Statistical Similarity of Test Scores

Written by: Nick Paternostro

Abstract

Alternative methods of learning to enhance the overall experience of attending school has been under investigation for many years. More recently different methods of learning have been utilized, such as online learning, due to the current state of the world. A mid-west university is interested in correlating different forms of learning by comparing pre-test and post-test scores of students. The first study analyzes the pre-test and post-test scores for 155 pre-service teachers split into two groups, those who scored less than the median pre-test score and those who scored greater than or equal to the median pre-tests score. Using a gain statistic, the gain scores of each student are compared amongst different groups to determine if there is a significant statistical difference between the mean gain scores. Using a bootstrapping technique for 250,000 cases, the likelihood of obtaining the difference in mean produced from the groups if the null were true was tested. The first study, which involved the gain score of pre-service teachers taking a college mathematics course had a relatively low p-value, meaning the likelihood of having the difference in scores shown in Table 2 is highly unlikely if there were no statistically significant difference in the gain scores of Group A and Group B. Therefore, we reject the null hypothesis since there is probabilistic evidence that the mean gain of Group A is statistically significantly greater than the mean gain for Group B. The second study analyzes two cohorts of students; one group taught by a traditional learning method and the other by pilot learning method. Similar to the last study, bootstrapping was utilized to find the p-value. The second study had a p-value of roughly 0.08, therefore there is not enough probabilistic evidence to reject the null hypothesis of which there is no statistical difference between the mean gain scores. Lastly, for each study, the p-value was found based on a conducted t-test (Table 5 and Table 10). This value was then compared with the p-value generated by using the bootstrapping method.

Introduction

Due to the current state of online learning, a popular question becomes, will online learning and some of the new tools which have come about be utilized when normality is reached once again. Teaching methods such as asynchronous learning have developed over the course of the COVID-19 pandemic and should be considered for future incorporation[1]. Two studies are examined in this report. The first study focuses on the gain in scores from pre-test to post-test scores for pre-serve teachers. While it may be expected for pupils with better pre-test scores to have better post-test scores, the gain in score may be smaller for students with higher post-test scores. The first study aims to determine if there is a

significant difference in the mean gain scores of the two groups of students. The second study involves students at a mid-west university who are experiencing different forms of learning. One group of students is taught using the traditional learning style. The traditional style of learning utilizes existing texts and is taught by a lecture-style method. The other group of this study is taught by using the pilot method which utilizes a modified text that emphasizes greater student engagement. Educators should be interested in results produced in this study since a conclusion will be made whether students have larger gains in their scores (pre-test to post-test) when the student has a lower score initially and which style of teaching would be best to increase the gain of students.

Methods

Both data sets were imported from excel into Mathematica. Since excel files are usually tab-delimited files (.xls), the data had to be converted to a comma-delimited (.csv) file to import the pre-test and post-test scores for each student individually. This data is generated by an unknown source however it would be useful for educators at all levels to understand which methods of teaching would be optimal. The same series of events were conducted for both studies, therefore to avoid re-iterating the same steps, the general process will be described and then directed for specific cases. After the data was imported, the data was examined for repetitions. To avoid error further in the calculations, specifically during the bootstrapping process, numbers ranging from one to the length of the list containing the entire cohort of students were assigned to an individual case (student). By doing this, all grades are considered as an individual data set, therefore, when separating into pseudo-groups all students' scores will be considered even if there is repetition. Since the entire cohort of students is considered for the previous step, depending on the study, the data must be separated into cases. For the first study, the data is separated into two groups: students with pre-test scores lower than the median pre-test score for the entire cohort of 155 students, and the second group is made up of students with pre-test scores at or above the median pre-test score. For the second study, the data is separated depending on the learning method of the student. The data is imported with two methods of learning, traditional and pilot. Since the data for each student is imported in a way to take into account the style of learning, manual callings of elements in the list can be used to re-separate the data. Using equation (1), the gain scores of each group can be found. The gains are the focal point of this study and to understand the gains, descriptive statistics of the gains were examined for each group. Using Mathematica, the mean, median, standard deviation, skewness, kurtosis were found for each student's gain score. By utilizing descriptive statistics, the distribution of gain scores by each group of students can be characterized. The mean and median gain scores are first and foremost used to analyze which group of students had the larger rise in score. Alternatively, the mean and median gain scores could be used to interpret the difference between the mean gain scores. The standard deviation, skewness, and kurtosis are additional parameters used to characterize the distribution. However, in these studies, these parameters serve a greater purpose. To utilize a t-test, conditions must be satisfied. A t-test will return a p-value, the likelihood of what one saw happen if the null were true. Although the t-test may seem helpful, there are built-in conditions that must be satisfied for the t-test to produce accurate results.

