

Effectiveness of the 2005 Edition of Scott Foresman Science  
for Science and Reading Achievement:  
A Report of a Randomized Experiment in Two West Virginia  
School Districts

Denis Newman  
Empirical Education Inc.

# Contents

<b>INTRODUCTION</b> .....	<b>4</b>
<b>METHODS</b> .....	<b>4</b>
RESEARCH DESIGN .....	4
MATERIALS .....	4
SITE DESCRIPTION.....	5
SAMPLE AND RANDOMIZATION .....	5
DATA COLLECTION .....	5
<i>Test Scores</i> .....	6
<i>Surveys and Interviews</i> .....	6
STATISTICAL ANALYSIS .....	6
<b>RESULTS</b> .....	<b>6</b>
FORMATION OF THE EXPERIMENTAL GROUPS .....	7
Table 1: Counts of Teachers and Students .....	7
Table 2: Counts of Teachers and Students .....	8
Table 3: Distribution of Science Pretest Scores between Pilot and Control .....	9
Table 4: Distribution of Reading Pretest Scores between Pilot and Control .....	9
Table 5: Distribution of SES Categories between Pilot and Control .....	9
Table 6: Distribution of “Years Teaching” between Pilot and Control .....	10
ATTRITION .....	10
CLASSROOM USAGE PATTERNS.....	10
<i>Time for Science</i> .....	10
Table 7: t-test of difference between pilot and control in time spent on science instruction .....	10
<i>Extent of Hands On Science</i> .....	11
Table 8: Frequency of hands-on science for pilot and control groups .....	11
Table 9: t-test of difference between pilot and control in % of science time spent on hands on activities ....	11
<i>Engagement of Students</i> .....	11
Table 10: Comparison of pilot and control teachers in their assessment of how engaged/interested students were.....	11
<i>Approaches to Reading the Textbook</i> .....	12
Table 11: comparison of pilot and control teachers in their use of the textbook in class.....	12
COMPARISONS OF TEACHER EVALUATIONS OF THEIR PROGRAMS .....	12
<i>Attractiveness</i> .....	13
Table 12: T-test of the comparison of pilot and control teachers on their assessment of their program’s attractiveness.....	13
<i>Ease of Use</i> .....	13
Table 13: T-test of the comparison of pilot and control teachers on their assessment of their program’s ease of use .....	13
<i>Pedagogical value</i> .....	14
Table 14: T-test of the comparison of pilot and control of pilot and control teachers on their assessment of their program’s pedagogical value .....	14
<i>Value in Addressing Major Science Standards Categories</i> .....	14
Table 15: T-test of the comparison of pilot and control teachers on their assessment of their program’s value for addressing major standards categories.....	15
<i>Value in Addressing Specific Science Standards Categories</i> .....	15
Table 16: T-test of the comparison of pilot and control teachers on their assessment of their program’s value for addressing specific standards categories .....	15
ANALYSIS OF THE WESTEST SCIENCE AND READING SCORES .....	16
<i>Science</i> .....	16
Table 17: t-test of the difference between the means for science outcomes .....	16
Table 18: Results for science outcomes using the model that included treatment and pretest score .....	16
<i>Reading</i> .....	17
Table 19: t-test of the difference between the means for science outcomes .....	17
Table 20: Results for reading outcomes using the model that included treatment and pretest score .....	17
<b>DISCUSSION</b> .....	<b>18</b>

**ACKNOWLEDGEMENTS** ..... 18

# Introduction

We were asked to investigate the effectiveness of *Scott Foresman Science (SFS)*, an elementary program that includes many elements to support science inquiry and reading. SFS was a new program that had not been released at the time of the research. We were also interested in the responses of the teachers to SFS especially as it compared to the district's current science textbook. Our research took place in West Virginia in the winter and spring of 2005 and took the form of a randomized experiment or randomized controlled trial. We randomly assigned thirty-eight 3rd, 4th and 5th grade teachers to field test the new program or to continue working with their current program. In this way, we could compare the achievement in science and reading of students in classes using the program to equivalent classes using the old program. While the experiment did not detect any differences in student achievement, the contrast in teacher assessment of their respective programs shows the new program is viewed favorably academically and pedagogically.

