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

Nathaniel Anozie, April Galyardt, and Elise Olson

May 4, 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 and the student take turns selecting stories. The Tutor is able to estimate the student's reading level, and will, on its turn, select a new story at this level. When it's the student's turn to choose a story, the tutor provides a list of stories to the student at that level. The student may pick a new story or one that they've previously read, they may also pick a story at a different level. According to the Project LISTEN website, "The Reading Tutor intervenes when the reader makes mistakes, gets stuck, clicks for help, or is likely to encounter difficulty."[2] For more than a decade, researchers have been testing Project LISTEN in various cities in the U.S. and Canada.

This study seeks to determine the effectiveness of the tutor as compared with other reading programs used in the Pittsburgh area. To do this we ask whether the Reading Tutor improves the fluency, reading, and spelling test scores of a sample of students. We also ask which other factors contribute to improvement.

In Section 2 we will discuss the data: how it was obtained and some exploratory data analysis. Section 3 will present our analysis, including how we decided to address issues with the data, our general findings, and the power of our tests. In Section 4 we will discuss the results, what conclusions can be drawn and conclude with recommendations for future studies.

2 Data

2.1 Study Design

This study was conducted during the 2005-2006 school year in two school districts. Fifteen third-graders and 12 fourth-graders from Fort Pitt Elementary in the Pittsburgh Public School Distric participated, along with 38 second-graders from White Oak Elementary in the McKeesport Area School District, for a total of 66 students.

[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/

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 were supposed to use the tutor 30 minutes daily; however acutal usage varied widely. (See Figure 1.)

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 were to use the automated tutor for 30 minutes each day during school. Students in the Control treatment particpated in alternate instruction.

Researchers assigned students at Fort Pitt to treatment group by matching on pre-test. However, White Oak insisted on extra protocols before agreeing to host the study. Researchers identified 50 weaker students as candidates for the study. White Oak selected 40 participants from this group; they also decided which students were assigned to each condition. White Oak was nervous about the Reading Tutor running smoothly, so they assigned the stronger readers among the particpants to use the Tutor during the Fall (Group 1). This would not be a significant obstacle to the original data analysis planned by the study designers; unfortunately that analysis proved unrealistic, and a different analysis had to be conducted. This will be discussed in detail later.

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

2.2 Measurements

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 can "sound-out" nonsense 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: total tutor usage, the date of test, the administer of the test, homeroom teacher, gender of the student, and ethnicity of the student.

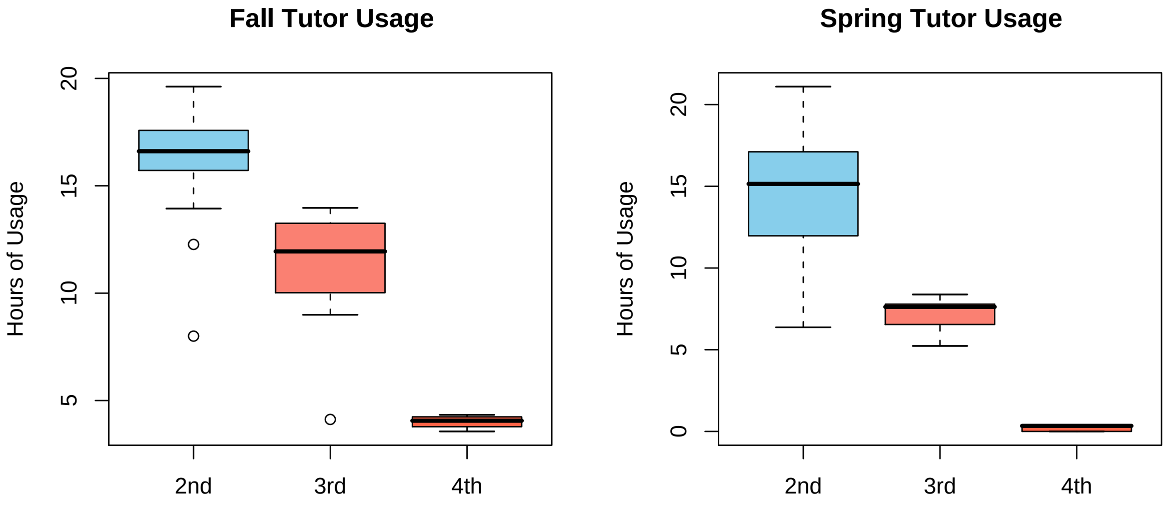


Figure 1: This is the total amount of Tutor usage for students assigned to use the Reading Tutor during each semester. If a student had used the tutor for 30 minutes 5 days a week for an 18 week semester, total usage should be around 40 hours. The 15 total hours for the second graders is consistent with using the Tutor about twice a week. The fourth graders effectively did not use the Tutor, in fact the total average for 4th graders in the Spring is 20 minutes – one session.

However, 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, grade, and total tutor usage are highly confounded, 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 Intended Analysis for the Two-Treatment Crossover Design

When this experiment was designed, the planned analysis was to compare each student's learning gain using the Reading Tutor to their gain in the control condition. In this manner, each student would serve as their own control and random assignment would not be necessary.

This within-subject comparison is only valid if we assume that students have a constant learning rate over time. Unfortunately this is not the case. For both groups, in all of the variables measured, learing in the Spring was significantly lower than learning in the Fall. Gains in Spelling for each group during each semester are shown in Figure 2.

We performed a regression analysis using Season as an additional variable to predict learning gain. In this analysis, Season was the only significant predictor. Since the Seasonal Effect overwhelms any effects for the Tutor, a student's learning in the Fall is a poor control for the same students learning during the Spring. This necessitates between-group comparisons during the same season.

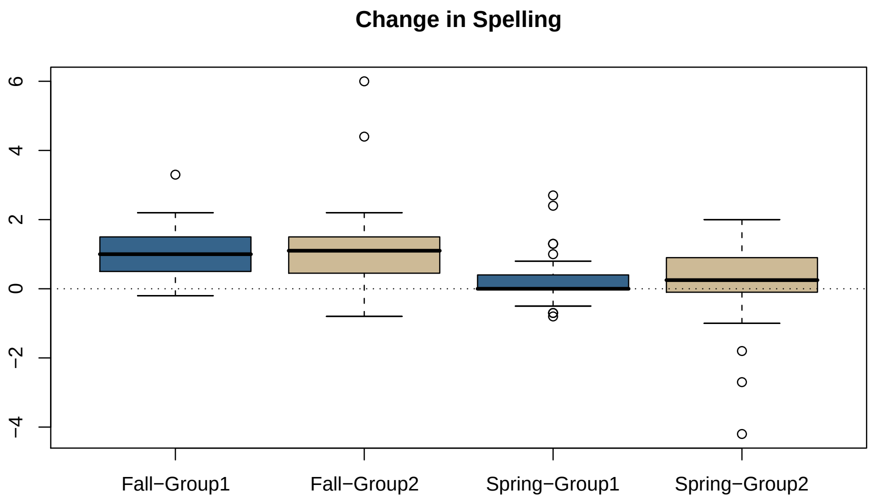


Figure 2: The Fall change is the difference between the pre-test and the mid-test. Likewise, the Spring change is the difference between the mid-test and the post-test. Group 1 indicates the Fall-treatment group, while group 2 used the Reading Tutor in the spring. Both groups had significantly lower performance in the Spring than in Fall. The difference between Fall and Spring is similar for Reading Composite and Fluency (Appendix 5.1). This plot includes all participating students.

