

BA830 Business Experimentation and Causal Methods

Collaboration Insights: Exploring the Impact of Collaboration on Learning Performance

Kunjingyi Chen (UID: U53672980)
Riris Grace Korolina (UID: U19024662)
Sricharan Mahavadi (UID: U13672047)
Shizuka Takahashi (UID: U04066554)
Yumeng Tang (UID: U90878530)



March 7, 2024

Abstract

Our study hypothesizes that collaborative learning, where individuals engage in shared problem-solving and discussion, leads to better comprehension and retention of information than individual learning. This hypothesis is grounded in the theory that collaboration facilitates deeper engagement with material, as learners can benefit from diverse viewpoints and clarification of complex concepts through peer interaction. The null hypothesis of the study is that the performance of the individual learning approach is the same as the performance of the collaborative learning group. We particularly expect that the advantage of collaborative learning over individual learning will be more evident when participants are challenged with difficult questions, presuming that group discussions might enhance understanding and lead to improved performance. This experiment aims to empirically test this hypothesis, providing insight into the potential educational benefits of collaborative learning environments. Through our study, it is indicated that there is a slightly higher score in the treatment group and more variability in scores among collaborative learners. The Average Treatment Effect (ATE) analysis suggested a moderate, positive impact of collaborative learning on scores. However, due to some limitations through the study, the conclusion suggests a plan for future research to enhance educational practices by exploring various learning materials, increasing sample sizes, and further investigating the nuances of collaborative learning. This approach aims to refine educational offerings and create a more effective learning environment.

Keywords: Collaboration | Group Experiment | Learning Performance

Introduction

The effectiveness of individual versus collaborative learning approaches has been a topic of debate in educational research. Despite the recognized advantages of collaborative learning, there is still controversy about the effectiveness of collaborative versus individual learning, especially in terms of knowledge assessment performance. Some studies have found individual learning to be more successful, while others have asserted that collaborative learning facilitates improved student achievement. Our experiment seeks to contribute to this ongoing debate by examining individual and group performance in a controlled learning environment.

Methodology

Participants

We designed an experiment in which 75 participants were divided into individual learners as the control group and collaborative groups as the treatment groups. The control group consisted of 25 participants who worked individually, while the treatment group had 50 participants who worked in pairs, forming 25 groups.

Procedure

All participants were required to watch the same 5-minute YouTube video titled "27 Facts That Will Make You Question Your Existence", which was related to space and astronomy¹. In the subsequent section, they needed to complete a quiz consisting of 15 questions: 5 easy, 5 moderate, and 5 hard difficulty levels, assessing comprehension of the interesting facts presented in the video. During this section, collaborative groups worked together to answer the post-assessment quiz, while individual groups completed the same quiz independently. The time spent on the quiz was also tracked for all participants. Moreover, we also displayed a 20-minute countdown timer on the screen so that participants could see how much time they spent on the quiz. However, we did not impose a strict time limit, allowing groups to work on the quiz for more than 20 minutes if needed. We will use this information as part of our analysis variables.

Exploratory Data Analysis (EDA)

Our dataset consists of various types including categorical, string, integer, and numeric. Before conducting EDA, we perform data cleansing, which involves dropping some personal information that is not needed in the analysis process and replacing null values.

i. Distribution of Score by Treatment

Based on [Figure 6](#), the mean score is 9.6. Approximately 56% of the participants scored above the mean, while around 44% scored below the mean. From the control and treatment group, we observed that the control group has a median score of around 9, with the middle 50% of scores (the interquartile range) falling roughly between 8 and 10. From the treatment group, we also had a median score of around 9, similar to the control group. However, the interquartile range is slightly broader, extending approximately from 7 to 10, indicating more variability in scores for this group. Overall, from [Figure 7](#), the treatment group has a higher score compared to the control group.

ii. Distribution of Score by Question Difficulty

Based on [Figure 8](#), we observe that the mean number of correct answers is highest for Easy questions, compared to medium and hard-level questions. The participants were more likely to answer Easy questions correctly, somewhat less likely to answer medium questions correctly, and least likely to answer Hard questions correctly. This pattern is what one would typically expect in a well-designed experiment where question difficulty is accurately gauged and reflects the participants' ability to respond correctly based on that difficulty.

