Empirical Education

RESEARCH REPORT

# Comparative Effectiveness of *Scott Foresman Science*:

## A Report of a Randomized Experiment in Reynoldsburg City Schools

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

### About Empirical Education Inc.

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

# Comparative Effectiveness of *Scott Foresman Science:*

## A Report of a Randomized Experiment in Reynoldsburg City 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 research project consists of a randomized experiment in a few of the Reynoldsburg City Schools. 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 *SFScience* is more effective than the current curriculum being used by the participating campuses in the Reynoldsburg City Schools. 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 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 Ohio state versions:

**Table 1. Scott Foresman Supplied Materials**

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

### District Science Materials

Not all teachers had textbooks for the students; when they did, there were two textbooks in use. In fifth grade, the teachers were using Harcourt Brace. In third grade, teachers were using the 2003 version of Scott Foresman. Some teachers used materials that they had developed.

## Site Descriptions

### Reynoldsburg, OH

The city of Reynoldsburg is located 12 miles East of Columbus and is generally considered part of the larger metropolitan area. With a population of 32,000 in 11 square miles, it is a small residential community.

**Table 2. Reynoldsburg Racial Makeup**

| Race/Ethnicity | % of Population |
|---|---|
| **White** | 84.0 |
| **African American** | 10.4 |
| **American Indian/Native Alaskan** | 0.2 |
| **Asian and Native Hawaiian or Pacific Islander** | 1.7 |
| **Other race** | 0.2 |
| **Two or more races** | 1.7 |
| **Hispanic origin (of any race)** | 1.8 |

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

### Reynoldsburg City Schools, OH

Reynoldsburg City Schools operate six elementary schools, three middle schools, and one high school; two elementary (K-4) schools and one middle school (fifth and sixth grades) participated in this study. The following tables summarize the demographic makeup of the school district.

**Table 3. Background of the Reynoldsburg City Schools**

| Reynoldsburg City Schools | |
|---|---|
| **Total schools** | 10 |
| **Total teachers** | 316 |
| **Student to teacher ratio** | 20.46 |
| **Grades** | PK -12 |
| **Student population** | 6948 |
| **Migrant students** | 0.5% |
| **ELL students** | 2.2% |

Source: Ohio Department of Education, 2005 and CCD Public School District Data for 2005-2006

**Table 4. Ethnic Makeup of the Reynoldsburg City Schools**

| Race/Ethnicity | % of population |
|---|---|
| **White, non-Hispanic (%)** | 66.5% |
| **Black, non Hispanic (%)** | 25% |
| **Hispanic (%)** | 2% |
| **Asian/Pacific Islander (%)** | 1.8% |
| **American Indian/Alaskan Native (%)** | 0.13% |

Source: Ohio Department of Education, 2005 and CCD Public School District Data for 2005-2006

## 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 Reynoldsburg City Schools occurred on September 22, 2005 with 16 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. One teacher decided not to participate and excused herself. Of the remaining 15 teachers, all filled out consent forms, but only 11 agreed to form participant pairs. Four teachers expressed concern over participation because their school was already conducting a curricular adoption. These four teachers needed to check with their school principal before they could commit to the study fully. Consequently, researchers only randomized 11 teachers, five pairs and one unpaired teacher. Two days after the randomization meeting, those four teachers at the school with the adoption program communicated with us that they would not be able to participate due to prior commitments and excused themselves from the study. Since these teachers did not participate in the randomization process, they were deemed non-participants unrelated to assignment conditions rather than attrition.

### Randomization

The unit of randomization at this site is the teacher. Eleven 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 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 years of teaching experience, resulting in a within grade-level randomization paired on teaching experience. 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 Reynoldsburg site**

| Teacher Assignment Status | Number Participating |
|---|---|
| ***SFScience*** | 5 |
| **Control** | 6 |
| **Total** | **11** |

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 10 classrooms assigned to the control condition and 7 classrooms to the *SFScience* condition. No one teacher taught more than three science classes.

### Sample Size

One concern we had was with sample size. Sample size (in this case, number of teachers) is one of the factors that determines how precisely we can measure an effect of a given size. With smaller samples we are usually able only to detect larger effects. We usually measure the size of an effect in terms of standard deviation units, which tells us how big the effect is, controlling for the spread in observed scores. Based on the available sample size, and certain assumptions about other parameters that affect the size of the effect that we can detect, we calculated that we can detect an effect size as small as .60. 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 .44. We emphasize that the matching design possibly lowers this value. From this we see that the experiment is not powered to detect a very small effect, which may be real but not discernable given the number of teachers in the study.

With 11 teachers total, we realized that we did not have as large a sample as was called for by our initial design. Because the importance of the information warranted gathering the available data even if the results ultimately proved inconclusive, the district in consultation with the researchers decided to move forward with the experiment.

## 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 September 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 October 26th, 2005. Classroom observations were conducted during the week of March 20th. All five teachers in the *SFScience* group were observed. Three of the six 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 74.75% for both groups, an 80.00% response rate for the *SFScience* teachers, and a 70.37% 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 | 60.00% | 66.67% | 63.64% |
| **Survey 2** | Jan. 16 – 20 | Resources | 80.00% | 66.67% | 72.73% |
| **Survey 3** | Jan. 23 – 27 | Interactions with materials/Students | 80.00% | 83.33% | 81.82% |
| **Survey 4** | Feb. 6 – 10 | More Interactions | 80.00% | 83.33% | 81.82% |
| **Survey 5** | Feb. 20 - 24 | Time & Preparation | 80.00% | 83.33% | 81.82% |
| **Survey 6** | Mar. 6 – 10 | Materials & Resources | 100% | 66.67% | 81.82% |
| **Survey 7** | Mar. 20 – 24 | Assessments | 80.00% | 50.00% | 63.64% |
| **Survey 8** | May 1 – 5 | More Interactions | 80.00% | 66.67% | 72.73% |
| **Survey 9T*** | May 26 | Final Survey | 80.00% | N/A | 80.00% |
| **Survey 9C**** | May 26 | Final Survey | N/A | 66.67% | 66.67% |

*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.

There were no anomalies reported in the administration of the assessments, but the pretest schedule was interrupted at the middle school because they had a fire in one of their buildings. The middle school closed a wing and classrooms were reorganized temporarily.

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. 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.

**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 | 5 | 7 | 59 | 47 | 35 | **141** |
| **Control** | 3 | 6 | 10 | 59 | 41 | 118 | **218** |

[1] In technical terms, randomization ensures lack of bias due to selection, 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

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Totals** | **3**[a] | **11** | **17** | **118** | **88** | **153** | **359** |

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

### Years of Teaching Experience

During the randomization process teachers identified themselves according to years of teaching experience. This was the second criterion by which we subdivided teachers in the pairing process. 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.

**Table 8. Distribution of Years Teaching Experience**

| | Number of Teachers | | |
|---|---|---|---|
| **Condition** | **0 to 3 years** | **4 or more years** | **Totals** |
| ***SFScience*** | 1 | 4 | **5** |
| **Control** | 2 | 4 | **6** |
| **Totals** | **3** | **8** | **11** |

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 the *SFScience* condition were relatively new to teaching at their grade level.

Additionally, we noted that three teachers (two *SFScience* and one control) did not teach science in the previous academic year.

