Project LISTEN: Evaluation of an Automated Reading Tutor in Pittsburgh Public Schools

April Galyardt, Nathaniel Anozie, and Elise Olson

March 9, 2007

1 Introduction

An individual's ability to read is fundamental to his or her success in life. Unfortunately, as reported by the 2003 National Assessment of Adult Literacy, an estimated 30 million - 14 percent of the the total US population - are below a basic literacy level.[1] "Literacy Innovation that Speech Technology ENables" (Project LISTEN) is an inter-disciplinary research project at Carnegie Mellon University attempting to addressing this problem by providing guided reading practice for students through an automated tutor.

When using the tutor, children wear a headset and read the stories aloud as the tutor listens. The Reading Tutor then automatically selects stories at the student's reading level and allows the child to choose from these stories. "The Reading Tutor intervenes when the reader makes mistakes, gets stuck, clicks for help, or is likely to encounter difficulty."[2] Project LISTEN has been tested in several cities in the U.S. and Canada and has recently been tested in Pittsburgh.

The primary objective of this study is to determine the effectiveness of the tutor as compared with other reading programs in the Pittsburgh Public School System. To do this we ask whether the Reading Tutor improves the fluency, reading, and spelling test scores of a sample of Pittsburgh Public School students. We also ask which other factors contribute to improvement.

In Section 2 we will discuss the data: how it was obtained and potential problems. Section 3 will present several forms of analyses of increasing complexity and discuss our plans for further analysis. In Section 4 we will discuss our results so far and the difficulties of the analysis that remains, as well as implications for the project as a whole.

2 Data

Sixty-six students in Pittbsurgh, PA, USA from two public schools participated in the study during the 2005-2006 school year: 38 second-graders from White Oak Elementary, 15 third-graders from Fort Pitt Elementary, and 12 fourth-grader also from Fort Pitt Elementary.

Students were divided into 2 treatment groups. Group 1 used the tutor in the Fall and had alternative reading instruction during the Spring; Group 2 had the same treatment with the order reversed. (See Table 1.) Students using the tutor spent 30 minutes reading with the tutor daily. However, the two schools used different programs for their alternative reading instruction. The second graders at White Oak used the Read Naturally program, while the third and fourth graders

[1] Proliteracy Worldwide. "The State of Adult Literacy 2006," webpage: http://www.proliteracy.org/downloads/stateoflit06pdf.pdf 2006. pg 10

[2] Project LISTEN webpage: http://www.cs.cmu.edu/ listen/

at Fort Pitt participated in a slightly different program consisting of a mix of individual reading, group reading, and journaling.

Group 1	Group 2
September: Pre-Test	
Reading Tutor	Control
January: Mid-Test	
Control	Reading Tutor
May: Post-Test	

Table 1: Study Design Diagram. Students assigned to the Reading Tutor will use the automated tutor for 30 minutes each day during school. Students in the Control treatment will particpate in alternate instruction.

Literacy gains were measured by tests in fluency, reading, and spelling before the experiment, at the cross-over point, and at the conclusion of the experiment (Table 1). Fluency scores are counts of the number of words a student correctly read aloud in one minute from a passage corresponding to his or her grade level. The reading score is a composite of four subtests of the Woodcock Reading Mastery Test: word identification, which tests how well students read words; word attack which tests how well students read non-words; word comprehension, a measure of vocabulary; and passage comprehension. These composite scores are then placed on a grade equivlancy scale so that the end score is an indication of the grade level at which the student can read. For example, if a student has a score of 3 on the reading scale, this means he or she can read words that an average third grader can read. Spelling scores come from a test of written spelling and are likewise scaled to reflect grade equivalency. In addition to grade, school, and test scores, we have further information on students which may be helpful in predicting improvement including: the date of test, the administer of the test, homeroom teacher, gender of the student, and ethnicity of the student.

There are a few problems of note with the data. First, there are many students for whom we are missing data, including a large number of students for whom we lack only indication of when they did or did not use the tutor. We also know that several students left the study before its completion for various reasons leaving complete data for only 65 students. Furthermore, the students at White Oak Elementary were grouped according to the time that best fit their class schedules and not by random assignment. This has lead some of the investigators to suspect that those in Group 2 had lower ability levels than those in Group 1. All of this brings into question the comparability of the 2 groups.

Second, many of the covariates are highly correlated. All of the third and fourth graders are African-Americans attending Fort Pitt, while all of the second graders attend White Oak and the majority of them are White. As a result, school, ethnicity, and grade give very similar information, making it difficult to determine which is truly the best predictor. We will further address these and other related issues in the analysis and discussion sections.

