Under Pressure:

Are your peers impacting your grade?

A study on the effects a peer proctor has on the mathematical performance of a test taking student.

Table of Contents:

I. Introduction

We are trying to determine the effect of peer observers on preparatory high school students' performances on a simple mathematical quiz. Students at independent boarding high schools find themselves in a high pressure environment that requires them to excel both socially and academically. Oftentimes, those two things are interconnected; there may be social consequences to high academic performance, and vice versa. We chose this topic because we wanted to investigate whether social pressure has an immediate influence on academic performance. Do prep school students strive harder and perform better when they are being observed by peers? Or do they become nervous and falter under pressure? We chose a basic arithmetic quiz, comprised of addition, subtraction, multiplication, and division calculations, because we wanted all the students to be capable of answering each question. The number of questions completed on the arithmetic quiz within a set amount of time provided us with quantitative matched-pairs data that we used to compare a student's academic performance alone and in the presence of peer observers. We also collected demographic data, like gender, grade, and math level, to see if different demographics responded differently to peer pressure.

II. Data Collection

Collection

In order to investigate this question, we conducted an experiment with a sample size of 31 volunteers from the student body of an independent boarding high school in the northeast. Students were approached in the school dining hall during lunch periods and asked to participate in a statistical study. When receiving responses from students, we conducted the experiment in a quiet room of the dining hall to control for distractions and removed all calculators and paper to control for unfair advantages. We surmised that students could be affected by sounds and supplementary resources.

Based on the results of a coin flip, the volunteers were randomly assigned to one of two Groups, each of which received a different order of treatment. Subjects who flipped heads were assigned to treatment Group A, and those who flipped tails were assigned to treatment Group B. Students in each group were asked to take a digital test on basic arithmetic, composed of basic addition, subtraction, multiplication, and division problems (refer to appendix for the online test). For addition and subtraction, the range is (2 to 100) ± (2 to 100). For multiplication, the range is (2 to 9) x (2 to 50). For division, the range is (2 to 50) / (2 to 9). The test gave each student 120 seconds to solve as many questions as possible, and it required that the taker answer each question correctly before moving onto the next question. Each test was administered individually, with experimental units receiving treatments one at a time.

Students assigned to Group A first took the test without a proctor in the room while Group B first took the test with proctors in the room monitoring the process. After the first test was finished for both groups, each group was given a new randomized assortment of questions and prompted again to solve as many correctly as possible within the same time frame. Group B took the second test without proctors and Group A took the second test with proctors. The number of questions completed in a fixed amount of time for each participant's unproctored and proctored test were recorded as scores, and the difference between the two scores were calculated (Test score without proctors - Test score with proctors).

To reduce the possibility of confounding variables and to control the experiment, all testing occurred in the same room on the same computer. Throughout the course of the experiment, participants were read the same script, which is included below. Additionally, the proctors were composed of the same two student peers- a girl and boy- for the entirety of the experiment. After completing testing, each experimental unit was given a short survey to fill out, consisting of questions about gender, grade level, math level, and other aspects of identity (see Appendix).

When constructing the experiment, we did consider selecting participants through a simple random sample. However, because it was likely that randomly-selected students would not be willing to participate and would thus incur non-response bias, we decided to take volunteers. In order to maintain randomness, we randomly sorted volunteers into two treatment groups. In addition, each student was compared only with themselves, as we designed the experiment with a matched-pairs test in mind. It is also worth noting that all volunteers agreed to participate in the study before they were informed of their task, so any bias present from self-selection was minimized. The decision to take volunteers expedited the experiment and allowed us to perform treatments on more experimental units, increasing our sample size. We surmised that our sample size of 31 was large enough to control for chance variation and ensure replication.

