RESEARCH REPORT

Comparative Effectiveness of *Scott Foresman Science*:

A Report of a Randomized Experiment in Ogden City School District

Gloria I. Miller
Andrew Jaciw
Boya Ma
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 Ogden City School District 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 community assess new or proposed instructional and professional development programs through scientifically based pilot implementations. The company draws on the expertise of world-class researchers and methodologists assuring that the research is objective and takes advantage of current best practice in rigorous experimental design and statistical analysis. The company's findings let educators quantify the value of programs and help them partner with providers to implement those most effective for their students.

Comparative Effectiveness of *Scott Foresman Science:*

A Report of a Randomized Experiment in Ogden City School District

Table of Contents

Introduction

Pearson Education contracted with Empirical Education Inc. to conduct five randomized experiments to determine the effectiveness of its *Scott Foresman Science* (*SFScience*) curriculum and associated materials.

This research project consists of a randomized experiment in Ogden City School District. 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 *Scott Foresman Science* program is more effective than the current curriculum being used in the participating campuses in the Ogden City School 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: Science Concepts and Processes, and Reading Achievement.

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

Not all teachers had textbooks for the students; when they did, there were two textbooks in use, Harcourt Brace and older versions of Scott Foresman. Some teachers used materials that they had developed together with Utah State Core Curriculum.

Site Descriptions

Ogden, UT

The city of Ogden is located approximately 35 miles north of Salt Lake City and is Utah's sixth largest city, encompassing 27 square miles. It has an estimated population of 77,000 according to the 2000 census.

Table 2. Ogden Racial Makeup

Race/Ethnicity	% of Population
White	55.0
African American	2.3
American Indian/Native Alaskan	1.2
Asian and Native Hawaiian or Pacific Islander	1.6
Other race	12.9
Two or more races	2.9
Hispanic origin (of any race)	23.6

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

Ogden City School District, UT

Ogden City School District considers it self an inner-city district enriched by multi-cultural diversity. They operate a total of 16 elementary schools, four middle schools, and four high schools; four Title I schools (K-5) participated in this study. The following tables summarize the demographic makeup of the school district.

Table 3. Background of Ogden City School District

Ogden City Schools	
Total schools	24
Total teachers	595
Student to teacher ratio	21.8
Grades	PK -12
Student population	12,963
Economically disadvantaged	64.9%
ELL students	24.4%

Source: Utah Public Schools, 2005 and CCD Public School District Data for 2004-2005

Table 4. Ethnic Makeup of the Ogden City Schools

Race/Ethnicity	% of population
White, non-Hispanic (%)	53.9%
Black, non Hispanic (%)	3%
Hispanic (%)	39.7%
Asian/Pacific Islander (%)	1.9%
American Indian/Alaskan Native (%)	1.5%

Source: Utah Public Schools, 2005 and CCD Public School District Data for 2004-2005

Sample and Randomization

Recruiting

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 Ogden City Schools occurred on May 25, 2005 with 21 teachers from four different schools. 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. Three teachers decided not to participate and excused themselves. Of the remaining 18 teachers, all filled out consent forms, and two additional teachers that could not be present for the meeting were represented by their Principal. We contacted these two teachers via email, and they filled out consent forms. Twenty teachers were formed into pairs.

Randomization

The unit of randomization at this site is the teacher. Twenty 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). There are various ways to randomize teachers to conditions. We used a matched-pairs design whereby we first identified pairs of teachers at similar grade levels, we randomized one teacher to treatment and the other to control. Matched pairs were based on schools and grade level taught and on years of teaching experience, resulting in a within schools and grade-level randomization paired on teaching experience. Only two pairs broke these rules. One pair was formed within schools with one teacher at 4th grade and one at 5th. A second pair was at the same grade level, but at different schools. A pairing strategy 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 numbers of teachers are displayed in the table below. In some of the schools, science is considered a "specialty" subject. Teachers can specialize in science instruction and teach other students not assigned to their self-contained classroom. In these cases of "departmentalized" instruction, all students under the teacher's science instruction are considered part of the study.

Table 5. Participating teachers at Ogden site

Teacher Assignment Status	Number Participating
SFScience	10
Control	10
Total	**20**
Note: Later in the study, one control teacher dropped from the study.	

Because specialization causes some teachers to have more than one group of students for instruction, the number of classes involved in the study exceeds the total number of teachers participating. There were a total of 11 classrooms assigned to the control condition and 13 classrooms assigned to the treatment condition.

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 affect the size of the effect that we can detect, we calculated that we could detect an effect size as small as .45. This is calculated 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 .33. 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 that may be real but not discernable given the number of teachers in the study.

Data Sources and Collection

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

Observational and Interview Data

In general, observational data are used to inform the description of the learning environment, instructional strategies employed by the teachers, and student engagement. These data are minimally coded. Our observation of the initial training in the use of *Scott Foresman Science* materials was conducted on September 15th, 2005. Classroom observations were conducted during the week of April 19th. Eight of the 10 teachers in the *SFScience* group were observed. Six of the nine control group teachers were observed.

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 interviews of both groups were conducted throughout the timeframe of the study.

Survey Data

Surveys were deployed to both *SFScience* and control group teachers beginning on December 5, 2005 and continuing on a bi-weekly basis until late May of 2006. Response rates were calculated using a simple percentage calculation based on the ratio of actual received responses to the number of expected responses. There were five teachers in the *SFScience* group and six teachers in the control group. 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 89.47% for both groups, an 87.78% response rate for the *SFScience* teachers, and a 91.36% response rate for the control teachers.

