Student's Preferences of Learning Methods

—, —,
—, —, Yuta Kondo

December 2, 2024

1 Abstract

In this research paper, we explore whether online technologies such as generative AI and online resource platforms have surpassed traditional academic resources like lecture slides and textbooks as the preferred study resource. Our group aimed to answer whether students' demographic background such as gender and native language have any different preferences for a study resource/learning method. We elected to use the stratified random sampling method to collect 60 unique data points, 30 for native English speakers and non-native English speakers separately, based on volunteered inputs from the students in the course STA304. Although the sampled students' demographic backgrounds are found to be independent from their academic achievements, we have found that the overwhelming majority of them prefer textbook and lecture slides significantly less than generative AI and online resource platforms such as YouTube as a means of study resource.

2 Introduction

Learning methods can vary. However, an interesting question is: "Is there a best learning method?" In this study, we want to find if there is an association between different learning methods and their academic achievement or the association between the different learning methods and their time spent on study.

After the pandemic and the emergence of generative AI, students have more study tools, not limited to traditional tools like textbooks and lecture slides. Students may prefer to use the online learning platform as supplementary material to the course material. In addition to these learning methods, people have recently liked to use generative AI to receive quick answers. Therefore, in this study, we will typically focus on three groups of learning methods: traditional learning methods. (e.g., textbooks, lecture slides, etc.), online learning platforms. (e.g., YouTube, Coursera, etc.), and generative AI (e.g., ChatGPT, Gemini).

Since most of the members in the group are international students who are not native English speakers, we believed non-native English speakers prefer to use generative AI or YouTube compared with the textbook in English. Therefore, we decided to use the stratified sampling method.

In October 2024, we provided a ten-question survey to the students in the STA304 class and collected data such as demographics (e.g., sex, first language), their preferred learning method, their CGPA, and their time spent on study based on their preferred learning method. As the setup, there were two groups of data, one for the group of non-native English speakers and the other for the group of native English speakers. We then conducted the following research questions:

- **(RQ1)** Is there any association between the preferred learning method and their demographic, such as gender, and whether their first language is English?
  - There are two hypotheses in this case, where one is focused on the preferred learning method and gender, and the other is focused on the first language.

  - **(RQ1A)** Is there any association between preferred learning methods and gender?
    - Null Hypothesis: A student's preferred learning method is independent of their gender.
    - Alternate Hypothesis: A student's preferred learning method isn't independent of their gender.
  - **(RQ1B)** Is there any association between the preferred learning method and first language in English?
    - Null Hypothesis: A student's preferred learning method is independent of their first language in English.
    - Alternate Hypothesis: A student's preferred learning method isn't independent of their first language in English.

- **(RQ2)** Is there a mean difference in preference level (i.e. 1 to 5) in the different learning methods?
  - Null Hypothesis: There is no mean difference in the preference level for different learning methods.
  - Alternative Hypothesis: There is at least one group that differs from the others in the preference level.
- **(RQ3)** Is there a mean difference in CGPA (i.e. 1 to 4) in the preferred learning method chosen by the student?
  - Null Hypothesis: There is no mean difference in the CGPA for the preferred learning method chosen by the student.
  - Alternative Hypothesis: There is at least one group that differs from the others in the CGPA.
- **(RQ4)** Is there a mean difference in time spent on study (i.e., 1 to 24 hours per day) in the preferred learning method chosen by the student?
  - Null Hypothesis: There is no mean difference in the time spent on study for the different preferred learning methods chosen by students.
  - Alternative Hypothesis: There is at least one group that differs from the others in the time spent on study.

In the remaining section of the report, section 3 will focus on the methodology we used for collecting the data. In section 4, we will show the results of the data analysis through these data on R. After that, section 5 will discuss the meaning of the result of the analysis. In the remaining two sections, section 6 will discuss the limitations we have found, and section 7 will drop the conclusion of this research. The appendix will be followed after section 7.