Script

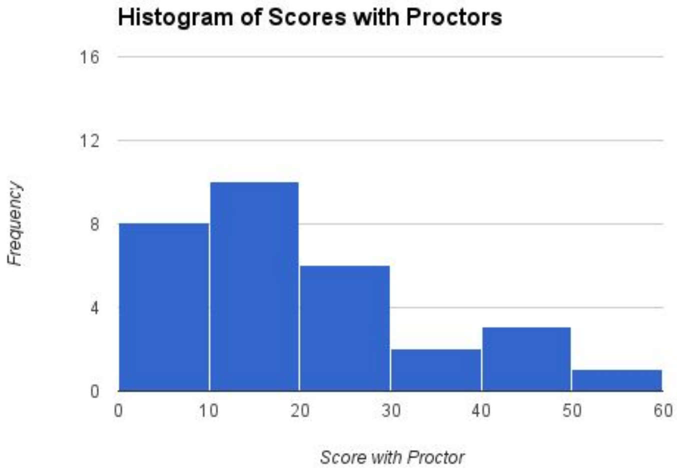
<u>Group A</u>

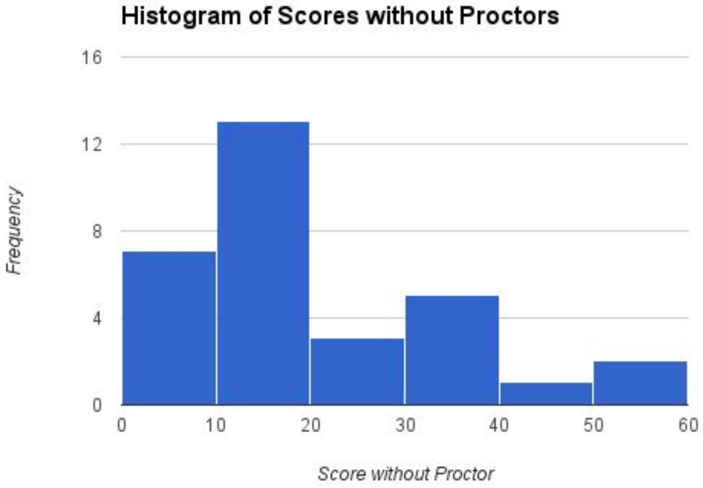
"Your task is to complete this mathematical quiz and get as many correct answers in the allotted time period, which is 120 seconds. In order to move on to the next question, you must answer the question correctly. When you are finished, call us in and we will record your score. (Person takes the test, we come in). You will now take the mathematical quiz again, and once again you will try to answer as many questions correct as you possibly can. We will then record your score again and ask you to fill out a short survey. After that you will be free to go. Thank you."

<u>Group B</u>

"Your task is to complete this mathematical quiz and get as many correct answers in the allotted time period, which is 120 seconds. In order to move on to the next question, you must get answer the question correctly. (Person takes the first test) We will now leave the room and once we have left, take the test again. Once again you will try to answer as many questions correct as you possibly can. When you are finished, call us in and we will record your score. After that, we will ask you fill out this short survey, you will be free to go. Thank you."