Additionally, it may be difficult to interpret the t-test results without understanding the process. Therefore, an alternative method for finding the p-value will be presented. However, to use a t-test, the built-in assumptions must be satisfied and if they are violated, it will make the results questionable. The three main assumptions of a t-test are as follows: the groups of data should be independent, the variances of the two data sets should be equal, and lastly, the distributions should be approximately normal. By analyzing the standard deviation of each group, the variance can be found and compared as well. In addition, by comparing the skewness and kurtosis of the groups in each respective study, a conclusion can be made regarding the normality of the study. Throughout this process, it is important to evaluate how the current steps will contribute to the goal of this study; finding the likelihood of seeing the difference in means of the groups (Table 2, first study and Table 7, second study) given the null hypothesis were true. Before moving onto the bootstrapping method for finding the p-values of the data, for a better understanding of whether the gain scores of the two groups, for each respective study, are significantly different, the proportion of Group A gain scores for which the gain lies above the median gain for Group B was found. Also, the proportion of group B gain scores for which the gain lies below the median gain for Group A was found. If the distributions were similar in shape we would expect this calculation to yield around 0.5. The higher it is above that proportion, the more we have an indication that the distribution for Group A has shifted to the right of the median for the group B gains, meaning there is a higher possibility of the two groups of gain scores being significantly different.

gain = ([post-test score]-[pre-test score])/(1-[pre-test score]) (eqn 1)

As mentioned above, there are quite a few assumptions involved when computing the p-value by using a t-test. To thoroughly understand the information which a p-value provides, bootstrapping technique is used. The idea of bootstrapping is as follows:

1. With the given, divided groups of each respective study, find the gain relating to each student's pre-test and post-test score. Then, find the mean gain score for each group

2.Pool all gain scores for both groups of each study, a cohort of pre-service teachers and a cohort of mid-west university students, respectively. Then randomly select as many students as were in Group A and assign them a new group called, pseudo-group A. After that, the students remaining who were not selected to pseudo-group A are assigned to pseudo-group B. For example, if the total number of students were 155 and there were 73 students in Group A, then 73 randomly selected students from the entire cohort of students would be placed in pseudo-group A and the remaining 82 students would be assigned to pseudo-group B. It should be noted that since the pseudo-groups are randomly generated, there will be different means, hence a different difference in mean gain scores for each time it is calculated.

3. Calculate and Record the mean scores(Table 3 and Table 8) difference in mean scores for pseudogroup A and pseudogroup B (Table 4 and Table 9)

4. After we pool and randomly select the points for "pseudo-group A" and the remaining points for "pseudo-group B" and then compare the difference in mean gains between pseudo-group A and pseudo-group B to see if it is smaller or larger than the difference in mean gains for Group A and Group

B, to obtain a comparative sample for this difference in means, we repeat the above steps for 250,000 times and see how often we get a difference in mean gains smaller than that between Group A and Group B.
NOTE: The total number of ways the data could be sampled is roughly 1*10^45. Being that this computation would take a longer amount of time to compute than the universe has been around, only a small sample is taken. Unfortunately, since all the possibilities are not tested, there is still room for error.
5. Compute the p-value. Assuming there was no difference, how likely is it that we will see a statistical difference the way we see it? That is, what is the probability of seeing the difference in means we do see under the null hypothesis of no statistical difference? How many cases out of the 250,000 were the mean gain of the pseudo-groups smaller than the difference between Group A and Group B? For visualization, plot the distribution of differences in means between the pseudo-groups A and B and locate the difference mA-mB on that histogram.
Lastly, conclude whether there is a statistical difference between the two gain scores for each respective study.

