Empirical Education®

RESEARCH REPORT

Comparative Effectiveness of *Scott Foresman Science*:

A Report of a Randomized Experiment in Federal Way Public Schools

Gloria I. Miller
Andrew Jaciw
Xin Wei
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 Federal Way Public Schools for their assistance and cooperation in conducting this research. The research was sponsored by Scott Foresman, which provided Empirical Education Inc. with independence in reporting the results.

About Empirical Education Inc.

Empirical Education Inc. was founded to help school districts, publishers, and the R&D communities 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 Federal Way Public Schools

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* (*SFScience*) curriculum and associated materials. This report addresses the experiment in Federal Way Public Schools in Washington State.

The primary purpose of this research is to produce scientifically based evidence of the comparative effectiveness of the *Scott Foresman Science* program. The question being addressed by the research is whether the *SFScience* is more effective than the current curriculum being used by the participating campuses in the Federal Way Public Schools District. The research focuses on 3rd, 4th, and 5th grade students. The outcomes were measured by student achievement on standardized tests administered at the beginning and end of the project. Two test areas were selected as the outcome measures: the Northwest Evaluation Association's Science Concepts and Processes, and Reading Achievement assessments.

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

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

Methods

Research Design

Our study is a comparison of outcomes for classes taught using the *Scott Foresman Science* curricular materials (*SFScience* group) and classes taught with the current materials used in the district (control group). Teachers volunteered for participation and, from 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 Activity Flip Chart Vocabulary Cards (set) Teacher’s Edition Package Teacher’s Resource Package 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	Student Edition Activity Book Workbook Science Kits (one for each of the four units, sufficient supplies for a class of 32, eight groups of four) Leveled Readers Super Kit: includes six copies of each of 12 Below-Level, On-Level, and Advanced Leveled Readers).

District Science Materials

The Federal Way Public Schools District continues to develop science materials. They have developed several science “kits” that are shared and rotated among the teachers in the district on approximately two month basis. The kits contain a teacher’s guide, hands-on activities, short reading materials, and worksheets. Few teachers have textbooks for students; when they did, the textbooks in use were older versions of Scott Foresman and a smattering of other materials. Teachers report that they did not have entire class sets of any textbooks.

Site Descriptions

Federal Way, WA

The city of Federal Way is located 25 miles south of Seattle and just 8 miles north of Tacoma. It is the sixth largest city in Washington State with a population of almost in 86,000 people in a 22 mile square area.

Table 2. Federal Way Racial Makeup

Race/Ethnicity	% of Population
White	68.8
African American	7.9
American Indian/Native Alaskan	0.9
Asian	12.3
Native Hawaiian or Pacific Islander	1.0
Other Race	3.7
Two or more Races	5.3
Hispanic Origin (of any race)	7.5

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

Federal Way Public Schools, WA

Federal Way Public Schools serve a larger area than the city of Federal Way including two other cities located in King County. Federal Way Public Schools operate 37 schools, 23 elementary schools, seven middle schools, Public Academy, Internet Academy (K-12), and five high schools; three elementary (K-6) schools participated in this study. The following tables summarize the demographic makeup of the school district.

Table 3. Background of the Federal Way Public Schools

Federal Way Public Schools	
Total schools	37
Total teachers	1147
Student to teacher ratio	19.6
Grades	PK -12
Student population	22,449
Migrant students	0%
ELL students	9.2%

Source: CCD Public School District Data for 2005-2006

Table 4. Ethnic Makeup of the Federal Way Public Schools

Race/Ethnicity	% of Population
White, non-Hispanic (%)	51.5
Black, non Hispanic (%)	13.7
Hispanic (%)	14.6
Asian/Pacific Islander (%)	15.8
American Indian/Alaskan Native (%)	1.6

Source: Office of Superintendent of Public Instruction, Washington State Report Card, 2006 – Federal Way Public Schools only.

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 Federal Way Public Schools 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 Federal Way Public Schools occurred on June 14, 2005 with 20 teachers, three principals, and two district-level administrators. 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. One teacher decided not to participate and excused herself. Of the remaining 19 teachers, all filled out consent forms, and principals filled-out contact information for an additional three teachers, who could not be present for the meeting because of previous engagements. These three teachers were contacted first by email and then with a phone call to explain the particulars of the study. In total 22 teachers signed consent forms and agreed to participate.

Randomization

The unit of randomization at this site is the teacher. Twenty-two 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 June, teaching assignment was not confirmed until September. At that time, 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 grade level taught and on whether they taught Gifted and Talented students, resulting in a within grade-level randomization paired on student groups and schools. 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 are displayed in the table below.

Table 5. Participating Teachers

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

Note: Twenty-two teachers were originally randomized, one control teacher left the district.

Sample Size

Sample size (in this case, number of teachers) 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 impact 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 further lowers this value. 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 June 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 15th, 2005. Classroom observations were conducted during the week of February 21st, 2006. Teachers at this district strongly preferred to be interviewed rather than be observed and so only two teachers in the *SFScience* group were observed, and one teacher from the control group. Nine *SFScience* and eight 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 6 summarizes the topics and response rate by survey number. A total of nine surveys were deployed with an overall response rate of 79.37% for both groups, an 87.88% response rate for the *SFScience* teachers, and a 70.00% 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 6. 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	63.63%	40.00%	52.38%
Survey 2	Jan. 16 - 20	Resources	100%	80.00%	90.48%
Survey 3	Jan. 23 - 27	Interactions with materials/Students	81.81%	80.00%	80.95%
Survey 4	Feb. 6 - 10	More Interactions	100%	80.00%	90.48%
Survey 5	Feb. 20 - 24	Time & Preparation	100%	90.00%	95.24%
Survey 6	Mar. 6 - 10	Materials & Resources	100%	90.00%	95.24%
Survey 7	Mar. 20 - 24	Assessments	72.72%	60.00%	66.67%
Survey 8	May 1 - 5	More Interactions	72.72%	20.00%	47.62%
Survey 9T*	May 26	Final Survey	100%	N/A	100%
Survey 9C**	May 26	Final Survey	N/A	90.00%	90.00%

*Asked only of *SFScience* teachers.

**Asked only of Control teachers.

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. Researchers provided teachers with a one-hour review of the testing procedures and given a Proctor manual. Researchers provided additional support by pre-packaging all testing materials on an individual teacher basis.

These tests are scored on a Rasch unIT (RIT) scale, a measurement scale developed to simplify the interpretation of test scores. This scale is used to measure student achievement and student growth on an equal-interval scale so that a change of one unit indicates the same change in growth, regardless of the actual numerical values. RIT scores 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 November 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.