3.2 Tutor Effect on Fourth Graders

As indicated earlier, the fourth graders in the tutor condition barely used the tutor at all. They had an average of 4 sessions in the Fall, and an average of 1 session in the Spring. This yields some very odd results when all 3 grades are included in a regression analysis. In this case, we find that the best model for predicting gains is one which has only group and grade as predictors. The following model was the best preditictive and interpretive model for predicting gain in fluency scores in the fall. We again refer to fluency as our example of the results of this analysis.

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

The corresponding plot is shown in Figure 3. (See Appendix 5.2 for the complete set of plots). 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. This model indicates that Grade 2 students benefit from the tutor, while Grade 3 students show no benefit in either condition. However, Grade 4 students appear to be adversely affected by the tutor.

This trend is apparent in all other measurement variables in the Fall. When there were discernable differences between the groups in the Spring, they followed the same pattern. Specifically, gain in scores for fourth graders in the tutor group are always less than or equal to gains in the control group. Results from two-sample t-tests indicated no significant differences, but these tests had low power since there were only 12 fourth graders in the study. This is still a noteworthy trend that could perhaps have been verified with a higher sample size. The full table of t-test results is contained in Appendix 5.3

The reason for this potentially negative effect among fourth graders is unclear since we know

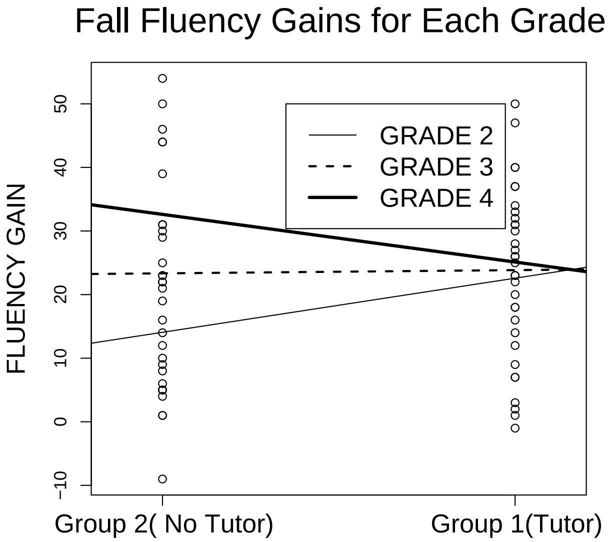


Figure 3: The lines indicate the predicted average improvement for students during the Fall by group and grade, using Equation 1. The average improvement for second graders using the tutor is higher than the control group; however for third graders there is no difference, and fourth graders using the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

that grade is highly confounded with pre-test, school, and tutor usage. However, one plausible explanation is that is being assigned to the tutor group, but not actually using it (as indicated by the very low usage hours for the fourth graders) is a major contributing factor to these results. We can only speculate as to why these students had such low usage hours. Perhaps they were sent into the computer lab to use the tutor, but due to lack of strict monitoring, the students chose to do other activities instead. It is also possible that the teacher never sent these students to the computer lab. Whatever the reason for the low usage, student learning appears to have suffered as a result.

3.3 Linear Regression Models for Fall and Spring Gain in Scores

To address our main question regarding the effectiveness of the tutor, we restrict our focus to the second graders at White Oak, since, as indicated by Figure 1, tutor usage at Fort Pitt was minimal.

We estimate the effect of the tutor by fitting linear regression models with Gain in learning predicted by Tutor and pre-test. Since the group using the tutor in the Fall was selected to have stronger students, pre-test and mid-test scores are strongly correlated with Tutor condition. Therefore, pre-test in the Fall and mid-test in the Spring must be included as covariates in the model if we hope to isolate the effect of the tutor. Other variables, such as Gender and Ethnicity, were added to this base model when selected by PRESS as important predictors. Since there are 7 different tests for 2 semesters, we fit a total of 14 different models. We also considered the possibility of an interaction term between condition and pre-test.

In the Fall, Tutor usage was significantly positively correlated with gain in Passage Comprehension ($p = 0.040$) and Spelling ($p = 0.024$), and was moderately significant in Fluency ($p = 0.064$). However, in the Spring, students who had used the Tutor in the Fall, and were now in the Control condition outperformed students in the Spring Tutor condition in Passage Comprehension ($p = 0.096$), Word Comprehension($p = 0.096$), and Total Reading Comprehension ($p = 0.093$). These Spring results are only moderately significant, but may support the hypothesis that the differences we observe are due to the fact that the Group who used the Tutor in the Fall were simply better students than the Group using the Tutor in the Spring. This necessitates a careful examination of the models for which Tutor usage was an important predictor.

3.3.1 Word Attack

Word Attack is not discussed further because it was not strongly correlated with any predictor during the Fall semester, and was strongly negatively correlated with Midtest scores during the Spring semester. We believe it to be a suspect variable for other reasons as well, for example several second graders had Word Attack grade equivalency scores in the 9-12 range.

3.3.2 Passage Comprehension

The two groups had significantly different gains in Passage Comprehension during both semesters. Moreover, in Passage Comprehension, prior knowledge was not a significant predictor of gain in either semester, but PRESS did select gender and ethnicity as important predictors. A table of coefficients with their p-values is given in Appendix 5.4, Table 3.

In Figure 4 we see that the Fall Tutor Group had significantly higher learning gains during both semesters. Recall that this group was selected by White Oak Elementary as the better students. Thus, this result indicates that what we are seeing is the "Matthew Effect": students who are already better students usually have higher gains (i.e. the rich get richer).

