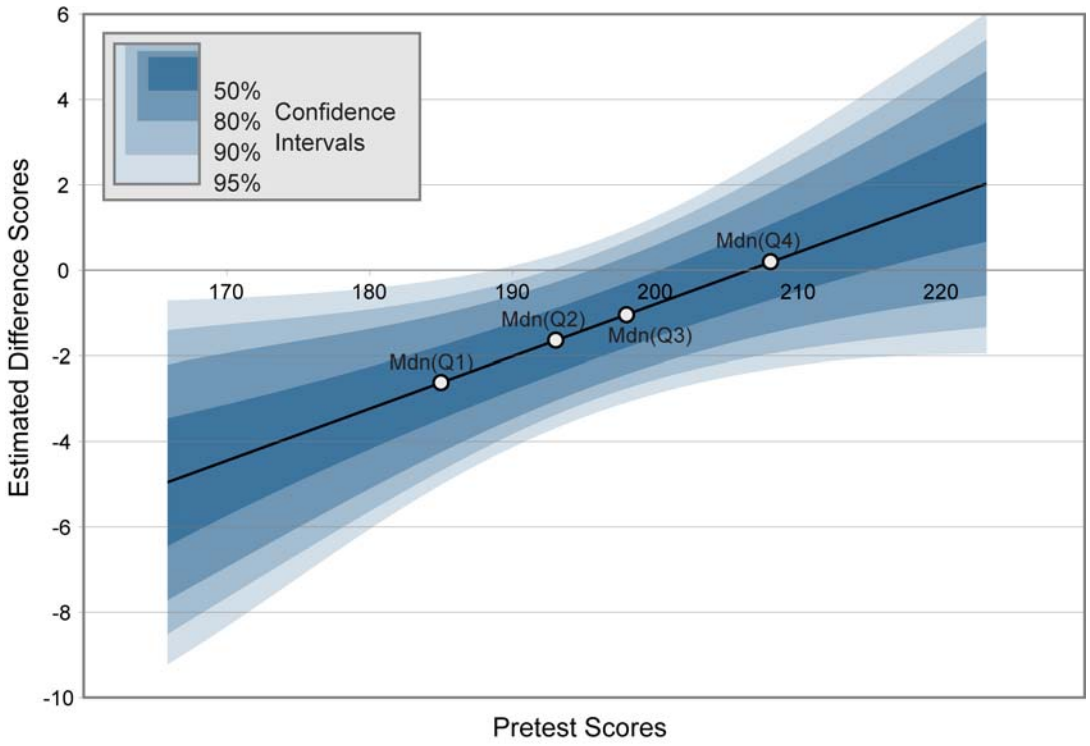

Figure 3. Differences between *SFScience* and Control Group Science Achievement: Median Pretest Scores for Four Quartiles Indicated

Figure 4 presents the same information represented in Figure 3 but this time in the form of a bar graph showing the estimated posttest scores and the difference between *SFScience* and control conditions for students at the medians of the first and fourth quartiles as identified by the pretest measure. The bar graph includes the 80% confidence interval as a marker at the top of the bars. This marker is an alternative representation of the 80% band in Figure 3 and is meant to be interpreted as: for either *SFScience* -control comparison, we are 80% sure that the true difference between conditions would place the tops of the bars simultaneously within the confidence interval markers. We see that for a student scoring at the median of the first quartile the estimated performance in the treatment condition is lower than that in the control condition and there is no overlap in the confidence intervals. There is little difference in estimated performance between treatment and control for a student at the median of the fourth quartile.