Teachers did report that 3rd grade students had some difficulty in completing the tests and some students took 2 or more hours finishing each test. Other teachers reported that some of their higher achieving 5th grade students took long periods of time with each test. All teachers perceived that the tests were not necessarily easy and that students were not accustomed to being tested in this way (two test administrations each with a locator test component.)

Statistical Analysis and Reporting

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

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

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

Effect sizes. We also translate the difference between treatment and control into a standardized effect size by dividing the difference by the amount of variability in the outcome (also called the standard deviation). This allows us to compare the results with results we find 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. 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 size 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* <=.05. (This is the level of confidence conventionally referred to as "statistical significance.")
2. We have some confidence when .05< *p* <=.15.
3. We have limited confidence when .15 < *p* <=.20.
4. We have no confidence when *p* > .20.

Results

Formation of the Experimental Groups

Groups as Initially Randomized

The randomization process guarantees that there is no intentional or unintentional bias in the selection of teachers and students into the treatment or the control condition. It does not, however, guarantee that the groups will be perfectly matched. It is important to inspect the two groups to determine whether, in spite of randomization, there are any significant differences on factors that affect the outcome[1]. The following tables address the nature of the groups. Table 7 shows the distribution of teachers, classes, grades, and students between *SFScience* and control conditions. This is the complete number of students in the experiment at the time that the experiment began in September 2005.

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

Table 7. Distribution of the *SFScience* and Control Groups by Schools, Teachers, Grades, and Counts of Students

	No. of schools	No. of teachers	No. of classes	Students in Grade 3	Students in Grade 4	Students in Grade 5	Total students
SFScience	3	11	11	114	89	75	**278**
Control	3	10	10	100	52	100	**252**
Totals	**3**[a]	**21**	**21**	**214**	**141**	**175**	**530**

[a] Each of the 3 schools participated in both conditions.

Teacher Variables

Years of Teaching Experience

During the randomization process we paired teachers according to additional factors such as the grade level they taught, whether or not they taught regular self-contained classrooms, and years of teaching experience. We did ask teachers to indicate years of teaching experience and other background information. We stratified according to this variable, which we believed affected student scores, to avoid a potential imbalance in outcomes due to chance discrepancies between conditions in years of teaching experience. As can be seen from the tables below, years of teaching experience is not a differentiating factor since most teachers are established in their careers.

Table 8. Distribution of Years Teaching Experience

	Number of Teachers		
Condition	**0 to 3 years**	**4 or more years**	**Totals**
SFScience	1	10	**11**
Control	1	9	**10**
Totals	**2**	**19**	**21**

The following tables further describe the background characteristics of the teachers in the study. In general, most of the teachers in the study are established in their careers and hold college degrees with no particular emphasis on science coursework. One difference noted is the number of years teaching at the current grade level. Many of the teachers in both the *SFScience* condition and control were relatively new to teaching at their grade level.

Additionally, we noted that some teachers alternate teaching grade levels because they have a looping schedule that allows them to teach the same group of students for two years.

Table 9. Years Teaching Experience

		Early career (0-3 years)	Emerging professional (4-6 years)	Mid-career professional (7-15 years)	Highly experienced professional (15+ years)
	Number of teachers	%	%	%	%
SFScience	11	9.1%	18.2%	36.6%	36.6 %
Control	10	10%	10%	20%	60%

Table 10. Years Teaching in Grade Level

		0-3 years	4-6 years	7-15 years	15+ years
	Number of teachers	%	%	%	%
SFScience	11	63.6%	9.1%	9.1%	9.1%
Control	10	50%	10%	30%	10%

Table 11. Years Teaching Science

		0-3 years	4-6 years	7-15 years	15+ years
	Number of teachers	%	%	%	%
SFScience	11	9.1%	18.2%	63.6%	9.1%
Control	10	10%	10%	40%	40%

Table 12. Science Coursework in College

		None	Some	Minor	Major
	Number of teachers	%	%	%	%
SFScience	11	0%	100%	0%	0%
Control	10	0%	100%	0%	0%

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

		Attended PD in last two years	No PD in the last two years
	Number of teachers	%	%
SFScience	11	9.1%	90.9%
Control	10	40%	60%

Post Randomization Composition of the Experimental Groups

In checking for balance in the composition of the experimental groups, we examine student characteristics such as ethnicity and gender and student pretest outcomes. We use a chi-square test when testing for balance or a Fisher's exact test if expected group counts are less than 10.

Student Variables

Ethnicity

Table 14 summarizes the distribution of student ethnicity. The predominant ethnic group within our sample is Caucasian, followed by Asian. This coincides with the general ethnic composition of the city of Federal Way, which implies that this sample is a good representation of the community. As a result of random assignment, the ethnicity of the students is evenly distributed across the *SFScience* and control groups. The result of the statistical test is consistent with this assertion.

Table 14. Ethnicity for *SFScience* and Control Groups

	Ethnicity						
Condition	Asian	Hispanic	Native American	Multi-racial	Black	White	Totals[a]
SFScience	73	21	5	2	48	128	277
Control	59	27	4	0	47	114	251
Totals	132	48	9	2	95	242	528
Statistics		**Value**	***p* value**				
Fisher's Exact Test		0.00	.63				

Note. Since some of the cells have expected counts less than 10, Chi-square may not be a valid test. Therefore, Fisher's exact test is used here.

[a] There are 2 students who are missing ethnicity information.

Gender

Table 15 summarizes the distribution of gender. As a result of random assignment, the balance of males and females is 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	146	132	278
Control	133	119	252
Total	279	251	530
Statistics	**DF**	**Value**	***p* value**
Chi-square	1	0.01	.95

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

NWEA Science

Table 16. Difference in Science Pretest Scores between *SFScience* and Control Students

Descriptive statistics: Pretest outcomes	Raw group means	Standard deviation	Number of students[b]	Standard error	Effect size[a]
SFScience	198.28	9.62	194	0.69	.18
Control	196.68	8.09	179	0.60	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
Condition (*SFScience* – control)	1.60		371	-1.73	.08

[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.

[b] 157 students are missing fall science test scores.

The *SFScience* and control groups had slightly different average pretest scores on NWEA Science, 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 0.72, indicating that the difference we are seeing is very likely due to chance.

NWEA Reading

Table 17. Difference in Reading Pretest Scores between *SFScience* and Control Students

Descriptive statistics: Pretest outcomes	Raw group means	Standard deviation	Number of students[b]	Standard error	Effect size[a]
SFScience	203.29	14.05	197	1.00	.25
Control	199.98	12.67	181	0.94	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
Condition (*SFScience* – control)	3.31		376	-2.40	.02

[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.

[b] 152 students are missing fall reading test scores.