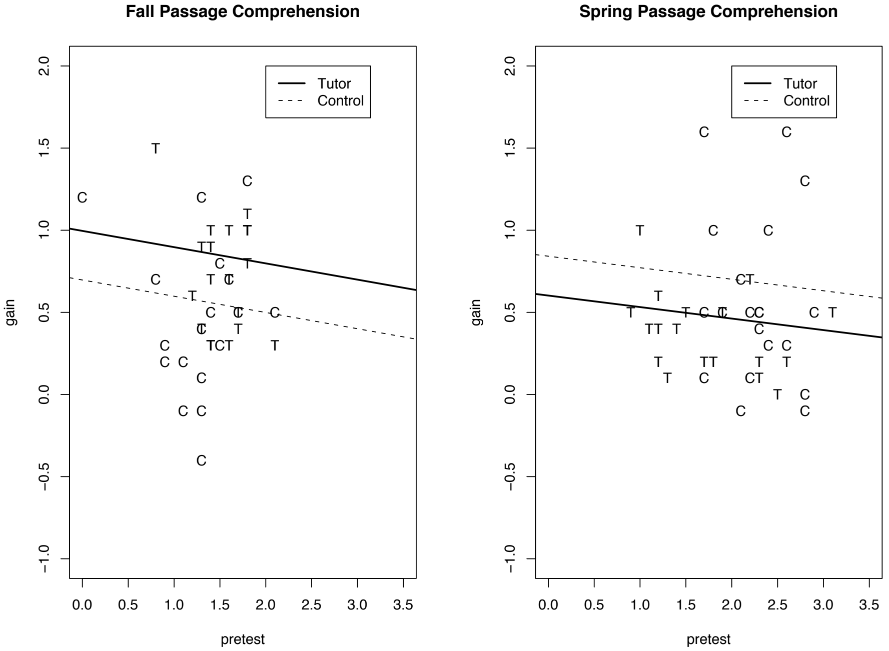


Figure 4: Learning Gains in Passage Comprehension by Treatment Condition and Pretest. For the Spring Semester, the "Pre-test" is the January test score. Recall that the students using the Reading Tutor in the Fall are the same students in the Control condition during the Spring. This indicates that even though Pre-test is not a significant predictor of gain, the Fall Tutor Group did better during both semesters.

3.3.3 Fluency and Spelling

In the Spring, there were no discernible differences in Fluency between the students based on any measurements, so we will only consider the Fall. Now, during the Fall, both Tutor Condition and Ethnicity were significant at the $\alpha = 0.05$ level. Results for Spelling are similar to those for Fluency. In the Fall, Tutor and Pre-test were significant predictors of Gains. During the Spring, the only important predictor was Mid-Test.

Unfortunately, because of the selection differences between the two groups, we are unable to determine if positive effect associated with Tutor usage is caused by the Tutor.

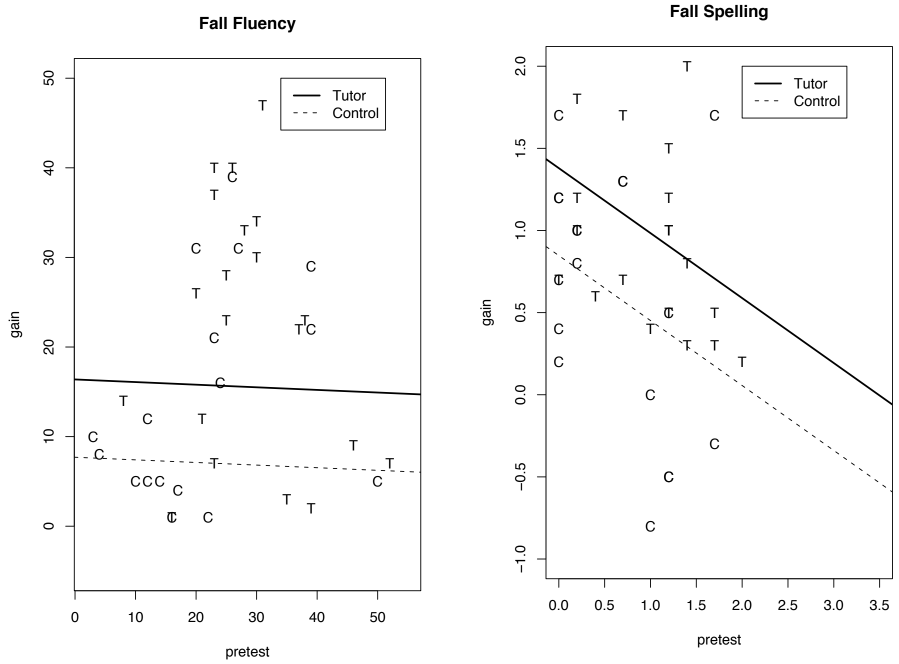


Figure 5: Learning gains in Fluency and Spelling by Treatment Condition and Pre-test for the Fall. We see that there is a significant difference in gains between the two groups for both tests, although we cannot make any statements about causality.

3.3.4 Total Reading Composite and Word Comprehension

During the Fall, there were no significant differences between the groups in Word Comprehension or Total Reading Composite. However, in the Spring, the Control was better than the Tutor Group, with moderate significance ($\alpha = 0.1$). Once again, the better students were the students that White Oak assigned to the Fall Tutor Condition, specifically chosen as better students.

3.3.5 Word Identification

Word Identification is also a special case. In both the Fall and the Spring, the regression models without interaction terms did not reveal any effects associated with the Tutor. However, when we included an interaction term, we found that both the Treatment Group and the interaction term were moderately significant. During both semesters, the Group using the Tutor had a moderately negative interaction effect, suggesting that in this area, the Tutor may counteract the "Matthew Effect".

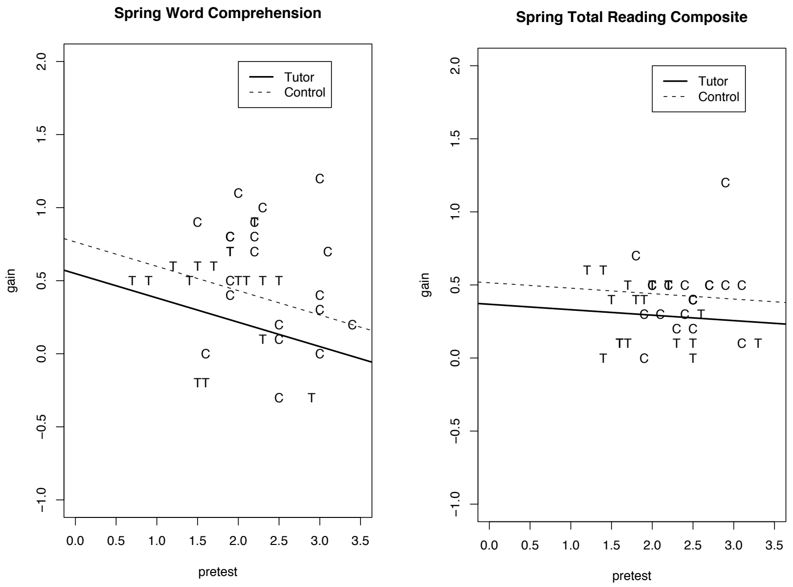


Figure 6: For both Word Comprehension and Total Reading Composite, the tutor was only a significant predictor of gains in the Spring. However, in both cases, it was the control group, the group selected to have stronger students, that had higher gains.

