Test Score Gains From Pre-Test to Post-Test Assessments: Comparing The Top of The Class to The Bottom and Traditional to Higher Engagement Instruction

By Alana R. McGraw

Abstract

Pre-tests are an education tool that help teachers assess students prior knowledge and inform students on what will be expected of them and what they will need to know. In this report, we investigated whether the gains in test scores from pre-test to post-test were significantly different depending on whether the student had one of the higher or lower scores on the pre-test and whether the students received traditional or higher engagement instruction. Using Mathematica, we calculated the difference in how much the test scores increase from pre-test to post-test depending on how they performed on the pre-test or what kind of instruction they received and compared that to 250,000 differences in test score gains from groups of the same size that are randomly sampled from their data sets. We determined that the difference in average test score gains were statistically significant for the individuals that scored higher on the MTH 130 pretest compared to those that scored lower with a difference of 16.03% and a p-value of 0.0, and those that were instructed from a traditional text and a higher engagement text with a difference of 4.22% and a p-value of 0.043. This could indicate that teacher are not addressing the needs of the students performing at the lower end of the spectrum on their pre-tests well enough and that learning might be improved when students are taught in a way that increases engagement.

Introduction

Pretests are a tool used by educators to assess their students pre-existing knowledge. They can be administered in the beginning of a course or before presenting new material in order to determine the needs of their students and provide instruction tailored to those needs. Teachers can determine what prerequisite information their students have and what information they must review before they can introduce new material. Pre-tests can help teachers create effective teams for group work by helping them pair people with varying understandings of the material (1).

These pre-tests can also help students. They allow students to relate what they already know to what they need to know for their summative assessments. The pre-test gives the students an idea of what material will be covered and and the depth of knowledge that will be required for exams. They can also act as a motivational tool that can help improve student attention (1).

There have been studies that indicate pre-tests may improve student performance, but there are a number of factors that can influence how effective the pre-tests are. These factors include the type of instruction, the effort put forth by the teachers and students to ensure their success and the ability of the students (1).

Information on the effects of these factors on the success of pre-tests can be helpful for developing best practices for educators to improve learning for all students. In this report we will be addressing

whether the test score gains are significantly greater in the students that scored at or above the pretest median compared to those who scored below it. We will also determine whether there is a statistically significant difference in the mean test score gains between the traditional and high engagement groups.

Methods

Data Collection

The data was obtained from the MTH 332 Website; the first data set was named "mth130prepost" and the second was named "pilot_versus_tradition". The first set of data was collected from a cohort of 155 pre-service teachers taking a college mathematics course called MTH 130. The information from the second data sheet was collected from a cohort of students in a mid-west university. The students were randomly split into two group that received "traditional" instruction and a "pilot" group that received instruction from a text that focused on more student engagement.

Data Processing

Once the data was downloaded, preprocessing began in the excel files. In the second data set, this started by rearranging the data so that all pre-test and all post-test scores were in the same column and an additional column was added to distinguish between the scores that came from the traditionally instructed class and those that came from a new "pilot" text. Those in the traditional group were given a 0 in that column and those in the pilot group was given a 1. The median pre-test score for the first data set was calculated and a new column was added to the data sheet in order to sort the samples into groups; those that scored at or above the median received a 1 in this column and those that scored below the median on the pre-test were given a 0. For both data sets, a sample ID was added for each pre-test and post-test score. The pre-test and post-test scores were multiplied by 100 to convert them into percentages. The test score gains were calculated for each sample using the following formula:

$$\text{gain} = (\text{[post-test score]} - \text{[pre-test score)}/(1-\text{[pre-test score]})$$

The first and second data sets were imported into Mathematica for analysis. All entries without data were removed from both data sets. Each data set was broken into two groups, with those that scored at or above the median and those that scored below the median on the pre-test for the first data set and those that were instructed with the traditional text or the pilot text for the second data set.