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

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

Table 6. Survey Response Rates

Survey number	Date	Topic	Treatment response rate	Control response rate	Overall response rate
Survey 1	Dec. 5 - 9	Science Schedule & Instructional Time	80.00%	55.56%	68.42%
Survey 2	Jan. 16 - 20	Resources	80.00%	100%	89.47%
Survey 3	Jan. 23 - 27	Interactions with materials/Students	100%	88.89%	94.74%
Survey 4	Feb. 6 - 10	More Interactions	100%	100%	100%
Survey 5	Feb. 20 - 24	Time & Preparation	100%	88.89%	94.74%
Survey 6	Mar. 6 - 10	Materials & Resources	90.00%	100%	94.74%
Survey 7	Mar. 20 - 24	Assessments	90.00%	100%	94.74%
Survey 8	May 1 - 5	More Interactions	80.00%	88.89%	84.21%
Survey 9T*	May 26	Final Survey	70.00%	N/A	70.00%
Survey 9C**	May 26	Final Survey	N/A	100%	100%

*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. We refer to these tests when reporting science achievement and reading achievement throughout the report. In the fall of 2005, the NWEA Science and Reading tests were administered to the students at the various schools as a pretest measure. As a posttest measure, the Science and Reading tests were administered in the spring of 2006. The paper-and-pencil versions of these tests are referred to as ALT tests and all sites were provided these materials. Both of these tests are adaptive and comprehensive, and are designed to measure growth over time. The sets of tests consist of multiple levels, with overlapping degrees of difficulty. Several different levels are given within the same classroom. To ensure a good match of student to test, there are five test levels for Science and eight test levels for Reading. The first time a student is tested, the appropriate test level is determined by use of placement tests, referred to as locator tests. During the second and subsequent administrations, the student is automatically assigned to a level based on previous results. 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.

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

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

In addition to examining impacts and interactions where we anticipate effects, to better understand unexpected results, we use other demographics, teacher characteristics, and supplementary observational data in exploratory analyses to generate additional hypotheses about which factors 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 hasn't. Thus a *p* value of .1 gives us a 10% probability of that happening. We can also think of it as the level of confidence, or the level of belief we have that the outcome we observe is not simply due to chance. While ultimately depending on the risk tolerance of the user of the research, we suggest the following guidelines for interpreting *p* values:

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

Results

Formation of the Experimental Groups

Groups as Initially Randomized

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

[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

control conditions. This is the complete number of students in the experiment at the time that the experiment began in September 2005.

Table 7. Distribution of the Experimental 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	4	10	13	62	146	90	298
Control	4	9	11	60	57	132	249
Totals	**4**[a]	**19**	**24**	**122**	**203**	**222**	**547**

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

Mid-way through the study in the February timeframe, one control teacher reported that she had not had the opportunity to teach any science and was unlikely to find time in her schedule to teach science in the future. She requested to have her name removed from the roster of participating teachers. This teacher was marked as inactive by request, but the attrition is noted as unrelated to assignment since in either condition, science instruction is required. This reduced the total teacher count to 10 *SFScience* and 9 control teachers.

Teacher Variables

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 stratified according to these variables, which we believed affected student scores, to avoid a potential imbalance in outcomes due to chance discrepancies between conditions in years of teaching experience.

Table 8. Distribution of Years Teaching Experience

	Number of Teachers		
Condition	0 to 3 years	4 or more years	Totals
SFScience	1	8	**9**
Control	1	7	**8**
Totals	**2**	**15**	**17**

Note: One *SFScience* and one control teacher did not provide this information.

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

	Number of teachers	Early career (0-3 years)	Emerging professional (4-6 years)	Mid-career professional (7-15 years)	Highly experienced professional (15+ years)
SFScience	9	10%	40%	30%	10 %
Control	8	10%	30%	20%	20%

Note: One *SFScience* and one control teacher did not provide this information.

Table 10. Years Teaching in Grade Level

	Number of teachers	0-3 years	4-6 years	7-15 years	15+ years
SFScience	10	40%	40%	10%	0%
Control	9	67%	11%	0%	11%

Table 11. Years Teaching Science

	Number of teachers	0-3 years	4-6 years	7-15 years	15+ years
SFScience	10	30%	40%	20%	0%
Control	9	30%	20%	20%	10%

Table 12. Science Coursework in College

	Number of teachers	None	Some	Minor	Major
SFScience	9	10%	80%	0%	0%
Control	5	10%	40%	0%	0%

Note: One *SFScience* and four control teachers did not provide this information.

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

	Number of teachers	Attended PD in last two years	No PD in the last two years
SFScience	9	20%	70%
Control	8	40%	40%

Note: One *SFScience* and one control teacher did not provide this information.

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.

From the previous tables, we see that 547 students enrolled in the fall. Of these, 62 students have been designated as requiring special education support; we will not include those students in the analysis. Hence, the following analyses are based on a sample size of 485 students.

Student Variables

English Proficiency

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

Table 14. English Learner Status for *SFScience* and Control Groups

Condition	English Proficiency		
	Not proficient	Proficient	Totals
SFScience	64	201	**265**
Control	69	151	**220**
Totals	**133**	**352**	**485**
Statistics	**DF**	**Value**	***p* value**
Chi-square test	1	3.14	.08

Ethnicity

Table 15 summarizes the distribution of student ethnicity. Results of Fisher's Exact Test demonstrate that ethnicity was not distributed evenly between the conditions in spite of randomization. There are proportionally less Caucasian students in the control group than in the *SFScience* group. The imbalance may lead the estimate of the impact to depart from its true value.

Table 15. Ethnicity for *SFScience* and Control Groups

	Ethnicity				
Condition	Multi-racial	Hispanic	Black	White	Totals
SFScience	8	155	12	90	**265**
Control	6	158	4	52	**220**
Totals	**14**	**313**	**16**	**142**	**485**

Statistics	Value	*p* value
Fisher's Exact test	<.01	.01

Note: Due to the small number of cases, we combined students who are Asian and Native American as Multi-racial. Since some of the cells have an expected number of cases less than 10, Fisher's exact test is reported.

Socio-Economic Status

Table 16 shows the distribution of the socio-economic status (SES) of the students in each group, as determined by participation in the National School Lunch program.