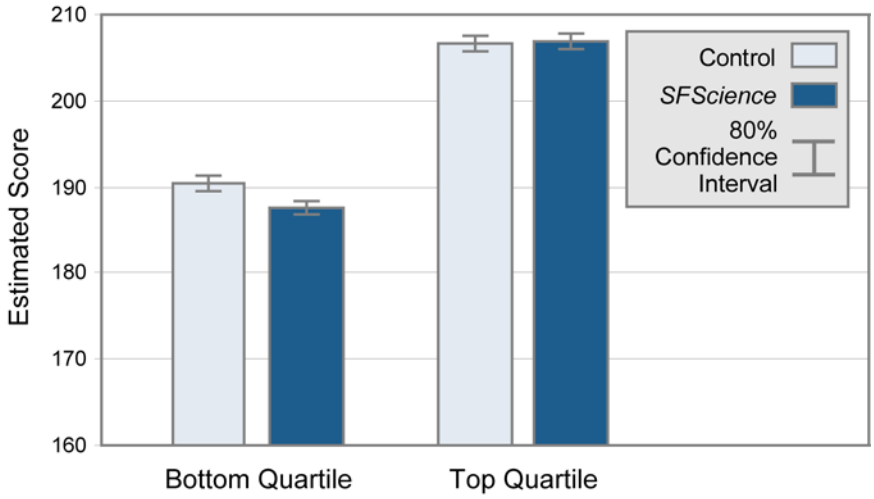

Figure 4. Difference between *SFScience* and Control Group Science Achievement: Median Students in Top and Bottom Quartiles

Analysis Including Gender as a Moderator

We were also interested in whether *SFScience* was differentially effective for males and females because much of the research literature indicates that gender differences exist in students' performance on science outcomes.[6] Table 29 shows the moderating effect of gender on students' performance on the NWEA Science test. The advantage of being in the *SFScience* condition is greater for girls than it is for boys. The *p* value of .19 gives us limited confidence that the actual differential impact is different from zero.

Table 29. 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	195.79	1.00	17	195.85	<.01
Estimated change in outcome for each unit increase on the pretest	-0.61	1.13	17	-0.54	.60

[6] As a rule, we decide on moderator variables before the beginning of the trial. We declare which variables we will examine the moderating effects of in advance to demonstrate that we are not mining results post hoc. The moderators are variables of theoretical interest that potentially affect how strongly the treatment impacts the outcome. We report only the result for which we have at least limited confidence that the true effect is different from zero.

Average *SFScience* effect for girls	1.57	0.85	426	1.84	.07
Difference (boys minus girls) in average performance in the control condition	0.78	0.03	426	24.27	<.01
Difference (boys minus girls) in the average *SFScience* effect	-1.49	1.13	426	-1.32	.19

Random effects[b]	Estimate	Standard error	*z* value	*p* value
Teacher mean achievement	3.24	1.67	1.95	.03
Within-teacher variation	33.65	2.31	14.59	<.01

[a] Schools were modeled as a fixed factor but the estimated effects are not included in this table; because we estimated fixed effects for schools, the estimated value for a control student with an average pretest applies to a particular school.

[b] Teachers were modeled as a random factor.

In Figure 5, the overlap of the confidence intervals shows that the differential impact of *SFScience* for male and female students can easily be due to chance. This is consistent with our finding in Table 29 in that we have limited confidence that the actual differential achievement between females and males in the *SFScience* group is different from zero. However, we observe that male students achieve slightly higher science scores than female students in the control group.