Data Summary

Lists were made for each subgroup of both data sets containing only the gain in test score for each student so that they could be used to summarize the data. Descriptive statistics were calculated for the groups to describe the data center, spread, and distribution of the test score gains. These calculations include the mean, standard deviation, skewness, kurtosis, and five-point summary (Table 1 & Table 3). Histograms for each of these groups was created as a graphical representation of the test score gains (Figure 1 and Figure 5). Smooth histograms were generated for each of these groups; the data set 1 groups were displayed together and the data set 2 groups were displayed together to compare the distributions between groups (Figure 2 and Figure 6) . A side by side box and whisker plot was constructed for these data sets to graphically represent data spread, dispersion, and outliers (Figure 2 and Figure 6). Quantile plots were generated for these data sets to graphically compare the data distribution to the normal distribu-

tion (Figure 3 and Figure 7).

Bootstrapping

The mean gains for the at or above and below were calculated above, and the difference between those means was calculated. Then the entire first data set was considered so that new groupings could be generated. From the data set, individuals were randomly sampled and added to a pseudo group that is as long as the at or above group. The remaining students were added to another pseudogroup. The mean test score gains were calculated, and the difference in the means was recorded. This was repeated 250,000 times. The distribution of the differences in means from the boot strapping random sampling were graphically displayed in a histogram with line at the observed difference in means between groups for data set 1. The p-value was calculated as the fraction of times that the difference in means was greater than the observed difference in means over the total number of calculated differences in means. The statistical significance of the difference in means was also assessed using a t-test as a comparison. This was repeated for the second data set groups.

Results

The mean, standard deviation, skewness, kurtosis, and five-point summary are described for the test score gains from those that scored at or above the median and those that scored below the median on the pretest assessment for the MTH 130 class (Table 1). The histograms and a side-by-side box and whisker plot for the gains from each group are visually describe the distribution of the data (Figure 1 and Figure 2). The maximum, minimum, median, upper quartile, and lower quartile numerically summarize the distribution of test score gains (Table 1). The skewness is positive for the test score gains of those that scored at or above the median and negative for those that scored below the median on the pretest. The below median group has a kurtosis closest to 3 at 3.12 while the at or above group is at 2.74. The quantile plot for the below the median group curves downward at the top and bottom but the middle follows the theoretical quantile. The quantile plot for the at or above group curves downward at the bottom and upwards at the top while the middle follows the theoretical quantile (Figure 3).

The difference between the mean gains of the at or above group and the below group is 16.03%. The p-value obtained through bootstrapping is 0.0 and the p-value from the student t-test is 3.65*10^-6 (Table 2). A histogram visually describes the difference in test score gains that would be expected if the the average test score gains were the same for each group (Figure 4).

Table 1. Description of the gains in test scores from the pre-test to the post-test assessments including mean, standard deviation, skewness, kurtosis and a five-point summary. Data was divided into those that scored at or above the median pretest score and those that scored below the median pretest score in the MTH 130 class.

	At or Above Median Pretest (%)	Below Median Pretest (%)
Mean	60.37	44.34
Standard Deviation	13.67	26.22
Skewness	0.067	−0.54
Kurtosis	2.73	3.12
Five−Point Summary		
1st Quartile	32.5	−25
2nd Quartile	50.7	26
Median	61.2	46.2
3rd Quartile	68.8	62.5
4th Quartile	92.9	97.3

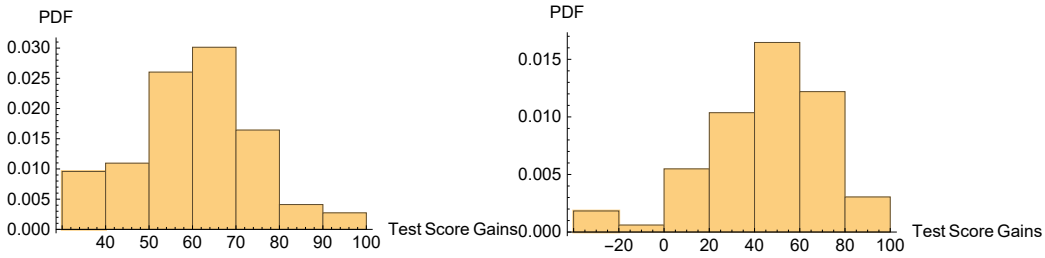

Figure 1. The gains in scores from the pre-test and post-test assessments from the MTH 130 class. Those who scored less than the median on the pre-test are on the left, and those who scored at or above the pre-test median are on the right.