Randomization resulted in SES being evenly balanced between *SFScience* and control. We tested this formally, and the *p* value of .25 indicates that the small imbalance that we see is easily due to chance.

Table 16. SES for *SFScience* and Control Groups

Condition	In the Free Lunch program		
	No	Yes	Totals
SFScience	30	235	**265**
Control	18	202	**220**
Totals	**48**	**437**	**485**

Statistics	DF	Value	*p* value
Chi-square test	1	1.33	.25

Gender

Table 17 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 17. Gender for *SFScience* and Control Groups

Condition	Gender		
	Male	Female	Totals
SFScience	130	135	**265**
Control	105	115	**220**
Totals	**235**	**250**	**485**
Statistics	**DF**	**Value**	***p* value**
Chi-square test	1	0.08	.77

Characteristics of the Experimental Groups Defined by Pretest

We also checked whether randomization resulted in balance on pretest scores, a variable that we include in most of our analyses to increase the precision of our estimates. Table 18 shows the results of students without disabilities in grades 3 to 5 for whom pretests were available.

NWEA Science Test

Table 18. Difference in Pretest Scores between Students in the *SFScience* and Control Groups

Descriptive statistics: Pretest outcomes	Raw group means	Standard deviation	Number of teachers	Standard error	Effect size[a]
SFScience	191.99	9.53	179	0.71	0.05
Control	191.56	8.01	153	0.65	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
Condition (*SFScience* – control)	0.43		330	-0.44	.66

[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.

Randomization resulted in science pretest scores being evenly balanced between *SFScience* and control. We tested this formally, and the high *p* value of .66 indicates that the small imbalance that we observe is easily due to chance.

NWEA Reading Test

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

Descriptive statistics: Pretest outcomes	Raw group means	Standard deviation	Number of teachers	Standard error	Effect size[a]
SFScience	193.19	13.48	181	1.00	-0.06
Control	194.02	13.67	157	1.09	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
Condition (*SFScience* – control)	-0.83		336	0.56	.57

[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.

From Table 20, randomization also resulted in reading pretest scores being evenly balanced between *SFScience* and control. The high *p* value of .57 indicates that the small imbalance that we observe is easily due to chance.

Attrition After the Pretest

NWEA Science Test

Out of a total enrollment of 485 based on the cases of students without disabilities in grades 3, 4, and 5 on fall class rosters, 127 students (26%) did not take the posttest and 153 students (32%) did not have pretest scores. Of these 332 students who have pretest scores, no one is missing posttest scores. Twenty-six students who have posttest scores did not have a pretest score.

NWEA Reading Test

Similarly, 122 students (25%) did not take the NWEA Reading posttests, and 147 students (30%) did not have pretest scores. Of these 338 students who have pretest scores, no students are missing posttest scores. Twenty-five students who have posttest scores did 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

Four elementary schools participated in the study; all covered grades Kindergarten through 5th grade.

Classroom Settings for Instruction

The classroom setting was observed during the week of April 19th, 2005. The classroom observations were conducted once during the length of the intervention. Most teachers were 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 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. Additionally, teachers reported that room to work with hands-on activities was also a problem, extra planning had to done in order to accommodate the activities.

Some teachers had a few computer stations in the classroom, but not enough for every student. Televisions and video playback/recorder systems were in evidence or accessible by both teacher groups. Almost half 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.

The control group teachers had fewer packaged materials to teach science and consequently had to buy or bring materials from home. Overall, most teachers had the materials they needed to teach science, but storage and working space were at a premium for the hands-on activities. Almost all teachers supplemented instruction with some sort of Internet activities.

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 5th science kits were delivered in November. All grades were missing the audio versions of the textbooks and the activity flip charts. Two teachers reported having insufficient materials for all the students.

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

For the self-contained classroom teacher, science is taught as part of all subjects taught to the students. Teachers typically alternate science instruction with social studies. An alternating schedule allows the teacher to plan and gather resources to provide instruction for three weeks at a time. Not all teachers followed this scheduling pattern. Some teachers scheduled science instruction for approximately 30 minutes a day, but they noted that it was difficult to incorporate labs into the existing schedule and so *SFScience* teachers shifted to 3 days per week for 50 minutes to an hour.

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. One control and one *SFScience* teacher did not report instructional times on a consistent basis, and so we averaged times for all other teachers and did not include data from teachers missing more than 2 data points out of the five times they were asked to report. *SFScience* teachers reported an average of 24.0 total hours and control teachers reported an average of 24.8 total hours of instruction for the length of the implementation. As we observe later in Table 37, we have no confidence that the actual difference is different from zero.

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 20. At least two teachers reported not having any textbooks for their students. Five teachers practiced whole class reading of science for more than half of the time they spent on science. Three teachers reported using whole class reading activities less than half, leading us to conclude that this was a very common activity. Many teachers have students that are learning English and so vocabulary is stressed in all class activities.

Table 20. 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	44.4%	44.4%	22.2%	22.2%	44.4%	44.4%

For conducting laboratory activities, control teachers indicated that they have no set pattern of usage because of the planning required to incorporate activities. Teachers did agree that students found the activities fun, but had to work at making the connection with the concepts. Teachers named teaching schedule, classroom size and configuration, and limited availability of laboratory materials as serious barriers to conducting hands-on activities.

Table 21. 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	11.1%	11.1%	11.1%	22.2%	44.4%

Planning time for science instruction is also an important factor for implementing curriculum. All nine control teachers responded that they spent approximately 30% (50 minutes per week) of their total available planning time on science instruction. All teachers in the *SFScience* group report spending from 20% to 40% (approximately 25 to 40 minutes) planning for science.