iii. Distribution of Collaboration Scores by Treatment Group

We ask the feedback from the participant to rate their collaboration performance during the test. From the result on the on the [Figure 9](#), we find that the distribution of the collaboration scores indicates that both individuals or groups in the treatment group received a collaboration score of

¹ "27 Facts That Will Make You Question Your Existence," YouTube, October 22, 2018, <https://www.youtube.com/watch?v=FkQWpQd9Zdo&themeRefresh=1>

5.0, which could be interpreted as a very positive outcome if a high score is desirable. This suggests that the treatment was effective in promoting collaboration among participants if that was the goal.

Experimental Results

A. Treatment Effect Results

A. i) Average Treatment Effect :Treatment Impact Estimation

Based on scores, the ATE is 0.48, which means the group averaged 0.48 points higher than individuals. Based on time, the ATE is 141.94 equal to 2.37 minutes, which shows the group was on average 2.37 minutes longer in completing the experiment.

A. ii) Conditional Average Treatment Effect: Moderated Treatment Analysis

The Conditional Average Treatment Effect (CATE) analysis revealed heterogeneous treatment effects across gender and age subgroups. Females experienced a slight negative effect (-0.07), while males had a small positive effect (0.18) meaning in a group setting, females scored less compared to males in group setting.. The treatment effect varied substantially across age groups, with the 20-23 and 28-31 groups showing negative effects (-0.37 and -1.00, respectively), while the 24-27 and 32+ groups exhibited positive effects (0.40 and 2.25, respectively). These findings underscore the importance of considering subgroup-specific treatment effects and can inform targeted interventions. However, these results do not hold statistical significance.

A. iii) Power of Experiment : Experimental Power Assessment

For the score-based analysis, the power of the experiment was 0.66, indicating a moderate ability to detect true effects. For the time-based analysis, the power was 1.00, indicating an excellent ability to detect any true effects on time analysis.

A. iv) Proportion Test : Randomization Calculation

To validate the integrity of the randomization process, proportion tests were performed to compare the treatment and control groups across score and completion time measures. The resulting p-values for both outcomes exceeded the conventional 0.05 significance level. This lack of statistically significant differences provides reassuring evidence that the randomization successfully balanced the groups at baseline, supporting the internal validity of the study design and enabling unbiased comparisons between the treatment and control conditions.

A. v) Cohen's D : Effect Size Estimation

For the score analysis, Cohen's d was around 0.22, indicating a small effect size and a small difference between the treatment and control groups. However, for the time analysis, Cohen's d was approximately 0.57, suggesting a large effect size and indicating a substantial difference in completion times between the treatment and control groups.. These effect size measures provide insight into the practical significance of the observed differences between the collaborative and individual learning groups.

B. Regression-Based Analysis

B. i) Analysis of Treatment Impacts on Score Outcomes

Based on the regression treatment on Scores quiz scores, we observed that the control group (treatment = 0) had an average score of 9.360. This figure sets the baseline for comparing the effect of the treatment. When we introduced the treatment, which represents a specific intervention or condition, participants on average scored 0.480 points higher than the control group. However, this slight improvement in scores for the treatment group is not backed by strong statistical evidence; it's not significant enough for us to be confident that the treatment was responsible for the increase. This is reinforced by an R-squared value of just 0.011, which means the treatment explains only about 1.1% of the variation in scores between the participants, indicating that other factors likely play a much larger role. Additionally, an adjusted R-squared of -0.009 tells us that the model, with the treatment as the only predictor, isn't quite right for predicting how well participants score. The F-statistic value of 0.555, with a high p-value, aligns with these findings, further suggesting that the treatment does not have a statistically significant effect on the quiz scores. In summary, our analysis indicates that while there may be a difference in scores between the control and treatment groups, we cannot conclusively attribute this difference to the treatment effect based on the current data.