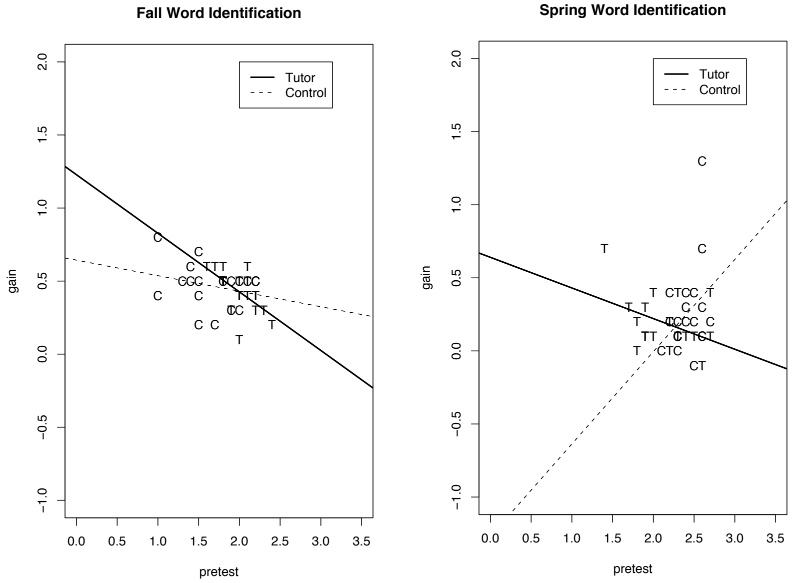


Figure 7: Word Identification is unique in that when allowing for interaction between Tutor and Pre-test (or Mid-test) we see that the Tutor and interaction terms are at least moderately significant. And in fact, the students with lower pre-test scores benefit most from the tutor, suggesting that the tutor may counteract the Matthew Effect in Word Identification. (Note that although these plots may appear to be null plots, that is simply because they are represented on the same scale as the other variables, but Word Identification had smaller variance.)

3.4 Power

Now we consider the issue of power: What are the probabilities our tests correctly detect differences between the two groups? In each of these models when we determine the significance of the tutor coefficient we are running a t-test of whether or not we should reject the null hypothesis that the coefficient is zero, or equivalently that the tutor has no effect. In these tests, we base significance on the p-value, the probability of incorrectly rejecting that hypothesis. To find the power of the test we ask: If the true difference were some non-zero β what is the probability our test would (correctly) detect this? The power estimates for different possible values of the tutor coefficient, β are presented in Appendix 5.6.[3]

A complementary question is, what size coefficient could we detect with a power of 0.80? These values are given below in Figure 8. Except for spelling and word attack, which had higher variances, our tests could usually detect a difference in gains of approximately 0.4 in terms of grade equivalency. So, for most of the reading tests, if the groups differed by close to a semester's worth of learning, our tests would detect that 80% of the time. For spelling, we could only detect differences slightly larger than a semester's learning. And we could detect differences of about 13 words per minute for fluency tests. These are fairly high differences and one would prefer to be able to detect differences of a scale. Much of this is due to the small sample size and recommendations for gaining higher power will be considered in the discussion section.

Detectable Differences (Beta's) with a Power of 0.80

Test	Fall Fluency	Spring Fluency	Fall Spelling	Spring Spelling	Fall Total Reading	Spring Total Reading
Beta	12.89	11.69	0.640	0.583	0.225	0.242

Test	Fall Word ID	Spring Word ID	Fall Word Attack	Spring Word Attack	Fall Word Comp.	Spring Word Comp.	Fall Passage Comp.	Spring Passage Comp.
Beta	0.114	0.273	1.14	0.801	0.399	0.360	0.399	0.399

Figure 8: The linear regression models with no interaction terms have a power of 0.80 to detect an effect of the Tutor ($\beta \neq 0$) for the values of β given in the table. These results are for n=38.

4 Discussion and Conclusions

4.0.1 Further Examination of the Matthew Effect

In several cases we saw possible evidence of the Matthew Effect: the students in the Fall treatment group who were selected to be stronger students, often out-performed the Fall control group during both semesters. Although this may be present, we can still ask whether the tutor reduces this effect. One way to test this is by examining the coffecients for interaction terms. In most variables the interaction terms were not significant, although as we saw with Word Identification, the negative interaction effect suggested that the tutor may counteract the Matthew Effect. An alternative analysis is to conduct hypothesis tests to determine if the correlation between pre-test

[3] Lenth, R. V. (2006). Java Applets for Power and Sample Size [Computer software]. Retrieved May 1, 2007 from http://www.stat.uiowa.edu/ rlenth/Power.

(or mid-test) and gains is lower in the Tutor group than the Control group. The values of these tests are given in Appendix 5.5. The p-value was lowest for the Word Identfication tests, confirming the results found in the regressions. Additionally, in this study, all of the correlations between gain and pre-test (or mid-test) in the Tutor group had lower values than in the control group. However overall, none of the tests were statistically signficant after adjusting for multiple tests, suggesting that a higher sample size is needed for futher examination of possible reduction of the Matthew Effect.