Density of Science Inquiry Reflected in the Classroom

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

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

- questions are scientifically oriented
- learners use evidence to evaluate explanations
- explanations answer the questions
- alternative explanations are compared and evaluated
- explanations are communicated and justified

This framework is reflected in the sequenced activities of the SF science program as a continuum:

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

- Investigate (DI: students are given the procedures and materials to conduct an investigation; GI: students are given suggestions for procedures and materials; FI: students devise a plan for the investigation).

When we asked the teachers on the surveys, we asked about time spent doing these different activities. Both *SFScience* and control group teachers were asked these questions. The variable "science inquiry" is a composite of the time spent in six different aspects of the inquiry process as a percentage. Hence, it is on a scale of 0 to 100 and can be thought of as a measure of "inquiry process density" with 100 being an indication that the teacher and students were practicing the inquiry process every time science was taught. The average percentage density for the *SFScience* group was 30.33 and for the control group it was 28.33. While a greater amount of density is noticed for the *SFScience* condition, statistically we have no confidence that this difference between the groups is different from zero (p = 0.8).

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, assessment book, science kits, graphic organizers, and additional materials. Emphasis was placed on the using the development of inquiry skills by using the materials as sequenced from Directed Inquiry (DI) to Guided Inquiry (GI) and finally to Full Inquiry (FI). The 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. The audio tapes were not demonstrated because they were on backorder. The instructor distributed two Scott Foresman created handouts: *Scaffolded Inquiry* and *How Meaningful Science Learning Supports Reading Comprehension*. The district science specialist also handed out materials related to the Science grade-level standards to both *SFScience* and control teachers later in the day during another meeting.

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. The 5th grade science kits were backordered until late November. Several teachers reported missing other materials, such as the Activity Flipcharts and audio CDs. In some classrooms, because of the specialty subject method of scheduling, only one set of student editions were available for the first couple of months. All backordered and missing materials had arrived by December and full implementation began at all grade levels.

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. Textbooks were used most frequently of all of the materials provided. Many teachers use whole class reading to support students learning English. Two teachers reported having difficulty with the textbooks because the teacher's version did not match the students' version.

Six of the ten possible *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 22. Percent of Teachers Covering Each Chapter in Unit A-Life Science

		Chapter					
		1	**2**	**3**	**4**	**5**	**6**
Number of respondents	6	83.3%	16.7	16.7	83.3	50.0	50.0
Note: Four *SFScience* teachers did not provide information.							

Table 23. Chapters in Unit B-Earth Science Covered

		Chapter			
		7	**8**	**9**	**10**
Number of respondents	6	66.7%	66.7%	83.3%	50%
Note: Four *SFScience* teachers did not provide information.					

Table 24. Chapters in Unit C-Physical Science Covered

		Chapter				
		11	**12**	**13**	**14**	**15**
Number of respondents	6	50%	33.3%	50%	16.7%	33.3%
Note: Four *SFScience* teachers did not provide information.						

Table 25. Chapters in Unit D-Space & Technology Covered

		Chapter		
		16	**17**	**18**
Number of respondents	6	16.7%	33.3%	16.7%
Note: Some teachers did not teach any chapters in this unit. Four *SFScience* teachers did not provide information.				

Although alignment to standards continues to be a big issue and a challenge at all grades levels, teachers felt more challenged by the laboratory activities; small spaces and language needs were difficult to manage. No teacher had completed a Full Inquiry activity by the time of the classroom observation. In general all teachers thought that the textbook was too difficult. It's too vocabulary-rich and requires background information that their students don't have. Many teachers commented that they would like to see more foundational activities, such as pictures or video to 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. Although all units suffered from a lack of alignment to Utah state standards, Unit D, Space & Technology was the most problematic.

Table 26. Percent of Teacher Responses to Alignment of Unit A-Life Science

		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	6	16.7%	0%	16.7%	66.7%	0%
Note: Four *SFScience* teachers did not provide information.						

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

		Did not teach any of this unit	Taught too few to have an opinion	Aligned to standards well	Aligned somewhat	Aligned poorly
Number of respondents	6	0%	16.7%	33.3%	50%	0%
Note: Four *SFScience* teachers did not provide information.						

Table 28. For Unit C, How Well Was the Content Aligned to State Standards?

		Did not teach any of this unit	Taught too few to have an opinion	Aligned to standards well	Aligned somewhat	Aligned poorly
Number of respondents	6	50%	0%	33.3%	16.7%	0%
Note: Four *SFScience* teachers did not provide information.						

Table 29. For Unit D, How Well Was the Content Aligned to State Standards?

		Did not teach any of this unit	Taught too few to have an opinion	Aligned to standards well	Aligned somewhat	Aligned poorly
Number of respondents	6	66.7%	0%	16.7%	16.7%	50%
Note: Four *SFScience* teachers did not provide information.						

Many teachers incorporated the Leveled Readers into their science instruction and also used it successfully with their reading instruction. All of the teachers remarked that the Leveled Readers were very successful for their students. They noted two difficulties, the packaging ─ insufficient numbers at the lower end, and that still the vocabulary was 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. As noted before, scheduling sufficient time for science instruction with the hands-on materials was a challenge.

Few *SFScience* teachers used the assessments with their students consistently. Teachers used some of the materials to gauge when they needed to re-teach vocabulary.

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 Ogden, 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. When they did find the time to teach science, teachers used the reading materials and not many of the laboratory activities. The total time of the implementation was truncated at the 5th grade. For third and fourth grades, the length at best was 5 months; the implementation for fifth grade was three months. The second barrier was teachers' own level of knowledge regarding science concepts. Several teachers reported that they felt most comfortable with the Earth Science unit and some of the Life Science, but ill equipped to teach the Physical Science and the Space and Technology units. The *SFScience* curriculum lack of alignment to the state standards also contributed to the overall weak implementation.