The specific achievement-related question we addressed is whether *SFS* is more effective, as measured by the West Virginia State test, the WESTEST, than the science program the district already had in place. Teachers used the materials beginning in January 2005. The experiment ended with the WESTEST, the state test that included tests of science and reading. The research schedule allowed for only four months of usage and for some important components only a month. During this short period of time, we also conducted more detailed observations and surveys that are reported separately. The research had formative elements both in the sense of observing and interviewing teachers about the use and usability of the program for possible improvements and in the sense of a pilot in preparation for a larger set of experiments now being conducted in the 2005-2006 school year.

This research is part of a larger program of independent research that Pearson Education is supporting with the goal of producing scientific evidence of the effectiveness of their products. Our research design reflects the requirements of the No Child Left Behind Act, which directs schools to consult reports of rigorous research to guide their adoptions of instructional programs. The US Department of Education has been explicit in interpreting this requirement in terms of randomized controlled trials when it comes to determining effectiveness. The comparison of groups composed through random assignment, when combined with a good understanding of the implementation, is the most effective way to measure the difference that a new program is likely to make.

## Methods

### ***Research Design***

The research is a comparison of achievement outcomes for classes taught using the *SFS* program and classes taught using a science textbook adopted several years earlier. Researchers assigned equal numbers of teacher volunteers to the pilot and control groups. The outcome measures are student level test scores. The design includes two levels: the unit of random assignment is the teacher and the unit for the outcome data is the individual student. Within a multi-level model, analysis of covariance (ANCOVA) can be used to control for the influence of factors such as incoming science achievement.

### ***Materials***

*SFS* is an innovative program for science learning because it goes beyond the standard textbook to include hands-on activity kits for each unit. The kits are designed to support inquiry-based learning through structured experiments that proceed through three levels: directed inquiry, guided inquiry, and full inquiry. Students proceed through the first two levels of inquiry in each

chapter culminating in a “full inquiry” at the completion of every unit. Unit topics (i.e. Life Science, Earth Science) are consistent across grade levels and only the content changes. Additionally, *SFS* includes a library of leveled readers which support the text by providing three distinct levels of reading ability. The below grade level reader assures that struggling readers are exposed to the chapter content while the above grade level reader provides extensions of the content into related areas. .

*SFS* includes an extensive teacher’s guide that provides science background, lesson plans, detailed instructions for leading the inquiry activities, and links to other resources such as web and content and assistance.

Teachers received an hour long workshop led by a Scott Foresman staff member that reviewed the structure and components of the program. The workshop leader explained that the program is more extensive than can be accommodated in the normal school day but emphasized a focus on the inquiry activities and the reading activities since these were the components that the company wanted to have tested. Beyond the initial training with its emphasis on inquiry and reading, teachers were free to make use of the materials as best suited the needs of their classroom and students.

### ***Site Description***

We conducted the research in two school districts. Putnam County lies west of Charleston. Wood County is north of Charleston on the Ohio River border with Ohio. The demographics of the two districts were very similar. The Putnam school district has approximately 8,800 students, 97.6% are White with virtually no English language learners. Approximately 30% are classified as economically disadvantaged, compared to 50% for the State. The Wood school district is similar. Their 13,700 students are 97% White with approximately 40% economically disadvantaged.

Both districts had adopted Harcourt’s science textbook in the previous round of adoptions. This provided a consistent control condition for the experiment.

### ***Sample and Randomization***

The sites were initially identified by a Scott Foresman staff member as districts interested in the product and willing to conduct a structured pilot with a subset of their classes. The district science leaders sent out a general invitation to their elementary teachers. Meetings with teachers in the two districts were held separately. In Wood County 17 teachers attended an after school meeting in October. We assigned teachers at the meeting to pilot the *SFS* program or serve as control by first pairing them according to similarity on important characteristics and then tossing a coin to determine the assignment. Such pairing is a form of blocking that helps to assure the initial similarity of the two groups and tends to improve the statistical analysis. The basis for the pairing was, first the grade taught and second whether the school was designated as Title 1. For the Wood pilot group, we carried out an initial formative investigation during December 2004 in which they reviewed and provided feedback on the teacher guide. In Putnam County the meeting in November was attended by 24 teachers. We used the same procedure in randomly assigning them to pilot and control. For the purposes of analysis, we treated the teachers from Putnam and Wood as a single group. The experiment, itself, was scheduled to start January 2005. Training for the pilot teachers took place separately for each district on the 19th and 20th of January