3 Analysis

3.1 Simple Analysis: Is there a difference in Score Improvements Between the Two Groups?

In analyzing the data we first take a naive approach by testing whether there is a difference in mean gain in scores for the two groups each semester. We will analyze the two semesters separately due to a clear seasonal difference in performance gains. See Figure 1. This difference will be further addressed in the Discussion Section.

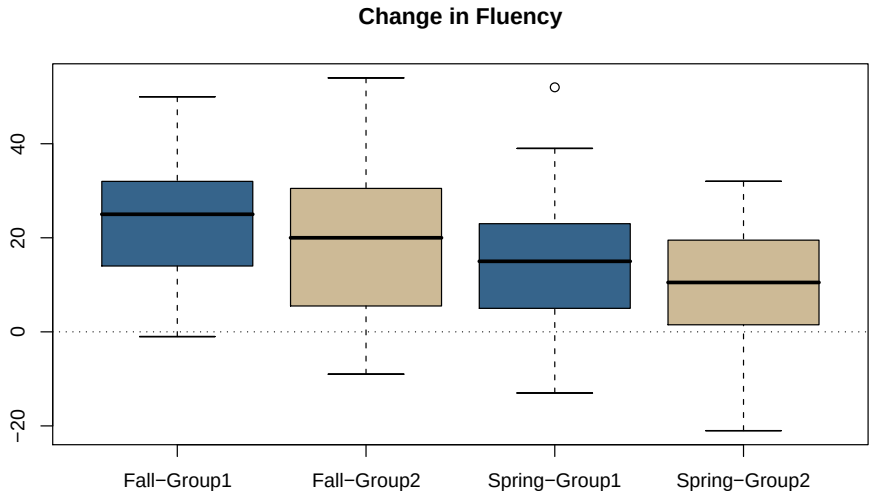

Figure 1: Change in fluency is measured by comparing fluency scores at the beginning and end of each period. The Fall change is the difference between the pre-test and the mid-test. Likewise, the Spring change is the difference between the mid-test and the post-test. Group 1 indicates the Fall-treatment group, while group 2 used the Reading tutor in the spring. You will notice that the students in both groups had much smaller improvements in the Spring. The difference between Fall and Spring is even more pronounced in Reading Comprehension and Spelling (Figure 5 in the Appendix).

However, we must first test whether the groups are comparable since as noted in the Data Section, the lack of random assignment does not assure us of this. If the distribution of pre-test scores is roughly equal between the two groups, we will assume the groups are sufficiently comparable. To test this, we run a series of F-tests for equal variances and t-tests for equal means between each group, breaking it up by grade. In the end, of the nine tests run for the three grades and three tests, only the reading pre-test scores in grade 2 were significantly different, with a p-value of .0437 for the test of equal means, after the Bonferroni correction for mulitiple testing. However, the lowest p-value obtained from all of the other tests was .1648, before the Bonferroni correction. (See Table 3 in the Appendix for the other p-values.) So, with the notable exception of second-grade reading comprehension, we cannot conclude that the groups are different in terms of prior knowledge.

Since the two groups are reasonably similar, we can conduct t-tests to determine whether there are differences in the mean change in scores for fluency, reading, and spelling between the two groups. The score gains for the Fall are the difference between the mid- and pre-test scores, while the Spring score gains are the difference between post- and mid-test scores. Here the lowest p-

value obtained is 0.1550, without the Bonferonni correction for multiple testing (see Table in the Appendix for all the p-values). So, from this simple analysis we cannot conclude that there is a difference in gain in scores between the two groups during either semester. However, this may still be inconclusive regarding the effectiveness of the tutor. Although the two groups appear to be fairly comparable, we saw that they were not entirely so and there could be other confounding factors causing differences (or lack of differences) between the two groups. Therefore we now proceed with more complex analysis using regression models.

3.2 Regression Models for Fall and Spring

Here our goal is to approximate the linear relationship between group and gain in scores while accounting for other factors. We fit six separate regression models for Fall and Spring for each of the three tests. The response variable is Gain in scores and the explanatory variables are Group, Grade, School, Gender, Ethnicity, and Pre-test scores. In the future we may also consider teacher, test administer, and date of test as factors, but for now we will consider only these six. The treatment variable of interest is Group as it indicates whether the tutor was used or not.