Quantitative Impact Results

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

In the following sections, our analysis of the quantitative results takes the same form. We present the results of statistical models where we estimate whether the impact of the intervention depends on the level of certain moderator variables. For instance, we show the results of a model that tests whether there is a differential impact across the prior score scale. That is, we test for the interaction of treatment with the prior score. The fixed factor part of the table provides estimates of the factors of interest, in particular, whether being in a *SFScience* or a control class makes a difference for the average student. At the bottom of the table we give results for technical review – these often consist of random effects estimates which are added to the analysis to account for the fact that the individual results that come from a common upper-level unit (e.g., class or teacher) tend to be similar (i.e., the observations are dependent.) In some cases, to account for these dependencies, we model fixed rather than random effects but do not present the individual fixed effects estimates. Modeling the dependencies results in a more conservative estimate of the treatment impact. We note that the number of cases used to compute the effect size will often be larger than the number used in the mixed model analysis because to be included in the latter analysis a student has to have both a pretest and a posttest score.

Science Outcomes

Analysis Including Pretest

Our first analysis addressed Science achievement using the NWEA Science Concepts and Processes scale. Table 30 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 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 31 through Table 33. 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 30. Overview of Sample and Impact of *SFScience* on Science Achievement

	Condition	Means	Standard deviation[a]	No. of students	No. of classes	No. of teachers	Effect size	*p* value[b]
Un-adjusted	***SFScience***	193.01	9.87	199	13	10	-0.08	.76
	Control	193.34	8.51	159	11	9		
Adjusted	***SFScience***	193.23	9.99	179	13	10	-0.01	.94
	Control	193.31[c]	8.36	153	11	9		

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

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

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

Figure 1 provides a visual representation of specific information in Table 30. The bar graphs represent average performance using the metric of NWEA Science.

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 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 30.) We can see that the two groups were essentially indistinguishable. The high *p* value for the treatment effect (.94) indicates we should have no confidence that the actual difference is different from zero. We added 80%

confidence intervals to the tops of the bars. The overlap in these intervals further indicates that any difference we see is easily due to chance.

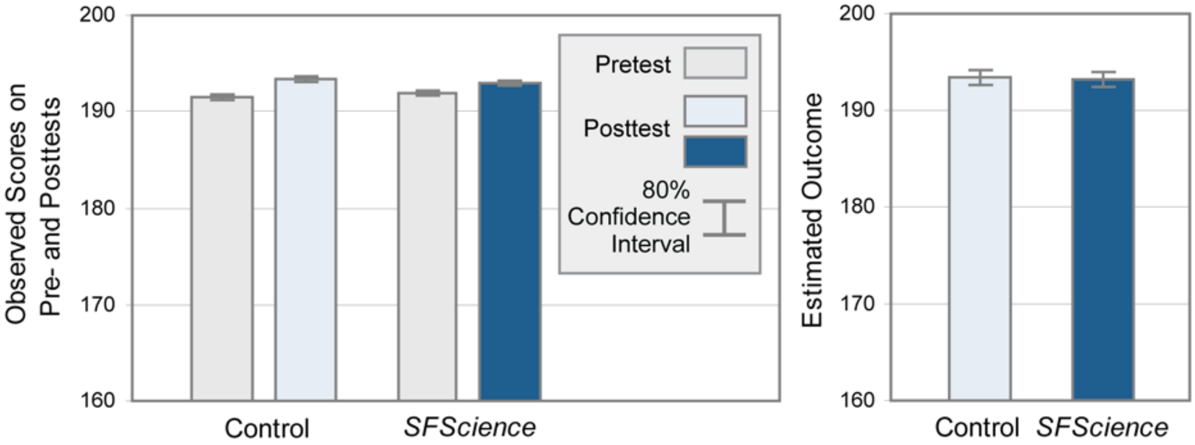

Figure 1. Impact on Science Achievement: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)

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.[2] We begin by examining the moderating effect of the prior score. Since the NWEA tests are on a continuous scale and the experiment involved three grades, we do not interpret low NWEA scores as indicating “low achieving” students within each grade. It is likely that third graders are more heavily represented in the lower range of the scores and fifth graders in the higher end of the scores. Table 31 shows the estimated impact of *SFScience* on students’ performance in science as measured by NWEA Science, as well as the moderating effect of the prior score.

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

Table 31. The Impact of *SFScience* on Student Performance on Science Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Estimated value for a control student with an average pretest	191.99	2.22	7	86.50	<.01
Impact of *SFScience* for a student with an average pretest	-0.06	1.06	7	-0.06	.95
Estimated change in control outcome for each unit increase on the pretest	0.71	0.06	307	10.96	<.01
Interaction of pretest and *SFScience*	-0.02	0.08	307	-0.26	.80
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	1.86	2.48		0.75	.23
Within-teacher variation	34.85	2.81		12.39	<.01

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

[b] Pairs were not modeled because fewer than 75% of teachers were paired. Teachers were modeled as a random factor.

Note. Of the 332 students we used to calculate the adjusted effect size, we removed 3 because they were influential points/outliers; as a result, 329 students were used in the impact model.

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 -0.06. This shows a small negative effect associated with *SFScience*. However, the *p* value of 0.95 gives us no confidence that the underlying effect is different from zero.

We also estimated the moderating effect of the pretest score on the impact of *SFScience* to see whether the intervention was differentially effective for students at different points along the pretest scale. The *p* value for this effect is .80. We have no confidence that the actual effect is different from zero.

As a visual representation of the results described in Table 31, we present a scatterplot in Figure 2, which shows student performance at the end of the year in science, as measured by the NWEA test, against their performance on NWEA Science in the fall. These graphs show where each student fell in terms of his or her starting point (horizontal x-axis) and his or her outcome score (vertical y-axis). Each point plots one student’s post-intervention score against his or her pre-intervention score. The darker points represent *SFScience* students; the lighter points, control students.

The two lines are the estimated values on the posttest for students in the *SFScience* and control conditions as determined using a simple model with no fixed effects.[3] Consistent with the results described above, we see that *SFScience* and the programs used in the control classrooms were equally effective as measured by the NWEA test.