### ***Data Collection***

The results are based on three sources of data. The outcome measure was the State science and reading tests. In addition to the previous year’s test score, the data on each student also included gender (note that ethnicity and English language proficiency were not a factor given the West Virginia population). Both quantitative and qualitative data were collected from the teachers through web-based surveys and telephone interviews. As part of a formative study, reported

separately, we put additional focus on the Wood pilot teachers through classroom observations and more extensive surveys to obtain feedback.

## **Test Scores**

The outcome measure was the student scores on the Science and the Reading subtests of the WESTEST. The test was developed with CTB/McGraw-Hill and is tied closely to the State standards. In each of the areas tested, the outcome is reported on a common scale so that growth from year to year can be measured and scores of students in different grades can be combined. Using a criterion referenced test tied to the specific state standards helps to assure that the test is aligned to what the teachers are required to do. Because testing starts in 3rd grade, we had pretest scores on all the 4th and 5th grade students. The State provides the test results in the form of paper documents that the district or the researchers hand entered. The State also provides individual item level "Right Response Records" which we will use in later analyses to provide a better focus on the specific content areas taught in each classroom.

## **Surveys and Interviews**

All teachers were surveyed 3 times during the experiment. An initial survey established the units to be taught and confirmed class assignments and other demographics. The survey at the midpoint of the experiment sampled the time spent on science, amount of inquiry activity, the engagement of the students and similar factors that may impact the results. These surveys also established for each teacher the specific content taught in that time period. The final survey provided detailed assessments of the science program (both pilot and control). The telephone interviews served mainly to verify the information received in the surveys and provided a barometer of the teacher morale, level of cooperation and a gauge of product acceptance.

## **Statistical Analysis**

We have two main outcome measures—the science and reading scores. We also have measures of implementation in the classrooms and measures of teacher appraisals of their respective science programs. For the analysis of the test scores, we develop a statistical model that takes into account the students scores from the previous year. An analysis of covariance (ANCOVA) allows us to look at the pretest variables (covariates) and the treatment variable simultaneously and to identify how the variables individually and in combination impact the outcome. We conducted 2 analyses, the first with the 4th and 5th grade students including the pretest covariate; the second with all the students omitting the covariate and depending on the random assignment for the assumption of equivalence. We use SAS PROC MIXED (from SAS Institute Inc.) as the primary tool for this work. This software is particularly appropriate in research on schools where the outcomes are measures of individual students but a whole class implementation is provided by the teacher. PROC MIXED allows us to account for the clustering of students in classes (the intraclass correlation) and provides a more accurate, and often more conservative, assessment of the confidence we should have in the findings.

For the teacher appraisals of the programs, we conduct a simpler analysis comparing, for each category of assessment a comparison of the mean values for the pilot and control classrooms. For the implementation variables, we examine them to identify areas where differences, e.g., in the time spent are large enough to have had an impact. In such cases, we consider using that teacher-level variable in the ANCOVA predicting the outcome measures. The analysis reported here is incomplete because the item level data is not yet available. When available, a refinement of the analysis of the science and reading scores will be possible.

## **Results**

The results of this experiment address differences between the randomly assigned pilot and control groups. These are both teacher or classroom level differences in usage and teacher

appraisal of their program and at the student level where we address differences in test score outcomes.

In most cases, the means for the pilot and control teachers or their students are compared using a t-test. Where a difference is substantial enough to have a low probability of having occurred by chance (i.e., a difference of that magnitude or greater is unlikely to occur if there was actually no difference) we also provide a standardized effect size estimate expressed as a multiple of a standard deviation.

### ***Formation of the Experimental Groups***

The randomizing process ensures that our estimates from the experiment are unbiased, but does not guarantee that the groups will be perfectly matched on all characteristics. Before conducting analysis, it is important to inspect the two groups to see if there are any significant differences between the groups that would impact that results and have to be controlled statistically. The following tables address the nature of the groups.