As with NWEA Science, the *SFScience* and control groups had slightly different average pretest scores on NWEA Reading. Again, when we accounted for the fact that outcomes for students of the same teacher tend to be related by modeling these dependencies, the *p* value increased to 0.48, again indicating that this difference is likely due to chance. In the analyses that follow, we add the pretest covariate in order to increase the precision of the impact estimate. (Still, we recognize that, with or without this covariate, the impact estimate is unbiased as a result of the randomization.)

Attrition after the Pretest

NWEA Science

A high percentage of students did not take the NWEA Science pretests. Out of a total enrollment of 530 based on fall class rosters, 157 students or (29.6%) do not have pretest scores. Of these remaining 373 students, no one is missing posttest scores. Fifty-two students who have posttest scores do not have a pretest score.

NWEA Reading

Similarly, a high percentage of students did not take the NWEA Reading pretests. Out of a total enrollment of 530 based on fall class rosters, 152 students or (28.6%) do not have pretest scores. Of these remaining 378 students, no one is missing posttest scores. Twenty-eight students who have posttest scores do not have a pretest score.

Implementation Results

In this section we describe more fully the aspects of the implementation that characterize this intervention. We used the following questions to guide our descriptions and analysis: What resources are needed to manifest the *SFScience* condition? Are there differences in the extent, quality, and type of implementation of the materials? We also studied the features of the implementation to identify possible variables related to the outcome measures. Our perspective takes into account three levels of resources needed to implement science instruction: those resources provided by either the district or by Scott Foresman, those provided by the individual schools, and those provided by the teacher.

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

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

Comparison of *SFScience* and Control Groups

Three elementary schools participated in the study; one school had a PreKindergarten, all covered grades Kindergarten through 6th grade.

Classroom Settings for Instruction

The classroom setting was observed during the week of February 21st, 2005. The classroom observations were conducted once during the length of the intervention. Most teachers were not observed in the classroom, but those that were, we observed for approximately 30 to 50 minutes, the length of the science instruction time period. Teachers were not asked to prepare specific lessons for observation, but we made an effort to coordinate the observation with the teacher prior to observation. Teachers reported that they felt pressured by administrators and peers alike to participate in the study. This is a noteworthy point because it raises the issue of how attitudes towards the science curriculum and/or the study may be impacted as a consequence of the “involuntary” volunteer. One teacher specifically indicated that he was peer-pressured into the study and so was apathetic to the implementation requirements. Since this condition of pressure seemed to exist equally in both groups as reflected by the teachers' choice of interview over observation, we feel that the motivational factors are equally distributed among the randomized conditions. Interviews also revealed that the 5th grade teachers were very interested in the curriculum because of the upcoming Washington Assessment of Student Learning (WASL), the state assessments. Only 5th grade students are given the science WASL test.

Although we did not observe many of the teachers teaching a lesson, all of these interviews took place at the schools, in the teachers' classrooms or in the library. 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. At one school, only the teacher had access to a computer in the classroom. Televisions and video playback/recorder systems were in evidence or accessible by both teacher groups. About a third of the control teachers supplemented instruction using videos. Other teachers reported that they rarely used videos but instead used the Internet. Every teacher had an overhead projector that they used periodically.

At each of the participating schools, some classrooms were organized as a Gifted and Talented (GATE) multiage groups. We had a total of five classrooms that were designated multiage GATE classrooms (three *SFScience* and two control).

The control group teachers had fewer packaged materials to teach science. In February, when the observations/interviews were conducted some control teachers had not received any of the district designed kits and so were using a variety of teacher-collected and developed materials.

Opportunities for Learning

Although this site was identified before the beginning of the 2005-2006 academic year in September, certain materials did not arrive until late November. Specifically, the science kits, graphic organizers and content transparencies were delivered in November for all of the grades. Additionally, 5th grade student editions were on backorder until November. In the interim teachers used the Leveled Readers, vocabulary cards, Workbooks, and Activity books for science instruction in the *SFScience* group.

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. An alternating schedule allows the teacher to plan and gather resources to provide instruction for two or three weeks at a time, teaching science everyday for those weeks and then switching to teach another subject for the next two weeks. The alternating pattern fit well with the teachers using the district 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 *SFScience* and 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 19.9 total hours and control teachers reported an average of 13 total hours of instruction for the length of the implementation. As we observe later in Table 35, we have some confidence that the actual difference is different from zero.

The pressure of state assessments was heavy for these teachers. By March, many teachers had stopped teaching science on a regular basis to focus on WASL tests. The *SFScience* group began to use the textbook and Leveled Readers as part of their reading instruction. Some teachers started using the WASL science test to teach science exclusively.

Control 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 18. At least two teachers reported not having any textbooks for their students. Only two teachers practiced whole class science reading for more than half of the time they spent on science. Three teachers reported using whole class reading activities less than half of the time, leading us to conclude that this was not a common activity.

Table 18. 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	77.8%	0%	11.1%	11.1%	33.3%	33.3%

For conducting laboratory activities control teachers indicated that they have no set pattern of usage because of the infrequent availability of district developed science kits. Teachers found the kits difficult to use in part because of the lack of coherence between activities and the organization of the materials. 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 19. 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	33.3%	33.3%	11.1%	11.%	11.1%

Planning time for science instruction is also an important factor for implementing curriculum. Six of the possible ten control teachers responded that they spent approximately 5% (10 minutes per week) of their total available planning time on science instruction. The other four control teachers report spending about 30% (30 to 40 minutes) of their time planning science instruction. Two teachers in the *SFScience* group report spending almost no time planning science instruction, the other nine teachers report percentages of planning time from 20% to 40% (approximately 25 to 40 minutes).

Density of Science Inquiry Reflected in the Classroom

Sections of the surveys were constructed to collect data on 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 SF science program as a continuum:

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

When we asked the teachers on the surveys, we asked about time spent doing these different activities. Both *SFScience* and control group teachers were asked these questions. The variable "science inquiry" is a composite of the time spent in six different aspects of the inquiry process as a percentage. Hence, it is on a scale of 0 to 100 and can be thought of as a measure of "inquiry process density" with 100 being an indication that the teacher and students were practicing the inquiry process every time science was taught. The average percentage density for the *SFScience* group was 18.42 and for the control group it was 27.84. While a greater amount of density is noticed for the control condition, the statistical test (*p* value of .22) gives us no confidence that this difference between the groups is different from zero.

Implementation of *SFScience*

Training and Support

The one-half day training took place on September 15, 2005 at the district offices. During the training, the Scott Foresman representative gave a demonstration of the science kits and the pedagogical method of hands-on inquiry. 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. Teachers also received an online log-in so that they could reference additional materials. Teachers also indicated that there was a lot of material to cover and it was difficult to digest all of the ideas in such a short period.