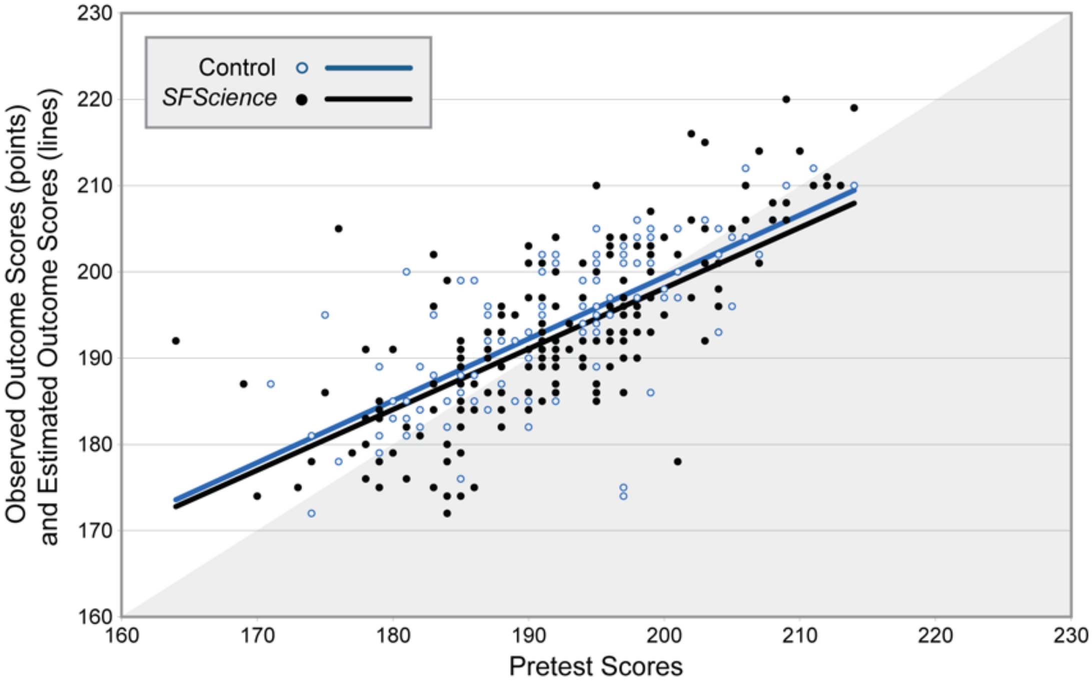

Figure 2. Comparison of Estimated and Actual Outcomes for *SFScience* and Control Group Students

Analysis Including Gender as a Moderator

We were also interested in whether *SFScience* was differentially effective for boys and girls in terms of Science achievement. Table 32 shows that there is no differential effect of *SFScience*

[3] Due to the complexity in estimating estimated values for models with fixed effects, we report a simpler model. We used the following criteria guide our decision; if 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 the *p* value does not go from ≥ .20 to <.20 (or from <.20 to ≥.20).

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.

depending on gender. In other words, boys and girls performed equally as well on NWEA Science when using the *SFScience* curriculum.

Table 32. Moderating Effect of Gender on Science Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Outcome for a girl with an average pretest in the control group	191.91	2.97	7	64.65	<.01
Estimated change in outcome for each unit increase on the pretest	0.64	0.05	306	13.85	<.01
Average *SFScience* effect for girls	0.19	1.54	7	0.12	.91
Difference (boys minus girls) in average performance in the control condition	1.46	1.09	306	1.34	.18
Difference (boys minus girls) in the average *SFScience* effect	-1.30	1.47	306	-0.89	.38
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	4.16	4.22		0.99	.16
Within-teacher variation	42.31	3.42		12.37	<.01

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

[b] Teachers were modeled as a random factor.

Note. Of the 332 students we used to calculate the adjusted effect size, we removed 3 because they were influential points/outliers; as a result, 329 students were used in the model.

Analysis Including English Proficiency as a Moderator

We were also interested in the moderating effect of student English proficiency on science achievement. In particular, we were interested in whether *SFScience* was differentially effective for English proficient and non-English proficient students. Table 33 shows the results of our analysis. We observe that there is no differential effect of *SFScience* depending on English proficiency status.

Table 33. Science Achievement Moderated by English Proficiency

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Average outcome for English learner in control	190.53	2.43	7	78.50	<.01
Estimated change in outcome for each unit increase on the pretest	0.67	0.04	305	15.32	<.01
Average *SFScience* effect for English learner	-0.64	1.57	7	-0.41	.70
Difference (score for English proficient minus English learner) in average performance in the control condition	1.53	1.09	305	1.40	0.16
Difference (score for English proficient minus English learner) in the average *SFScience* effect	0.67	1.56	305	0.43	.67
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	2.22	2.55		0.87	0.19
Within-teacher variation	33.17	2.68		12.36	<.01

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

[b] Teachers were modeled as a random factor.

Note. Of the 332 students we used to calculate the adjusted effect size, we removed 3 because they were influential points/outliers; as a result, 329 students were used in the model.

Reading Outcomes

Analysis Including Pretest

Our next set of analyses addresses reading achievement as measured by NWEA Reading. Table 34 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 35 and Table 36. 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 34. 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***	194.70	13.95	198	13	10	-0.11	.65
	Control	195.71	13.71	165	11	9		
Adjusted	***SFScience***	196.35	13.82	181	13	10	0.01	.92
	Control	196.16 [c]	13.72	157	11	9		

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

[b] The unadjusted *p* value is computed using a model that includes clustering of students in teacher but no other covariates. The adjusted *p* value is computed using a model that controls for clustering and that includes both the pretest and, where relevant, indicators for upper-level units in which the units of randomization are nested.

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

Figure 3 provides a visual representation of the information in Table 34. These graphs are interpreted the same way as Figure 1. This figure and the table preceding it provide an overview of the sample and overall impact of *SFScience* on student performance on NWEA Reading. The data indicates that any difference between the *SFScience* group and the control group in reading is easily due to chance.