4.0.2 Conclusions and Recommendations for Further Research

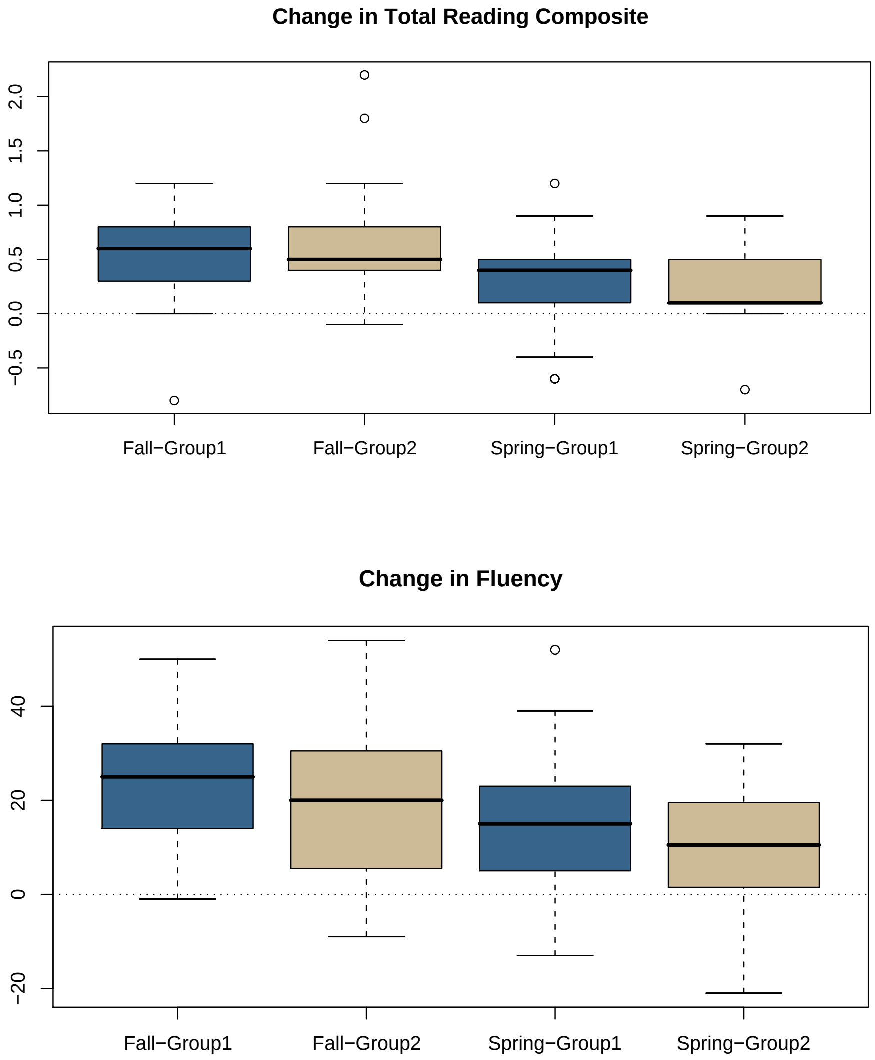
As stated above, there is reason to suspect that the tutor may help students with lower abilities learn better in some areas, especially in Word Identification. However, we would need greater sample sizes to detect this. Moreover, overall learning may be positively affected by use of the tutor, but unfortunately, due to an overwhelming seasonal effect and the Matthew Effect, we are unable to isolate any effect of the tutor from this sample.

Nevertheless, lessons learned here can be applied to make further studies more meaningful. First, since the seasonal effect in year-long studies is so large, between-group comparisons will be more accurate, which means random assignment is all the more necessary. Random assignment will also allow researchers to separate the Tutor effects from the Matthew effect.

Finally, to give more power to all analyses, greater final sample sizes are needed. In fact the original sample size planned for this study would have sufficiently powered nearly half of the regression analyses. (See Appendix 5.7. However, due to unknown reasons, the students from Fort Pitt did not participate fully in the study, necessitating their exclusion from the analysis. The lesson learned here is the importance of fidelity to the study design. Teachers will not faithfully implement a treatment if they do not ”buy-in” to a study. Therefore, we strongly recommend investing time in building relationships with the teachers. We expect that this will result in for accurate measurements on which to base conclusions.

5 Appendix

5.1 Changes in Reading and Spelling.


Change in Total Reading Composite

Change in Fluency

Figure 9: As with Spelling in Figure 2, the differences between Fall and Spring far outweigh the differences between Group 1 and Group 2.

5.2 Linear Regression Models Including all 3 Grades.

All models in this section are of the form:

$$\text{GAIN.FALL} = \beta_0 + \beta_1\text{GROUP} + \beta_2\text{GRADE} + \beta_3\text{GROUP} * \text{GRADE} \qquad (2)$$

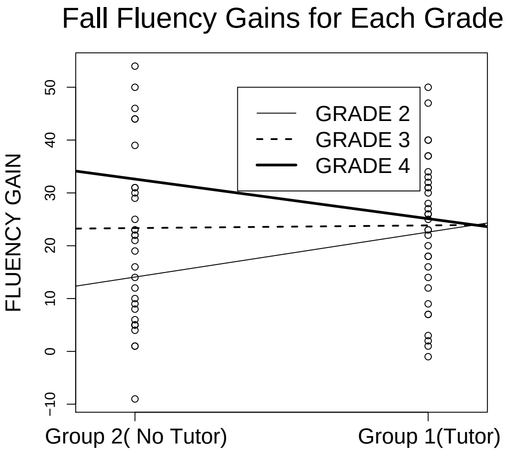


Figure 10: Second graders appear to show a positive effect, while 4th graders appear to show a negative effect.

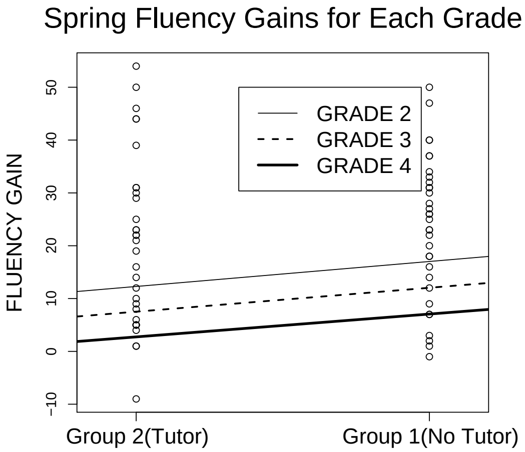


Figure 11: The average gain for students was significantly non-zero, but no other coefficients were significant.

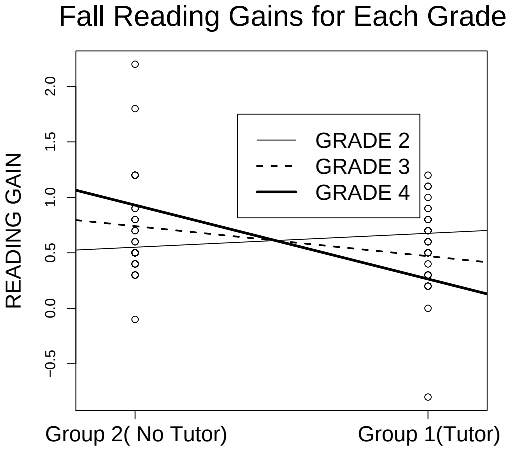


Figure 12: Second graders appear to show a positive effect, while 4th graders appear to show a negative effect.