Results

First Study

Descriptive Statistics of Mean Gain Scores

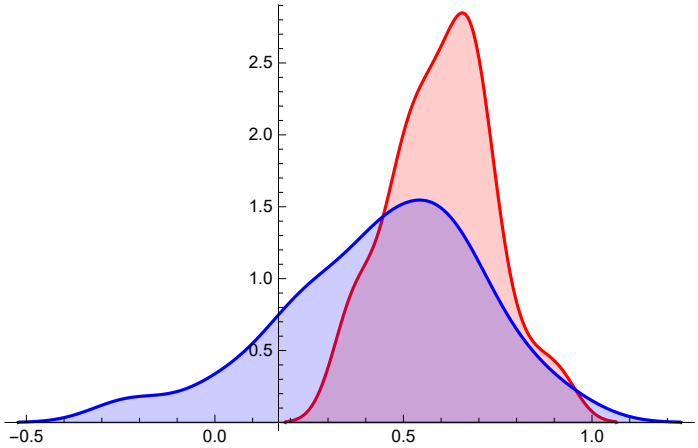
The pre-test and post-test scores of two groups of pre-service teachers were analyzed to determine the effect of the mean gain scores. Two groups were formed from the data based on the pre-test scores, students with pre-test scores lower than the median pre-test score of the entire cohort and then students who scored at or above the median pre-test score. The students who scored below the median pre-test score of the entire cohort were sampled to Group A and the student who scored at or above the median pre-test score were sampled to Group B. Using equation (1), the statistical gain for each student was calculated. Table 1 lists the descriptive statistics of the gain scores for each group. The mean of Group A is 0.6037 and the mean of Group B is 0.4434, hence the student with lower scores, n average, had a slightly larger gain than the students who originally scored higher on the pre-test. The median of Group A is 0.6119 and the median of Group B is 0.4621, therefore a similar conclusion can be made. The standard deviation of the gain score of Group A is 0.137 whereas the gain scores of Group B have a standard deviation of roughly 0.26. Since the standard deviations are not equal, the variances are not equal, therefore already one built-in assumption of the t-test has been violated. The skewness of the gain scores of Group A is 0.067 which is positively skewed; on the other hand, the gain scores of Group B are negatively skewed with skewness of -0.5386. Lastly, the kurtosis of each group is reported. For the gain scores of Group A, the kurtosis is 2.729, but for the gain scores of Group B, the kurtosis is 3.124. Based on these values, it is proven the gain scores deviate from a normal distribution, hence breaking another built-in assumption of the t-test. Table 2 highlights the difference in mean gain scores of Group A and Group B.

Table 1: Descriptive Statistics of Gain Scores for Group A and Group B

Out[•]=

	Group A Gain Score	Group B Gain Score
Length	73	82
Mean	0.6037	0.4434
Median	0.6119	0.4621
Standard Deviation	0.1367	0.2622
Skewness	0.0669	−0.5386
Kurtosis	2.729	3.124

Figure 1: Smooth Histogram of Group A (red) and Group B (blue) Gain score distributions



Based on the data presented in Figure 1 and Table 1, there seems to be a significant statistical difference. This will be further elaborated upon in the discussion.

Table 2: Difference in Mean Gain Score of Group A and Group B

Out[•]=

	Difference in Mean Gain Scores
Mean(Group A) – Mean(Group B)	0.1603

Bootstrapping

As mentioned earlier, there are quite a few built-in assumptions about the distribution of data when using a T-test. An alternative approach to ascertain if the means for two groups of data are statistically significantly different is to use the idea of bootstrapping. Using 250,000 cases, the steps of a bootstrapping process, listed in the methods section, were completed. Produced was the ability to see the difference in mean gain scores saw in Table 2, based on the assumption that the null hypothesis, there is no statistical difference, was true. Table 5 shows the resulting p-values from bootstrapping and using a t-test. However, using a t-test is not recommended in this scenario as quite a few built-in assump-

tions are violated. For a visual representation, the difference in mean gain scores for Group A and Group B compared with the differences of all 250,000 iterations of pseudo-group A and pseudo-group B, a histogram is a plot with a line corresponding to the difference in mean gain scores for Group A and Group B.