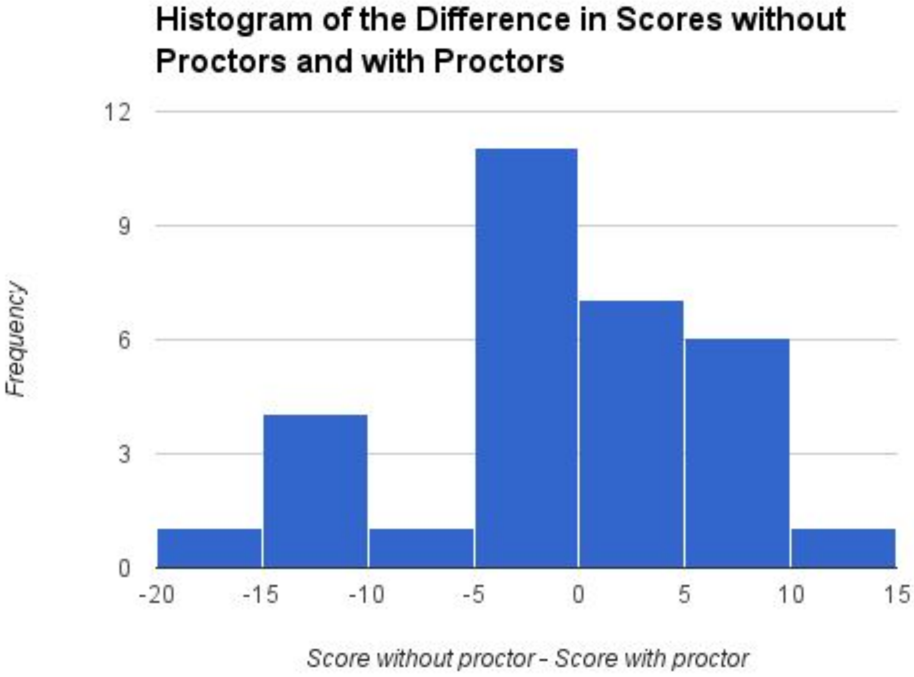
III. Data Display



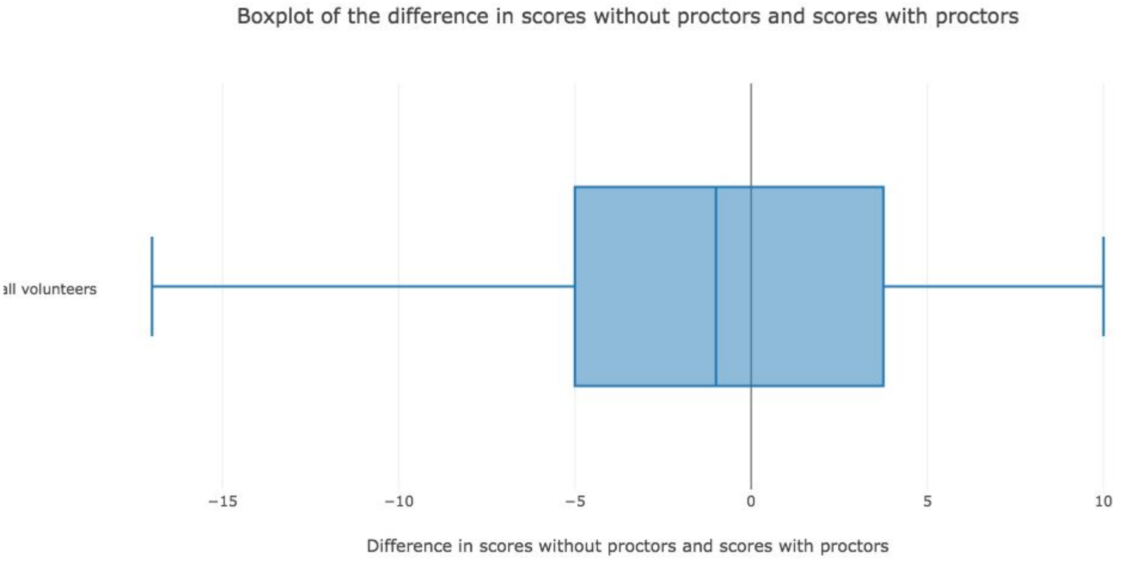


As seen in the histograms above, the overall shapes of both distributions are very similar: their centers are both in between 10 and 20, they both are skewed to the right, are roughly unimodal, and the variability of the two distributions is also roughly equal. However, the distribution of scores without proctors has one outlier of a score of 56 while the distribution of scores with proctors has no outliers. These graphs by themselves do not indicate a major difference between the two distributions.

Therefore, in order to further analyze the data, we took a histogram of the difference in scores.

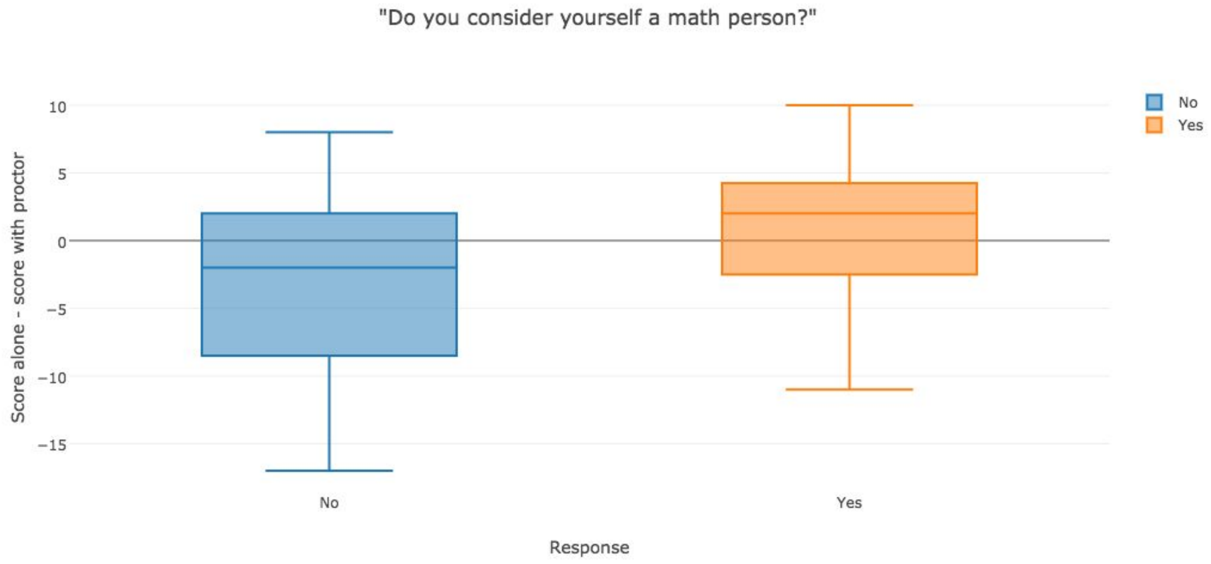


From this histogram, we can see that the distribution of the difference in scores without proctors and scores with proctors appears to be roughly symmetric, with a gap in the data occurring in between -10 and -5 points. The distribution is unimodal, with the most participants experiencing a difference in scores within the range of -5 to 0. The median of the distribution is at -1, reflecting the fact that a slight majority of the participants, 17 out of 31 or 54.83% percent had a negative difference in scores, meaning that they performed worse with proctors in the room.