To determine which other variables to include in the models we use Mallow's Cp to determine which set of covariates minimizes the prediction error. Although our purpose is not necessarily prediction, this method will help us find a simpler, more interpretable model. While necessarily including Group, each of the six models give slightly different sets of covariates that minimize prediction error. However, there were similarities between the models: each of them is smaller than the full model which includes all of the covariates and none of the sets include Gender as an important predictor. School was the most commonly selected variable, chosen in 4 out of the 6 models. Since, as noted before, school, ethnicity, and grade give very similar information we think that using grade as the additional variable will be most helpful in giving a model of relatively minimal prediction error that is also consistent and interpretable for each of the six models. Pre-test was also selected as the best predictor in one case, but as shown in Figure 2, Pre-test and Grade are also highly correlated (See Appendix 5.2 for similar reading and spelling plots). So overall, by using grade as an additional explanatory variable in the regression model we also including prior knowledge, school, and ethnicity which is strongly tied to school.

For the results of our analysis we will use Fluency as the example. The regression model for the Fall model is as follows.

$$\text{GAIN.FALL} = -4.47 + 24.51\text{GROUP} + 9.27\text{GRADE} - 8.00\text{GROUP} * \text{GRADE} \tag{1}$$

$$\begin{aligned}
\text{GRADE 2 : GAIN.FALL} &= -4.47 + 9.27 * 2 + (24.51 - 8.00 * 2)\text{GROUP} \\
\text{GRADE 3 : GAIN.FALL} &= -4.47 + 9.27 * 3 + (24.51 - 8.00 * 3)\text{GROUP} \\
\text{GRADE 4 : GAIN.FALL} &= -4.47 + 9.27 * 4 + (24.51 - 8.00 * 4)\text{GROUP}
\end{aligned}$$

The last three equations are the separate regression lines for each grade derived from the first equation. The corresponding plot is shown below in Figure 3. Except for the intercept, all of the coefficients were moderately significant with the lowest p-value at 0.07920. The significance of the interaction term here shows that the effect of the tutor depends upon grade. In the Fall, Group 1 had the tutor while Group 2 did not. So from this model we can conclude that during the Fall, Grade 2 students appeared to benefit from the tutor while students from Grades 3 and 4 did not (in fact score gains for grade 4 students are lower with the tutor than without). For the Fall, the models for Reading and Spelling are very similar, so the same general results apply. (See Appendix 5.5 for Reading and Spelling plots).

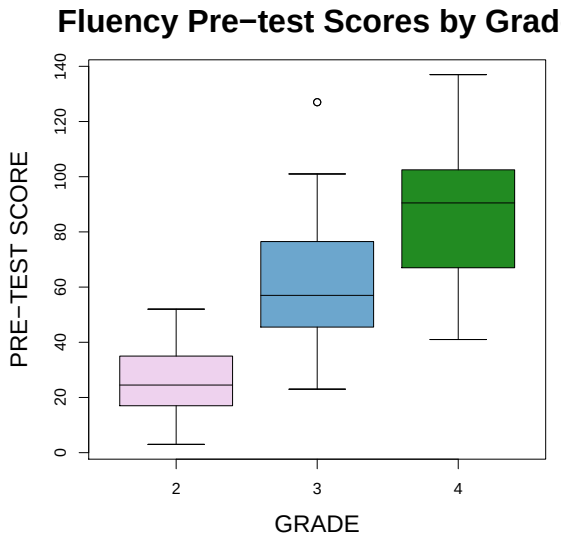

Figure 2: Each grade has a distinct range of pre-test scores, this high correlation makes grade and pre-test equivalent as predictors of student's improvment.

The Spring model gives different results. Here Group 2 as the tutor while Group 1 does not. The fitted model is

$$\text{GAIN} = 21.84 + 5.13\ \text{GROUP} - 4.77\ \text{GRADE} - 0.20\ \text{GROUP} * \text{GRADE} \tag{2}$$

However, except for the intercept, none of the coefficients are significant. As the plot shows, neither group nor grade appear to explain any significant difference in gain scores. This again highlights the major difference between the semesters.

The Spring Reading model is again very similar to the Spring Fluency model. However, the results for Spelling are more consistent with the Fall models, where second graders benefited from the tutors, third graders showed no difference and fourth graders did worse.

Overall, this analysis tells us that the students in Grade 2 at White Oak had higher Gains using the tutor than their control counterparts during the Fall for each of the three tests. This is true also for Spelling during the Spring, but there are no significant differences in Gains for Fluency and Reading between the groups during the Spring. This is likely due to the already noted seasonal effect. So, it appears that the reading tutor explains some of the difference in Gains, but direction of the association depends upon Grade. However, due to the high correlation between Grade, School, and Ethnicity we will not be able to tell which of these is truly affecting the students' learning in relation to the tutor.