Dependent variable: score	
(1)	
Intercept	9.360*** (0.454)
treatment	0.480 (0.644)
Observations	50
R ²	0.011
Adjusted R ²	-0.009
Residual Std. Error	2.278 (df=48)
F Statistic	0.555 (df=1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 1 Regression of treatment impacts on score outcomes

B. ii) Analysis of Treatment Effect on Scores with Covariates

This analysis aimed to understand how certain factors, such as age, gender, study time, and a specific treatment, influence quiz scores. The model starts with an intercept of 9.473, which represents the average score for a baseline group of participants who are 18-23 years old, female, and study between 10 to 15 hours per week. When examining age, we observed that participants in the age groups 24-27, 28-31, and 32+ did not show a statistically significant change in scores compared to this baseline group, with score changes of +0.428, +1.357, and +0.036, respectively. Notably, gender did appear to play a role; male participants scored on average 1.398 points lower than their female counterparts, and this was significant at the 10% level. Looking at study time, no significant score differences were found across the various categories when compared to the reference group. Similarly, the treatment, our variable of interest, only showed a non-significant score increase of 0.115 for those who received it. The R-squared value tells us that only 13.3% of the score variance could be attributed to these factors combined, and the negative adjusted R-

Dependent variable: score	
(1)	
Intercept	9.473*** (1.105)
age_group[T.24-27]	0.428 (0.836)
age_group[T.28-31]	1.357 (1.210)
age_group[T.32+]	0.036 (1.142)
gender[T.Male]	-1.398* (0.786)
hours_per_week[T.5 - 10 hrs]	0.088 (1.142)
hours_per_week[T.< 5 hrs]	0.820 (1.224)
hours_per_week[T.> 15 hrs]	1.011 (1.230)
treatment	0.115 (0.752)
Observations	50
R ²	0.133
Adjusted R ²	-0.036
Residual Std. Error	2.309 (df=41)
F Statistic	0.925 (df=8; 41)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 2 Regression of treatment effect on scores with covariates

squared value of -0.036 indicates that the model might not be the best fit for predicting quiz scores. Moreover, the F-statistic of 0.925 with a p-value above 0.1 reinforces the conclusion that the model's variables, when taken together, do not significantly predict quiz scores, suggesting that other factors not included in the model might be influencing quiz performance.

B. iii) Analysis of Treatment Effect on Test Completion Time

In the regression analysis focusing on the time participants spent to complete a test, the average completion time for the control group, who did not receive any specific treatment, was found to be approximately 551.720 seconds when other variables were held constant. When we introduced a certain treatment to participants, we found that, on average, it took them significantly longer—about 141.940 additional seconds—to finish the test compared to the control group. This finding was statistically significant, which means it is unlikely to have occurred by chance, and we can be reasonably confident that the treatment influenced the increase in completion time. This effect of treatment on completion time is not trivial but is not overwhelmingly large either, as indicated by an R-squared value of 0.075. This suggests that the treatment accounts for 7.5% of the variability in the completion times of participants. The relationship between the treatment and the completion time is clearly present, but other factors not included in this analysis also play a role in how long participants take to complete the test.

Dependent variable: completion_time	
(1)	
Intercept	551.720*** (49.671)
treatment	141.940** (71.904)
Observations	50
R ²	0.075
Adjusted R ²	0.056
Residual Std. Error	254.217 (df=48)
F Statistic	3.897* (df=1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 3 Regression of treatment effect on test completion time

B. iv) Effect of Dependent Variable on Independent Variables

In the statistical analysis of quiz performance across various groups, the intercept is significantly distinct from zero, signaling that participants who are not categorized by specific age, gender, or study hours—and who did not receive any special treatment—can expect an average quiz score of 9.4731. This significant intercept value implies the presence of other influential factors on quiz scores beyond those captured in the model. Additionally, the analysis reveals that gender differences are nearing significance, with male participants scoring on average 1.3982 points lower than their female counterparts, a finding with a p-value of 0.066 that borders