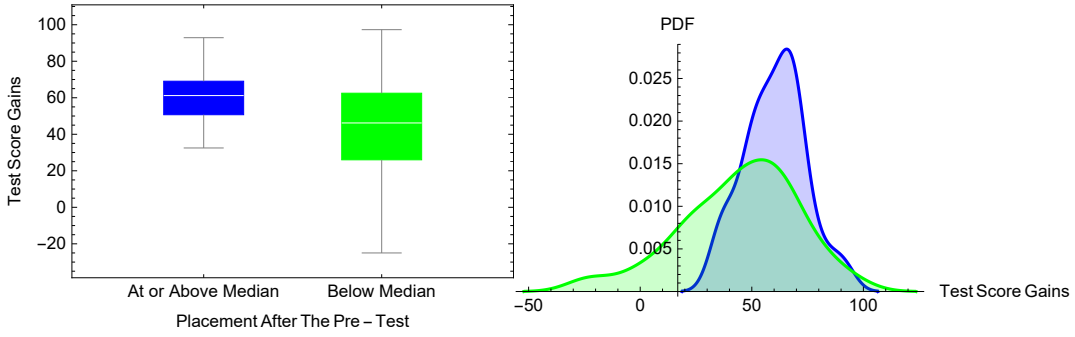

Figure 2. The test score gains from the pretest to post-test for those that scored above and below the median pretest score visually represented with a box-plot (left) and a smooth histogram (right). The at or above the median group gains are illustrated in blue and the below median group gains are in green.

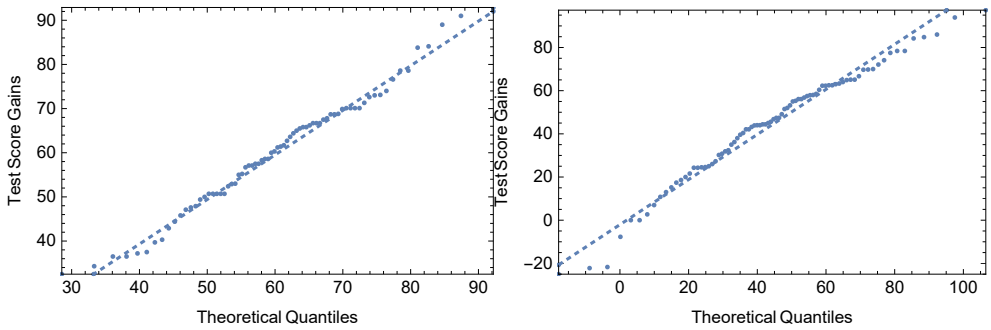

Figure 3. Quantile plot for the test score gains of the individuals that scored at or above the median (Left) and less than the median (Right) of the MTH 130 class on the pretest.

Table 2. The difference in the gains of test scores from the pretest to the post-test in those that scored at or above the median on the pretest and those that scored below the median on the pretest and the p-values produced from bootstrapping and t-tests.