3 Methodology

In October 2024, we provided a survey link on the Piazza discussion platform to collect data for our research. The research focuses on exploring the students' preferences for learning methods and the benefits of these preferred learning methods for the students enrolled in STA304H5 courses at the University of Toronto Mississauga campus.

We intend to use the stratified sampling method to collect these data. We want to sample the native English speaker versus the non-native English speaker.

The survey contains the following questions: the demographic question, their CGPA, asking for their preferred learning method (e.g. an online platform or generative AI), and the number of hours spent studying.

However, obtaining the list of all students in the course was difficult due to privacy; we did not enforce the random sampling in the first place. We utilized the volunteering sample method to put the survey link on the Piazza. To ensure the randomness of our data and eventually align our intentions, we first cleaned the data received from the volunteering sample method. After that, we split the received data (i.e., 94 in total) into groups of native English speakers and non-native English speakers. We wanted these two groups to have the same proportion during the analysis. To ensure that we have the same proportion of people in this research, we randomly sampled 30 people in each group and ended up with 60 samples. So, the sample size of each group is 30. There are 30 non-native English speakers and 30 native English speaker's data will be used in the analysis. The detailed computation for sample size is described in the analysis section.

4 Analysis

4.1 Sample Size:

In our study, most of the test requires the estimator of mean value. We want to determine the sample size so that we can have a good estimator to estimate the population mean. As mentioned in the section 3 methodology, we intend to use the stratified random sample method. There are two strata, one represent the English speaker group ; the other represents the non-English speaker group. We assume the proportion of these two group are equal at STA304 class. And, in STA304, we have total people $N = 200$. Moreover, we want to have the bound of error to be $B \approx 1.5991$. Based on these information, we still need a variance to compute the sample size for each stratum. We selected the RQ4 as a question to estimate the variance through the empirical estimation; where we have $\sigma^2 \approx \frac{\text{Range}}{4} \approx \frac{\text{21 hours for study time - 1 hour for study time}}{4}$ for both group. We then be able to compute the sample size of each stratum as following:

$$n = \frac{\sum_{i=1}^{L} N_i^2 * \sigma^2 * \frac{1}{a_i}}{N^2 * \frac{(B)^2}{4} + \sum_{i=1}^{L} N_i * \sigma^2} = \frac{100^2 * (\frac{19}{4})^2 * 2 + 100^2 * (\frac{19}{4})^2 * 2}{200^2 * \frac{1.5991^2}{4} + 100 * (\frac{19}{4})^2 * 2} \approx 29.9996$$

Therefore, we only need around 30 samples for each stratum.

4.2 RQ 1:

To be able to answer to RQ1, we split RQ1 into two sub-questions: RQ1A and RQ1B. For RQ1A, a Fisher's test was conducted to detect the relationship between gender and preferred learning method. We did not use the chi-square test because the expected value in each cell of contingency table is not all greater than or equal to 5. But, all of these data were randomly sampled. Based on the way we sampled, we can use the Fisher's test. In this test, we obtained a p-value of 0.5777.

For RQ1B, a chi-square test was conducted to detect the relationship between first language and preferred learning method. We can use the chi-square test because the expected value in each cell of contingency table are all greater than or equal to 5. Also, the data are retrieved through random sampling. We obtained a chi-square statistic of 1.1206 with a degree of freedom of 2 and a p-value of 0.571.

4.3 RQ 2:

To be able to answer to RQ2, we elected to use the Kruskal-Wallis test to determine whether there is a mean difference in preference level (i.e. 1 to 5) in the different learning methods.

The Kuskal-Wallis test works because only 2 assumptions can be satisfied (the results of the assumption tests are shown in the appendix section at the end of the document):

- Independence of Observations: Since the data were collected independently from STA304 students, the data should be independent.
- Homogeneity of Variance: This condition is satisfied through the Bartlett test.