OLS Regression Results						
=====						
Dep. Variable:	score	R-squared:	0.133			
Model:	OLS	Adj. R-squared:	-0.036			
Method:	Least Squares	F-statistic:	0.7855			
Date:	Thu, 07 Mar 2024	Prob (F-statistic):	0.618			
Time:	17:52:40	Log-Likelihood:	-107.82			
No. Observations:	50	AIC:	233.6			
Df Residuals:	41	BIC:	250.8			
Df Model:	8					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	9.4731	1.141	8.301	0.000	7.168	11.778
treatment	0.1147	0.714	0.161	0.873	-1.328	1.557
age_group_24-27	0.4279	0.767	0.558	0.580	-1.122	1.978
age_group_28-31	1.3566	1.878	0.722	0.474	-2.436	5.149
age_group_32+	0.0359	1.051	0.034	0.973	-2.087	2.159
gender_Male	-1.3982	0.740	-1.889	0.066	-2.893	0.096
hours_per_week_5 - 10 hrs	0.0879	1.044	0.084	0.933	-2.020	2.196
hours_per_week_< 5 hrs	0.8202	1.086	0.755	0.454	-1.373	3.014
hours_per_week_> 15 hrs	1.0112	1.081	0.935	0.355	-1.172	3.195
=====						
Omnibus:	2.594	Durbin-Watson:	2.276			
Prob(Omnibus):	0.273	Jarque-Bera (JB):	1.449			
Skew:	-0.006	Prob(JB):	0.485			
Kurtosis:	2.166	Cond. No.	8.67			
=====						

Figure 4 Regression on effect of dependent variable on independent variables

the conventional significance level. These results highlight the potential impact of gender on learning outcomes, suggesting a trend that, while not definitively significant in this model, may warrant further investigation. However, the model's low R-squared value indicates that the treatment and the variables considered in this study explain a relatively small fraction of the variability in the quiz scores, pointing to the existence of other factors affecting quiz performance that are not included in the current analysis.

B. v) Analysis of Question Difficulty Effects

In the regression analysis spanning three difficulty levels of quiz questions—easy, medium, and hard—the treatment's impact on correct scores was evaluated. For easy questions, while the intercept was statistically significant, indicating a baseline average score of 3.960 without treatment, the treatment itself did not significantly affect scores, as shown by its p-value. The model explains very little variance in scores, evidenced by a low R-squared value. Similarly, for medium-difficulty questions, the intercept was significant, with a baseline score of 3.120, yet the treatment's effect remained insignificant, and the model's explanatory power was marginally higher but still low. In the hard questions model, the significant intercept suggested a baseline average score of 2.280, and the treatment showed a negative coefficient; however, this was not significant, leaving the effect of treatment on harder questions inconclusive. Across all models, the treatment did not present a statistically significant influence on correct scores, and the R-squared values were consistently low, indicating the treatment's limited role in explaining score variability.

	(1)	(2)	(3)
Intercept	3.960*** (0.178)	3.120*** (0.254)	2.280*** (0.187)
treatment	0.200 (0.290)	0.480 (0.343)	-0.200 (0.273)
Observations	50	50	50
R ²	0.010	0.039	0.011
Adjusted R ²	-0.011	0.019	-0.010
Residual Std. Error	1.024 (df=48)	1.213 (df=48)	0.967 (df=48)
F Statistic	0.477 (df=1; 48)	1.957 (df=1; 48)	0.535 (df=1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01		

Figure 5 Regression of question difficulty effects

Experiment Limitations

The experiment may have some limitations that should be acknowledged. One potential limitation is the relatively small sample size, which could have limited the study's statistical power to detect significant differences between groups or conditions. A larger sample size would increase the likelihood of obtaining statistically significant results if they exist. Another limitation related to the collection of more comprehensive pre-assessment data like participants' GPA, education level, and other relevant academic or demographic indicators could strengthen our subsequent analyses for the experiment. The additional information could serve as covariates to consider for variation in incoming knowledge and aptitudes that may influence the participants' performance. Additionally, a longer experimental period might be necessary to gather more conclusive evidence and observe long-term effects. The duration of the current experiment, 2-week, may not have been sufficient to capture significant impacts or changes in participants' performance.

Conclusion

Although our study did not provide conclusive evidence that collaboration is definitively more effective than individual work in learning, we observed a trend towards improved performance,


especially on more challenging questions, in the collaborative group. This trend suggests the potential benefits of collaborative learning but warrants further investigation. Future research could address the limitations identified in this study by employing larger sample sizes, exploring different learning materials and assessment formats, and incorporating objective pre-assessment measures.