Difference in Mean Gains	16.03 (%)
Boot Strapping p−value	0.0
T−Test p−value	3.65∗10^−6

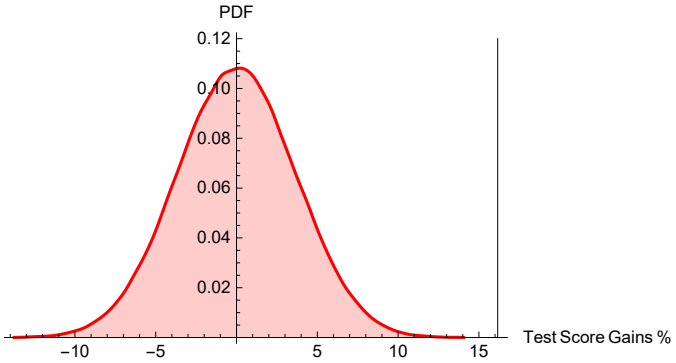

Figure 4. The difference in the average gains from pretest to post test when the MTH 130 class is randomly assigned into groups the same size as the at or above median group and the below median group 250,000 times. The line is the difference in average gains from the experimental at or above median group and the below median group.

The mean, standard deviation, skewness, kurtosis, and five-point summary are described for the test score gains from those that were instructed with the traditional text and new pilot study text (Table 3). The histograms and a side-by-side box and whisker plot for the gains from each group are visually describe the distribution of the data (Figure 5 and Figure 6). The maximum, minimum, median, upper quartile, and lower quartile numerically summarize the distribution of test score gains (Table 3). The skewness is slightly positive for the test score gains of both groups. The traditional instruction group has a kurtosis closest to 3 at 3.22 while the pilot instruction group is at 3.44. The quantile plot for the traditional group

curves downward at bottom and upward at the top while the middle follows the theoretical quantile but not very closely. The quantile plot for the pilot group curves upward at the bottom and top while the middle somewhat follows the theoretical quantile (Figure 7).

The difference between the mean gains of the at or above group and the below group is 4.22%. The p-value obtained through bootstrapping is 0.042 and the p-value from the student t-test is 0.085 (Table 4). A histogram visually describes the difference in test score gains that would be expected if the the average test score gains were the same for each group (Figure 8).

Table 3. Description of the gains in test scores from the pre-test to the post-test assessments including mean, standard deviation, skewness, kurtosis and a five-point summary. Data was divided into those that were instructed with a traditional text and those that were instructed with a newer modified text that emphasizes greater student engagement.

	Traditional Instruction (%)	New Instuction (%)
Mean	20.44	24.66
Standard Deviation	17.75	16.42
Skewness	0.23	0.12
Kurtosis	3.22	3.44
Five−Point Summary		
1st Quartile	−19.4	−20.7
2nd Quartile	8.9	14.3
Median	19.4	24.35
3rd Quartile	30.9	37.0
4th Quartile	66.2	64.9

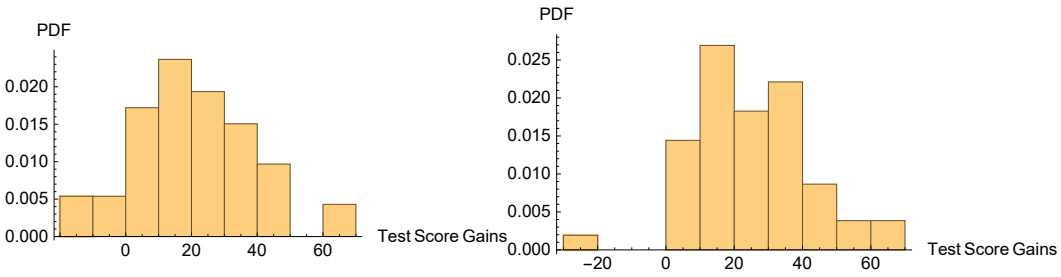

Figure 5. The gains in scores from the pre-test and post-test assessments from those taught with traditional instruction (Left) and those taught with the pilot instruction (Right).