**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*** | 5 | 20% | 20% | 0% | 60 % |
| **Control** | 6 | 17% | 0% | 50% | 17% |

Note. 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*** | 5 | 60% | 20% | 20% | 0% |
| **Control** | 6 | 33% | 0% | 33% | 17% |

Note. One control teacher did not provide this information.

**Table 11. Years Teaching Science**

| | Number of teachers | 0-3 years % | 4-6 years % | 7-15 years % | 15+ years % |
|---|---|---|---|---|---|
| ***SFScience*** | 5 | 20% | 60% | 20% | 0% |
| **Control** | 6 | 17% | 0% | 50% | 17% |

Note. One control teacher did not provide this information.

**Table 12. Science Coursework in College**

| | Number of teachers | None % | Some % | Minor % | Major % |
|---|---|---|---|---|---|
| ***SFScience*** | 5 | 0% | 100% | 0% | 0% |
| **Control** | 6 | 0% | 83% | 0% | 0% |

Note. One control teacher 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*** | 5 | 40% | 60% |
| **Control** | 6 | 33% | 67% |

Note. 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 teacher experience, student characteristics such as ethnicity, social economical status, and gender, and student pretest outcomes.

From the previous tables, we see that 359 students were enrolled in the study. Of these, 20 students were designated as needing special education supports; we will not include those students in the analysis. Hence, the following analyses are based on a sample size of 339 students.

### Student Variables

#### Ethnicity

Table 14 summarizes the distribution of student ethnicity. We see that ethnicity was not distributed evenly between conditions in spite of randomization. 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. Ethnicity for *SFScience* and Control Groups**

| | Ethnicity | | | | | | |
|---|---|---|---|---|---|---|---|
| **Condition** | **Asian** | **Hispanic** | **Native American** | **Black** | **White** | **Multi-racial** | **Totals** |
| ***SFScience*** | 1 | 5 | 1 | 29 | 88 | 1 | 125 |
| **Control** | 3 | 10 | 0 | 65 | 119 | 8 | 205 |
| **Totals** | 4 | 15 | 1 | 94 | 207 | 9 | 330 |
| **Chi-square statistics** | | | | | | | ***p* value** |
| **Fisher's Exact Test** | | | | | | | .11 |

Notes. Fisher's exact test is reported due to the fact that one or more of the expected cell counts is smaller than 10. Information about ethnicity is missing for 9 students.

#### Socio-Economic Status

We were missing SES data for 44% of the students so we were unable to use this variable in our analysis of the experimental groups.

### Gender

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

**Table 15. Gender for *SFScience* and Control Groups**

| Condition | Gender | | |
|---|---|---|---|
| | Male | Female | Totals |
| ***SFScience*** | 69 | 61 | **130** |
| **Control** | 118 | 91 | **209** |
| **Total** | **187** | **152** | **339** |
| **Chi-square statistics** | **DF** | **Value** | ***p* value** |
| | 1 | 0.37 | .54 |

### Characteristics of the Experimental Groups Defined by Pretest

We also checked whether randomization resulted in balance on science pretest scores, a variable that we include in most of our analyses to increase the precision of our estimates. Table 16 shows the results of students in grades 3 to 5 for whom pretests were available.

**Table 16. Difference in Pretest Scores Between Students in the *SFScience* and Control Groups**

| Descriptive statistics: Pretest outcomes | Raw group means | Standard deviation | Number of students | Standard error | Effect size[a] |
|---|---|---|---|---|---|
| ***SFScience*** | 196.71 | 11.38 | 106 | 1.11 | -.28 |
| **Control** | 199.65 | 9.28 | 169 | 0.71 | |
| ***t* test for difference between independent means** | **Difference** | | **DF** | ***t* value** | ***p* value** |
| **Condition (*SFScience* – control)** | -2.94 | | 273 | 2.34 | .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.

The *SFScience* and control groups had slightly different average pretest scores on the NWEA test, as shown in Table 16. However, when we accounted for the fact that outcomes for students of the same teacher tend to be related by modeling these dependencies, the p-value increased to .92. 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

Out of the eligible enrollment of 339 in grades 3, 4, and 5 on fall class rosters, 20.6% of students did not take the posttests. Sixty-four students (19%) did not have pretest scores. Of these remaining 275 students, no one is missing posttest scores.

### NWEA Reading

Out of the eligible enrollment of 339 in grades 3, 4, and 5 on fall class rosters, 26.7% of students did not take the posttests. One hundred one students (29.7%) did not have pretest scores. Of these remaining 238 students, no one is missing posttest scores.

## Implementation Results

In this section we describe more fully the aspects of the implementation that characterize this intervention. We used the following questions to guide our descriptions and analysis: What resources are needed to manifest the *SFScience* condition? Are there differences in the extent, quality, and type of implementation of the materials? We also studied the features of the implementation to identify possible variables related to the outcome measures. Our perspective takes into account three levels of resources needed to implement science instruction: those 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 schools participated in the study, two elementary schools (K-4) and one middle school (5$^{th}$ & 6$^{th}$). There are some minor differences between the middle school and elementary school environment.

#### Classroom Settings for Instruction

The classroom setting was observed during the week of March 20$^{th}$, 2006. The classroom observations were conducted once due to 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.

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. Other teachers had students arranged in small clusters of 4 students with no specific designation of the front of the classroom. Whiteboards were distributed around the room.

The middle school setting had a separate room for storing laboratory supplies. We noted district science kits and textbooks stored there, but much was in disarray and incomplete. One teacher

reported that they did not have time to organize the materials in storage and so were not used as often as intended. Teachers had some storage cupboards in their classrooms as well, but these were filled with other materials. In general, all teachers commented that storage of materials is a common problem.

Most teachers had some computer stations in the classroom, but not enough for every student. Groups of students worked at the computers for 10 to 20 minutes at a time then returned to their desks. The activities at the computer were Web quests and scavenger hunts or looking up science facts to supplement instruction. The computer activities were practiced by both *SFScience* and Control groups. Televisions and video playback/recorder systems were in evidence or accessible by both teacher groups. Some teachers liked to supplement instruction using videos of Bill Nye the Science Guy, other teachers reported that they rarely used videos but instead used the Internet. Every teacher had an overhead projector that they used periodically.

One teacher in the *SFScience* group used a voice amplification system to ensure that all students heard her voice. She taught a gifted 5$^{th}$ grade group that needed constant supervision and direction.

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. All teachers supplemented instruction with some sort of Internet activities.

### Opportunities for Learning

This site was identified after the school year had started and materials did not arrive until late November. Teachers had already begun using other materials and methods to teach science. The implementation of the SFScience materials began in early December.

At both the elementary schools and middle school, 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. Two 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 23.8 total hours and Control teachers reported an average of 21.9 total hours of instruction for the length of the implementation. As we observe later in Table 34, this difference is not statistically distinguishable.

During the timeframe of the study, the participating middle school experienced two events that disrupted the science instruction schedule. There was a fire in one of the buildings in November. This caused disruption to the pretest administration schedule and to all instruction.

Middle school teachers were asked to suspend teaching subjects other than mathematics and reading because the school was failing to meet their Academic Yearly Progress (AYP) on state exams. Science instruction was suspended for a total of three and half months from mid-November to March.

### Control Materials

As noted before, there were two types of textbooks in evidence, Harcourt Brace and Scott Foresman Science 2003. When asked about materials usage some control teachers responded as shown in Table 17. At least two teachers reported not having any textbooks for their students.