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

Availability and Use of Materials

Every teacher assigned to the *SFScience* group received sufficient materials to use with the number of students that they taught. At all grades the science kits were backordered until late November. Several teachers reported missing other materials, such as the graphics organizers and vocabulary cards. Student textbooks were backordered for all of the 5th grade.

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

All 11 of the *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 20. Percent of Teachers Covering Each Chapter in Unit A-Life Science

		Chapter					
		1	2	3	4	5	6
Number of Respondents	11	81.8%	63.6%	36.4%	81.8%	72.7%	45.5%

Table 21. Chapters in Unit B-Earth Science Covered

		Chapter			
		7	8	9	10
Number of Respondents	11	36.4%	36.4%	18.2%	27.3%

Table 22. Chapters in Unit C-Physical Science Covered

		Chapter				
		11	12	13	14	15
Number of Respondents	11	18.2%	18.2%	0%	9.1%	0%

Note: Most teachers did not teach any chapters in this unit.

Table 23. Chapters in Unit D-Space & Technology Covered

		Chapter		
		16	17	18
Number of Respondents	11	18.2%	18.2%	0%

Note: Most teachers did not teach any chapters in this unit.

As can be seen, teachers mostly chose to teach from the Life Science and Earth Science units.

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". The four 3rd grade teachers thought that the textbook was too difficult for their students. In general all teachers thought that the textbook was too difficult. It's too vocabulary-rich and requires background information and experiences that their students don't have. 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 24. For Unit A, How Well Was the Content Aligned to State Standards?

		How Well Aligned				
		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	11	0%	27.3%	27.3%	36.4%	9.1%

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

		How Well Aligned				
		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	11	36.4%	27.3%	27.3%	9.1%	0%

Table 26. For Unit C, How Well Was the Content Aligned to State Standards?

		How Well Aligned				
		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	11	63.6%	18.2%	18.2%	9.1%	0%

Table 27. For Unit D, How Well Was the Content Aligned to State Standards?

		How Well Aligned				
		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	11	72.7%	0%	27.3%	0%	50%

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: 1) the packaging — there weren't enough copies of readers at the lower end, and 2) the vocabulary was still too difficult for their English Language Learners.

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. A specific case noted is the onion skin experiment with the microscopes. The teacher said, "The microscopes were worthless, not kid friendly at all".

Few *SFScience* teachers used the assessments because they determined that they were too difficult for their students. Two teachers used an open-book method to help the students with the assessments and 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 Federal Way, implementation was much weaker than the desired ideal model.

Summary of Implementation

Certain factors emerged as barriers to a smooth implementation. Perhaps first among those is the actual time of science instruction. Because of backordered materials actual implementation did not begin until December for the inquiry portion of the curriculum. Science reading was implemented earlier but the core design element, inquiry, was missing. Lack of alignment to the standards also contributed to the overall implementation. And lastly, the focus on state assessments inhibited teachers from continuing science instruction past March. The length of implementation at best was four months.

Quantitative Impact Results

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

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

Science Outcomes

Analysis Including Pretest

Our first analysis addressed Science outcomes using the NWEA Science Concepts and Processes scale.   Table 28 provides a summary of the sample we used in the analysis and the results for the comparison of *SFScience* and control. This shows the means and standard deviations as well as a count of the number of students, classes and teachers in each group. The last two columns provide the effect size, that is, the size of the difference between the means for *SFScience* and control in standard deviation units, and the *p* value, indicating the probability of arriving at a difference as large or larger than the absolute value of the one observed when truly there is no difference. The “Unadjusted” row is based on all students with a posttest with or without a pretest and the estimated effect size takes into consideration the clustering of students in upper-level units (i.e., that students are grouped within teachers.) The “Adjusted” row is based on the students who have both pretests and posttests. This is the sample that we use in the analyses on which we base our results reported in Table 29 through Table 31. 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 28. Overview of Sample and Impact of *SFScience* on Science Achievement

	Condition	Means	Standard[a] deviations	No. of students	No. of classes	No. of teachers	Effect size	*p* value[b]
Un-adjusted	***SFScience***	201.35	10.25	224	11	11	0.17	.09
	Control	199.63	9.38	201	10	10		
Adjusted	***SFScience***	199.82	10.35	194	10	10	-0.01	.95
	Control	199.92[c]	9.22	179	10	10		

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

[b] The *p* value for the unadjusted effect size is computed using a model that 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.

[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 28. The bar graphs represent average performance in science achievement using the NWEA Science Concepts and Processes as the metric.

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

The panel on the right shows estimated performance on the posttest for the two groups based on a model that adjusts for students' pretest scores and other fixed effects (i.e., this is a visual display of results from the row labeled 'adjusted' in Table 28.) We can see that the two groups were essentially indistinguishable. The high *p* value for the treatment effect (.95) 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. The overlap in these confidence intervals further indicates that any difference we see is easily due to chance.

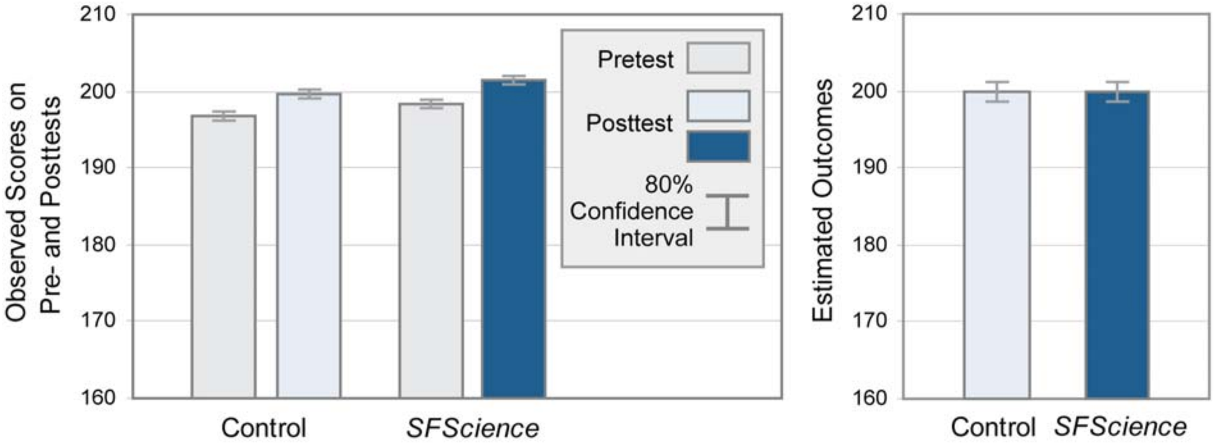

Figure 1. Impact on Science Achievement: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)

Analysis Including Pretest as a Moderator

We now report on the analyses that examine not just the overall impact of *SFScience* but also the moderating effects of other variables. 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 29 shows the estimated impact of *SFScience* as measured by NWEA on the performance of students with an average score on the pretest as well as the estimated moderating effect of the prior score.