Tables 1 and 2 shows the distribution of teachers and students for the pilot and control groups and the distribution of classrooms in among grades and schools. The classrooms are well distributed between control and pilot groups with 9 control classrooms and 10 pilot classrooms.

**Table 1: Counts of Teachers and Students**

<b>SF Science Group</b>				
<b>Grade</b>	<b>Schools (ID#)</b>	<b>Teacher (ID#)</b>	<b>Students</b>	
3	550	633	23	
	551	634	19	
	556	661	21	
	557	646	22	
	558		656	20
			658	18
	560	654*	7	
	562		648	19
			652	24
	563	668	22	
	565	650	19	
	N for 3rd grade		11*	214
	4	548	637	22
554		639	16	
560		654*	11	
563		667	21	
566		662	22	
N for 4th grade		5*	92	
5	549	641	23	

	553	642	25
	554	645	17
N for 5th grade		3	65
Total N		18*	371

\* Note: teacher 654 had a mixed 3-4 class

**Table 2: Counts of Teachers and Students**

grade	schoolid	teacherid	N Obs
3	549	632	18
	552	635	21
	556	653	20
	557	669	20
	559	660	24
	560	655	15
		659	16
	561	649	22
		657	21
	562	647	24
		651	23
N for 3rd grade		11	224
4	553	636	20
	554	640	17
	558	666	18
	560	663	19
	562	664	22
N for 4th grade		5	96
5	553	643	18
	554	644	21
N for 5th grade		2	39
Total N		18	359

We can also compare the control and pilot students on variables that may be relevant to the analysis as shown in the following tables. Tables 2 and 3 compare the groups on the pretest scores in science and reading. This comparison addresses only the 4th and 5th grade students who had taken the WESTEST the previous year.

**Table 3: Distribution of Science Pretest Scores between Pilot and Control**

Group	n	Science Pretest Score		
		Mean	SD	SE
Pilot	267	640.640	29.729	1.819
Control	231	640.394	27.148	1.786
<b>Difference between means</b>		0.247		
<b>t value</b>		0.10		
<b>p value</b>		0.924		

**Table 4: Distribution of Reading Pretest Scores between Pilot and Control**

Group	n	Science Pretest Score		
		Mean	SD	SE
Pilot	268	641.933	32.509	1.986
Control	233	641.219	33.015	2.163
<b>Difference between means</b>		0.714		
<b>t value</b>		0.24		
<b>p value</b>		0.808		

On both measures the two groups are remarkably well matched. Socioeconomic status as indicated by participation in the free or reduced lunch program is another source of variability reported in Table 5.

**Table 5: Distribution of SES Categories between Pilot and Control**

Group	In free or reduced lunch programs		
	No	Yes	Total
Pilot	193	136	329
Control	207	110	317
Total	400	246	646
<b>X<sup>2</sup> statistic</b>		2.74	
<b>p value</b>		0.0978	

In this case, the pilot group had somewhat more students in the program. Teacher differences are also important to consider. Years teaching is an area of considerable variation in the teacher sample. Table 6 shows that the pilot teachers were a somewhat more experienced group.

**Table 6: Distribution of “Years Teaching” between Pilot and Control**

Group	Years teaching			
	n	Mean	SD	SE
Pilot	18	21.526	9.192	2.109
Control	18	15.211	10.042	2.304
<b>Difference between means</b>		6.316		
<b>t value</b>		2.02		
<b>p value</b>		0.051		

### ***Attrition***

Of the 38 teachers who were assigned to pilot or control in the initial meetings, two dropped out prior to the training, one from pilot and one from control. Both were veteran teachers with 30 plus years of teaching. One unexpectedly retired. The other dropped out for unknown reasons. The student population was stable with an attrition rate of only 4.7%.

### ***Classroom Usage Patterns***

An impact of a new curriculum program may be to change the amount of time spent on various activities, the kinds of activities undertaken and the attitudes or engagement of the students. The surveys probed these issues in order to determine if implementation of science was different between the two groups.