3.3 Mixed Effects Models

The regression models have revealed some very definite patterns in the data; however, since we are gathering data on the same indivuduals over time, these measurements are not independent. A hierarchical linear model will better model the dependencies between measurements on the same student, and the dependencies between Grade, School, Ethnicity, and individual students.

This analysis has yet to be conducted.

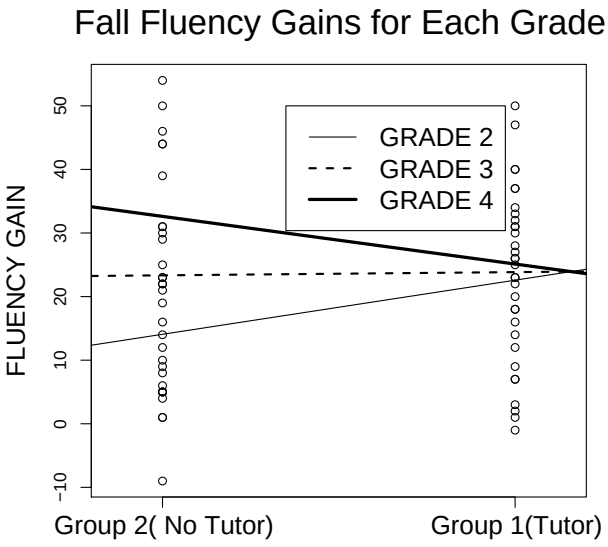

Figure 3: The lines indicate the predicted average improvement for students during the Fall by group and grade, using Equation 1. The average improvement for second graders using the tutor is higher than the control group; however for third graders there is no difference, and fourth graders using the tutor show less improvement than those in the control condition.

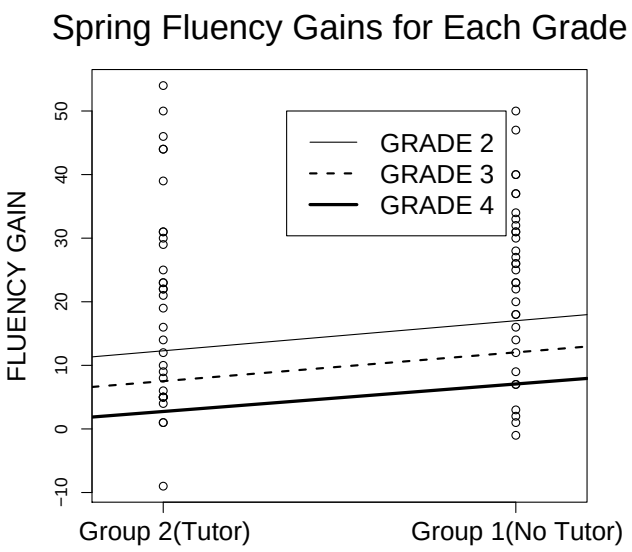

Figure 4: The lines indicate the predicted average improvement for students during the Spring by group and grade, using Equation 2. However, during the Spring, none of the coefficients are significant, except the intercept, so we cannot assume there is any difference between the tutor condition and control condition during the Spring.

4 Discussion

4.1 The Seasonal Effect

This study was originally designed with the idea that each student would be used as their own control: A student's learning rate using the Reading Tutor would be compared to the same student's learning rate participating in conventional instruction. This idea proved completely infeasible because all students, regardless of condition showed less improvement in the Spring than in the Fall. In fact, in Spelling, students had an average gain of a full grade level in the Fall and an average gain of zero in the Spring (Figure 5 in the Appendix).

One important, albeit inadvertant, implication of this experiment is that students do not learn at a constant rate. Moreover, learning rate varies less from student to student than it varies from month to month.

4.2 Conclusions

The Reading Tutor had a significant positive effect for second graders during the Fall. However, there was almost no effect for third graders, and a negative effect for fourth graders during the Fall. During the Spring, average gains were much lower overall. The two treatment groups showed no significant difference in Reading and Fluency in the Spring. However, in Spelling, we see the same treatment effect in the Fall and in the Spring; the second graders using the tutor showed a greater improvement than their counterparts in the control group, while the tutor had a negative effect on fourth graders.