Table 29. Mixed Model Estimating the Impact of *SFScience* on Science Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Estimated value for a control student with an average pretest	206.39	6.77	6	30.5	<.01
Impact of *SFS*cience for a student with an average pretest	0.04	1.8	6	0.02	.98
Estimated change in control outcome for each unit increase on the pretest	0.71	0.06	351	11.45	<.01
Interaction of pretest and *SFScience*	0.16	0.08	351	1.84	.07
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	11.01	7.54		1.46	.07
Within-teacher variation	34.61	2.61		13.25	<.01

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

[b] Teachers were modeled as a random factor.

Note. We took out 1 student in the impact analysis since he/she was considered an outlier.

The row in the table labeled “Impact of *SFScience* for a student with an average pretest” tells us whether *SFScience* made a difference in terms of student performance on NWEA Science for a student who has an average score on the pretest. The estimate associated with *SFScience* is 0.04. This shows a small difference associated with *SFScience*. However, the *p* value of .98 gives us no confidence that the underlying effect is different from zero.

We also estimated the moderating effect of the pretest score on the impact of *SFScience* to see whether it was differentially effective for students at different points along the pretest scale. The *p* value for this effect is .07. 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 positive estimate of .16 indicates that treatment is more beneficial for students at the upper-end of the pretest scale. We explore the nature of this interaction in detail below.

As a visual representation of the results described in Table 29, 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 Science 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 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.[2] We see that the slopes of the two lines are different, an indication of the interaction effect.[3]

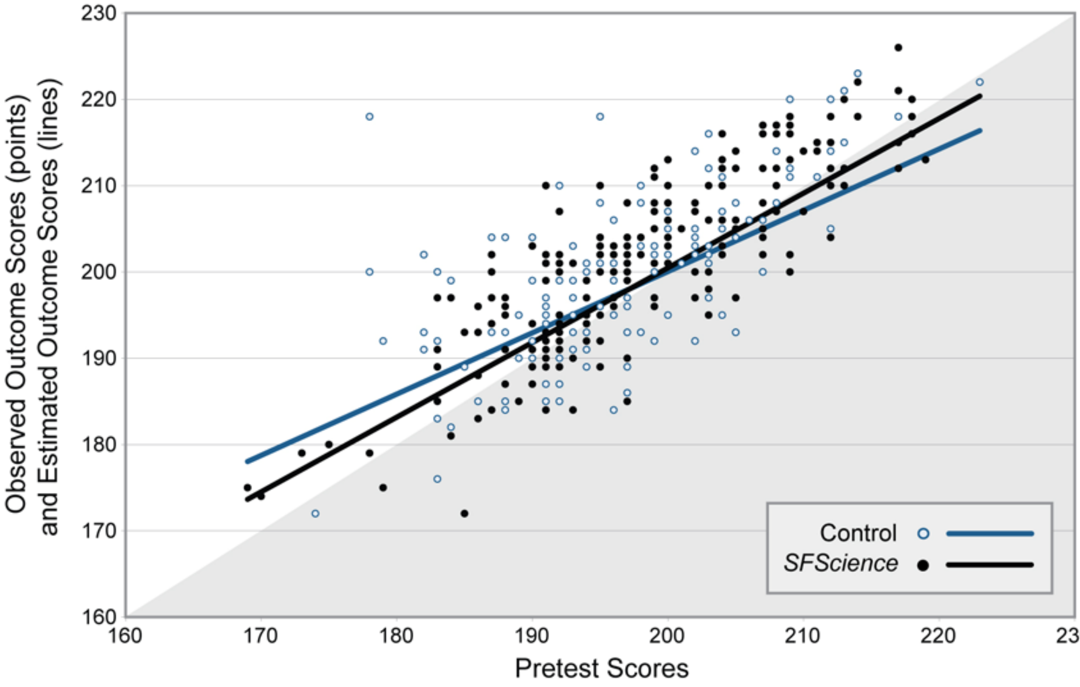

Figure 2. Comparison of Estimated and Actual Outcomes for *SFScience* and Control Students

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

[2] 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).

[3] 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 observe where *SFScience* had its greatest impact.[4] 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 29, there is evidence of a differential impact of the intervention across the prior score scale as measured by NWEA Science. In spite of the positive interaction effect, the impact for students at the medians of the top and bottom quartiles is not large enough to warrant concluding that the effect for these students is different from zero.

[4] 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).

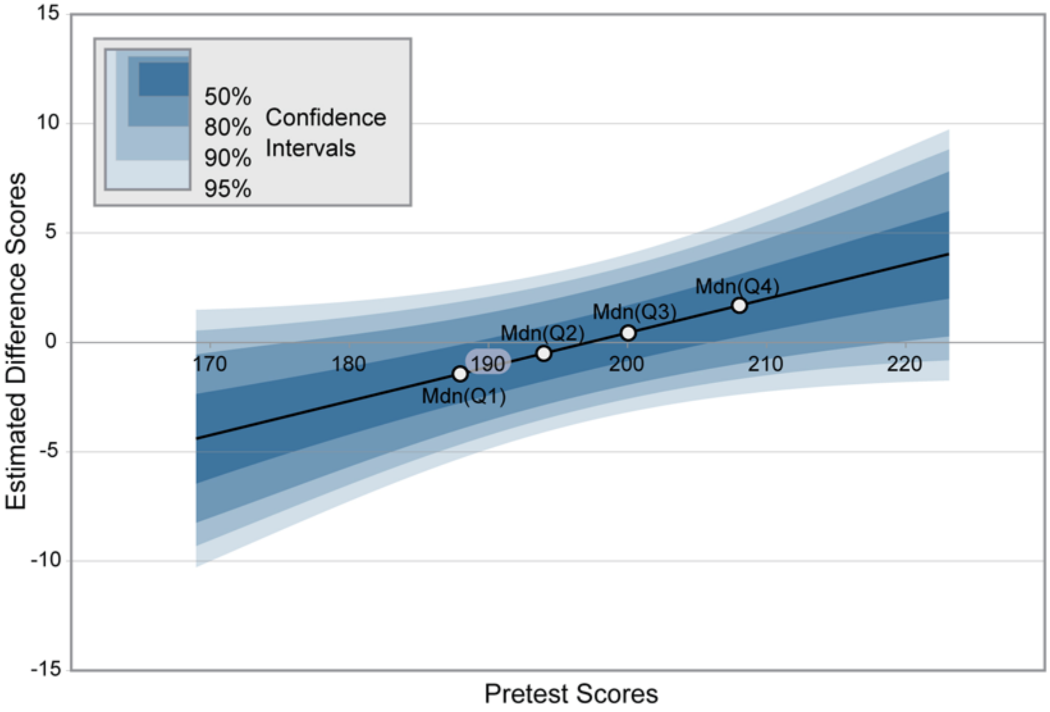

Figure 3. Differences between *SFScience* and Control Group Science Achievement: Median Pretest Scores for Four Quartiles Indicated

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 pre-test 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 there is little difference in the estimated outcomes in the two conditions and there is a substantial amount of overlap in the confidence intervals. The same applies to a student at the median of the fourth quartile.