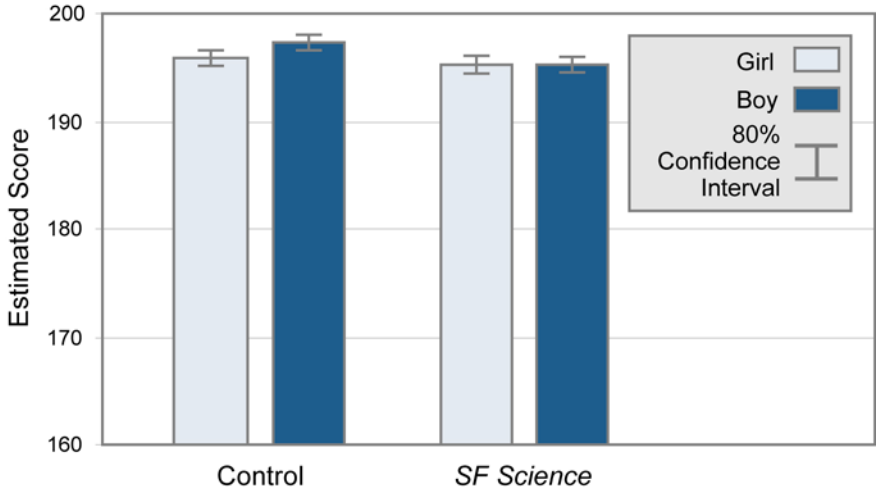

Figure 5. Moderating Effect of Gender on Science Achievement

Analysis Including English Proficiency as a Moderator

We were also interested in the moderating effect of student English proficiency on science achievement. In particular, we were interested in whether *SFScience* was differentially effective for English proficient and non-English proficient students. Table 30 shows the results of our analysis. We observe that there is no differential effect of *SFScience* depending on English proficiency status.

Table 30. Science Achievement Moderated by English Proficiency

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Outcome for English learner in control group with an average pretest	197.05	1.17	17	168.31	<.01
Estimated change in outcome for each unit increase on the pretest	0.78	0.03	426	23.78	<.01
Average *SFScience* effect for English learner	-2.30	1.47	17	-1.57	.13
Difference (score for English proficient minus English learner) in average performance in the control condition	-0.75	1.02	426	-0.74	.46
Difference (score for English proficient minus English learner) in the average *SFScience* effect	1.25	1.40	426	0.89	.37

Random effects[b]	Estimate	Standard error	*z* value	*p* value
Teacher mean achievement	3.20	1.66	1.93	.03
Within-teacher variation	33.86	2.32	14.59	<.01

[a] Schools were modeled as a fixed factor but the estimated effects are not included in this table; because we estimated fixed effects for schools, the predicted value for a control student with an average pretest applies to a particular school.

[b] Teachers were modeled as a random factor.

[c] The prior score was centered at the mean; therefore, the effect estimates apply to a student who had an average score on the pretest.

Reading Achievement

Analysis Including Pretest

Our next set of analyses addressed reading achievement as measured by NWEA Reading. Table 31 provides a summary of the sample we used in the analyses and the results for the comparison of *SFScience* and control. 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 pre- and posttests. The adjusted effect size is 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. This is the sample that we use in the analyses on which we base our results reported in Table 31 and Table 32. The means, and therefore the effect size, are adjusted to take into account the student pretest scores. The adjusted effect size is based on a model that includes fixed effects for schools as well as pairs within which we randomized. It also figures in the effect of students being grouped within teachers.

Table 31. 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***	203.26	14.00	265	11	11	-0.12	.30
	Control	202.61	14.18	210	9	9		
Adjusted	***SFScience***	201.56	13.20	167	11	11	-0.03	.64
	Control	201.91[c]	13.53	154	9	9		

[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 figures in clustering and includes the pretest as a covariate, as well as other fixed effects, as needed.(Copied from Federal Way)

[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 6 provides a visual representation of specific information in Table 31. The bar graphs represent average student performance on 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 31.) We can see that the two groups were essentially indistinguishable. The *p* value for the treatment effect indicates we 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 illustrates there is a .64 probability that the difference is due to chance.