As was noted in Section 3.2 Regression Models, model selection sometimes chose Ethnicity or School as a more important predictor than Grade. It is very dangerous to try to interpret this because School, Grade, Ethnicity, Pre-test scores and the alternate instruction used in the control condition are completely confounded: All of the second graders attended White Oak Elementary, which is a suburban, predominantly White school with high standardized test scores. All of the third and fourth graders attend Fort Pitt Elementary, which is an urban, 97% African-American school with 73% of third graders at or below the Basic level in reading.[3]

Something is preventing the third and fourth grade students at Fort Pitt from benefitting from the automated tutor in the same way that the second grade students at White Oak benefitted. We suggest three possible interpretations of this difference, each as likely as the next. One interpretation would be that the tutor was better than the Read Naturally program that the second graders used in the control condition, but not better than the mix of individual reading, group reading, and journaling used in the control condition by the third and fourth graders. Another interpretation that must be considered is that the tutor is poorly suited to the environment at Fort Pitt, perhaps for disciplinary reasons. A final possibility is that the tutor is more efficient at teaching beginning readers and less well suited to provide the practice that readers at a higher level require.

[3]Full results for third grade performance on the 2005-2006 Pennsylvania System of School Assessment (PSSA) are given in Table 2 in the appendix.

5 Appendix

5.1 Changes in Reading and Spelling.

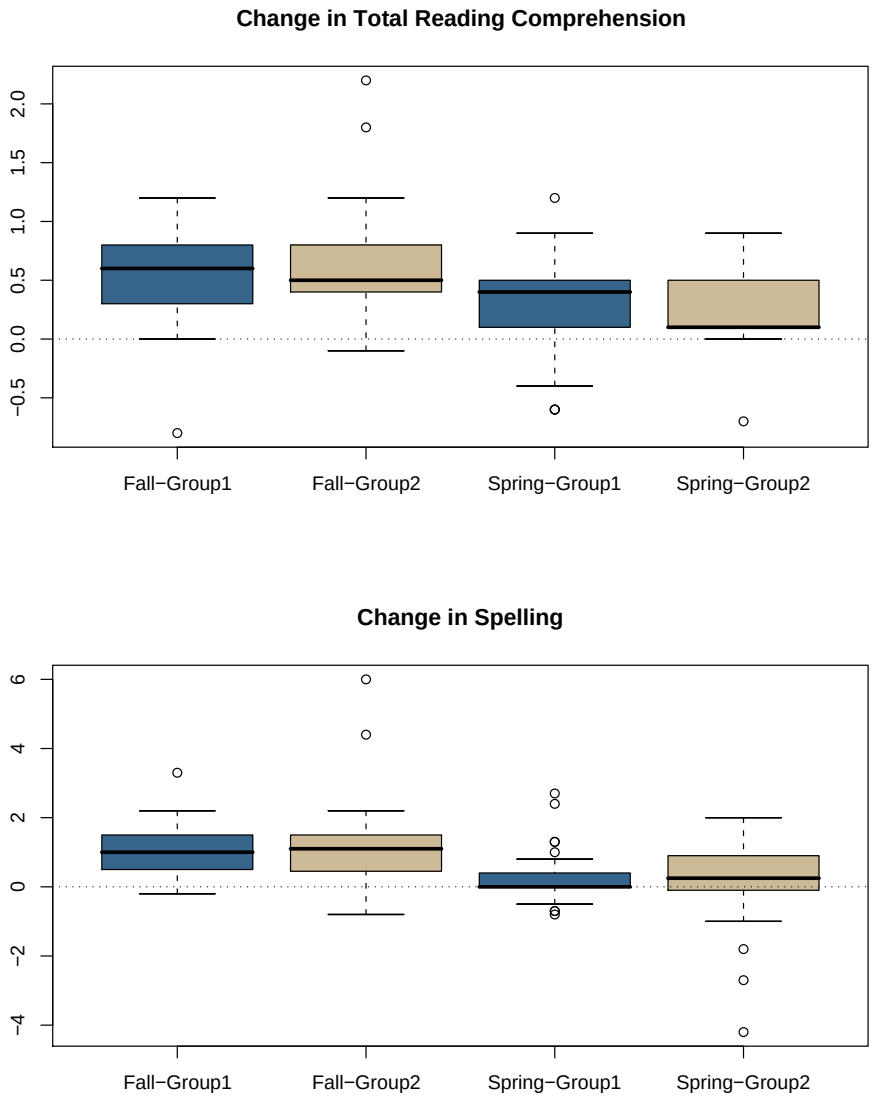

Figure 5: As with Fluency in Figure 1, the differences between Fall and Spring far outweigh the differences between Group 1 and Group 2.