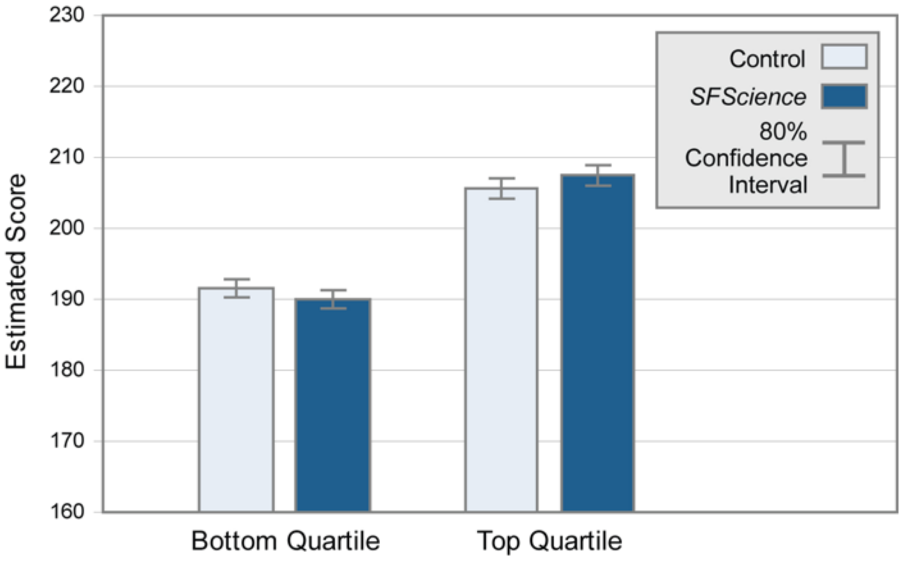

Figure 4. Difference Between *SFScience* and Control Group Science Achievement: Median Students in Top and Bottom Quartiles

The overlap of the confidence intervals shows that the contrast between *SFScience* and control for the lower and high scoring students can easily be a matter of chance. Even though the differences between *SFScience* and control for these students at the median of the bottom and top quartile are small (and the confidence markers overlap), we can see that the direction of the difference changes for the two pairs of bars. The first quartile predictions show that the students in the control group benefit, whereas the fourth quartile predictions indicate that the *SFScience* students benefit. The information in Table 29 gives us some confidence that this reversal is not due to chance.

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. Table 30 shows the moderating effect of gender on students' performance on NWEA Science. The advantage of being in the *SFScience* condition is greater for girls than it is for boys. The *p* value of .05 means we have high confidence that the actual differential impact is different from zero. Figure 5 illustrates this graphically.

Table 30. Moderating Effect of Gender on Science Achievement

Fixed effects	Estimate	Standard error	DF	*t* value	*p* value
Outcome for girl in control group with an average pretest	205.64	6.58	6	31.25	<.01
Estimated change in outcome for each unit increase on the pretest	1.37	2.02	6	0.68	.52
Average *SFScience* effect for girls	0.83	0.04	349	20.65	<.01
Difference (boys minus girls) in average performance in control condition	1.76	0.84	349	2.11	.04
Difference (boys minus girls) in the average *SFScience* effect	-2.22	1.15	349	-1.93	.05
Random effects[a]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	13.29	8.68		1.53	.06
Within-teacher variation	29.95	2.27		13.21	<.01

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

[b] Teachers were modeled as a random factor.

[C] The prior score was centered at the mean, therefore, the effect estimates apply to a girl or boy who had an average score on the pretest.

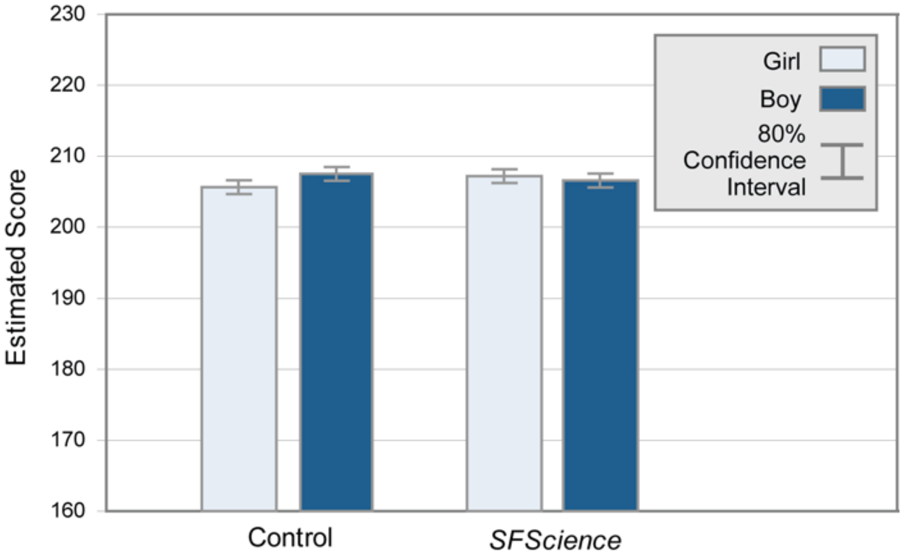

Figure 5. Moderating Effect of Gender on Science Achievement

Analysis Including Ethnicity as a Moderator

We were also interested in whether *SFScience* was differentially effective for students of different ethnicities because of the demographic composition of the district. Table 31 shows estimates from the model that tests the moderating effect of ethnicity on students' performance on NWEA Science. In the absence of treatment, a White student who came into the experiment with an average pretest score performs better than a Black student with the same pretest score. However, *SFScience* is not differentially effective for Black and White students.