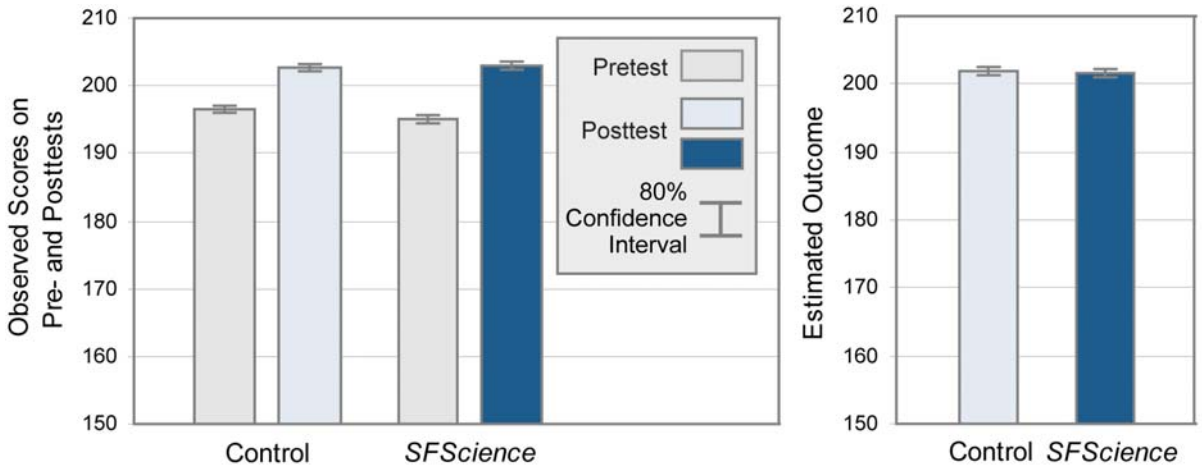

Figure 6. Impact on Reading Achievement: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)

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. 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 32 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 32. The Impact of *SFScience* on Student Performance on Reading Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Estimated value for a control student with an average pretest	198.72	0.65	17	306.14	<.01
Impact of *SFScience* for a student with an average pretest	-0.36	0.75	17	-0.48	.64
Estimated change in control outcome for each unit increase on the pretest	0.88	0.04	297	22.85	<.01
Interaction of pretest and *SFScience*	0.01	0.05	297	0.23	.81
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	0.25	0.88		0.28	.39
Within-teacher variation	38.57	3.14		12.28	<.01

[a] Schools were modeled as a fixed factor but the estimated effects are not included in this table; because we estimated fixed effects for schools, the estimated value for a control student with an average pretest applies to a particular school.

[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 NWEA Reading for a student who has an average score on the pretest. The estimate associated with *SFScience* is -0.36. This shows a small negative difference associated with *SFScience*. However, the *p* value of .64 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 .81. We have no confidence that the true effect is different from zero.

As a visual representation of the results described in Table 32, we present a scatterplot in Figure 7, which shows end-of-year student performance, as measured by NWEA Reading, against their performance on NWEA Reading in the fall. This graph shows where each student stands 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 two lines are the estimated values on the posttest for students in the *SFScience* and control conditions as determined using a simple model with no fixed effects.[7] We see that the slopes of

[7] 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

the two lines are parallel, an indication that the moderating effect of the pretest score on the impact of *SFScience* was not differentially effective for students at different points along the pretest scale.