5.2 Pre-Test Boxplots by Grade for Reading and Spelling.

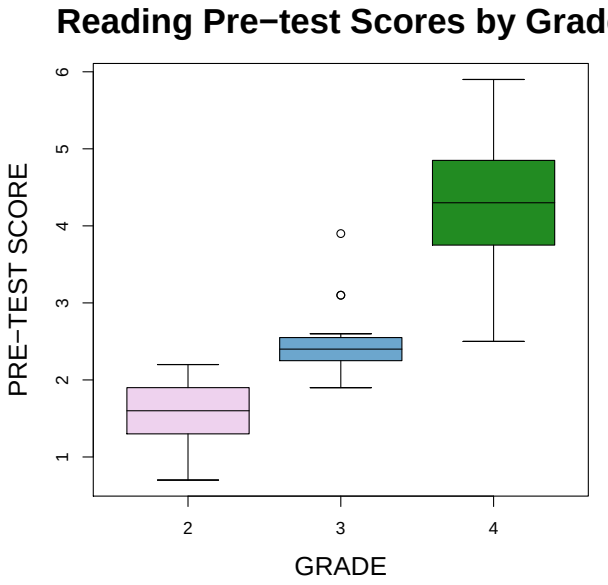

Figure 6: The side-by-side boxplots of Reading Pre-test scores show that Grade and Pre-test scores are highly correlated.

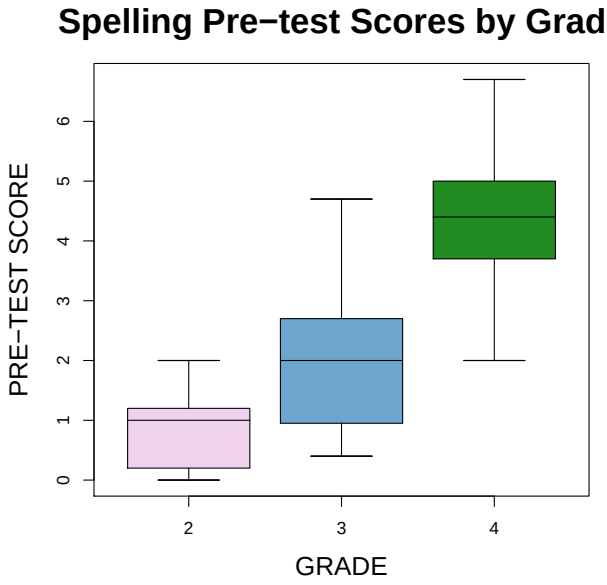

Figure 7: The side-by-side boxplots of Spelling Pre-test scores show that Grade and Pre-test scores are highly correlated.

5.3 Pensylvania System of School Assessment (PSSA)

	White Oak	**Fort Pitt**
Advanced in Reading	43%	8%
Proficient in Reading	31%	19%
Basic in Reading	13%	30%
Below Basic in Reading	13%	43%
Advanced in Math	57%	11%
Proficient in Math	31%	41%
Basic in Math	7%	35%
Below Basic in Math	4%	14%

Table 2: Percent of third graders at White Oak Elementary and Fort Pitt Elementary scoring in each perfomance category on the PSSA during the 2005-2006 school year, as reported by the Pennsylvania Board of Education.

5.4 P-values for t-tests and F-tests

	Grade 2	Grade 3	Grade 4
Fluency: Test for Equal Variances	0.1648	0.3063	0.2035
Fluency: Test for Equal Means	0.9745	0.5472	0.9770
Reading: Test for Equal Variances	0.0134	0.8642	0.6739
Reading: Test for Equal Means	0.0024	0.6148	0.7976
Spelling: Test for Equal Variances	0.3642	0.6663	0.1251
Spelling: Test for Equal Means	0.9811	0.5265	0.6733

Table 3: P-values for F-tests for equal variances and t-tests for equal means for Pre-test scores between each group, by grade. The values given are before the Bonferonni correction for multiple tests. Note that except for Reading in Grade 2, none of the p-values are significant, suggesting that the 2 groups are relatively comparable in terms of prior knowledge.

	Fluency	Reading	Spelling
Fall	0.4016	0.2869	0.7455
Spring	0.1550	5974	0.7481

Table 4: P-values for t-tests on gain in scores between the two groups. The values given are before the Bonferonni correction for multiple tests. Note that none of the p-values are significant and will therefore not be significant after a Bonferroni correction. These results show that we cannot conclude a difference in improvement between the two groups.