Table 3: Mean Gain Scores for Pseudo-Groups (randomly generated groups)

Out[•]=

	Mean Gain
Pseudo–Group A	0.5197
Pseudo–Group B	0.5181

Table 4: Difference in Mean Gain Scores for Pseudo-Groups

Out[•]=

	Difference in Mean Gain Scores for PsuedoGroups
Mean(pseudo–group A) – Mean(pseudoGroup B)	0.00167

Figure 2: Differences in Mean Gains Distribution for 250,000 Pseudo-Groups and a line which represents the differences in mean gain scores of Group A and Group B

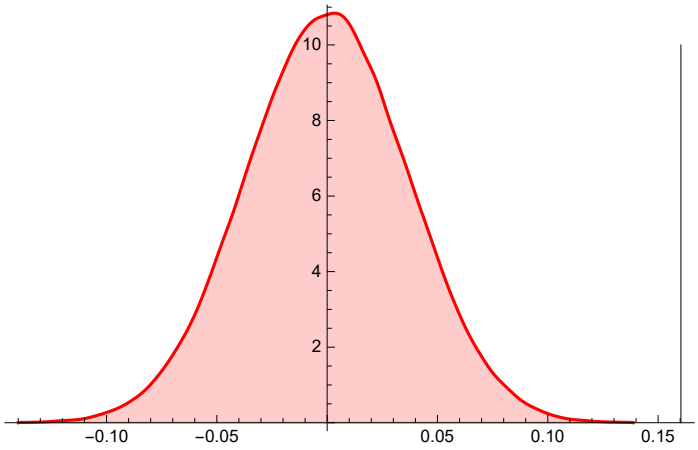


Table 5: P-Value Chart

Out[•]=

P–Value (Bootsrapping)	0
P–Value (T–Test)	3.64e–06

Second Study

The second study consists of pre-test and post-test scores of two groups of students who experienced different methods of teaching. The students who were taught using the traditional method scored a mean gain score of 0.204 while the pilot learning group had a mean score of 0.246. Therefore, the students who experienced the pilot learning method have a larger gain in their post-test score on

average. The median of the traditional learning group is approximately 0.19 wheres and the median of the pilot learning group is about 0.24. A similar conclusion as the means can be made with the given medians. The standard deviation of gain scores of the student in the traditional learning group is 0.1776 whereas the gain scores of students in the pilot learning group have a standard deviation of roughly 0.164. Although the standard deviations are not equal and the variances are not equal, these standard deviations are similar, so the results are not as questionable as the previous study. The skewness of the gain scores of students in the traditional learning group is 0.235 which is positively skewed and similarly, on the other hand, the gain scores of students in the pilot learning group are also positively skewed with skewness of 0.122. Lastly, the kurtosis of each group is reported. For the gain scores of students in the traditional learning group, the kurtosis is 3.225, but for the gain scores of students in the pilot learning group, the kurtosis is 3.439. Unlike the previous study, these values do not seriously deviate from the assumptions of a t-test, therefore, the results from a t-test will not be completely inaccurate. Table 7 highlights the difference in mean gain scores of students in the traditional learning group and students in the pilot learning group.

Table 6: Descriptive Statistics of Gain Scores for Students in the Traditional Learning Group and the Pilot Learning Group

Out[•]=

	Traditonal Learning Group Gain Score	Pilot Learning Group Gain Score
Length	93	104
Mean	0.2043	0.246
Median	0.1944	0.2437
Standard Deviation	0.1776	0.1641
Skewness	0.2355	0.1215
Kurtosis	3.225	3.439

Figure 3: Smooth Histogram of Gain Scores for Students in the Traditional Learning Group and the Pilot Learning Group

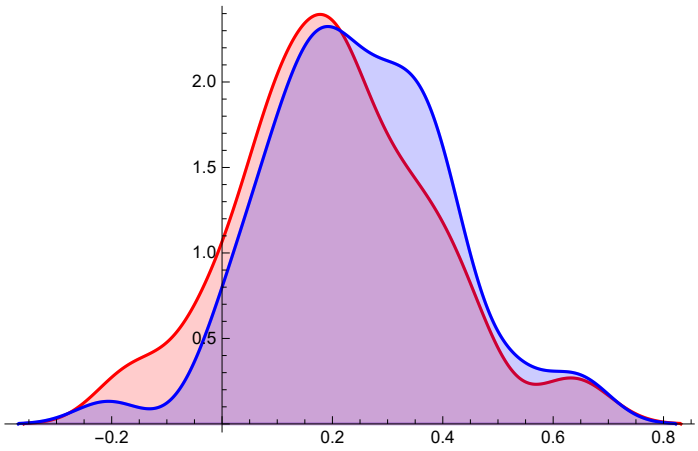


Table 7: Difference in Gain Scores for Students in the Traditional Learning Group and the Pilot Learning