Entire Sample
Sample Size: 31 Median: -1 Minimum: -17 Maximum: 10 First Quartile: -5 Third Quartile: 4 Interquartile Range: 9 Outliers: None

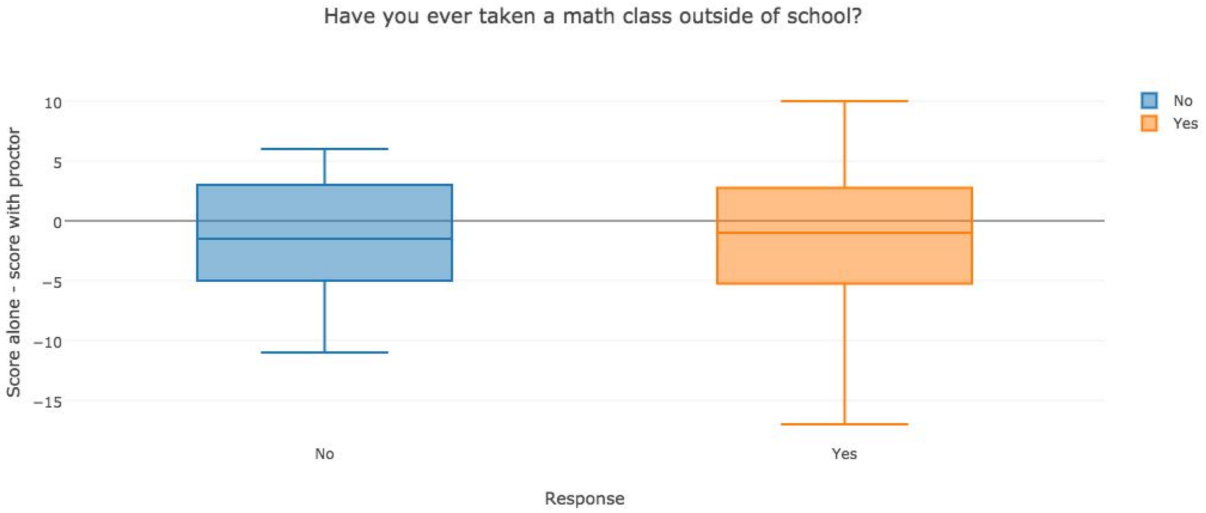
Another way of representing the distribution of the difference in scores without proctors and scores with proctors is through a box plot, shown above. The distribution shows some left skew, with the middle 50% of the data falling between -5 and 4 (Med=-1), and the first 25% falling between -17 and -5. The first quartile range of 12 is greater than the interquartile range of 9 (and certainly greater than the range of any other quartile), which tells us that the extreme negative data is more spread out and less concentrated than the rest of the data.



For “non-math” person	For “math” person
Sample Size: 20 Median: -2 Minimum: -17 Maximum: 8 First Quartile: -8.5	Sample Size: 11 Median: 2 Minimum: -11 Maximum: 10 First Quartile: -4

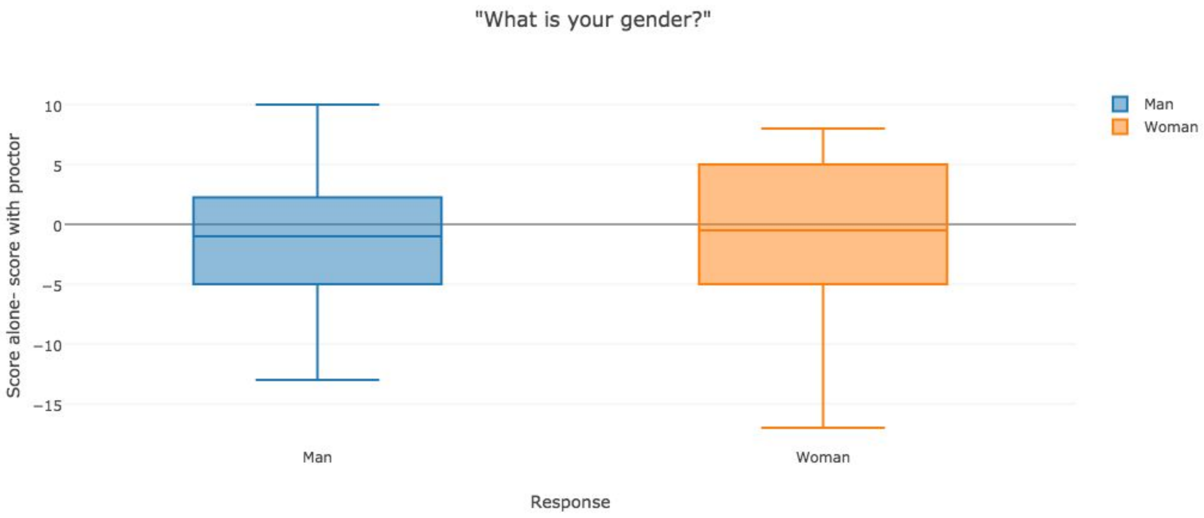
Third Quartile: 2 Interquartile Range: 10.5 Outliers: None	Third Quartile: 5 Interquartile Range: 9 Outliers: None

The box plot above shows that the distribution of the difference in scores (without proctor- with proctor) for people who self-identified as “math” people is approximately normal while that of self-identified “non-math” people is quite left-skewed. The center of the distribution of the difference in scores for “math” people (Med=2) is higher than that of “non-math” people (Med=-2). There is more variability in the difference in test scores for “non-math” people (Range=25, IQR=10.5) than of “math” people (Range=21, IQR=9).