**Table 17. Primary Sources for Science Instruction**

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

For conducting laboratory activities control teachers indicated that they have no set pattern of usage. It depends on the topic and the availability of materials. For at least one teacher every lesson included an inquiry activity. Teachers felt that they needed an average of 90 minutes of instruction to include hands-on inquiry activities into their lessons. Additionally they indicated that needed better designed classrooms to accommodate science inquiry because they felt that the physical aspects of their rooms were limiting their choices of activities. They also wished for materials better aligned to the standards. One teacher specifically reported using Project Jason materials developed by the National Science Teachers Association for her hands-on science unit on planets. Teachers are providing their own materials brought from home approximately 20% of the time.

**Table 18. Percentage of Time Devoted to Hands-on Science Activities**

| How much time was spent on hands-on science activities (where students practiced science inquiry steps: investigation, hypothesis, observation and data collection, presentation of results)? | | | | | | |
|---|---|---|---|---|---|---|
| | | 90-100% | 50-89% | 30-49% | 10-29% | Less than 10% |
| **Number of Respondents** | 4 | 25% | 25% | 25% | 25% | 0% |

Planning time for science instruction is also an important factor for implementing curriculum. Five of the possible seven Control teachers responded that they spent approximately 20% (30

to 60 minutes per week) of their total available planning time on science instruction, but that they could use more time. Only one teacher reports planning instruction with other members of her grade level group which takes place every six weeks.

### 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 43.75 and for the control group it was 33.15. While a greater amount of density is noticed for *SFScience* condition, statistically we have no confidence that this difference between the groups is different from zero.

## Implementation of *SFScience*

### Training and Support

The one-half day training took place on October 26, 2006 at the district office boardroom. 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 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.

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

One teacher noted that it had been several years since she had last taught science and much of her knowledge was “rusty”. Another teacher commented that she would like more time to research current topics of interest to make the curriculum relevant to her students. Yet, another teacher noted that she did not have the background knowledge to teach any of the components of Unit C – Physical Science. It is clear that teachers would benefit from getting additional support to model other components of the curriculum and from science instruction.

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

### Availability and Use of Materials

Every teacher assigned to the *SFScience* group received sufficient materials to use with the number of students that they taught whether they taught in a self-contained classroom or in a specialty subject classroom. Several teachers reported missing some materials, the Leveled Reader Teacher’s Guide, workbooks and the assessment CD.

*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.

Four of the five *SFScience* teachers responded to the survey questions regarding the content covered in their classrooms. Teachers could select as many chapters within a unit that they covered. Note that content presented in chapters vary by grade level. This data is presented as an overall idea of what was used by the teachers and not specific to any one grade level.

**Table 19. Percent of Teachers Covering Each Chapter in Unit A-Life Science**

| | | Chapter | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| **Number of Respondents** | 4 | 100% | 75% | 25% | 50% | 25% | 0% |

**Table 20. Chapters in Unit B-Earth Science Covered**

| | | Chapter | | | |
|---|---|---|---|---|---|
| | | 7 | 8 | 9 | 10 |
| **Number of Respondents** | 4 | 50% | 50% | 50% | 25% |

**Table 21. Chapters in Unit C-Physical Science Covered**

| | | Chapter | | | | |
|---|---|---|---|---|---|---|
| | | 11 | 12 | 13 | 14 | 15 |
| **Number of Respondents** | 4 | 0% | 0% | 50% | 25% | 0% |

Note. Teachers did not teach any other chapters in this unit and not all teachers taught chapters in this unit.

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

| | | Chapter | | |
|---|---|---|---|---|
| | | 16 | 17 | 18 |
| **Number of Respondents** | 4 | 50% | 50% | 50% |

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. One teacher reported she got in trouble with her principal for teaching a chapter not in the state standards. So it is not just a question of whether it is required, but also a question of not teaching any unit not required. Teachers were very vocal about needing texts that strictly align with the standards because it takes much more planning time to make changes. Each chapter had applicable activities but then the DI-GI-FI sequence was destroyed. No teachers completed the inquiry sequence so that they could give the students a Full Inquiry experience. 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 normally “A” students were getting “B’s”. This dynamic was stressful for both teachers and students as parents became more aware and began complaining.

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

**Table 23. Percent of Teacher Responses to Alignment of Unit A-Life Science**

| | | 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 | 4 | 0% | 0% | 50% | 0% | 50% |

**Table 24. 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 | 4 | 25% | 0% | 0% | 75% | 0% |

**Table 25. 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 | 4 | 25% | 25% | 0% | 0% | 50% |

**Table 26. 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 | 4 | 25% | 0% | 25% | 0% | 50% |

Unit C, Physical Science did not apply for third grade and Unit D, Space and Technology did not apply for fourth grade. Overall, the teachers were disappointed that the content did not align better to their state standards when the books were supposed to be geared for Ohio.

Many teachers had trouble incorporating the Leveled Readers into their science instruction, but used it successfully with their reading instruction. Those teachers that were using the "specialty subject" model of instruction had fewer opportunities to use the Readers. Many teachers report that their students asked for the Readers specifically. The higher performing students seemed to really "run" with the Readers.

As for the Science Kits, storage was a problem and the racks did not help much. The 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.

Five of the six *SFScience* teachers used at least one assessment with their students. One teacher thought students were performing poorly on the assessments because the assessment had one chapter's worth of material and in the teacher's opinion was too much. She started to give shorter section tests throughout the chapter with only questions from that current section. Students still struggled. The teachers thought that the CD containing assessment materials were very helpful and much more useful because of the flexibility it allowed to formulate questions. All of the teachers thought that the assessments were too difficult. One teacher doctored them so that the multiple choice questions only had two options. Even then, students had a tough time and teacher needed to curve grading.

### 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 Reynoldsburg, 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 alignment of the curricular content to the state standards for each of the grade levels. This lack of alignment led teachers to spend more time planning and to skip whole sections of a chapter which in turn destroyed the designed sequence of activities. Students did not get the scaffolded steps of the inquiry process central to the design of this curriculum in an orderly sequence. The second factor that must be considered is the total length of implementation. For third and fourth grades, the length at best was 5 months; the implementation for fifth grade was three months.

## 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.

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 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 often will be larger than the number used in the mixed model analysis because to be included in the latter analysis a student has to have both a pretest and a posttest score.

### Science Outcomes

#### Analysis Including Pretest

Our first analysis addressed Science achievement using the NWEA Science Concepts and Processes scale. Table 27 provides a summary of the sample we used in the analysis and the results for the comparison of *SFScience* and control. The "Unadjusted" row gives information about all the students in the original sample for whom we have a posttest. This shows the means and standard deviations as well as a count of the number of students, classes and teachers in that group. The last two columns provide the effect size, that is, the size of the difference between the means for *SFScience* and control in standard deviation units, and the *p* value, indicating the probability of arriving at a difference as large or larger than the absolute value of the one observed when there truly is no difference. The "Unadjusted" row is based on all students with a posttest and the estimated effect size takes into consideration the clustering of students in upper-level units (i.e., that students are grouped within teachers.) The "Adjusted" row is based on the students who have both pretests and posttests. This is the sample that we use in the analyses on which we base our results reported in Table 28 and Table 29. The means, and therefore the effect size, are adjusted to take into account the student pretest scores. The adjusted effect size is based on a model that includes fixed effects for schools as well as pairs within which we randomized. It also figures in the effect of students being grouped within teachers.