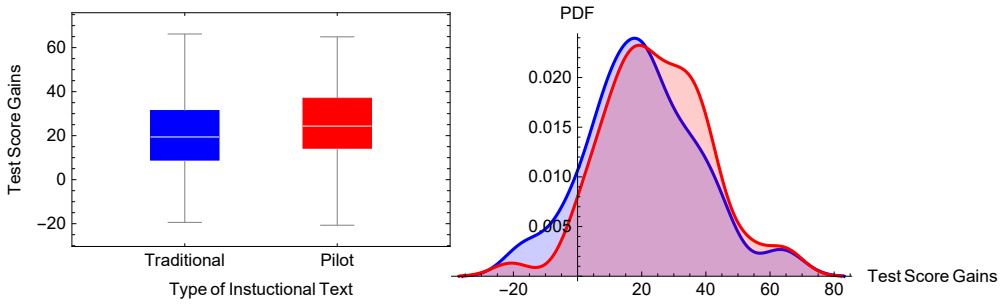

Figure 6. The test score gains from the pretest to post-test for those that were taught with traditional instruction and those taught with the pilot instruction visually represented with a box-plot (left) and a smooth histogram (right). The traditionally instructed are illustrated in blue and the pilot are in red.

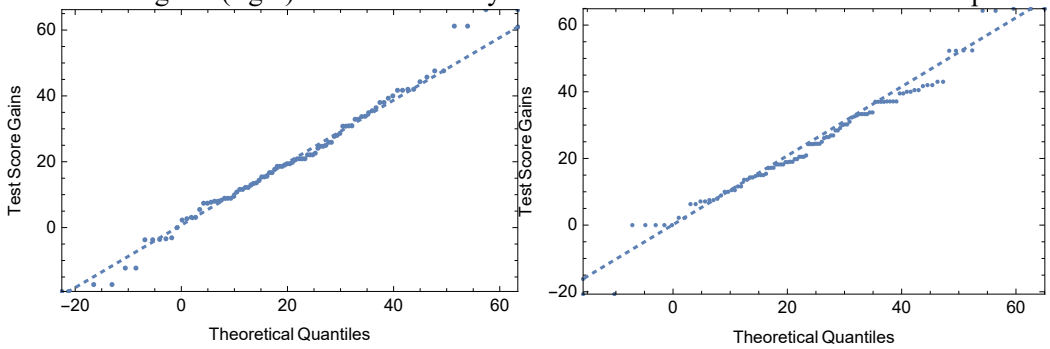

Figure 7. Quantile plot for the test score gains of the individuals that received traditional instruction (Left) and pilot instruction (Right).

Table 4. The difference in the gains of test scores from the pretest to the post-test in those that were instructed with the traditional text and those that were instructed with the modified text. The p-values produced from bootstrapping and t-tests are also summarized here.

Difference in Mean Gains	4.22%
Boot Strapping p−value	0.042
T−Test p−value	0.085

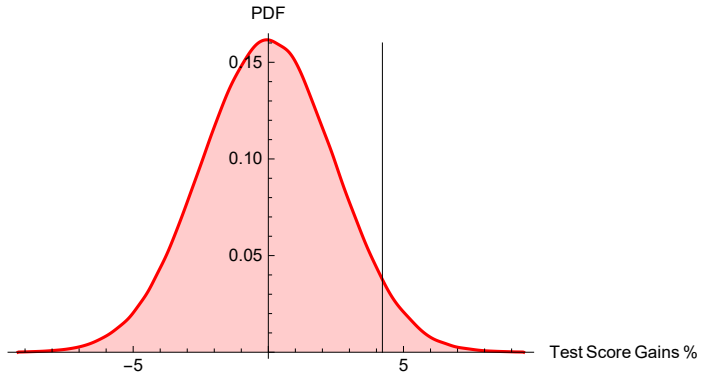

Figure 8. The difference in the average gains from pretest to post test when the sample from the second data set is randomly assigned into groups the same size as the traditionally instructed group and the pilot instructed group 250,000 times. The line is the difference in average gains from the experimental tradition-