Future Plan

Future research should explore diverse learning materials and assessments to validate collaboration benefits across contexts. Deeper investigation into collaborative dynamics can optimize strategies for enhanced learning. Increasing sample sizes will strengthen statistical findings. Implementing objective pre-assessments allows isolating learning gains through covariate analysis, controlling for prior knowledge variance.

Appendix A

Table 1 Question Quiz

Easy Question Section	
<div> <div>19:40</div> <div>Video Section</div> <p>Please take notes while watching the video. There will be 15 questions afterwards. Kindly refrain from using ChatGPT or any external resources. This is a FUN QUIZ!!!</p> <p>Note : Please respond to the inquiries based on the content of the video. It's important not to make any assumptions. Answers are based on the video and not on general knowledge/assumptions.</p> <p>Please watch this video before answering this quiz. https://www.youtube.com/watch?v=FkQWpOd9Zdo</p>  </div>	<div> <div>19:21</div> <div>Question Section 1</div> <p>Please respond to the inquiries based on the content of the video. It's important not to make any assumptions</p> <p>Which object in our solar system is approximately twice as big as Earth? * 1 point</p> <p> <input type="radio"/> A) Saturn <input type="radio"/> B) Jupiter's Great Red Spot <input type="radio"/> C) The Sun <input type="radio"/> D) VY Canis Majoris </p> <p>How far apart are the Earth and the moon at their farthest point? * 1 point</p> <p> <input type="radio"/> 252,088 miles <input type="radio"/> 1 million miles <input type="radio"/> 4 billion miles <input type="radio"/> 621 quadrillion miles </p> </div>
<div> <div>19:07</div> <p>How many times wider than Earth is Saturn? * 1 point</p> <p> <input type="radio"/> 2 times <input type="radio"/> 4 times <input type="radio"/> 9 times <input type="radio"/> 11 times </p> <p>What is the biggest star we know of? * 1 point</p> <p> <input type="radio"/> VY Canis Majoris <input type="radio"/> Sun <input type="radio"/> Earth <input type="radio"/> Jupiter </p> <p>How wide is the Milky Way galaxy? * 1 point</p> <p> <input type="radio"/> 252,088 miles <input type="radio"/> 621 quadrillion miles <input type="radio"/> 100,000 miles </p> </div>	
Medium Question Section	