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

| | Condition | Means | Standard deviations[a] | No. of students | No. of classes | No. of teachers | Effect size | *p* value[b] |
|---|---|---|---|---|---|---|---|---|
| **Un-adjusted** | ***SFScience*** | 200.38 | 10.62 | 110 | 7 | 5 | 0.05 | 0.92 |
| | **Control** | 202.36 | 10.15 | 175 | 10 | 6 | | |
| **Adjusted** | ***SFScience*** | 202.43 | 10.41 | 106 | 7 | 5 | 0.04 | 0.74 |
| | **Control** | 202.09[c] | 10.13 | 169 | 10 | 6 | | |

[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 specific information in Table 27. 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 29.) We can see that the two groups were essentially indistinguishable. The high *p* value for the treatment effect (.74) 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.

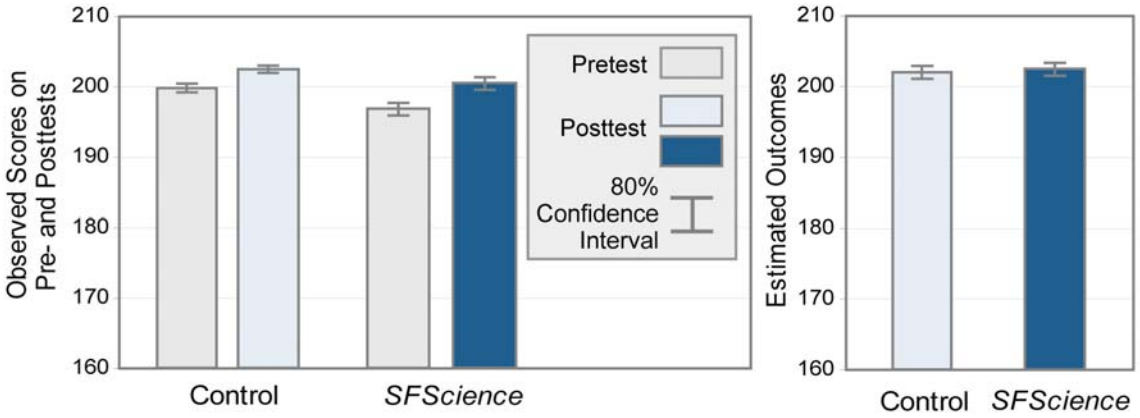


**Figure 1. Impact on Science Achievement: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)**

### 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. We begin by examining the moderating effect of 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 28 shows the estimated impact of *SFScience* on students' performance in science as measured by NWEA Science, as well as the moderating effect of the prior score.

**Table 28. Mixed Model Estimating the Impact of *SFScience* on NWEA Science Outcomes**

| Fixed effects[a] | Estimate | Standard error | DF | *t* value | *p* value |
|---|---|---|---|---|---|
| **Predicted value for a control student with an average pretest** | 200.56 | 1.58 | 4 | 126.78 | <.01 |
| **Impact of *SFScience* for a student with an average pretest** | 0.36 | 1.37 | 4 | 0.26 | .80 |
| **Predicted change in control outcome for each unit increase on the pretest** | 0.81 | 0.06 | 260 | 13.63 | <.01 |
| **Interaction of pretest and *SFScience*** | -0.24 | 0.08 | 260 | -3.11 | <.01 |
| **Random effects[b]** | **Estimate** | **Standard error** | | ***z* value** | ***p* value** |
| **Teacher mean achievement** | 1.95 | 3.7 | | 0.53 | .3 |
| **Within-teacher variation** | 31.72 | 2.78 | | 11.41 | <.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 predicted value for a control student with an average pretest applies to a particular school and assignment pair.

[b] Teachers were modeled as a random factor.

The row in the table labeled "Impact of *SFScience* for a student with an average pretest" tells us whether *SFScience* made a difference in terms of student performance on the NWEA test of science for a student who has an average score on the pretest. The estimate associated with *SFScience* is 0.36. This shows a small difference associated with *SFScience*. However, the *p* value of .80 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 smaller than 0.01. We have strong confidence that the actual effect is different from zero. In other words, the effect of *SFScience* was different depending on where the student started on the pretest.

As a visual representation of the results described in Table 28 we present a scatterplot in Figure 2, which shows student performance at the end of the year in science, as measured by NWEA Science, against their performance on the 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 shaded area in the lower right of the graph is the area of negative change (i.e., where students lost ground).

The two lines are the predicted values on the posttest for students in the *SFScience* and control conditions as determined using 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.

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

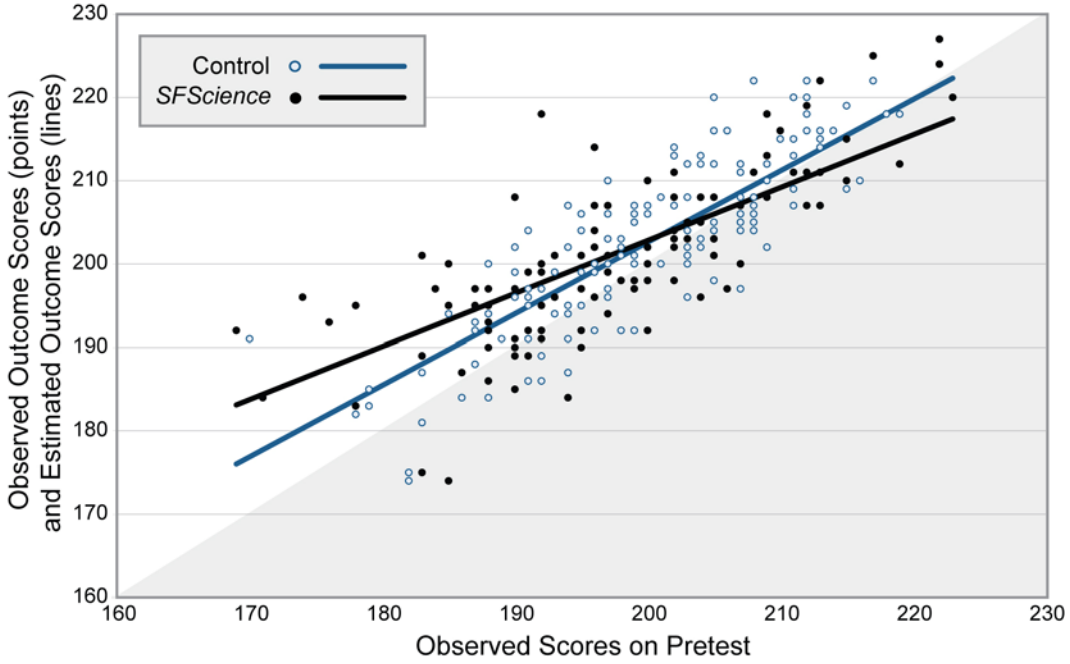


**Figure 2. Comparison of Predicted and Actual Outcomes for *SFScience* and Control Group Students**

Figure 3 illustrates the interaction in terms of the predicted difference between the *SFScience* and control groups for different points along the prior score scale. This display of the results allows us to see where *SFScience* had its greatest impact.[3] 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 predicted outcome for a *SFScience* student minus the predicted outcome for a control student. Around the difference line, we provide gradated bands representing confidence intervals. These confidence intervals are an alternative way of expressing uncertainty in the result. The band with the darkest shading surrounding the dark line is the "50-50" area, where the difference is considered equally likely to lie within the band as not. The region within the outermost shaded boundary is the 95% confidence interval—we are 95% sure that the true difference lies within these extremes. Between the 50% and 95% confidence intervals we also show the 80% and 90% confidence intervals. We also add points along the middle line to mark what the estimated treatment effect is for the median student for each quartile of the pretest. Consistent with the results in Table 28, there is evidence of a differential impact of the intervention across the prior score scale as measured by NWEA Science. Considering the points representing the median student in the bottom and top quartiles, it appears that *SFScience* is beneficial for the students scoring at the lower range of

[3]As with the scatter plot, 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).

the test. We have no confidence that there is an actual benefit of *SFScience* for students scoring at the upper range of the test.