### **Time for Science**

An important question is whether the adoption of a new program results in more time being spent on the subject area. Table 7 reports the teachers’ report of approximate number of minutes (number of sessions per week times the length of the typical period) spent on science instruction.

**Table 7: t-test of difference between pilot and control in time spent on science instruction**

Group	Weekly Minutes of Science			
	n	Mean	SD	SE
Pilot	16	174.500	60.658	15.1644
Control	14	153.214	59.569	15.9205
<b>Difference between means</b>		21.286		
<b>t value</b>		-0.97		
<b>p value</b>		0.3419		

Pilot teachers reported spending somewhat more time but the size of the difference does not reach an adequate level of confidence. Amount of time spent is not likely to be an important factor in differential achievement outcomes.

## Extent of Hands On Science

Since the SF program provides a lot of materials appropriate for hands on and inquiry based teaching. One measure of the extent of use of this approach is the frequency of science lessons that using these materials. Table 8 shows that both groups were similar on this dimension.

**Table 8: Frequency of hands-on science for pilot and control groups**

Group	Frequency				Total
	Almost every session	Once every 3 to 5 sessions	Once every 6-8 sessions	Once every 9 to 12 sessions	
Pilot	1	10	4	1	16
Control	2	11	1	0	14
<b>Total</b>	<b>3</b>	<b>21</b>	<b>5</b>	<b>1</b>	<b>30</b>
<b>X<sup>2</sup> statistic</b>	3.06				
<b>p value</b>	0.3823				

Another measure of hands on science is the percent of class time devoted to it. Table 9 shows that the control teachers reported spending more of their science lesson time on this mode of instruction. In this case, the effect size estimate suggests that this is a substantial difference.

**Table 9: t-test of difference between pilot and control in % of science time spent on hands on activities**

Group	% of Science Class Time			
	n	Mean	SD	SE
Pilot	16	20.938	10.680	2.6700
Control	13	31.538	11.252	3.1207
<b>Difference between means</b>		-10.601		
<b>t value</b>		-2.60		
<b>p value</b>		0.0151		
<b>Effect size</b>		-0.96639		

## Engagement of Students

Survey questions asked teachers to rate their student's level of engagement in the science instruction. Table 10 shows that pilot teachers rated their students as slightly more engaged.

**Table 10: Comparison of pilot and control teachers in their assessment of how engaged/interested students were**

Level of Student Engagement and Interest				
Group	n	Mean	SD	SE
Pilot	18	4.000	0.542	0.1278
Control	17	3.676	0.706	0.1712
Difference between means		0.324		
t value		1.53		
p value		0.1366		

### Approaches to Reading the Textbook

With the emphasis on reading in the SF Science program, we were interested in whether it had an impact on the way teachers used the text. Table 11 illustrates that both groups made use of their textbooks in very similar ways.

**Table 11: comparison of pilot and control teachers in their use of the textbook in class**

	Group	Occasionally or never	Frequently or always
Teacher read aloud in class, as students read along	Pilot	14	3
	Control	11	7
Students read aloud in class	Pilot	3	15
	Control	4	13
Students read silently in class	Pilot	16	2
	Control	15	2
Students read text as homework	Pilot	16	2
	Control	15	2

### Comparisons of Teacher Evaluations of Their Programs

A survey at the end of the trial period asked teachers to rate the science programs they had been using on a variety of dimensions. In each case, the rating was on a 5 or on a 10 point scale. The results are summarized here in five clusters of results. Each cluster uses the average score for each teacher across the dimensions in the cluster. The specific dimensions are listed in each case below.

## Attractiveness

Three questions addressed very general appraisals of attractiveness. These were:

- Educational value of overall curriculum
- Educational approach
- Visual appeal

The ratings are reported in Table 12 which shows that the difference was very small and there is a good probability that there was no difference between the groups.

**Table 12: T-test of the comparison of pilot and control teachers on their assessment of their program's attractiveness**

Group	n	Rating of Attractiveness		
		Mean	SD	SE
Pilot	18	3.389	0.873	0.20567
Control	18	3.278	0.8023	0.18912
<b>Difference between means</b>		0.111		
<b>t value</b>		0.40		
<b>p value</b>		0.6934		

## Ease of Use

Ease of use for classroom implementation was addressed in two dimensions.