After checking the necessary assumptions, we can conduct the Kruskal-Wallis test and end up with the following results: Kruskal-Wallis chi-squared = 11.603, df = 2, and p-value = 0.003023. This indicates that there is at least one group that differs from the others in terms of preference level.

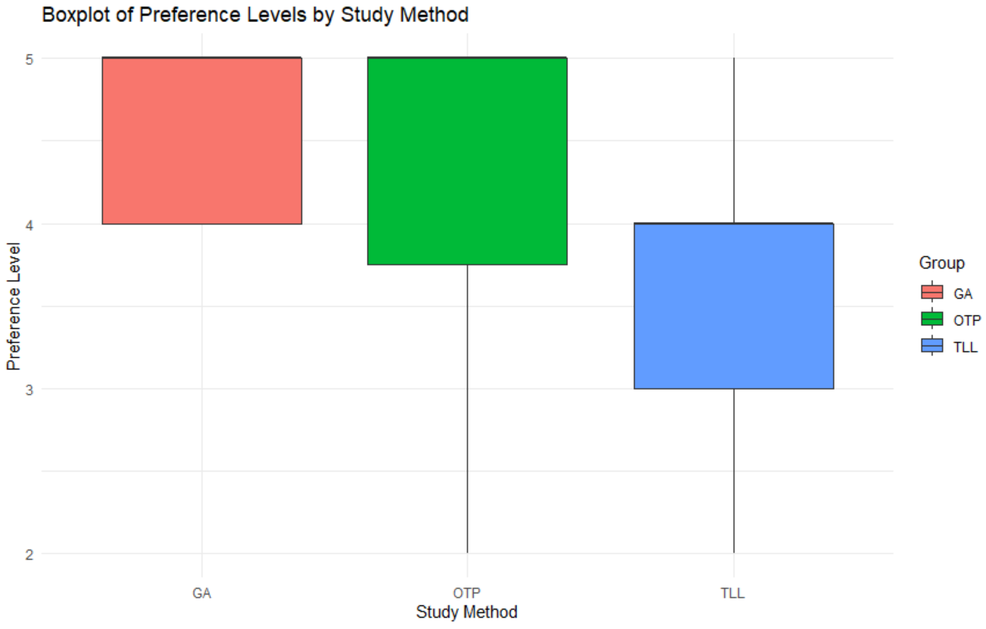

Figure 1: Boxplot for different study methods/resources based on preference

Code	Representation
TLL	Textbook/Lecture Slides/Lectures
OTP	Online tutoring platforms (e.g. YouTube, Coursera)
GA	Generative AI (e.g. ChatGPT, Claude)

Table 1: Code and corresponding representations

Category	Min	1st quantile	Median	Mean	3rd quantile	Max
Preference level of TLL	2.000	3.000	4.000	3.559	4.000	5.000
Preference level of OTP	2.000	3.750	5.000	4.375	5.000	5.000
Preference level of GA	4.0	4.0	5.0	4.6	5.0	5.0

Table 2: Summary of Preference Level for Each Group

Category	IQR	Standard deviation	95% Confidence Interval
Preference level of TLL	1.00	1.078472	(3.182527, 3.935120)
Preference level of OTP	1.25	1.024695	(3.828979, 4.921021)
Preference level of GA	1.00	0.5163978	(4.230591, 4.969409)

Table 3: Summary of Preference for Each Group

Assumption	Checking method	p-value
Equal Variance	Bartlett test	0.06289
Normality	Shapiro-Wilk test	4.224e-05

Table 4: Assumption Results

4.4 RQ 3:

To be able to answer to RQ3, we also chose the Kruskal-Wallis test to determine whether there is a mean difference in CGPA (i.e. 1 to 5) in the preferred learning method chosen by the student. Since the normality assumption fails, the ANOVA test does not work here.