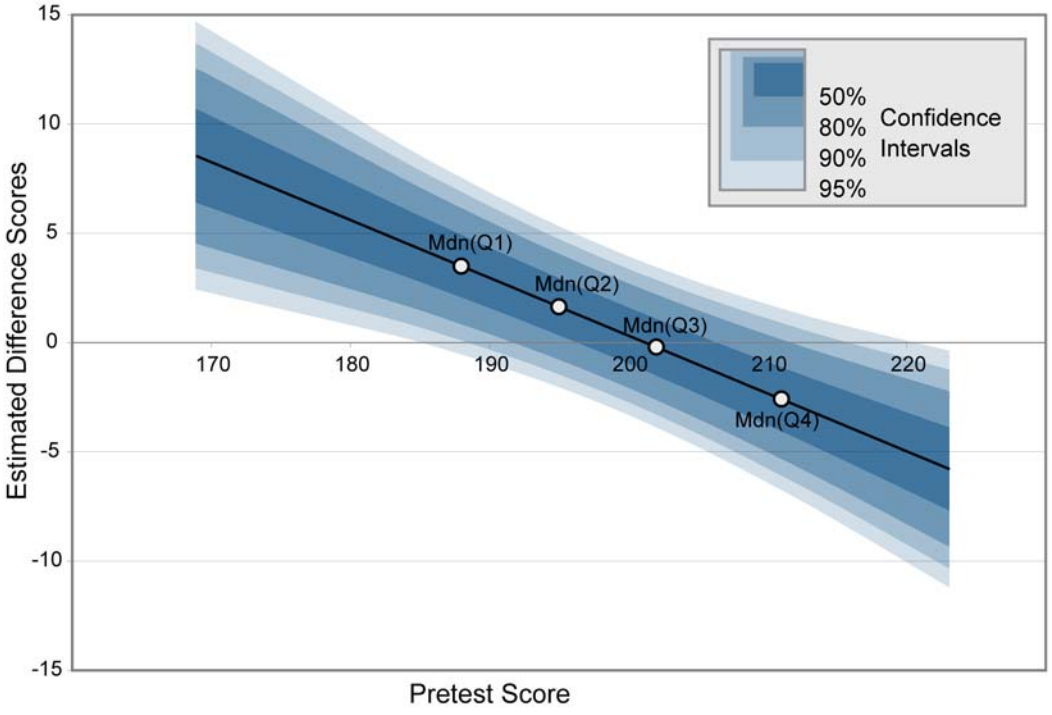


**Figure 3. Differences between *SFScience* and Control Group Science Achievement: Median Pretest Scores for Four Quartiles Shown**

Figure 4 presents the same information represented in Figure 3 but this time in the form of a bar graph showing the predicted difference between *SFScience* and control conditions for students at the medians of the first and fourth quartiles of the pre-test measure. The bar graph includes the 80% confidence interval as a marker at the top of the bars.[4] 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 at the median of the first quartile there is a substantial difference in the predicted outcomes in the two conditions, with the *SFScience* group scoring higher than the control group, and there is no overlap in the confidence intervals. The same does not apply to a student at the median of the fourth quartile.

[4]As with the scatter plot, 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).

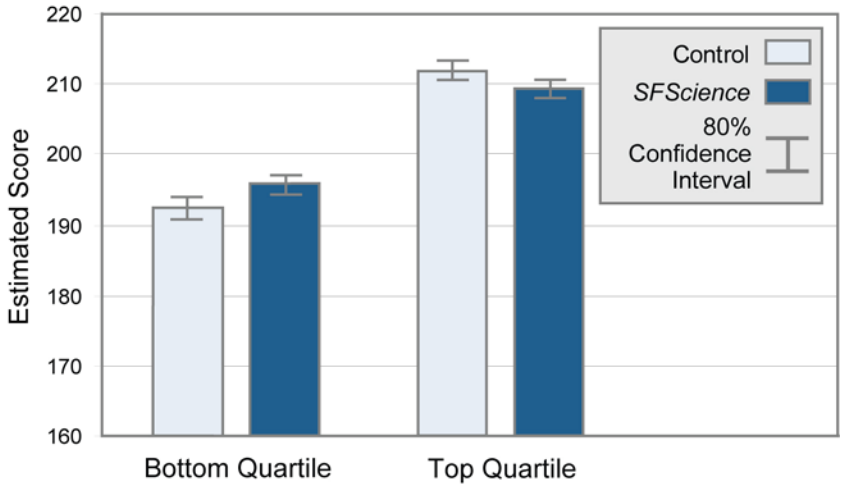


**Figure 4. Difference between *SFScience* and Control Group on NWEA Science Achievement: Median Students in Top and Bottom Quartiles**

### Analysis Including Gender as a Moderator

We were also interested in whether *SFScience* was differentially effective for boys and girls. Table 29 shows the moderating effect of gender on students' performance on NWEA Science. The advantage of being in the *SFScience* condition is greater for boys than it is for girls. The *p* value of .20 gives us limited confidence that the actual differential impact is different from zero.

**Table 29. Moderating Effect of Gender on Science Achievement**

| Fixed effects[a] | Estimate | Standard error | DF | *t* value | *p* value |
|---|---|---|---|---|---|
| **Average outcome for a girl in the control group** | 200.46 | 1.22 | 4 | 164.38 | <.01 |
| **Average *SFScience* effect for girls** | -0.76 | 1.39 | 4 | -0.54 | .62 |
| **Predicted change in outcome for each unit increase on the pretest[c]** | 0.65 | 0.05 | 260 | 13.74 | <.01 |
| **Difference (boys minus girls) in average performance in the control condition** | 0.99 | 0.95 | 260 | 1.04 | .30 |
| **Difference (boys minus girls) in the average *SFScience* effect** | 1.94 | 1.52 | 260 | 1.28 | 0.20 |
| **Random effects[b]** | **Estimate** | **Standard error** | | ***z* value** | ***p* value** |
| **Teacher mean achievement** | 0.48 | 2.04 | | 0.24 | .41 |
| **Within-teacher variation** | 36.5 | 3.19 | | 11.43 | <.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 predicted value for a control 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 girl or boy who had an average score on the pretest.

Figure 5 illustrates that boys have higher outcomes in the *SFScience* group as compared to the control group.