5.5 Models and Plots for Reading.

Fall Reading

$$\text{GAIN.FALL} = 0.17 + 0.92\text{GROUP} + 0.19\text{GRADE} - 0.40\text{GROUP} * \text{GRADE} \quad (3)$$

Spring Reading

$$\text{GAIN.SPRING} = 0.43 + 0.34\text{GROUP} - 0.07\text{GRADE} - .0.07\text{GROUP} * \text{GRADE} \quad (4)$$

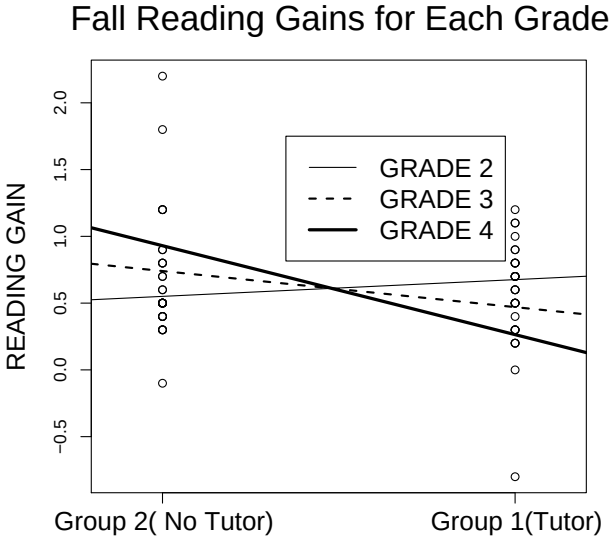

Figure 8: Relationship between Fall Reading Gain Score and Group by Grade using Equation 3. As indicated by the significance of the interaction term the slopes vary by Grade.

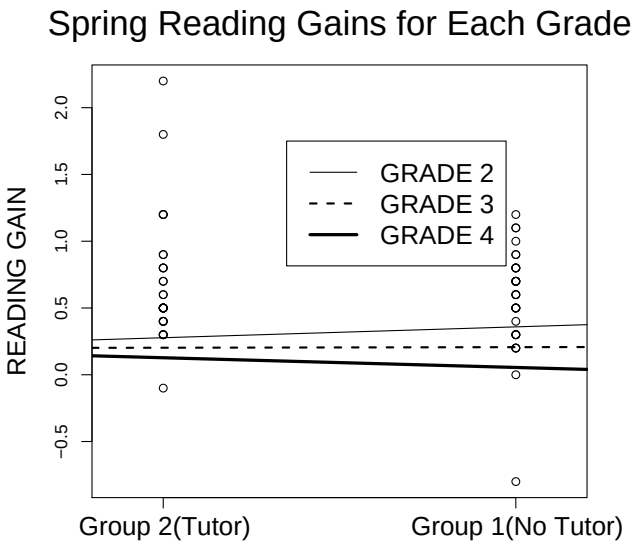

Figure 9: Relationship between Spring Reading Gain Score and Group by Grade using Equation 4. As indicated by the lack of significance of each of the estimates, no clear relationship can be seen between Gain, Group or Grade.

5.6 Models and Plots for Spelling.

Fall Spelling

$$\text{GAIN.FALL} = -1.19 + 2.21\text{GROUP} + 0.88\text{GRADE} - 0.86\text{GROUP} * \text{GRADE} \quad (5)$$

Spring Spelling

$$\text{GAIN.SPRING} = 1.54 - 1.68\text{GROUP} - 0.51\text{GRADE} + 0.67\text{GROUP} * \text{GRADE} \quad (6)$$

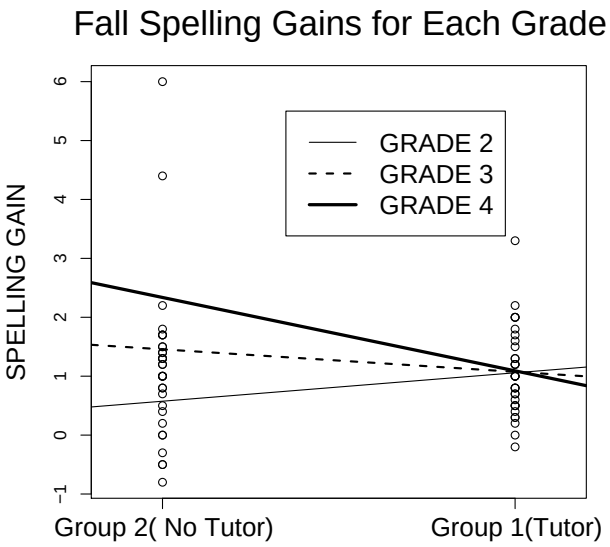

Figure 10: Relationship between Fall Spelling Gain Score and Group by Grade using Equation 5. As indicated by the significance of the interaction term the slopes vary by Grade.