Table 31. Science Achievement Moderated by Ethnicity

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Average outcome for Whites in the control group	206.97	6.61	6	31.33	<.01
Estimated change in outcome for each unit increase on the pretest	0.81	0.04	340	19.89	<.01
Average *SFScience* effect for Whites	-0.04	2.02	6	-0.02	.99
Difference (Asians minus Whites) in average performance in the control condition	-1.20	1.03	50	-1.16	.25
Difference (Hispanics minus Whites) in average performance in the control condition	-1.44	1.44	50	-1.00	.32
Difference (Blacks minus Whites) in average performance in the control condition	-2.44	1.21	50	-2.01	.05
Difference (Asians minus Whites) in the average *SFScience* effect	0.70	1.41	340	0.50	.62
Difference (Hispanics minus Whites) in the average *SFScience* effect	-1.72	2.22	340	-0.77	.44
Difference (Blacks minus Whites) in the average *SFScience* effect	0.63	1.67	340	0.37	.71
Mixed model: Technical details for random components	**Estimate of Variance Component**	**Standard Error**		**z value**	***p* value**
Teacher mean achievement	13.15	8.60		1.53	.06
Within teacher mean variation	29.35	2.25		13.04	<.01

[a] Schools and pairs of teachers used for random assignment were modeled as a fixed factor but the estimated effects are not included in this table.

[b] Teachers were modeled as a random factor.

[c] The prior score was centered at the mean, therefore, the effect estimates apply to members of each subgroup who have an average score on the pretest.

Note. We excluded results for the category of Native American since there were only 9 cases that belonged to this group.

Reading Outcomes

Analysis Including Pretest

Our next set of analyses addresses Reading achievement as measured by NWEA Reading. Table 32 provides a summary of the sample we used in the analyses and the results for the comparison of *SFScience* and control. This shows the means and standard deviations as well as a count of the number of students, classes and teachers in each group. The last two columns provide the effect size, that is, the size of the difference between the means for *SFScience* and control in standard deviation units, and the *p* value, indicating the probability of arriving at a difference as large or larger than the absolute value of the one observed when there truly is no difference. The “Unadjusted” row is based on all students with a posttest and the estimated effect size takes into consideration the clustering of students in upper-level units (i.e., that students are grouped within teachers.) The “Adjusted” row is based on the students who have both pretests and posttests. This is the sample that we use in the analyses on which we base our results reported in Table 33 and Table 34. 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 32. 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***	208.01	14.20	216	11	11	0.17	.41
	Control	204.82	12.10	190	10	10		
Adjusted	***SFScience***	205.83	13.16	197	11	11	0.08	.27
	Control	204.79[c]	12.22	181	10	10		

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

[b] The *p* value for the unadjusted effect size is computed using a model that 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.

[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 32. The bar graphs represent average student performance on NWEA Reading. The panel on the left shows average pre- and post-test 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 32.) We can see that the two groups were essentially indistinguishable. The *p* value for the treatment effect (.27) indicates we have no confidence that the actual difference is different from zero. We added 80% confidence intervals to the tops of the bars.

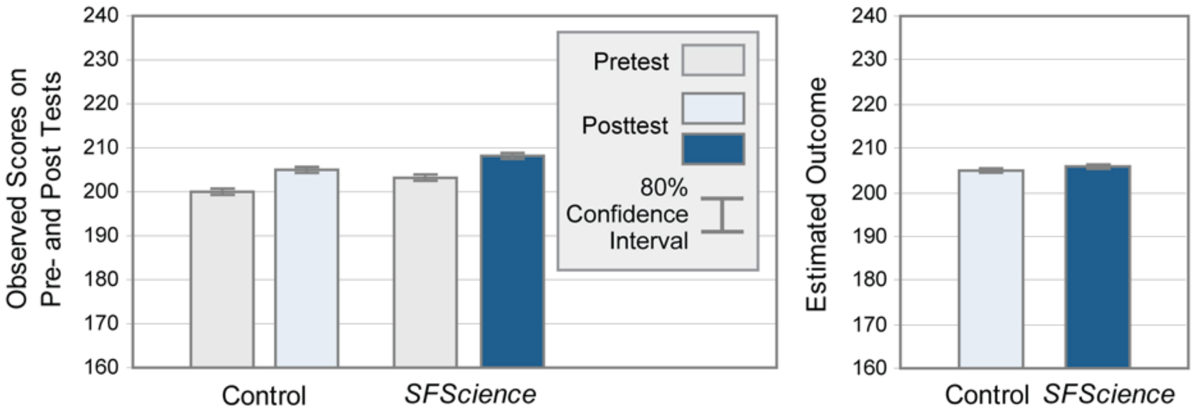

Figure 6. Impact on Reading Achievement: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)

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 33 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 33. Mixed Model Estimating the Impact of *SFScience* on Reading Achievement

Fixed effects	Estimate	Standard error	DF	*t* value	*p* value
Estimated value for a control student with an average pretest	206.17	0.65	19	315.82	<.01
Impact of *SF*Science for a student with an average pretest	1.18	0.9	19	1.31	.21
Estimated change in control outcome for each unit increase on the pretest	0.86	0.03	354	27.59	<.01
Interaction of pretest and *SFScience*	-0.02	0.04	354	-0.56	.58
Random effects[a]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	2.75	1.42		1.93	.03
Within-teacher variation	25.72	1.93		13.29	<.01

[a] 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 1.18. This shows a positive effect of *SFScience*. However, the *p* value of .21 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 .58. We have no confidence that the true effect is different from zero.

As a visual representation of the results described in Table 33, 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.[5] The graph confirms the findings described above: there is no average effect and no interaction effect.