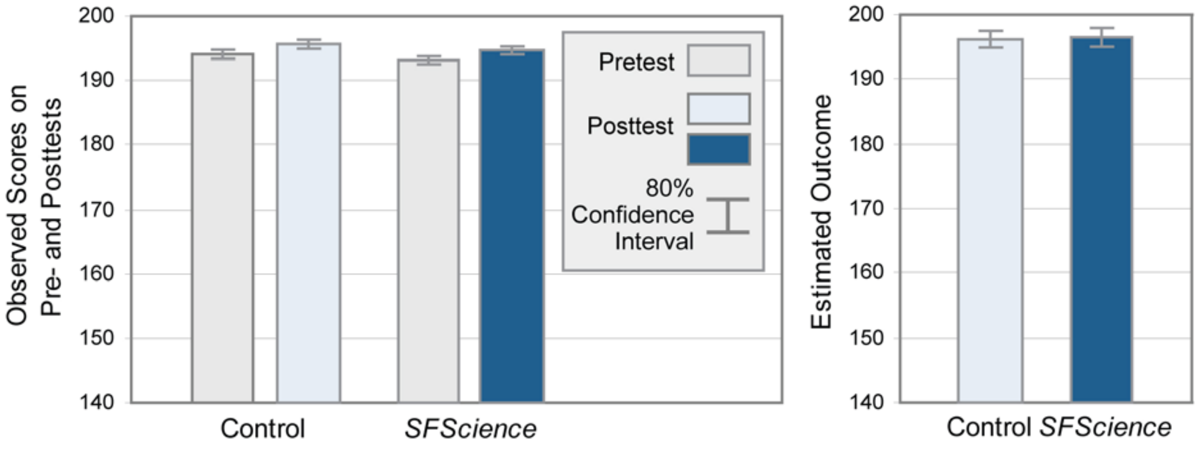

Figure 3. Impact on Reading Achievement: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)

Analysis Including Pretest as a Moderator

We now report on the analyses that examine both the overall impact of *SFScience* as well as the moderating effects of other variables.[4] 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 35 shows the estimated impact of *SFScience* on students' performance in science as measured by NWEA Reading, as well as the moderating effect of the prior score.

Table 35. The Impact of *SFScience* on Reading Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Estimated value for a control student with an average pretest	191.16	4.06	7	47.10	<.01
Impact of *SFScience* for a student with an average pretest	0.27	1.93	7	0.14	.89
Estimated change in control outcome for each unit increase on the pretest	0.88	0.04	314	20.91	<.01
Interaction of pretest and *SFScience*	-0.06	0.06	314	-1.04	.30
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	10.40	7.73		10.40	7.73
Within-teacher variation	44.64	3.56		44.64	3.56

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

[b] Pairs were not modeled because fewer than 75% of teachers were paired. Teachers were modeled as a random factor.

Note. Of the 338 students we used to calculate the adjusted effect size, we removed 2 because they were influential points/outliers; as a result, 336 students were used in the impact model.

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

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

score on the pretest. The estimate associated with *SFScience* is 0.27. This shows a small positive effect of *SFScience*. However, the *p* value of .89 gives us no confidence that the 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 .30. We have no confidence that the actual effect is different from zero.

As a visual representation of the results described in Table 35, we present a scatterplot in Figure 4, which shows student performance at the end of the year in reading, as measured by the NWEA test, against their performance on NWEA Reading in the fall. These graphs show where each student fell in terms of his or her starting point (horizontal x-axis) and his or her outcome score (vertical y-axis). Each dot 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. Consistent with the results described above, we see *SFScience* and the programs used in the control classrooms were equally effective as measured by NWEA Reading.

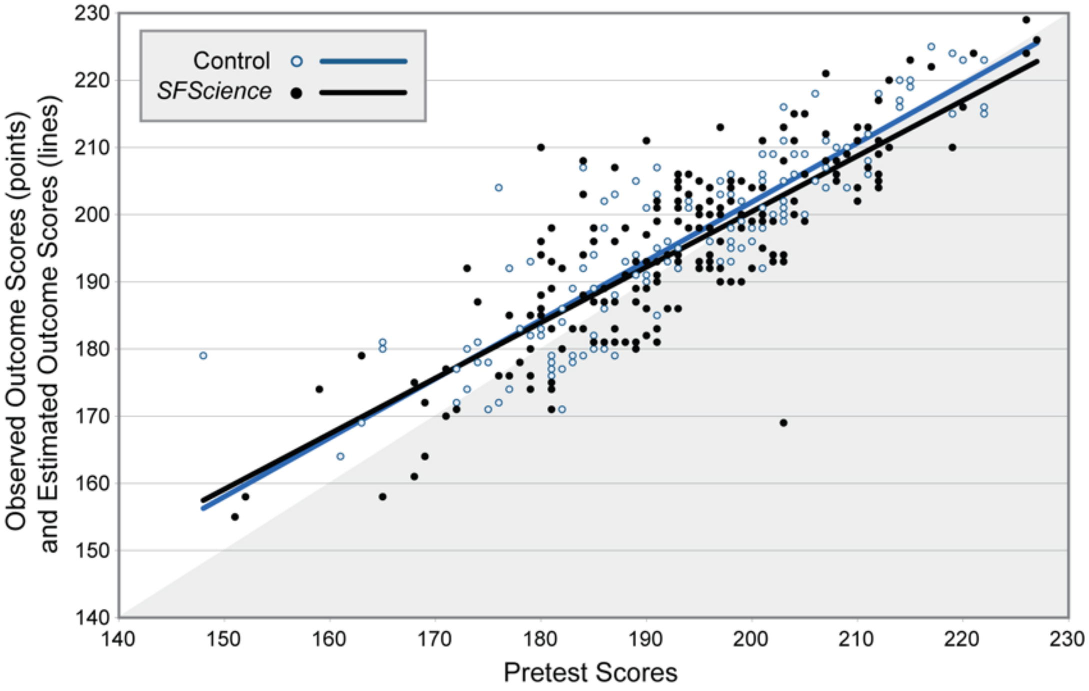

Figure 4. Comparison of Estimated and Actual Outcomes for *SFScience* and Control Group Students

Analysis Including English Proficiency as a Moderator

We were also interested in whether *SFScience* was differentially effective for English proficient and non-English proficient students. Table 36 shows that there is no differential effect of *SFScience* depending on English proficiency status.