- Ease of integration
- Classroom management

The ratings are reported in Table 13 which shows that the difference was very small and there is a good probability that there was no difference between the groups.

**Table 13: T-test of the comparison of pilot and control teachers on their assessment of their program's ease of use**

Group	n	Ease of Use		
		Mean	SD	SE
Pilot	18	3.639	0.782	0.184
Control	18	3.472	0.866	0.204
<b>Difference between means</b>		0.167		
<b>t value</b>		0.61		
<b>p value</b>		0.5485		

## Pedagogical value

An important cluster of ratings concerned various aspects of pedagogical value. The ratings for each teacher were averaged across the following dimensions.

- Breadth of information in student text
- Depth of information in student text
- Teacher-led science demonstrations
- Students' hands-on investigations
- Students collecting and analyzing data
- Students predicting, discussing & interpreting results from investigations
- Use of hands-on models of scientific concepts
- Use of authentic scientific phenomena, objects & situations
- Connections between science & the real world
- Connections between science & other curriculum areas

The ratings are reported in Table 14, which shows that the difference in this case was substantial. The ratings by pilot teachers of all of the above dimensions, individually, were more favorable. An effect size of 1.137 is considered very large. On these dimensions, the SF Science program was well ahead of the existing program.

**Table 14: T-test of the comparison of pilot and control of pilot and control teachers on their assessment of their program's pedagogical value**

Group	Pedagogical Value			
	n	Mean	SD	SE
Pilot	18	18	7.765	1.059
Control	18	18	6.053	1.848
<b>Difference between means</b>		1.712		
<b>t value</b>		3.41		
<b>p value</b>		0.0017		
<b>Effect size</b>		1.137		

## Value in Addressing Major Science Standards Categories

The survey asked how well the teacher's science program addressed West Virginia's six major science standards categories.

- History and nature of science
- Science as inquiry
- Unifying themes
- Science subject matter/concepts
- Scientific design and application

- Science in personal and social perspectives

As with the questions of pedagogical value, teachers using the SF Science program rated it much higher than the teachers using the district's existing program. The results are shown in Table 15.

**Table 15: T-test of the comparison of pilot and control teachers on their assessment of their program's value for addressing major standards categories**

Group	n	Value for Addressing Major Standards		
		Mean	SD	SE
Pilot	18	8.102	1.610	0.3795
Control	18	6.402	1.527	0.3598
<b>Difference between means</b>		6.316		
<b>t value</b>		3.25		
<b>p value</b>		0.0026		
<b>Effect size</b>		1.08		

### Value in Addressing Specific Science Standards Categories

The survey asked: how well does your science curriculum enable students to learn the following:

- examine the careers & contributions of men and women of diverse cultures
- Recognize scientific knowledge is subject to modification...
- demonstrate curiosity, initiative & creativity by designing & conducting experiments & investigations
- Interpret data presented in a table etc to answer questions & make predictions
- Use scientific instruments and everyday materials to investigate...
- apply mathematical skills and use metric units in measurement
- understand that systems are made of parts that interact with one another
- group or order a set of objects according to an established scheme
- knowledge & understanding of scientific facts, concepts, etc.
- apply knowledge, understanding & skills...to daily life experiences
- understand the interdependence of science & technology
- evaluate personal & societal benefits when examining health, etc.

As shown in Table 16, the teachers using the SF program consistently gave it substantially higher ratings. Again, the standardized effect size of over one standard deviation is considered very large.

**Table 16: T-test of the comparison of pilot and control teachers on their assessment of their program's value for addressing specific standards categories**

Years teaching	
----------------	--

Group	n	Mean	SD	SE
Pilot	18	7.945	1.157	0.2727
Control	18	6.297	1.911	0.4505
<b>Difference between means</b>		1.648		
<b>t value</b>		3.13		
<b>p value</b>		0.0036		
<b>Effect size</b>		1.04		

## ***Analysis of the WESTEST Science and Reading Scores***

A major reason for undertaking a randomized experiment is to determine if the introduction of the new science program results in higher achievement levels. We were interested in the science outcomes as well as the reading outcomes since readability and differentiated reading is an important feature of the program. In each case, we look first at the mean outcomes for all the students using a simple comparison of means. As a check on these results we also develop a statistical model using ANCOVA for the 4th and 5th graders for whom we have the pretest measures from the prior year.