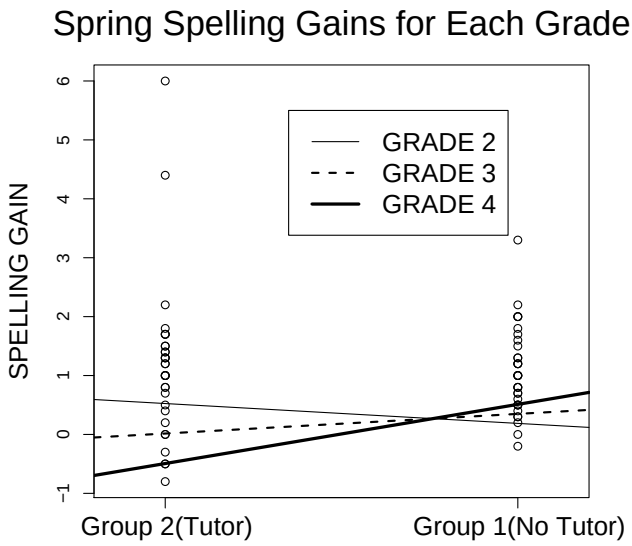

Figure 11: Relationship between Spring Spelling Gain Score and Group by Grade using Equation 6. Unlike the Spring Fluency and Reading models, the coefficients for this model were found to be significantly different from zero. Since now the groups are reversed in terms of tutor usage, the directions are reversed, but we still see that Grade 2 benefits, Grade 3 stays the same, and Grade 4 does worse with the tutor.

5.7 Model Coefficients for Fluency, Spelling and Reading

	Fluency Fall			
	Estimate	Se	t.statistic	p.value
(Intercept)	-4.47	8.459	-0.528	0.599
GROUP	**24.51**	**12.149**	**2.017**	**0.048**
GRADE	**9.271**	**3.045**	**3.045**	**0.003**
GROUP:GRADE	-8.001	4.482	-1.785	0.079
	Fluency Spring			
	Estimate	Se	t.statistic	p.value
(Intercept)	**21.835**	**8.53**	**2.56**	**0.013**
GROUP	5.132	12.251	0.419	0.677
GRADE	-4.773	3.07	-1.555	0.125
GROUP:GRADE	-0.202	4.519	-0.045	0.965

Table 5: Table of Regression Coeficients for Fluency Fall and Spring Model. Significant coefficients at the $\alpha = 0.05$ level are in bold. GROUP is one if student received Listen in the Fall, zero other wise. GRADE is 2,3, or 4. The outcome is the gain in test score from Pre-test to Mid-test for Fall, and from Mid-test to Post-test for Spring.

	Reading Fall			
	Estimate	Se	t.statistic	p.value
(Intercept)	0.171	0.237	0.723	0.473
GROUP	**0.917**	**0.34**	**2.696**	**0.009**
GRADE	**0.19**	**0.085**	**2.226**	**0.03**
GROUP:GRADE	**-0.396**	**0.125**	**-3.154**	**0.002**
	Reading Spring			
	Estimate	Se	t.statistic	p.value
(Intercept)	**0.427**	**0.215**	**1.983**	**0.052**
GROUP	0.235	0.309	0.761	0.45
GRADE	-0.075	0.078	-0.967	0.338
GROUP:GRADE	-0.077	0.114	-0.675	0.502

Table 6: Table of Regression Coeficients for Reading Fall and Spring Model. Significant coefficients at the $\alpha = 0.05$ level are in bold. GROUP is one if student received Listen in the Fall, zero other wise. GRADE is 2,3, or 4. The outcome is the gain in test score from Pre-test to Mid-test for Fall, and from Mid-test to Post-test for Spring.

	Spelling Fall			
	Estimate	Se	t.statistic	p.value
(Intercept)	**-1.186**	**0.568**	**-2.086**	**0.041**
GROUP	**2.211**	**0.816**	**2.708**	**0.009**
GRADE	**0.881**	**0.205**	**4.304**	**0**
GROUP:GRADE	**-0.864**	**0.301**	**-2.869**	**0.006**
	Spelling Spring			
	Estimate	Se	t.statistic	p.value
(Intercept)	**1.543**	**0.618**	**2.495**	**0.015**
GROUP	-1.678	0.888	-1.89	0.064
GRADE	**-0.509**	**0.223**	**-2.288**	**0.026**
GROUP:GRADE	**0.671**	**0.328**	**2.047**	**0.045**

Table 7: Table of Regression Coeficients for Spelling Fall and Spring Model. Significant coefficients at the $\alpha = 0.05$ level are in bold. GROUP is one if student received Listen in the Fall, zero other wise. GRADE is 2,3, or 4. The outcome is the gain in test score from Pre-test to Mid-test for Fall, and from Mid-test to Post-test for Spring.