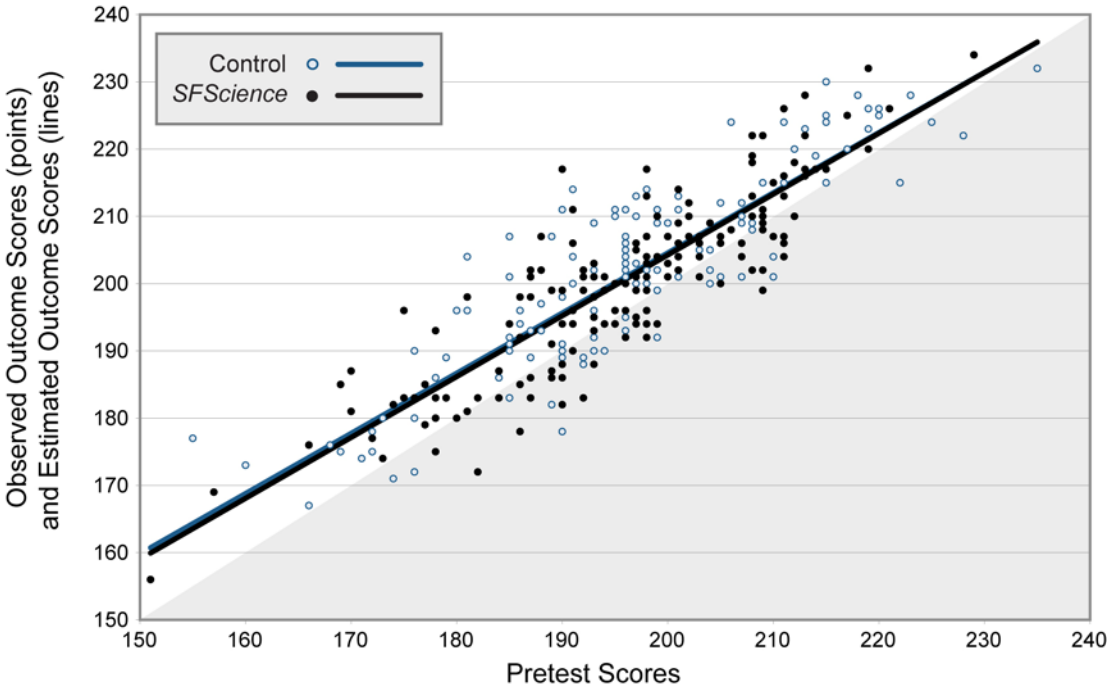

Figure 7. Comparison of Estimated and Actual Reading Achievement for *SFScience* and Control Students

Analysis Including English Proficiency as a Moderator

We were also interested in the moderating effect of student English proficiency on reading achievement. Table 33 shows the results of our analysis. We observe that there is no differential effect of *SFScience* depending on English proficiency status.

---

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 33. Reading Achievement Moderated by English Proficiency

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Outcome for English learner in control group with an average pretest	197.98	1.09	17	182.29	<.01
Estimated change in outcome for each unit increase on the pretest	0.89	0.03	295	32.51	<.01
Average *SFScience* effect for English learner	-0.01	1.52	17	0.00	.99
Difference (score for English proficient minus English learner) in average performance in the control condition	1.13	1.19	295	0.95	.34
Difference (score for English proficient minus English learner) in the average *SFScience* effect	-0.56	1.66	295	-0.34	.73
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	0.76	1.11		0.69	.25
Within-teacher variation	37.52	3.07		12.22	<.01

[a] Schools were modeled as a fixed factor but the estimated effects are not included in this table; because we estimated fixed effects for schools, the predicted value for a control student with an average pretest applies to a particular school.

[b] Teachers were modeled as a random factor.

[c] The prior score was centered at the mean; therefore, the effect estimates apply to a student who had an average score on the pretest.

Exploratory Analysis of 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 defining a distinct path or link between the intervention and student-level achievement, as illustrated in Figure 8. Part of the impact of *SFScience* on student outcomes may be mediated by the intermediate variables. *SFScience* can have a direct impact on both student outcomes and on instructional time, a teacher-level outcome. The link from instructional time to the student outcome is correlational but an important relationship to explore.