### **Science**

A comparison of the means is reported in Table 17. The scores for pilot and control students are very close and the small difference does not approach statistical significance.

**Table 17: t-test of the difference between the means for science outcomes**

Group	n	Science Achievement		
		Mean	SD	SE
Pilot	366	650.008	57.172	2.9884
Control	349	651.871	30.799	1.6486
<b>Difference between means</b>		1.863		
<b>t value</b>		0.54		
<b>p value</b>		0.5903		

We also examined these results with a statistical model that controls for the pretest score and the fact of the clustering of students within classrooms. The p value for the treatment is also very large indicating that the tiny difference cannot be distinguished from zero in this case.

**Table 18: Results for science outcomes using the model that included treatment and pretest score**

Solution for Fixed Effects						
Effect	Standard Estimate	Error	DF	t value	p value	
Intercept	656.59	1.9396	13	338.51	<.0001	
Pretest score (centered at the mean)	0.8372	0.04072	252	20.56	<.0001	
Treatment (pilot = 1; control = 0)	-0.2058	2.6213	252	-0.08	0.9375	

Solution for Random Effects					
Cov Parm	Subject	Estimate	Error	Z value	p value
UN(1,1)	Class (School)	2.6685	10.7973	0.25	0.4024
Residual		407.30	36.3600	11.20	<.0001

## Reading

The results for reading are the same for the analyses reported in Tables 19 and 20.

**Table 19: t-test of the difference between the means for science outcomes**

Group	Reading Achievement			
	n	Mean	SD	SE
Pilot	368	649.995	45.580	2.3760
Control	349	650.825	33.582	1.7976
Difference between means		0.831		
t value		0.28		
p value		0.7822		

**Table 20: Results for reading outcomes using the model that included treatment and pretest score**

Solution for Fixed Effects						
Effect	Standard Estimate	Error	DF	t value	p value	
Intercept	657.67	2.6366	13	249.44	<.0001	
Pretest score (centered at the mean)	0.7642	0.03157	254	24.21	<.0001	
Treatment (pilot = 1; control = 0)	0.5295	3.5953	254	0.15	0.8830	

  

Solution for Random Effects					
Cov Parm	Subject	Estimate	Error	Z value	p value
UN(1,1)	Class (School)	34.6855	19.4614	1.78	0.0374
Residual		235.65	20.9358	11.26	<.0001

The experiment did not detect any impact of the new program on student achievement.

## Discussion

Teachers in our experiment voiced a very strong endorsement of the Scott Foresman Science program with respect to its alignment with the science standards that West Virginia asks them to address. Using a randomly assigned control group, we were able to measure the difference between these appraisals and the control teachers' appraisals of the district's currently adopted program. In terms of impact of the new program on classroom activities or on the science achievement of students, however, we detected little or no difference between the two groups.

This was a short trial and not all the materials were available to the teachers for what was effectively four months of classroom usage. In addition, the analysis of the science results was not able to take advantage of the item level data on the science tests because the districts provided it after the completion of this round of analysis. Additional work can use this data to focus specifically on items that addressed the content areas actually taught during those four months and to decompose the science score into strands related to higher level thinking versus factual knowledge. When these data are available, we can also examine potential impact of SES and teacher variables such as experience, which were not addressed in this analysis.

Differences in classroom activities that result from new programs may be an important mediating variable. The fact that little change was detected may be a result of the short duration of the experiment or the limited survey methods. For example, only the pilot classrooms were observed so that those results (reported separately) are not available for experimental comparison. Improving the precision of these measures of classroom activity will be important for future research since any such impact can be included in the explanatory model. Ultimately effectiveness is measured by achievement outcomes rather than changes in teacher behavior or student engagement. Additional research is needed to demonstrate such an impact.

## Acknowledgements

We are grateful to the people in the Putnam and Wood County School Districts for their cooperation and assistance in conducting this research.