<div style="background-color: #4a4a9a; color: white; padding: 5px; text-align: center;">18:48</div> <div style="background-color: #000080; color: white; padding: 5px;">Question Section 2</div> <div style="background-color: #e6e6ff; padding: 10px; margin-top: 10px;"> <p>On a same scale what would the diameter of Milky Way galaxy be if our solar system was shrunk to the size of a quarter? * 1 point</p> <p><input type="radio"/> USA</p> <p><input type="radio"/> Mercury</p> <p><input type="radio"/> Basketball</p> <p><input type="radio"/> Boston University</p> </div> <div style="background-color: #e6e6ff; padding: 10px; margin-top: 10px;"> <p>How long would it take in years for 1 billion seconds to pass? * 1 point</p> <p><input type="radio"/> 11 days</p> <p><input type="radio"/> 31 years</p> <p><input type="radio"/> 100 years</p> <p><input type="radio"/> 1,000 years</p> </div>	<div style="background-color: #4a4a9a; color: white; padding: 5px; text-align: center;">18:37</div> <div style="background-color: #e6e6ff; padding: 10px; margin-top: 10px;"> <p>What is the diameter of the Milky Way galaxy in light-years? * 1 point</p> <p><input type="radio"/> 10,000 light-years</p> <p><input type="radio"/> 25,000 light-years</p> <p><input type="radio"/> 50,000 light-years</p> <p><input type="radio"/> 100,000 light-years</p> </div> <div style="background-color: #e6e6ff; padding: 10px; margin-top: 10px;"> <p>What is the estimated age of some of the objects seen in the Hubble telescope picture? (Post BIG BANG) * 1 point</p> <p><input type="radio"/> 1 billion years old</p> <p><input type="radio"/> 3 billion years old</p> <p><input type="radio"/> 11 billion years old</p> <p><input type="radio"/> 13 billion years old</p> </div> <div style="background-color: #e6e6ff; padding: 10px; margin-top: 10px;"> <p>What is the name of the spiral galaxy similar to the Milky Way that is twice as wide? * 1 point</p> <p><input type="radio"/> NGC 6286</p> <p><input type="radio"/> NGG 1234</p> </div>
<h2 style="margin: 0;">Hard Question Section</h2>	
<div style="background-color: #4a4a9a; color: white; padding: 5px; text-align: center;">18:14</div> <div style="background-color: #000080; color: white; padding: 5px;">Question Section 3</div> <div style="background-color: #e6e6ff; padding: 10px; margin-top: 10px;"> <p>If VY Canis Majoris is 2,000 times the diameter of our Sun, how much bigger is its volume than the Sun's volume? * 1 point</p> <p><input type="radio"/> 2,000 times bigger</p> <p><input type="radio"/> 4,000,000 times bigger</p> <p><input type="radio"/> 8,000,000 times bigger</p> <p><input type="radio"/> 16,000,000 times bigger</p> </div> <div style="background-color: #e6e6ff; padding: 10px; margin-top: 10px;"> <p>What is the name of the period in the video described as one of the busiest star-forming periods? * 1 point</p> <p><input type="radio"/> A. The Early Middle Ages</p> <p><input type="radio"/> B. The Cambrian Explosion</p> <p><input type="radio"/> C. The Cosmic Noon</p> <p><input type="radio"/> D. The Big Bang</p> </div>	<div style="background-color: #4a4a9a; color: white; padding: 5px; text-align: center;">18:02</div> <div style="background-color: #e6e6ff; padding: 10px; margin-top: 10px;"> <p>How many stars are estimated to be in the universe, based on Carl Sagan's quote in the video? * 1 point</p> <p><input type="radio"/> More than all the grains of sand on all the beaches of Earth</p> <p><input type="radio"/> More than all the blades of grass on Earth</p> <p><input type="radio"/> More than all the trees on Earth</p> <p><input type="radio"/> More than all the people who have ever lived on Earth</p> </div> <div style="background-color: #e6e6ff; padding: 10px; margin-top: 10px;"> <p>What is the estimated diameter of the observable universe? * 1 point</p> <p><input type="radio"/> 621 quadrillion miles</p> <p><input type="radio"/> 93 billion light-years</p> <p><input type="radio"/> 100,000 light-years</p> <p><input type="radio"/> 4 billion miles</p> </div> <div style="background-color: #e6e6ff; padding: 10px; margin-top: 10px;"> <p>Compared to Saturn, how many times wider are the rings of Saturn? * 1 point</p> <p><input type="radio"/> The rings are nine times wider than Saturn</p> <p><input type="radio"/> The rings are half the width of Saturn</p> <p><input type="radio"/> The rings are equal in width to Saturn</p> </div>

Table 2 Dictionary

Column	Description	Values
user_id	Unique identifier for each participant	String
treatment	Indicator of study method (0 = individual, 1 = group)	0 (Individual), 1 (Group)
score	Participant's academic performance score	Integer
completion_time	Time taken to complete the test (in seconds)	Numeric (Seconds)
easy_correct	Number of easy questions answered correctly	Integer
medium_correct	Number of medium difficulty questions answered correctly	Integer
hard_correct	Number of hard questions answered correctly	Integer
knowledge_rating	Participant's self-rated knowledge on space and astronomy (scale of 1-5)	Integer
hours_per_week	Participant's study hours per week	Categorical (e.g., 0-5 hrs, 5-10 hrs)
age_group	Participant's age group	Categorical (e.g., 20-23, 24-27)
gender	Participant's gender	Categorical (e.g., Male, Female)
collaboration_score	Participant's collaboration score (only for group study)	Numeric (Scale of collaboration)
preferred_work_style	Participant's preferred work style	Categorical (e.g., Group, Individual)