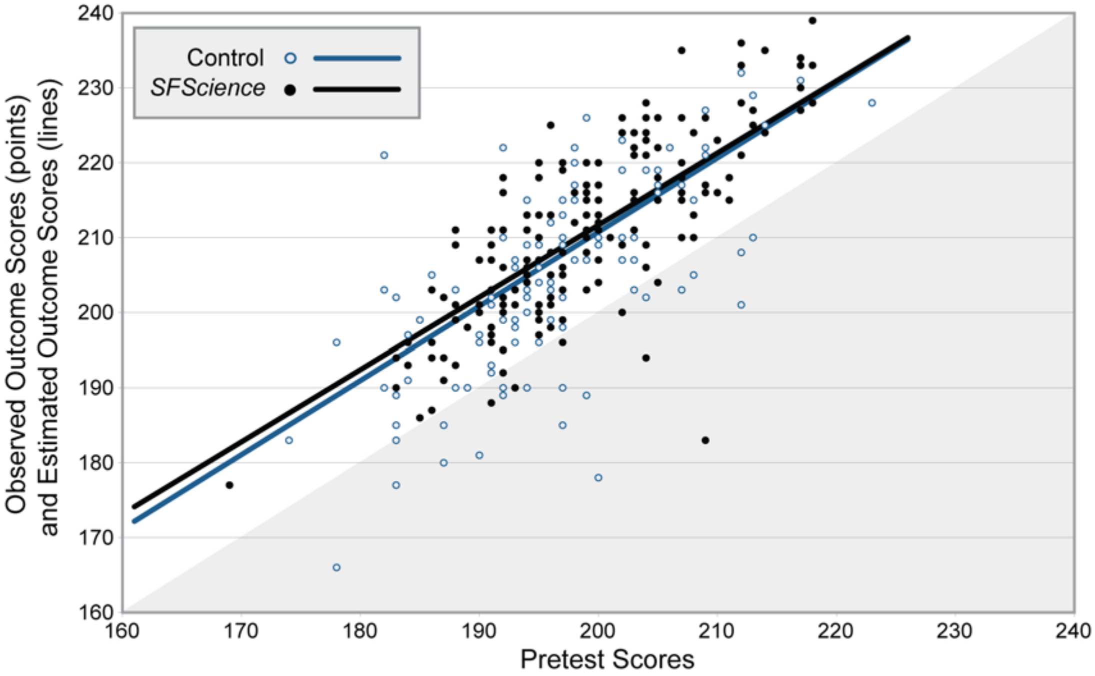

Figure 7. Comparison of Estimated and Actual Outcomes for *SFScience* and Control Students

Analysis Including Ethnicity as a Moderator

As with the results for science, we estimated the interactions of condition (*SFScience* versus control) with student ethnicity. We were interested in whether the condition's effect was differentially effective for students of different ethnic backgrounds. Table 34 shows estimates from the model that tests the moderating effect of ethnicity on students' performance on NWEA Reading. We don't see differences between White students and students of other ethnic categories in control group performance. Nor do we see differences between White students and students of other ethnic categories in the effect of *SFScience*. That is, we have no confidence that the true differences between the comparison groups are different from zero.

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

Table 34. Reading Achievement Moderated by Ethnicity

Fixed effects [a]	Estimate	Standard error	DF	*t* value	*p* value
Average outcome for Whites in the control group	37.04	4.82	19	7.69	<.01
Estimated change in outcome for each unit increase on the pretest	0.84	0.02	342	35.84	<.01
Average *SFScience* effect for Whites	1.00	1.15	19	0.87	.39
Difference (Asians minus Whites) in average performance in the control condition	-0.50	0.97	49	-0.51	.61
Difference (Hispanics minus Whites) in average performance in the control condition	-0.22	1.62	49	-0.13	.89
Difference (Blacks minus Whites) in average performance in the control condition	-0.70	1.08	49	-0.65	.52
Difference (Asians minus Whites) in the average *SFScience* effect	0.36	1.31	342	0.27	.79
Difference (Hispanics minus Whites) in the average *SFScience* effect	-2.09	2.19	342	-0.96	.34
Difference (Blacks minus Whites) in the average *SFScience* effect	-0.70	1.57	342	-0.45	.66
Mixed model: Technical details for random components	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	3.39	1.70		2.00	.02
Within teacher mean variation	27.03	2.07		13.06	<.01

[a] Schools and pairs of teachers used for random assignment were modeled as a fixed factor but the estimated effects are not included in this table

[b] Teachers were modeled as a random factor.

[c] The prior score was centered at the mean, therefore, the effect estimates apply to members of each subgroup 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.

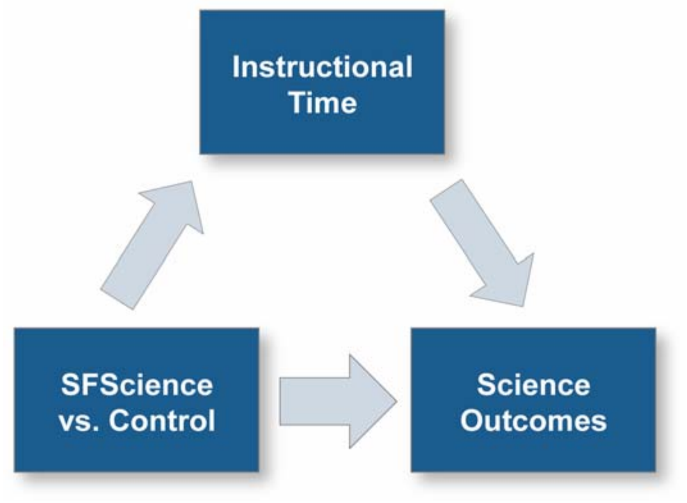

Figure 8. Relationships for Exploratory Analysis of Implementation Variables

Instructional Time

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

We look first at the impact on instructional time. Table 35 shows *SFScience* teachers taught approximately 10 more hours of science during the year. The *p* value of .08 gives us some confidence that the actual difference is different from zero. We have some confidence that *SFScience* causes an increase in the number of minutes used on science instruction.

Table 35. The Impact of *SFScience* on Weekly Minutes of Science Instruction Time

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Hours of science for a control teacher	13.04	4.69	1	2.78	<.01
Impact of *SFScience* on hours of science instruction	9.94	4.17	4	2.38	.08
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Residual teacher variance	43.45	30.73		1.41	.08

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

Additionally, 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 small negative relationship between *SFScience* usage and the student outcome. The *p* value for this effect is .23, which gives us no confidence that the true relationship is in fact different from zero.

Table 36. Relationship of Instruction Time to Student Outcome

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Estimated value for a student with an average pretest	208.01	6.22	3	33.44	<.01
Estimated change in outcome for each unit increase on the pretest	0.81	0.04	298	18.57	<.01
Estimated change in outcome for hour of science time	-0.11	0.07	3	-1.49	.23
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	0.35	1.85		0.19	.42
Within-teacher variation	35.98	2.95		12.21	<.01

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

[b] Teachers were modeled as a random factor.

[C] The prior score was centered at the mean, therefore, the effect estimates apply to a male or female who had an average score on the pretest.

Discussion

We began this research in Federal Way Public Schools with the question of whether *Scott Foresman Science* was as effective as or more effective than their existing programs we were comparing it to. Our question applied both to science achievement as well as to whether the science program made a measurable difference in reading achievement beyond the growth resulting from the core reading program.

We found no overall difference between the science or reading scores of students taught using *SFScience* as compared to the established program. However, in science, we found that *SFScience* tended to be more effective than the existing program for students initially scoring at the higher end of the pretest scale. Since the pretest we used is scored along a continuous growth scale, we might translate this finding into an expectation that the program may be more effective for students in the later 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 ethnicity. For gender, we found an effect, for which we have strong confidence, that *SFScience* improved the standing of girls and closed the initial gap between boys and girls in science achievement. We found no differential benefit for other ethnicities compared to White students.

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