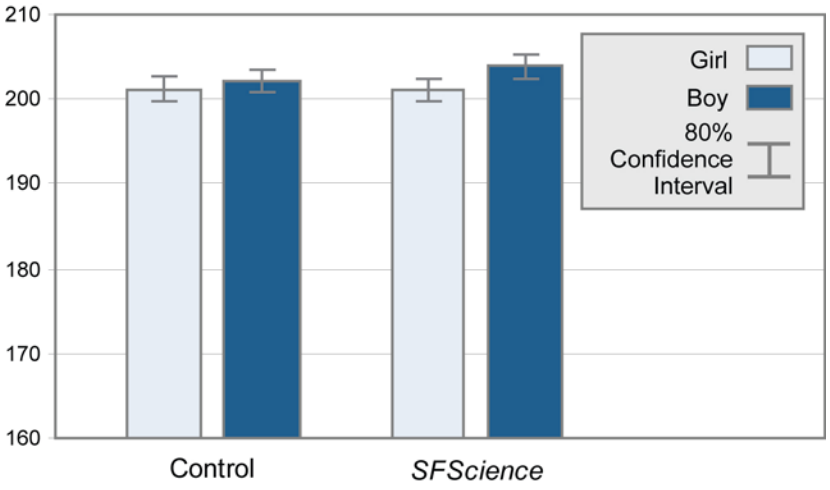


**Figure 5. Moderating Effect of Gender on Science Achievement**

### Analysis Including Ethnicity as a Moderator

We were also interested in whether *SFScience* was differentially effective for students from different ethnicities. Table 30 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 30. Science Achievement Moderated by Race**

| Fixed effects[a] | Estimate | Standard error | DF | *t* value | *p* value |
|---|---|---|---|---|---|
| **Average outcome for Black students in the control group** | 199.96 | 0.89 | 241 | 223.22 | <.01 |
| **Average *SFScience* effect for Black students** | 1.15 | 1.55 | 241 | 0.74 | .46 |
| **Predicted change in outcome for each unit increase on the pretest[b]** | 0.65 | 0.05 | 241 | 13.32 | <.01 |
| **Difference (White students minus black students) in average performance in the control condition** | 2.49 | 0.95 | 241 | 2.61 | .01 |
| **Difference (White students minus Black students) in the average *SFScience* effect** | -0.79 | 1.72 | 241 | -0.46 | .65 |
| **Random effects** | **Estimate** | **Standard error** | | ***z* value** | ***p* value** |
| **Residual teacher variance** | 30.19 | 2.75 | | 10.98 | <.01 |

[a] 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 predicted value for a control student with an average pretest applies to a particular school and assignment pair. Teachers were also modeled as a fixed factor.

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

## Reading Outcomes

### Analysis Including Pretest

Our next set of analyses addresses reading achievement as measured by NWEA Reading. Table 31 provides a summary of the sample we used in the analyses and the results for the comparison of *SFScience* and control. The "Unadjusted" row gives information about all the students in the original sample for whom we have a posttest. This shows the means and standard deviations as well as a count of the number of students, classes and teachers in that group. The last two columns provide the effect size, that is, the size of the difference between the means for *SFScience* and control in standard deviation units, and the *p* value, indicating the probability of arriving at a difference as large or larger than the absolute value of the one observed when there truly is no difference. The "Unadjusted" row is based on all students with a posttest and the estimated effect size takes into consideration the clustering of students in upper-level units (i.e., that students are grouped within teachers.) The "Adjusted" row is based on the students who have both pretests and posttests. This is the sample that we use in the analyses on which we base our results reported in Table 32. The means, and therefore the effect size, are adjusted to take into account student pretest scores, as well as fixed effects for schools and pairs within which we randomized. It also figures in the effect of students being grouped within teachers.

**Table 31. Overview of Sample and Impact of *SFScience* on Reading Achievement**

| | Condition | Means | Standard deviations[a] | No. of students | No. of classes | No. of teachers | Effect Size | *p* value[b] |
|---|---|---|---|---|---|---|---|---|
| **Un-adjusted** | ***SFScience*** | 203.07 | 15.24 | 90 | 7 | 5 | -.09 | .86 |
| | **Control** | 207.99 | 13.99 | 173 | 10 | 6 | | |
| **Adjusted** | ***SFScience*** | 209.44 | 14.93 | 75 | 6 | 4 | .06 | .69 |
| | **Control** | 209.10[c] | 13.86 | 163 | 10 | 6 | | |

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

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

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

Figure 6 provides a visual representation of specific information in Table 31. The bar graphs represent average performance in the metric of 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 high *p* value for the treatment effect (.69) 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.

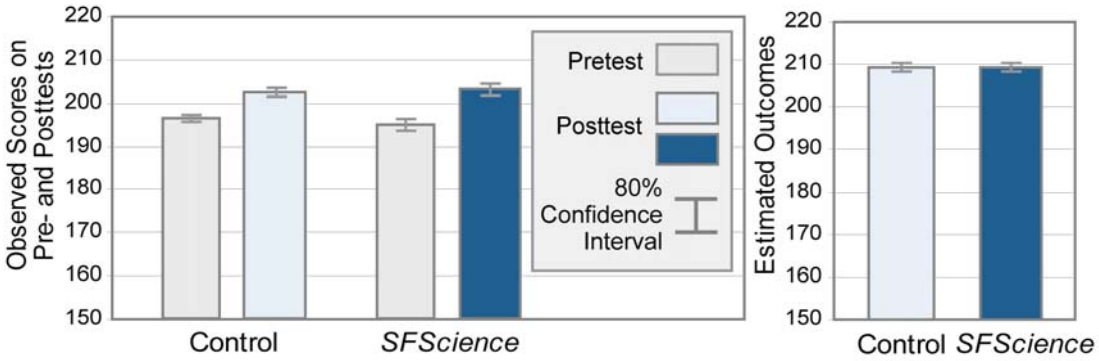


**Figure 6. Impact on Reading Achievement: Unadjusted Pre- and Posttest Means for Control and *SFScience* (Left); Adjusted Means for Control and *SFScience* (Right)**

### Analysis Including Pretest as a Moderator

We now report on the analyses that examine not just the overall impact of *SFScience* but also the moderating effects of other variables. We begin by examining the moderating effect of the prior score. Since the NWEA tests are on a continuous scale and the experiment involved three grades, we do not interpret low NWEA scores as indicating "low achieving" students within each grade. It is likely that third graders are more heavily represented in the lower range of the scores and fifth graders in the higher end of the scores. Table 32 shows the estimated impact of *SFScience* on students' performance in reading as measured by NWEA Reading, as well as the moderating effect of the prior score.

**Table 32. Mixed Model Estimating the Impact of *SFScience* on Reading Achievement**

| Fixed effects[a] | Estimate | Standard error | DF | *t* value | *p* value |
|---|---|---|---|---|---|
| **Predicted value for a control student with an average pretest** | 207.27 | 0.98 | 6 | 211.52 | <.01 |
| **Impact of *SFScience* for a student with an average pretest** | 1.69 | 1.42 | 6 | 1.19 | .28 |
| **Predicted change in control outcome for each unit increase on the pretest** | 0.77 | 0.04 | 223 | 20.35 | <.01 |
| **Interaction of pretest and *SFScience*** | -0.08 | 0.06 | 223 | -1.34 | .18 |
| **Random effects[b]** | **Estimate** | **Standard error** | | ***z* value** | ***p* value** |
| **Teacher mean achievement** | 1.39 | 2.19 | | 0.63 | .26 |
| **Within-teacher variation** | 42.99 | 4.07 | | 10.56 | <.01 |

[a] Schools are modeled as a fixed factor but not included in this table. Unlike adjusted effect size calculation, we were not able to model pairs of teachers used for random assignment as fixed effects.