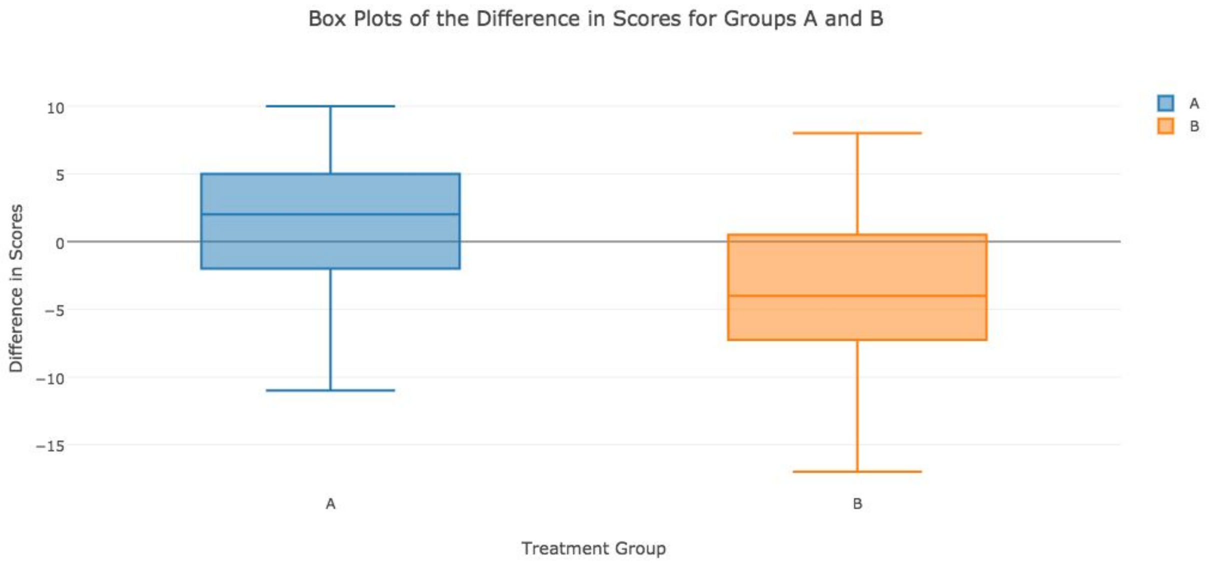
For people who have not taken a math class outside of school	For people who have taken a math class outside of school
Sample Size: 14 Median: -1.5 Minimum: -11 Maximum: 6 First Quartile: -5 Third Quartile: 3 Interquartile Range: 8 Outliers: None	Sample Size: 13 Median: -1 Minimum: -17 Maximum: 10 First Quartile: -5.25 Third Quartile: 2.75 Interquartile Range: 8 Outliers: None

For people who have taken a math class outside of school, the distribution of the difference in scores (without proctors-with proctors) is skewed to the left, while that of people who have not taken an extra-scholastic math class is quite normal. The medians of both distributions are not much different: the medians were -1.5 for people who have taken a math class outside of school and -1 for the distribution of people who have not taken a math class outside of school. Both distributions have similar variabilities since they both have the same interquartile ranges (IQR=8).



For men	For women
Sample Size: 17 Median: -1 Minimum: -13 Maximum: 10 First Quartile: -5 Third Quartile: 2.25 Interquartile Range: 7.25 Outliers: None	Sample Size: 17 Median: -0.5 Minimum: -17 Maximum: 8 First Quartile: -5 Third Quartile: 5 Interquartile Range: 10 Outliers: None

These box plots show that for men, the distribution of the difference in scores when a subject did the test without and with the proctors is approximately normal. However, for women, the distribution of the difference in scores when a subject did the test without and with the proctors is strongly left-skewed (negative). There are more women who scored worse with proctors than men. The medians of both distributions are not much different (Med_{men} = -1 and Med_{women} = -0.5), and the variability of the distribution of difference of scores for men is similar to that for women ($Range_{men} = 23$ and $Range_{men} = 25$).



From the box plots we can see that subjects in group A, who took the test without a proctor first, performed substantially better than subjects in group B, who took the test with a proctor first. Both distributions are roughly normal, and the median score for subjects in group A is substantially higher (and positive) compared to the score for subjects in group B ($Med_{Group\ A}$ = 2 and $Med_{Group\ B}$ = -4). The variability of both distributions are similar, group subjects in group A having a slightly higher variability ($Range_{Group\ A} = 25$ and $Range_{Group\ A} = 21$ and $IQR_{Group\ A} = 7.5$ and $IQR_{Group\ B} = 7$).