Group

Out[•]=

	Difference in Mean Gain Scores
Mean(Traditional Learning Group) – Mean(Pilot Learnign Group)	−0.0423

Bootstrapping

Similar to the previous study, a bootstrapping method was used to find the probability of observing the difference in mean gain scores for the students of the traditional group and pilot group that we see if the null hypothesis, there is no statistical difference in the mean gain scores, were true. Using 250,000 cases, the steps of a bootstrapping process was completed. Figure 4 and Table 8 offer insight into the mean gain scores of randomly generated groups. Table 10 shows the resulting p-values from bootstrapping and using a t-test. Unlike the previous study, there are likely fewer violations of the built-in assumptions that correspond to a t-test. However, to validate the use of a t-test, independence must be tested. For a visual representation, the difference in mean gain scores for students in the traditional learning group and the pilot learning group compared with the differences of all 250,000 iterations of the pseudo-groups a histogram is plotted with a line corresponding to the difference in mean gain scores for students in the traditional learning group and the pilot learning group.

Table 8: Mean Gain Scores for Pseudo-Groups (randomly generated groups) for Second Study

Out[•]=

	Mean Gain
Pseudo–Traditional Learning Group	0.239
Pseudo–Pilot Learning Group	0.215

Table 9: Difference in Mean Gain Scores for Pseudo-Groups for the Second Study

Out[•]=

	Difference in Mean Gain Scores
Mean(pseudo–Traditonal Learning Group) – Mean(pseudo–Pilot Learning Group)	0.0246

Figure 4: Differences in Mean Gains Distribution for 250,000 Pseudo-Groups
and a line which represents the differences in mean gain scores of
Students in the Traditional Learning Group and the Pilot Learning Group

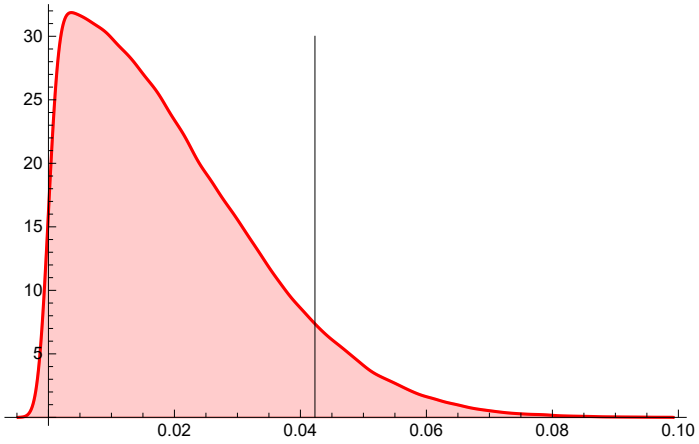


Table 10: P-Value Chart for Second Study

Out[•]=

P–Value (Bootsrapping)	0.0843
P–Value (T–Test)	0.0838

Discussion and Conclusion

Educators are interested in investigating different teaching methods and understanding the impact each method has on his/her students. To understand the effects of different teaching methods, educators are interested in using basic inference. By comparing the average difference of gain scores for the different groups, educators will not necessarily prove that the two teaching methods generate a statistically significant difference in average gain scores. Instead, probabilistic evidence as to whether the difference in mean gain scores between two groups of students will be provided to conclude whether or not it is highly likely that the learning methods will generate statistically similar gain scores. When analyzing the results provided above, it is important to consider how pilot learning and traditional learning are currently implemented into curriculums and why there may or may not be a statistical difference between the mean gain scores.

The first study investigates a cohort of 155 pre-service teachers taking a college mathematics course. As mentioned earlier, these pre-service teachers were divided into two groups based on how their pre-test scored compared with the median pre-test score for the entire cohort of 155 students. Using the gain equation presented in the methods section, the gain scores were found for pre-service teachers with a pre-test score below the median and then separately for pre-service teachers with a pre-test score at or above the median. Comparing the descriptive statistics for the mean gain scores is shown in Table 1; the mean of Group A is significantly larger than the mean for Group B. Additionally, the median of Group A is larger than that of Group B. When investigating for a statistical difference, since the mean and median of Group A is larger than that of Group B, there is a possibility for statistical difference between the two groups. The proportion of Group A data for which the gain lies above the median gain