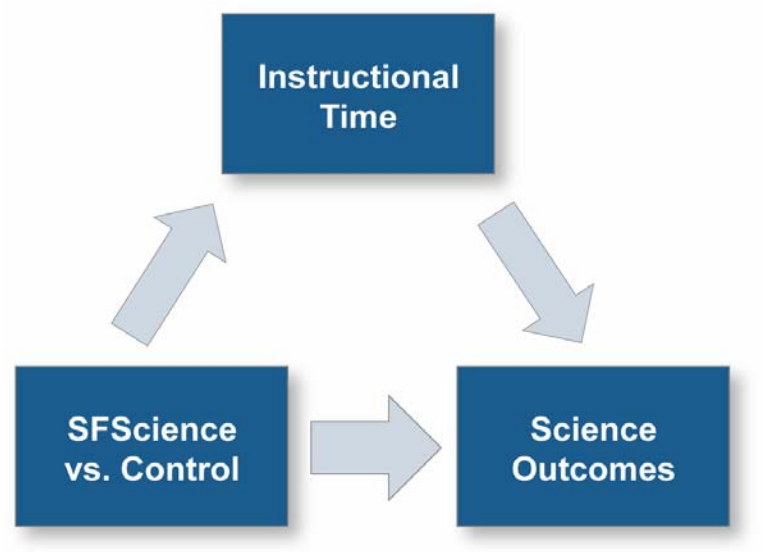

Figure 8. Relationships for Exploratory Analysis of Implementation Variables

Instructional Time

We wanted to explore the relationship between how much time was spent teaching science and science achievement. The surveys provided data on this variable. Our measure is the total hours spent teaching science during the experiment. Instructional time was measured by each teacher's self-report of the number of minutes she or he spent using *SFScience* per week, 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 34 shows *SFScience* teachers taught approximately two less hours of science during the year; However, the high *p* value of <.84 gives us a no confidence that the actual difference is different from zero. In other words, we have a no confidence that *SFScience* has any effect on the number of hours used on science instruction.

Table 34. 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	58.98	10.35	17	5.70	<.01
Impact of *SFScience* on hours of science instruction	-2.35	11.17	17	-0.21	.84
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Residual teacher variance	617.33	211.74		2.92	.02

[a] Schools are modeled as fixed factors. The estimates of these effects are not included in this table.

The table above shows that there is no impact of *SFScience* on Science Instruction Time. Even with little difference between the two groups, it is useful to explore whether there is a relationship between amount of science instructional time and student achievement. The result of this analysis is purely correlational – we have not assigned teachers to levels of instructional time with *SFScience* so 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 positive relationship between *SFScience* usage and the student outcome. However, the *p* value for this effect is .62, which gives us no confidence that the true relationship is in fact different from zero.

Table 35. Relationship of Instructional Time to Student Outcome

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Estimated value for a student with an average pretest	195.09	1.48	17	131.51	<.01
Estimated change in outcome for each unit increase on the pretest	0.77	0.03	428	24.01	<.01
Estimated change in outcome for each unit increase in instructional time	0.01	0.02	17	0.50	.62
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	3.58	1.78		2.01	.02
Within-teacher variation	33.79	2.3		14.62	<.01

[a] Schools were modeled as a fixed factor but the estimated effects are not included in this table; because we estimated fixed effects for schools, the estimated value for a control student with an average pretest applies to a particular school.

[b] Teachers were modeled as a random factor.

Discussion

We began this research in Visalia Unified School District 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. However, in science, we found that *SFScience* tended to be slightly less effective than the existing program for students initially scoring at the lower end of the pretest scale. Since the pretest we used (NWEA Science) is scored along a continuous growth scale, we might translate this finding into an expectation that the program may have less benefits for students in the earlier grades (within the third to fifth grade range of our experiment).

We did not find any difference in the value for reading depending on the student's initial reading 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 English proficiency status. We found an effect, for which we have limited confidence, that *SFScience* closed the initial gap between boys and girls in science achievement. We did not find any differential effect of *SFScience* on science or reading achievement between English speakers and English learners.

Our experiment in Visalia was small, involving only 20 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. In this case, for example, the opportunities for working with *SFScience* were limited because of a late delivery of some of the materials and the fact that teachers perceived the program as having a poor alignment to the state standards. This lack of alignment led teachers to skip sections disrupting the sequence of activities and the steps in the scaffolded inquiry process. An otherwise effective program has little chance to prove itself without a tight alignment to the goals set for instruction at the school.

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.