IV. Conditions for Inference

1. Random: While our experimental units were volunteers, and not randomly selected, they were randomly assigned to two groups who took the treatment in different orders. Thus, the data meets the random condition. Additionally, all the volunteers agreed to participate in the study before they were informed that they would be taking a math quiz, so any bias present from self-selection was minimized.

2. The second condition that had to be satisfied was the 10% condition. Because our sample size of 31 students is much less than 1/10th of the total student body population of 1,131, we can assume independence. 31 < 113.1

3. Normal: The distribution of the difference between score without proctors and score with proctors is roughly symmetric with no outliers and an only slight right-skew. The Normal condition requires an experiment to have a large enough sample size to apply inference procedures to it, because a distribution taken from a large enough sample will be very close to normal. Our sample

size of 30 is large, so we can use the Central Limit Theorem to assume for the purposes of our study that the sampling distribution of the sample mean is approximately normal.

V. Inference Procedures

We will be taking a matched pairs t-test at a 5% significance level for the mean of the difference in scores without proctors and with proctors to determine whether or not there was a significant difference in the mean score between a test taken without proctors and one taken with proctors from the population of students at this independent boarding school. We checked the conditions previously, so we will define our hypotheses and proceed with inference.

H_0: $\mu_{\text{without proctors-with proctors}} = 0$
($\mu_{\text{without proctors-with proctors}}$ represents the true mean of the difference in test scores when students take the test without proctors and when they take it with proctors)
Our null hypothesis states that the true mean of the difference in test scores when students take the test without proctors and when they take it with proctors is zero, or in other words the presence of a peer proctor has no impact on test performance.

H_a: $\mu_{\text{without proctors-with proctors}} \neq 0$
Our alternative hypothesis states that the true mean of the difference in test scores when students take the test without proctors and when they take it with proctors is not equal to zero, or in other words the presence of a peer proctor has an impact on test performance.

Test statistics:

$$t = \frac{\bar{x} - \mu}{s_x / \sqrt{n}}$$

$$t = \frac{-1.387-0}{6.627/\sqrt{31}} = -1.165$$

Degrees of Freedom: 30
P-value = 2 * P(t < -1.092) = 0.284

Since the p-value of 0.284 is greater than the significance level of 0.05, we fail to reject the null hypothesis and conclude that the mean difference in test scores when students take the test without proctors and when they take it with proctors is zero, suggesting that the presence of a peer proctor has no significant impact on student performance.

Since our initial investigation proved inconclusive, we performed several other two-sample t-tests by the demographic variables that we collected using our survey in order to investigate potential other avenues of inquiry.

First, we conducted a two-sample t-test for the difference of the means at 5% significance level to determine whether there was a significant difference in the mean of the difference in scores between students who self-identified as a "math" person and people who did not self-identify as a "math" person. This group still fulfills the random condition as explained above, as well as the 10% or independent condition. However, we cannot assume normality as both groups do not fulfill the central limit theorem since the sample sizes for both groups are less than 30. The box plots aforementioned show that both of the distribution of the difference in scores when a subject did the test without and with the proctors for people who self-identify as "math" people is approximately normal while that for those who do not self-identify as "math" people is quite left-skewed**.** Despite failing to fulfill the central limit theorem, we will progress with caution.

H_0: $\mu_{\text{non-math}} = \mu_{\text{math}}$
($\mu_{\text{non-math}}$ represents the true mean of the difference in test scores when students take the test without proctors and when they take it with proctors for people who do not self-identify as "math" people)
(μ_{math} represents the true mean of the difference in test scores when students take the test without proctors and when they take it with proctors for people who self-identify as "math" people)
Our null hypothesis states that the difference in the means of the difference in test scores where students self-identify as "math" people or do not is zero, or in other words whether or not one self-identifies as a "math" person does not influence their performance.

H_a: $\mu_{\text{non-math}} < \mu_{\text{math}}$
Our null hypothesis states that the difference in the means of the difference in test scores where students self-identify as "math" people and the means of test scores of students who do not is less than zero, or in other words if one self-identifies as a "math" person they will score better than someone who does not.

Test statistics: $t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$
$t = \frac{-3 - 0.533}{\sqrt{(7.16^2/16) + (5.55^2/15)}}$
$t = -1.54$
Degrees of Freedom: 28.045
P-value = P(t<-1.54) = 0.067

Since the p-value of 0.067 is greater than the significance level of 0.05, we fail to reject the null hypothesis and conclude that the difference in the means of test scores where students self-identify as "math" people or do not is zero, suggesting that whether or not people self-identify as a "math" person does not influence their scores. While our p-value was not small enough for us to reject the null, it was very close. Next time, a larger sample size would allow us to either be more confident in failing to reject the null, or increase our power and push our p-value below the significance level.