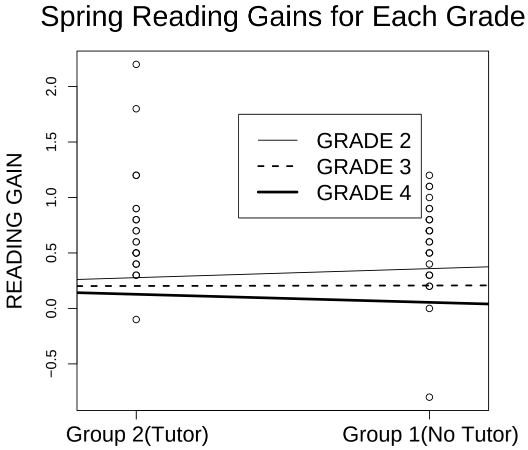


Figure 13: Spring gains in Total Reading Composite have no significant relationship with Group or Grade.

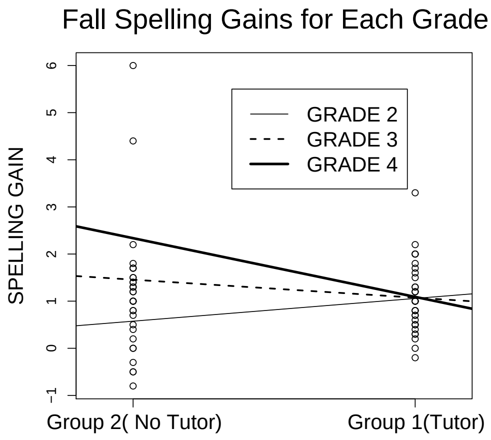


Figure 14: Second graders appear to show a positive effect, while 4th graders appear to show a negative effect.

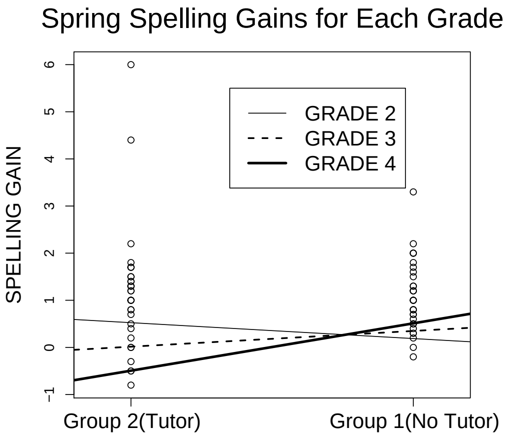
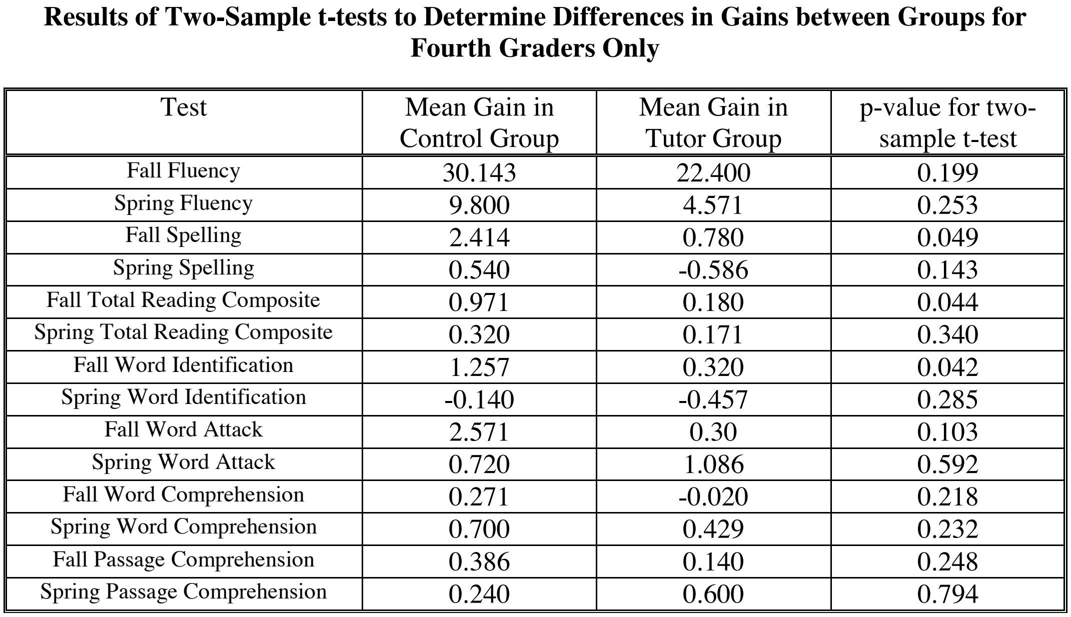


Figure 15: Note that the group using the Tutor is Group 2, now on the Left, so again we see that the tutor effect is positive for 2nd graders but negative for 4th graders.

5.3 T-tests for Differences Between Fourth Graders



Results of Two-Sample t-tests to Determine Differences in Gains between Groups for Fourth Graders Only

Test	Mean Gain in Control Group	Mean Gain in Tutor Group	p-value for two-sample t-test
Fall Fluency	30.143	22.400	0.199
Spring Fluency	9.800	4.571	0.253
Fall Spelling	2.414	0.780	0.049
Spring Spelling	0.540	-0.586	0.143
Fall Total Reading Composite	0.971	0.180	0.044
Spring Total Reading Composite	0.320	0.171	0.340
Fall Word Identification	1.257	0.320	0.042
Spring Word Identification	-0.140	-0.457	0.285
Fall Word Attack	2.571	0.30	0.103
Spring Word Attack	0.720	1.086	0.592
Fall Word Comprehension	0.271	-0.020	0.218
Spring Word Comprehension	0.700	0.429	0.232
Fall Passage Comprehension	0.386	0.140	0.248
Spring Passage Comprehension	0.240	0.600	0.794

Figure 16: The average learning gain for 4th graders using the tutor was always less than or equal to the gain of the students in the control condition. However, none of these differences were significant after correcting for multiple testing.

5.4 Linear Regression Models for 2nd Graders

Fall Model	Coefficient	Coefficient se	t	p-value
(Intercept)	0.697	0.245	2.748	0.010**
Reading Tutor	0.299	0.140	2.135	0.040**
Pre-Test	-0.099	0.179	-0.554	0.583
Male	-0.180	0.134	-1.340	0.189

Spring Model	Coefficient	Coefficient se	t	p-value
(Intercept)	0.602	0.245	2.372	0.024**
Control Group	0.239	0.140	1.713	0.096*
Mid-Test	-0.070	0.123	-0.570	0.572
Male	0.191	0.128	1.490	0.146
Ethnicity(White)	-0.338	0.135	-2.505	0.017**

Table 2: **Passage Comprehension** Coefficients and p-values for linear regression models of Learning Gain in Passage Comprhension. Gain as a function of Pre-Test and Treatment condition is plotted in Figure 4.