Table 36. Reading Achievement Moderated by English Proficiency

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Average outcome for English learner in control	189.99	4.21	7	45.10	<.01
Estimated change in outcome for each unit increase on the pretest	0.84	0.03	313	26.27	<.01
Average *SFScience* effect for English learner	1.67	2.32	7	0.72	.50
Difference (score for English proficient minus English learner) in average performance in the control condition	1.87	1.25	313	1.49	.14
Difference (score for English proficient minus English learner) in the average *SFScience* effect	-2.09	1.76	313	-1.19	.24
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	10.70	7.87		1.36	.09
Within-teacher variation	44.59	3.56		12.53	<.01

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

[b] Teachers were modeled as a random factor.

Note. Of the 338 students we used to calculate the adjusted effect size, we removed 2 because they were influential points/outliers; as a result, 336 students were used in the model.

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 5. 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.

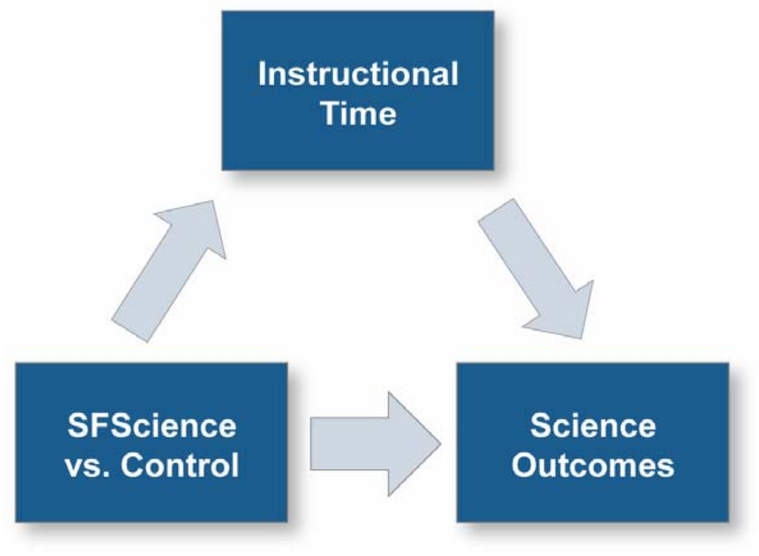

Figure 5. Relationships for Exploratory Analysis of Implementation Variables

Instructional Time

We wanted to explore the relationship between how much time was spent teaching science and science outcomes. The surveys provided data on this variable. Our measure is the total hours spent teaching science during the experiment. Instructional time was measured by each teacher's self-report of the number of minutes 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.

Table 37 shows *SFScience* teachers taught approximately 4 fewer hours of science during the year. However, the *p* value of .74 gives us no confidence that the actual difference is different from zero.

Table 37. The Impact of *SFScience* on Hours of Science Instruction Time

Fixed effects	Estimate	Standard error	DF	*t* value	*p* value
Hours of science for a control teacher	29.93	22.65	6	1.32	.23
Impact of *SFScience* on hours of science instruction	-3.72	10.68	6	-0.35	.74
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Residual teacher variance	398.94	230.33		1.73	.04

Given that there are differences in the amount of instructional time across the teachers in the experiment, we next explored whether there was a correlation between time spent and student achievement. The result of this investigation is purely correlational—we cannot be sure whether it is instructional time or some other variable which is correlated with instructional time (e.g., teacher enthusiasm) that is the true cause of the student outcome. A test of the correlation between instructional time and student performance in science reveals a slight negative relationship between *SFScience* usage and the student outcome. The *p* value for this effect is .12, which gives us some confidence that the true relationship is in fact different from zero.

Table 38. Relationship of Instructional Time to Student Outcome

Fixed effects	Estimate	Standard error	DF	*t* value	*p* value
Estimated value for student with an average pretest	-10.61	42.76	5	-0.25	.81
Estimated change in outcome for each pretest point	1.06	0.22	5	4.79	<.01
Estimated change in outcome for hour of science time	-0.07	0.04	5	-1.86	.12
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Residual teacher variance	3.74	2.36		1.58	.06

[a] Schools and pairs of grades used for random assignment are also modeled as a fixed factor but not included in this table.

Discussion

We began this research in Ogden City School District with the question of whether *Scott Foresman Science* was as effective as or more effective than their existing programs. 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.

Given the overall weak implementation of the *SFScience* program in the schools, the overall results are not surprising. We found no overall difference between the science or reading scores of students taught using *SFScience* as compared to the established program. We also found no differences in the effectiveness of the program that depending on whether the student a) started out at the higher range or lower range of the science or reading scale, b) was an English learner, c) was a boy or girl. Teachers in both groups reported spending approximately the same amount of time teaching science. *SFScience* and control classrooms were also not different in our measure in inquiry density of the science lessons.

Our experiment in Ogden was small, involving only 19 teachers. With small numbers we must caution that we have limited ability to detect with any statistical confidence small differences that may be important educationally. This experiment was part of a larger five-district national study but we recognize that the specific resources, demographics, and educational agendas make analyses of specific cases worthwhile, although often not applicable outside of the participating district. This report is not intended to provide widely generalizable results and the reader should consider the characteristics of this district to evaluate the applicability of the findings.