Next we conducted a two-sample t-test for the difference of the means at 5% significance level to determine whether there was a significant difference in the mean of the difference in scores between students were assigned to group A or group B. This group still fulfills the random condition as explained above, as well as the 10% or independent condition. However, we cannot assume normality as both groups do not fulfill the central limit theorem as the sample sizes for both groups are less than 30. Despite failing to fulfill the central limit theorem, the box plots indicate that the distributions are roughly normal so we will progress with caution.

H_0: $\mu_{\text{group A}} - \mu_{\text{group B}} = 0$

($\mu_{\text{group A}}$ represents the true mean of the difference in test scores between when students take the test without proctors before they take it with proctors)

($\mu_{\text{group B}}$ represents the true mean of the difference in test scores between when students take the test with proctors before they take it without proctors)

Null hypothesis: The difference of the means in the difference in test scores between when students take the test without proctors before they take it with proctors and when students take the test with proctors before taking it without proctors is 0.

H_a: $\mu_{\text{group A}} - \mu_{\text{group B}} > 0$

Alternative hypothesis: The difference of the means of the difference in test scores between when students take the test without proctors before they take it with proctors and when students take the test with proctors before taking it without proctors is greater than zero, meaning that if you take the test first without proctors you perform better than those who took the test with proctors first.

Test statistics: $t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$

$$t = \frac{1.571 - (-3.647)}{\sqrt{(5.19^2/14) + (6.79^2/17)}}$$

$t = 2.4216$

Degrees of Freedom: 28.045

P-value = P(t > 2.4216)= 0.01098

Since the p-value of 0.01098 is less than the significance level of 0.05, we reject the null hypothesis and conclude that there is a significant improvement in score between students that took the test without proctors first versus students that took the test with a proctor first. For this study, people who took the test without proctors first tend to perform better than those who took it with proctors first. This is a potential area of future study.

VI. Conclusion:

The above analyses of the differences between students' scores without proctors (peer observers), and their scores with proctors were divided into 3 different procedures: a t-test for the mean matched-pairs difference for the total sample size, a two-sample t-test for the difference in means between self identified "math" and "non-math" people, and another two-sample t-test for the difference in means between treatment groups A and B. Although we also collected other demographic data, which are included in the Data Display section, preliminary examination of those box plots led us to believe that any differences would be minimal. As a result, we did not pursue inference on those sets of data. In both the "math" person and treatment group procedures, we split up our 31 experimental units into two groups. Because of that, each of the sample sizes was less than 30, and we did not have large enough sample sizes to use the Central Limit Theorem to assume normality. Investigation of the box plot distributions of each set of data, however, led us to believe that the data were approximately normal and we proceeded with caution. After completing this analysis we can conclude that there is not a significant mean difference in test scores between students who take the test without proctors and students who take it with proctors. This suggests that the presence of a peer observer has no significant impact on student performance on a simple arithmetic quiz at an independent boarding school in the northeast. That said, there is a much smaller, though still insignificant, p-value for the two-sample difference in means for "math" people and non-"math" people. With further study and larger sample sizes, inference on this demographic may serve to either confirm our findings or push the p-value below the significance level of 0.05, allowing rejection of the null. The most conclusive inference procedure was the two-sample t-test for the difference in means between treatment groups A and B. We find that there is a convincing evidence that the order of treatments affects student performance on the tests. In the study, people who take the test without proctors first have a greater difference in scores without proctors and with proctors than those who take it with proctors first.

VII. Changes and Ideas for Further Study

Something that we would have liked to change about our experiment is the number of participants we were able to collect. We were limited by time, as it took a pretty significant amount of time to test each subject, and we were only able to test one subject at a time. This meant that we could not randomly select who would participate in our experiment. Even though we acquired enough participants to satisfy the Central Limit Theorem for the overall inference procedure, those participants were not equally distributed by grade, gender, or many of the other characteristics. With our relatively modest sample size of 31, our experiment was exposed to some chance variation in the coin flips and the demographics. For example, the majority of the group B flips were men. If we had tried to perform a two-sample inference procedure on the difference in means between men and women, we would not have known whether the difference was actually the result of gender, or of uneven distribution of treatment groups across genders. In an ideal experiment, the proportions of grades, gender, race, etc. would be closer to reflecting that of the population, and the results of the coin flip would be equally distributed across those demographics. This would have resulted in less bias and fewer confounding variables.