Fall-no interactions	Coefficient	Coefficient se	t	p-value
(Intercept)	7.70	5.97	1.29	0.206
Reading Tutor	8.68	4.53	1.92	0.064*
Pre-Test	-0.03	0.19	-0.15	0.880
White	9.09	4.68	1.94	0.061*

Fall-with interactions	Coefficient	Coefficient se	t	p-value
(Intercept)	0.92	7.16	0.129	0.898
Reading Tutor	25.15	11.05	2.277	0.029**
Pre-Test	0.23	0.25	0.928	0.360
Reading Tutor*Pre-Test	-0.63	0.39	-1.627	0.113
White	10.76	4.685	2.297	0.028**

Table 3: **Fluency** Coefficients and p-values for linear regression models of Learning Gain in Fluency. Gain as a function of Pre-test and Treatment condition is plotted in Figure 5.

Fall-model	Coefficient	Coefficient se	t	p-value
(Intercept)	0.847	0.191	4.432	0.00009***
Reading Tutor	0.532	0.225	2.363	0.02381**
Pre-Test	-0.395	0.184	-2.146	0.03892**

Table 4: **Spelling** Coefficients and p-values for linear regression models of Learning Gain in Spelling. Gain as a function of Pre-Test and Treatment condition is plotted in Figure 5.

Spring-model	Coefficient	Coefficient se	t	p-value
(Intercept)	0.368	0.174	2.114	0.04169**
Control	0.147	0.085	1.728	0.09284*
Mid-Test	-0.037	0.084	-0.445	0.65940

Table 5: **Total Reading Composite** Coefficients and p-values for linear regression models of Learning Gain in Total Reading Composite. Gain as a function of Pre-Test and Treatment condition is plotted in Figure 6.

Spring-model	Coefficient	Coefficient se	t	p-value
(Intercept)	0.548	0.217	2.526	0.01635**
Control	0.216	0.126	1.713	0.09589*
Mid-Test	-0.166	0.095	-1.758	0.08773*
White	0.238	0.127	1.874	0.06954*

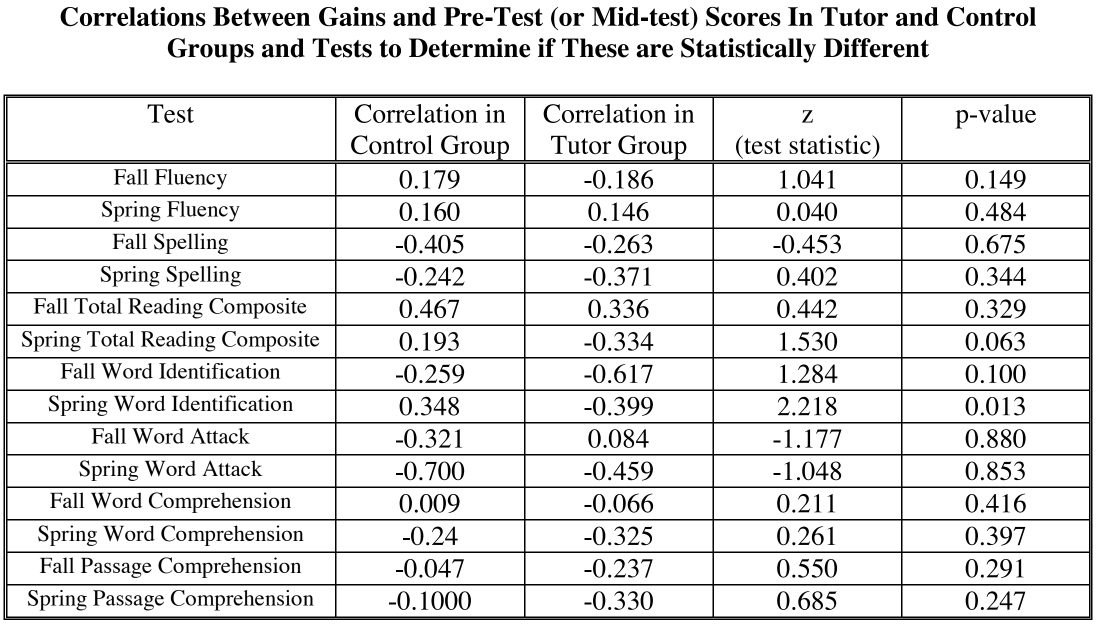
Table 6: **Word Comprehension** Coefficients and p-values for linear regression models of Learning Gain in Word Comprehension. Gain as a function of Pre-test and Treatment condition is plotted in Figure 6.

Fall-with interactions	Coefficient	Coefficient se	t	p-value
(Intercept)	0.643	0.146	4.406	0.00010***
Reading Tutor	0.585	0.321	1.819	0.07769*
Pre-Test	-0.106	0.086	-1.238	0.22406
Reading Tutor*Pre-Test	-0.295	0.167	-1.766	0.08643*

Spring-with interactions	Coefficient	Coefficient se	t	p-value
(Intercept)	0.641	0.352	1.818	0.07794*
Control	-1.909	0.892	-2.140	0.03963**
Mid-Test	-0.210	0.163	-1.283	0.20801
Control*Mid-Test	0.841	0.375	2.241	0.03163**

Table 7: **Word Identification** Coefficients and p-values for linear regression models of Learning Gain in Word-Identification. Gain as a function of Pre-test and Treatment condition is plotted in Figure 7.