Discussion

The objective of this report was to compare the test score gains of those that scored above the median and below the median on the pretest, and the gains from those that received traditional instruction compared to the "pilot" instruction intended to increase student engagement. In the summary of the MTH 130 data, we found that the average test score gain was higher in the at or above median group than the below median group. The below median group also had a much wider dispersion of test score gains, illustrated in the standard deviation, five-point summary and the visual representations of the datasets (Table 1 and Figure 2). Neither group has test score gains that fit the theoretical quantiles expected under the normal distribution (Figure 3). The above the median group is only slightly skewed to the right but have a kurtosis less than 3 indicating it has tails that are lighter than the normal distribution. The below the median group has a kurtosis slightly closer to 3 at 3.12, indicating that it might be slightly peaky or have slightly heavier tails than the normal distribution, and negative skewness shows that the test score gains is more skewed to the left than the normal distribution. This deviation from the normal distribution suggests that the assumptions for the t-test are not fulfilled.

In the summary of the mid-west university cohort, we found that the average test score gain was slightly higher in the pilot instruction group than the traditional instruction group. The pilot instruction and traditional instruction groups have similar dispersions of test score gains, illustrated in the standard deviation, five-point summary and the box-plot of the data (Table 3 and Figure 6). The smooth histogram shows that there is a lot of overlap but the pilot instruction groups are slightly to the right of the traditional instruction group's est score gains. Both groups do not fit the normal distribution very well, indicated by their quantile plots (Figure 7). Both groups have positive skewness showing that the distribution is slightly to the right of normal distribution, and the kurtosis above 3 for the traditional (3.22) and pilot (3.44) groups indicates that the data is either more peaky or had heavier tails than the normal distribution. This deviation from the normal distribution suggests that the assumptions for the t-test are not fulfilled.

The difference between the average test score gains of the above the median group and below median group is significantly higher than what would be expected if there was no difference between the average test scores of each group; this is also reflected in the p-value of 0.0 from the bootstrapping and the p-value of 3.65*10^-6 from the student t-test. The difference in the observed data is higher compared to what would be expected if both groups had the same mean test score gains under the assumptions of the null hypothesis; this is emphasized in figure 4. This could indicate the influence of students innate ability or it could be an indication that the teachers are teaching more towards the needs of the higher performing students.

The difference in test score gains of the traditional and pilot instruction groups was significantly higher than what would be expected if both groups had the same mean test score gain; the p-value from bootstrapping was 0.043. This p-value indicates that there is a 4.3% chance that the observed difference could occur under the assumptions of the null hypothesis, meaning it is unlikely but still possible. Assuming this difference did not occur by chance, this could indicate that the pilot instruction with higher levels of engagement leads to better learning outcomes. The student t-test resulted in a higher p-value at 0.085, which would indicate no significant difference between the mean test score gains of the two groups; this illustrates the fallibility of the t-test when the assumptions are not met. The difference in the mean test

score gains between the traditional and pilot groups observed is slightly higher than what is expected by chance if the mean test score gains were the same in both groups.

This information is limited in that it is only representative of the populations they are sampled from. The significant difference between the test score gains of the groups that scored above and below the median is only representative of the individuals in the MTH 130 class, it does not necessarily extend to other classes or even other universities with similar classes. The same is true for the significant difference in test score gains in the traditional and pilot study group because the results may vary depending on instructors and where the students are sampled from. In order to make studies that are generalizable to a larger population, the students must be randomly sampled from a broader population. The p-values from the t-tests were also reported in this study but it is important that these results are considered in the context that not all of their assumptions are met, making the results questionable. The bootstrapping method is also limited by how many sessions the computing power can manage in a reasonable amount of time. The more sessions that are performed, the more reasonable or unreasonable it will be to assume that the differences observed in groups did not occur by chance in the sampling of two groups with the same mean. In order to improve this study, the data should be sampled from a broader population and more sessions should be used in the boot strapping procedure.

In this report, we determined that the difference in average test score gains are statistically significant for the individuals that scored the best and worst on the MTH 130 pretest, and those that were instructed from a traditional test and a pilot text meant to increase student engagement.

References

1. Berry, Thomas. “Pre-test assessment.” American Journal of Business Education (AJBE) 1, no. 1 (2008): 19-22.