One of the limiting factors of this test was time. Each subject took a minimum of six minutes to complete both tests with interludes for scripted instructions, and this did not include the time it took to find the volunteer. Because we wanted to keep the same proctors for each test, we could only analyze one subject at a time. The efficiency of our test was quite low. If we were to do a further study with unlimited amounts of time and resources, we could have collected more experimental units. The increase in the number of subjects would have resulted in more normality and a less variability in the sampling distribution, which could have led to better and more clear conclusions with less confounding variables.

VIII. Appendix

Figure 1

Subject Number	Group	Score - With Proctor (Test I)	Score - No Proctor (Test II)	Difference (Test II Score - Test I Score)
1	A	37	43	6
2	B	8	16	8
3	B	16	12	-4
4	B	39	46	7
5	A	14	18	4
6	B	17	12	-5
7	A	12	18	6
8	B	23	23	0
9	A	11	21	10
10	B	50	37	-13
11	B	34	33	-1
12	B	9	11	2
13	A	11	8	-3
14	B	21	15	-6
15	B	13	16	3
16	A	17	22	5
17	A	12	14	2
18	B	30	27	-3
19	B	16	11	-5

20	A	41	43	2
21	B	8	3	-5
22	A	9	8	-1
23	B	20	9	-11
24	A	6	4	-2
25	B	57	56	-1
26	B	38	21	-17
27	A	4	5	1
28	A	17	6	-11
29	A	16	21	5
30	A	8	6	-2
31	B	17	6	-11

Figure 2

Subject Number	Grade	Gender	Current Math Class	Ethnicity
1	Lower	Woman	350	Chinese
2	Upper	Woman	360	Chinese
3	Senior	Man	-	-
4	Freshman	Man	-	-
5	Senior	Woman	-	-
6	Lower	Man	-	-
7	Senior	Woman	590	-

8	Freshman	Man	100	Asian
9	Lower	Man	350	European
10	Lower	Man	380	Chinese
11	Freshman	Man	595	Korean
12	Senior	Man	590	White
13	Lower	Woman	320	Trinidadian/Guyanese
14	Upper	Man	560	Hispanic
15	Upper	Man	560	White
16	Lower	Woman	350	Black
17	Senior	Man	575	White
18	Upper	Man	360	Black
19	Senior	Man	650	Ukrainian
20	Freshman	Woman	595	Korean
21	Freshman	Woman	100	Irish/Colombian
22	Senior	Man	575	French Canadian
23	Lower	Woman	560	Chinese
24	Lower	Woman	330	Irish
25	Senior	Man	650	Thai
26	Senior	Woman	650	Korean American
27	Senior	Woman	530	French
28	Lower	Man	330	White
29	Lower	Man	580	Chinese
30	Freshman	Woman	210	White
31	Freshman	Woman	320	White

Figure 3

Subject Number	Race	Country of Origin	Do you consider yourself a "math person?"	Math program outside of school?
1	Asian	China	Yes	No
2	Asian	Hong Kong/NYC	No	Yes
3	-	-	-	-
4	-	-	-	-
5	-	-	-	-
6	-	-	-	-
7	Asian / White	UK/USA	No	No
8	Asian	Taiwan	No	Yes
9	Caucasian	USA	Yes	Yes
10	Asian	USA	No	Yes
11	Asian	USA	Yes	Yes
12	Caucasian	USA	Yes	Yes
13	Black / African American	Canada	No	No
14	American Indian	USA	No	Yes
15	Caucasian	UK	No	No
16	Black /	USA	Yes	No

	African American			
17	Caucasian	Poland	Yes	No
18	Black/African American	USA	Yes	Yes
19	Caucasian	Ukraine	Yes	Yes
20	Asian	Korea	Yes	Yes
21	Caucasian	America	No	No
22	Caucasian	USA	No	No
23	Asian	USA	No	No
24	Caucasian	USA	No	No
25	Asian	Thailand	Yes	Yes
26	Asian	USA	No	Yes
27	Caucasian	France	No	No
28	Caucasian	USA	No	No
29	Asian	China	No	Yes
30	Caucasian	USA	No	No
31	Caucasian	USA	Yes	No

Figure 4

Survey for Participants

Grade Level: 9th 10th 11th 12th

Gender: Man Woman Other:____________________

Current Math Class: ______________________________

Ethnicity: ___________________________

Country of Origin:___________________________

Race:

Black/African American
Caucasian
American Indian/Alaska Native Native Hawaiian or Other Pacific Islander
Asian
Mixed: _____________________

Do you consider yourself a "math person?"   Yes   No

Have you ever participated in an outside of school math program?   Yes   No

IX. Bibliography

Alpert, Ben. "Arithmetic Game." Online Speed Drill. Zetamac, 18 July 2006. Web. 24 May 2016. <http://arithmetic.zetamac.com/>.

Girl Stressed While Taking A Test. Digital image. Resource Roundup: Standardized Testing. RemindBlog, 21 Apr. 2015. Web. 24 May 2016. <http://blog.remind.com/resource-roundup-standardized-testing/>.