5.5 Correlation between Gains and Prior Knowledge



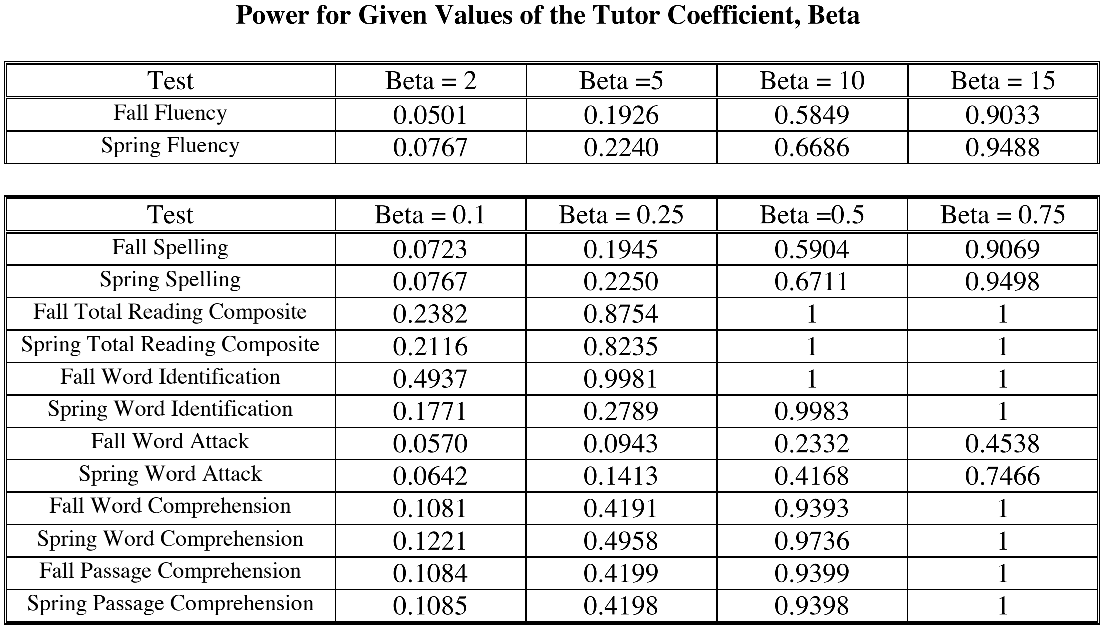
Correlations Between Gains and Pre-Test (or Mid-test) Scores In Tutor and Control Groups and Tests to Determine if These are Statistically Different

Test	Correlation in Control Group	Correlation in Tutor Group	z (test statistic)	p-value
Fall Fluency	0.179	-0.186	1.041	0.149
Spring Fluency	0.160	0.146	0.040	0.484
Fall Spelling	-0.405	-0.263	-0.453	0.675
Spring Spelling	-0.242	-0.371	0.402	0.344
Fall Total Reading Composite	0.467	0.336	0.442	0.329
Spring Total Reading Composite	0.193	-0.334	1.530	0.063
Fall Word Identification	-0.259	-0.617	1.284	0.100
Spring Word Identification	0.348	-0.399	2.218	0.013
Fall Word Attack	-0.321	0.084	-1.177	0.880
Spring Word Attack	-0.700	-0.459	-1.048	0.853
Fall Word Comprehension	0.009	-0.066	0.211	0.416
Spring Word Comprehension	-0.24	-0.325	0.261	0.397
Fall Passage Comprehension	-0.047	-0.237	0.550	0.291
Spring Passage Comprehension	-0.1000	-0.330	0.685	0.247

Figure 17: Results of hypothesis tests to determine if correlations between pre-test (or mid-test) and gains were different in each group. If the tutor did in fact counteract the Matthew effect, we would expect the correlations in the Tutor group to be lower. We see that for this sample, this is the case for all of the tests. Also, the p-values are especially low for Word Identification, confirming our regression analysis results. However, likely due to the small sample size, overall the results are not statistically significant when correcting for multiple testing.

5.6 Estimates of power

5.7 Recommended Sample Sizes for Future Studies

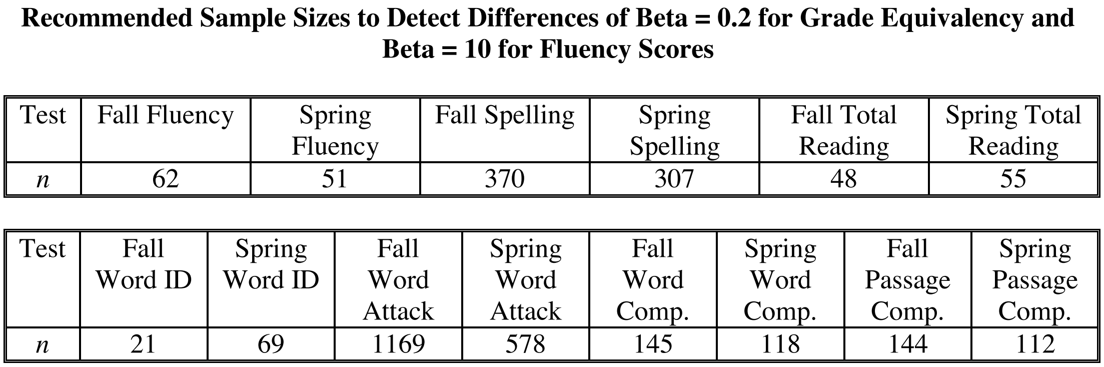


Power for Given Values of the Tutor Coefficient, Beta

Test	Beta = 2	Beta =5	Beta = 10	Beta = 15
Fall Fluency	0.0501	0.1926	0.5849	0.9033
Spring Fluency	0.0767	0.2240	0.6686	0.9488

Test	Beta = 0.1	Beta = 0.25	Beta =0.5	Beta = 0.75
Fall Spelling	0.0723	0.1945	0.5904	0.9069
Spring Spelling	0.0767	0.2250	0.6711	0.9498
Fall Total Reading Composite	0.2382	0.8754	1	1
Spring Total Reading Composite	0.2116	0.8235	1	1
Fall Word Identification	0.4937	0.9981	1	1
Spring Word Identification	0.1771	0.2789	0.9983	1
Fall Word Attack	0.0570	0.0943	0.2332	0.4538
Spring Word Attack	0.0642	0.1413	0.4168	0.7466
Fall Word Comprehension	0.1081	0.4191	0.9393	1
Spring Word Comprehension	0.1221	0.4958	0.9736	1
Fall Passage Comprehension	0.1084	0.4199	0.9399	1
Spring Passage Comprehension	0.1085	0.4198	0.9398	1

Figure 18: Power estimates based on linear regression models without interaction terms, and a sample size of 38. Spelling, Total Reading Composite, Word Identification, Word Attack, Word Comprehension, and Passage Comprehension are all on a grade equivalence scale. Thus, for these variables, $\beta = 0.5$ would be equivalent to a semester's difference between the group using the tutor and the control group. Fluency is measured in words read per minute, so the scale of this variable is different than the others.



Recommended Sample Sizes to Detect Differences of Beta = 0.2 for Grade Equivalency and Beta = 10 for Fluency Scores

Test	Fall Fluency	Spring Fluency	Fall Spelling	Spring Spelling	Fall Total Reading	Spring Total Reading
n	62	51	370	307	48	55

Test	Fall Word ID	Spring Word ID	Fall Word Attack	Spring Word Attack	Fall Word Comp.	Spring Word Comp.	Fall Passage Comp.	Spring Passage Comp.
n	21	69	1169	578	145	118	144	112

Figure 19: The original data set (including White Oak and Fort Pitt students) would have sufficiently powered 6 of the 14 tests.