[b] Teachers were modeled as a random factor.

The row in the table labeled "Impact of *SFScience* for a Student with an Average Pretest" tells us whether *SFScience* made a difference in NWEA Reading for a student who has an average score on the pretest. The estimate associated with *SFScience* is 1.69. This shows a positive effect of *SFScience*. However, the *p* value of .28 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 .18. We have limited confidence that the actual effect is different from zero.

As a visual representation of the results described in Table 32, we present a scatterplot in Figure 7, which shows student performance at the end of the year in reading, as measured by NWEA Reading, 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 point plots one student's post-intervention score against his or her pre-intervention score. The darker points represent *SFScience* students; the lighter points, control students. The shaded area in the lower right of the graph is the area of negative change (i.e., where students lost ground).

The two lines are the predicted 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 a weak interaction effect.

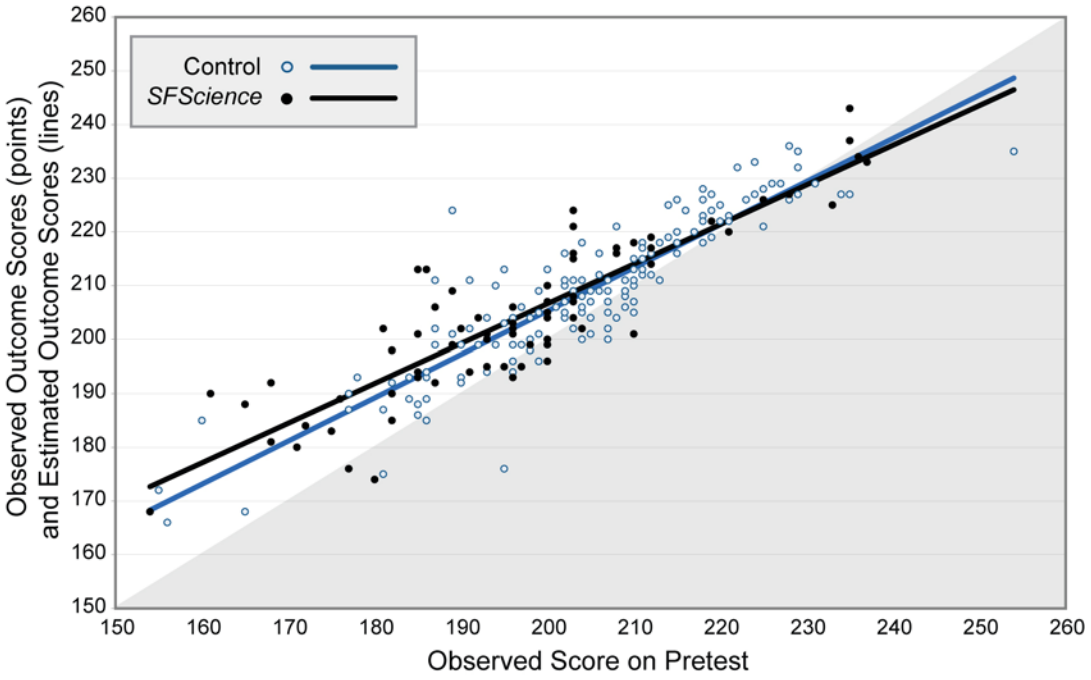


**Figure 7. Comparison of Predicted and Actual Outcomes for *SFScience* and Control Group Students**

Figure 8 illustrates the interaction in terms of the predicted difference between the *SFScience* and control groups for different points along the prior score scale[6]. This graph takes the same form as that presented in Figure 3. Consistent with the results in Table 32, there is evidence of a differential impact of the intervention across the prior score scale as measured by NWEA Reading. Considering the points representing the median student in the bottom and top quartiles, it appears that *SFScience* has more benefit for the students scoring at the lower range of the pretest. The median of the bottom quartile is sufficiently far from zero that we have some

[5] Displaying predicted values can be confusing when we model separate intercepts for upper-level units. The predicted 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)

[6] 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).

confidence that the actual difference in performance between *SFScience* and control groups is different from zero, with students in *SFScience* outperforming the controls. This advantage does not apply to students at the median of the top quartile.

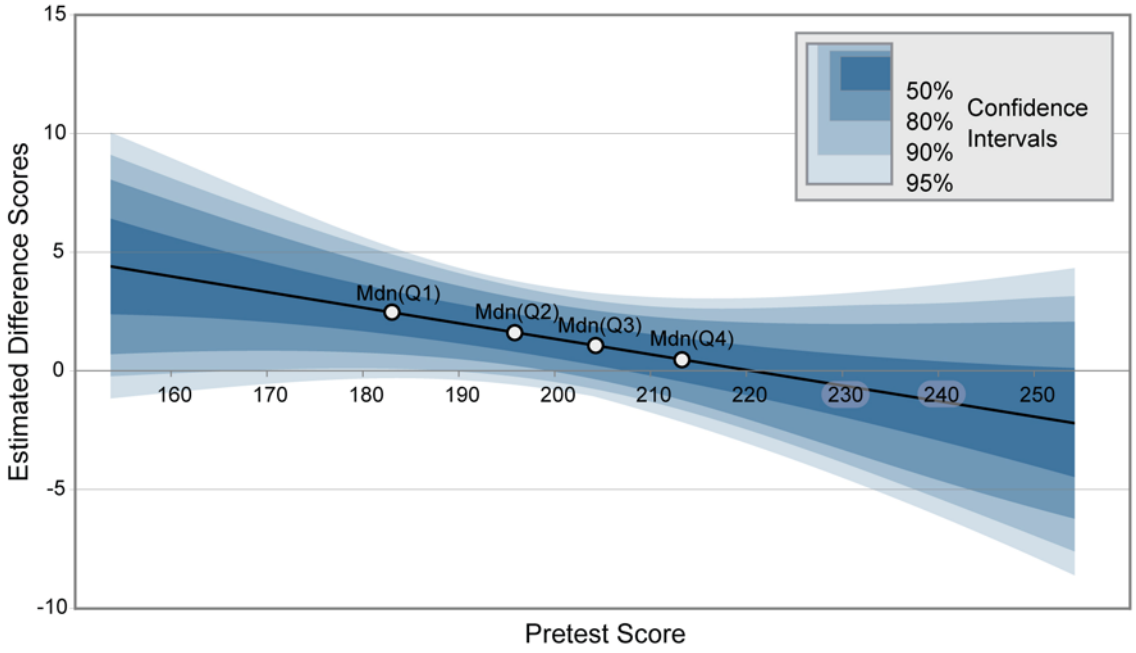


**Figure 8. Differences between *SFScience* and Control Group Reading Achievement: Median Pretest Scores for Four Quartiles Shown**

Figure 9 presents the same information represented in Figure 8 but this time in the form of a bar graph showing the predicted difference between *SFScience* and control conditions for students at the medians of the first and fourth quartiles of the pre-test measure[7]. The bar graph takes the same form and is interpreted in the same way as Figure 4. We see that for a student at the median of the first quartile there is a difference in the predicted outcomes in the two conditions, with the *SFScience* group scoring higher than the control group, and there is no amount of overlap in the confidence intervals. The same does not apply to a student at the median of the fourth quartile.