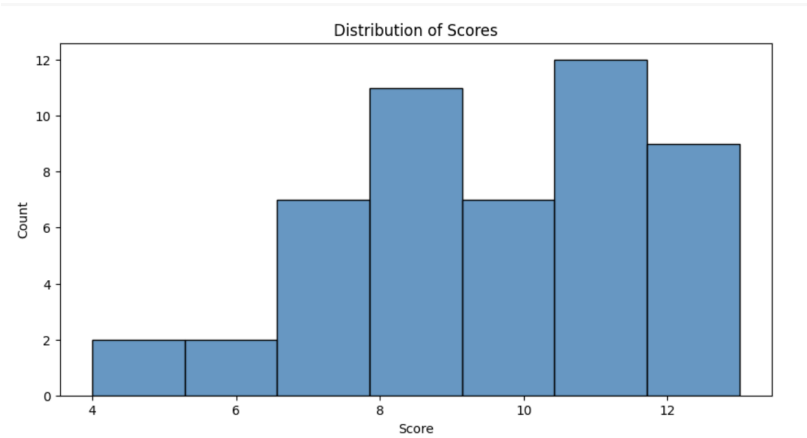


Figure 6 Distribution of scores

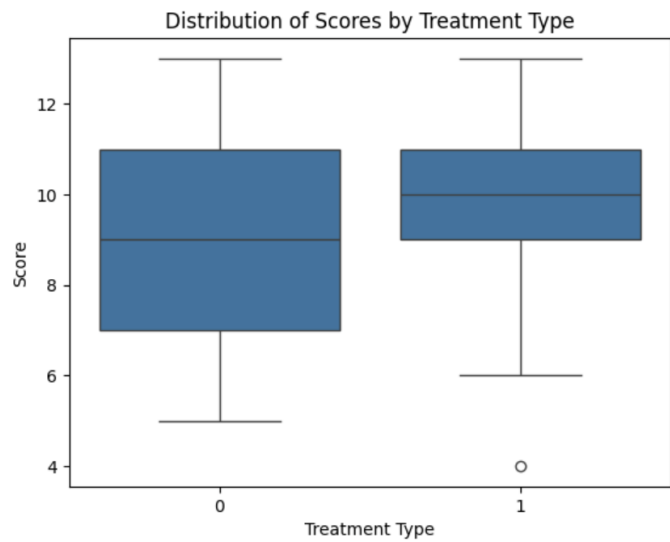


Figure 7 Distribution of score by group

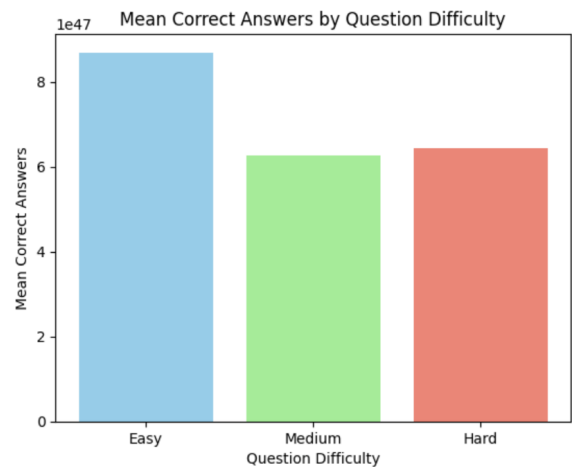


Figure 8 Distribution of scores by question difficulty

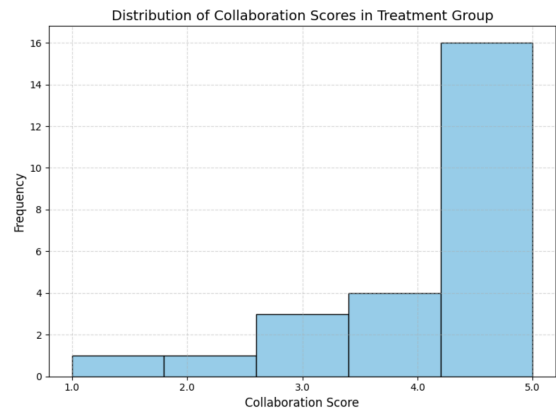


Figure 9 Distribution of collaboration score for treatment group