score for Group B is 0.849. Comparatively, the proportion of Group B for which the gain lies below the median gain of group A is 0.707. If the distributions were similar in shape, this calculation would yield around 0.5. Since both of these proportions are higher than 0.5, there is a larger indication that the distribution of gain scores for Group A has shifted right of the median for the Group B gains, and the distribution for Group B has shifted left of the median for the Group A gains. Furthermore, the standard deviation, skewness, and kurtosis are additional parameters that can be used for the comparison of the two groups of gain scores. The standard deviation offers little insight on how statistically similar the mean gain scores of each group are, however, the t-test relies on a similar variance as a built-in assumption. This is clearly broken as demonstrated in Table 1. The skewness of the two groups of gain scores are opposite, that is the skewness of group A (the gain scores for pre-service teachers with a pre-test score less than the median pre-test score of the entire cohort of students) is positive and skewed right, whereas the skewness of group B (the gain scores for pre-service teachers with a pre-test score at or above the median pre-test score of the entire cohort of students) is negative and skewed left. Lastly, although the distributions look different, the kurtosis of each distribution is similar. Both of these distributions vary quite significantly from a normal distribution. The high peak of Group A and the long tails of Group B corresponds to a kurtosis that slightly varies from a normal distribution. While the kurtosis and skewness of the groups can be compared and questioned to determine if there is a statistical difference between the mean gain scores of Group A and Group B, these parameters also offer insight as to why the built-in assumptions of a t-test are violated. The kurtosis and skewness offer quantitative parameters to demonstrate how these distributions do not follow a normal distribution. Figure 1 offers a visual representation of how the groups may have a significant difference in mean gain scores. Based on the descriptive statistics presented, there some evidence that the mean gain scores of the groups may be statistically significant. Also, these histograms demonstrate how the distribution varies from a normal distribution and no longer follows the assumptions of a t-test. It is important to consider the length of each group, as the pseudo-groups are groups containing randomly selected pre-service teacher scores to the same length of Groups A and B. It is further demonstrated how significant the difference between the mean gain scores is in Table 2. Assuming there was no difference, how likely is it that we will see a statistical difference the way we see it? Based on the null hypothesis, there is a very low likelihood that we would see that the distributions would be like this.

To ascertain if the means for the two groups of data are statistically significantly different, bootstrapping was used. For reasons mention above, the built-in assumptions of a T-test were violated which, if used, would produce questionable results. The method section of this report thoroughly describes the process of bootstrapping and the assumptions of a t-test. To determine if there is a statistically significant difference in mean gain scores of Group A and Group B, there must be a comparative study to determine the likelihood of the difference in mean gain scores we achieved. The question is asked, “Under the null hypothesis that there is no statically significant difference in the mean gain scores, how likely is it we would see the difference we see?” After utilizing the bootstrapping method and the mean difference of gain scores between pseudo-group A and pseudo-group B is found 250,000 times, the number of cases where the difference in mean gain score of the pseudo-groups was smaller than the

difference in mean score between Group A and Group B was 250,000. That is the probability of seeing the difference in mean gain scores between Group A and Group B assuming there was no statistical difference is 0. Since the difference in mean gain scores of Group A and Group B is so large, the likelihood of seeing this difference in gain scores of randomly generated data is 0. Therefore, there is probabilistic evidence to reject the null hypothesis and assume there is a statistical difference in gain scores between the two groups. Table 5 offers insight into the different p-values for the bootstrapping method compared to the t-test. As can be seen, the difference in results will still yield the same claim to reject the null hypothesis. For example, using bootstrapping, the p-value suggests that it is nearly impossible to get the suggested difference in means in Table 2 assuming the null were true. However, the t-test p-value is close to the p-value from bootstrapping value, but instead of having a likelihood of landing close to infinite heads in a row, the t-test p-value suggests the difference of means achieved (Table 2) given the null hypothesis were true is equivalent to scoring 18 heads in a row. Therefore, we reject the null hypothesis since there is probabilistic evidence that the mean gain of Group A is statistically significantly greater than the mean gain for Group B.