[7]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).

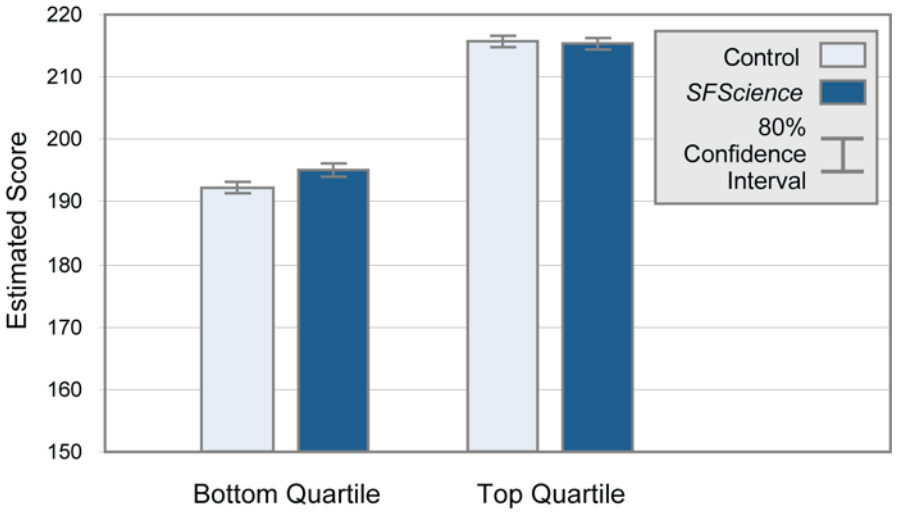


**Figure 9. Difference between *SFScience* and Control Group NWEA Reading Outcomes: Median Students in Top and Bottom Quartiles**

### 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 White and Black students. Table 33 shows estimates from the model that tests the moderating effect of ethnicity on students' performance on NWEA Reading. 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 White and Black students.

**Table 33. Reading Achievement Moderated by Race**

| Fixed effects[a] | Estimate | Standard error | DF | *t* value | *p* value |
|---|---|---|---|---|---|
| **Average outcome for a Black student in the control condition** | 205.43 | 1.13 | 3 | 182.31 | <.01 |
| **Average *SFScience* effect for a Black student** | -0.62 | 2.15 | 3 | -0.29 | .79 |
| **Predicted change in control outcome for each unit increase on the pretest[c]** | 0.73 | 0.04 | 198 | 19.98 | <.01 |
| **Difference (White students minus black students) in average performance in the control condition** | 2.06 | 1.1 | 198 | 1.88 | .06 |
| **Difference (White students minus Black students) in the average *SFScience* effect** | 0.54 | 2.29 | 198 | 0.24 | 0.81 |
| **Random effects[b]** | **Estimate** | **Standard error** | | ***z* value** | ***p* value** |
| **Teacher mean achievement** | 0.19 | 2.97 | | 0.06 | .47 |
| **Within-teacher variation** | 40.01 | 4.01 | | 9.98 | <.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 a girl or boy who had an average score on the pretest.

### 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 9. 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.

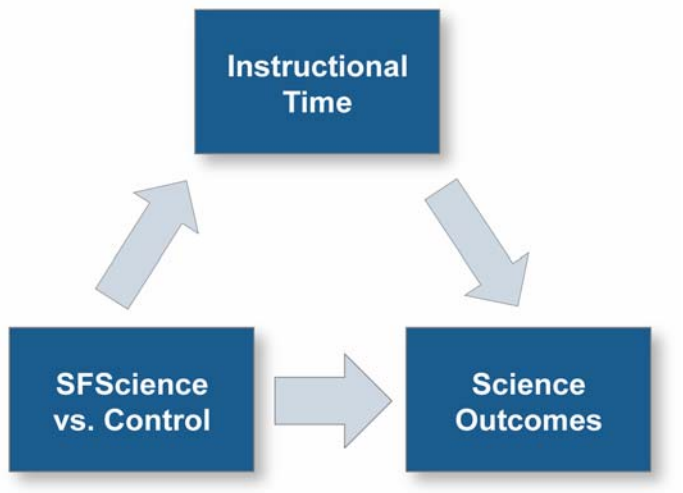


**Figure 10. Relationships for Exploratory Analysis of Implementation Variables**

#### 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 hours s/he spent using *SFScience* per year. Results were averaged across eight surveys that were administered every two weeks and adjusted for the number of weeks of implementation at that site.

We look first at the impact on instructional time. Table 34 shows that there was very little difference between the two groups of teachers on the amount of instructional time. The high *p* value gives us no confidence that the actual difference is different from zero.[8]

[8] We also confirmed this non-significant result using a non-parametric test (Mann-Whitney).

**Table 34. The Impact of *SFScience* on Hours of Science Instruction Time**

| Fixed effects[a] | Estimate | Standard error | DF | *t* value | *p* value |
|---|---|---|---|---|---|
| **Hours of science for a control teacher** | 21.88 | 6.65 | | 3.29 | 0.01 |
| **Impact of *SFScience* on hours of science instruction** | 1.93 | 10.52 | | 0.18 | 0.86 |

[a] The model did not include random effects.

Even with no difference between the two groups, it is useful to explore whether there is a relationship between amount of science 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, however, reveals no relationship between *SFScience* usage and the student outcome. Table 35 gives us a high *p* value and no confidence that the difference related to instructional time is different from zero.

**Table 35. Relationship of Instructional Time to Student Outcome**

| Fixed effects[a] | Estimate | Standard error | DF | *t* value | *p* value |
|---|---|---|---|---|---|
| **Predicted value for a student with an average pretest** | 198.45 | 3.514 | 243 | 56.46 | <.01 |
| **Predicted change in outcome for each unit increase on the pretest** | 0.67 | 0.05 | 243 | 14.40 | <.01 |
| **Predicted change in outcome for hour of science time** | 0.13 | 0.16 | 243 | 0.82 | .41 |
| **Random effects** | **Estimate** | **Standard error** | | ***z* value** | ***p* value** |
| **Residual variance** | 31.30 | 2.84 | | 11.02 | <.01 |

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

## Discussion

We began this research in Reynoldsburg City 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 both cases, we found that *SFScience* was more effective than the existing program for students initially scoring at the lower end of the pretest scales. Since the pretests we used (NWEA Science and NWEA Reading) were 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 earlier grades (within the third to fifth grade range of our experiment). We also note that fifth grade had considerably less time devoted to science. However, the fact that we found no relationship between amount of science instruction and outcome leaves open the possibility that for this district the program has greater value for the younger students.

We also looked at the relationship of *SFScience* to gender and ethnicity. For gender, we found that a tendency, for which we have limited confidence, that boys benefited more from *SFScience*. We found no differential benefit for Black or for White students.

Our experiment in Reynoldsburg was very small, involving only 11 teachers. With such small numbers we must be very cautious because of the increased possibility that chance differences in the make up of the *SFScience* and control may have influenced the result. We must also 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 start (teachers had already begun teaching using other materials) and the fact that science instruction was suspended for a time while the school focused on the subjects needed to make Adequate Yearly Progress under NCLB, which may have had an impact on the outcomes. An important factor in this district was the poor alignment of the curricular content to the state standards for each of the grade levels. This lack of alignment led teachers to spend more time planning and 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.