The second study consists of mathematics pre-test and post-test scores for the cohort of students in the mid-west university. At this university, students were split into two groups and each group experienced a different method of learning. Some students were taught using a traditional method which focused heavily on the lecture-style method using existing texts while the pilot group was taught with a modified text that emphasized greater student engagement. The goal of the second study was to determine if there is a statistically significant difference in the mean gain scores between the traditional and pilot groups. As in the previous study, the descriptive statistics of the distributions can be utilized for comparison amongst the groups as well as considering whether the gain scores for the two groups are statistically different. The mean gain scores vary much less than in the previous study. Table 3 lists the descriptive statistics for the gain scores of the two groups of students. The difference in mean gain score is only -0.0423 which does not provide evidence to reject the null hypothesis. The median for the traditional learning group is roughly 0.19 while the median for the pilot learning group is 0.24. There is a significantly larger difference between the median of the gains of each group compared to the mean, however, this still provides little evidence as to whether there is a statistically significant difference in the gain scores. The proportion of students in the traditional learning group for which whose gain scores gain lie above the median gain score for students of the pilot learning group is 0.586. Analogously, the proportion of students of the pilot learning group for which the gain lies below the median gain of students in the traditional learning group is 0.624. Since both of these proportions are slightly higher than 0.5 but overall close to this value, the distribution of gain scores for students in the traditional learning groups is similar to the distribution of gain scores for students of the pilot learning group. Unlike the previous study, the standard deviations of these groups are similar, hence a similar variance. This follows one of the built-in assumptions of the t-test. The skewness of the data is another parameter of interest. Both the students in the traditional learning group and the pilot learning group had positive, skewed right distributions, however, the traditional learning group had more significant skewness. Both distributions vary from normal distributions. The student who experienced

the traditional form of learning closer resembled a normal distribution, whereas the students in the pilot from of learning group had a slighter larger kurtosis since it is close to a bimodal distribution. In this scenario, the skewness and kurtosis of the distribution follow more closely to a normal distribution. However, there is still a deviation from a normal distribution for both of these groups. Therefore, it is unclear if the built-in assumption of a normal distribution for t-tests is violated, however, independence and similarity in the variance would have to be tested. Figure 3 provides some visual evidence that the gain scores for the traditional and pilot learning group may be similar, and little statistical difference exists. As of now, this is all speculation, as no comparison study has yet to be conducted.

Using bootstrapping yet again, the probability of observing a difference in the mean gain scores that were seen under the null hypothesis (Table 3) is estimated. Table 7 provides the means for each pseudo-group of the traditional learning and pilot learning methods. Additionally, Table 8 provides the difference in means of the pseudo-groups. By comparing the difference of multiple pseudo-groups, and then seeing where the difference in mean gain scores lies within this distribution. When analyzing the difference in mean gain score of the pseudo-groups and comparing these results with the mean gain scores for the actual groups 250,000 times, the number of cases where the difference in mean gain score of the pseudo-groups was smaller than the difference in mean score between the traditional learning group and pilot learning group was 228,918. The probability of seeing this difference in mean scores between the traditional learning group and the pilot learning group under the null hypothesis with no statistical difference is approximately 0.0843. Table 10 also offers insight into the different p-values for the bootstrapping method compared to the t-test for the second study. In this scenario, both methods will yield a p-value that corresponds to not rejecting the null hypothesis. For example, using bootstrapping, the p-value suggests that there is roughly an 8% chance of the suggested difference in means in Table 7 assuming the null were true. The t-test p-value is close to the p-value from bootstrapping value since the built-in assumptions were not seriously violated. The t-test p-value suggests the difference of means achieved (Table 7) given the null hypothesis were true is equivalent to scoring 4 heads in a row, the same as the bootstrapping method. Hence, when analyzing these results, there is not enough probabilistic evidence to reject the null hypothesis.

References

[1] ARCHAMBAULT, L., WEST, R. E., BORUP, J., & . LOWENTHAL, P. R. (2020). Thinking Beyond Zoom: Using Asynchronous Video to Maintain Connection and Engagement During the COVID-19 Pandemic. Journal of Technology and Teacher Education, 28(2)

[2]Dixson, M. D. (2010). Creating effective student engagement in online courses: What do students find engaging? Journal of the Scholarship of Teaching and Learning,, 10(2)

[3]The important assumptions for using the t-test. (n.d.).

[4]Gilbert, B. (n.d.). Online Learning Revealing the Benefits and Challenges. Fisher Digital Publications | St. John Fisher College Research.