The Kuskal-Wallis test works because only 2 assumptions can be satisfied (the results of the assumption tests are shown in the appendix section at the end of the document):

- Independence of Observations: Since the data were collected independently from STA304 students, the data should be independent.
- Homogeneity of Variance: This condition is satisfied through the Bartlett test.

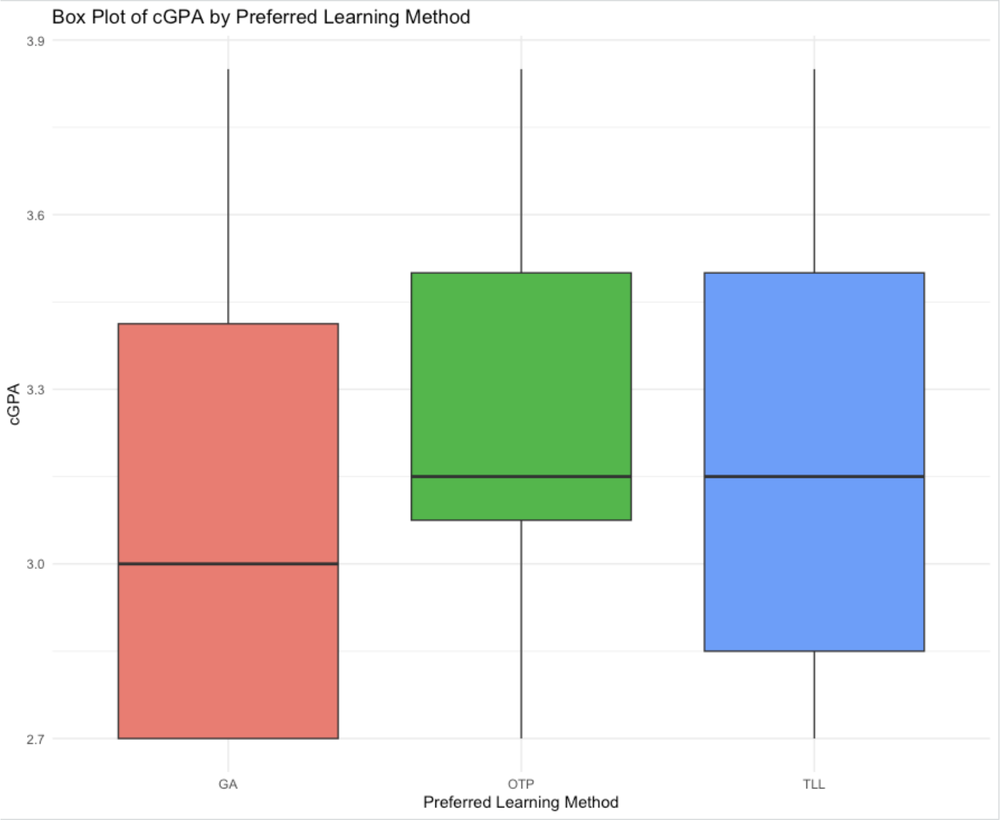

Figure 2: Boxplot for different study methods/resources based on cGPA

Category	Min	1st quantile	Median	Mean	3rd quantile	Max
CGPA of TLL	2.700	2.850	3.150	3.156	3.500	3.850
CGPA of OTP	2.700	3.075	3.150	3.244	3.500	3.850
CGPA of GA	2.700	2.700	3.000	3.080	3.413	3.85

Table 5: Summary of CGPA for Each Group

After checking the necessary assumptions, we can conduct the Kruskal-Wallis test and end up with the following results: Kruskal-Wallis chi-squared = 1.2012, df = 2, and p-value = 0.5485. This indicates that we fail to reject the null hypothesis that there is no mean difference in the CGPA for the preferred learning method chosen by the student.

Category	IQR	Standard deviation	95% Confidence Interval
CGPA of TLL	0.65	0.3930774	(3.018731, 3.293034)
CGPA of OTP	0.425	0.399531	(3.030855, 3.456645)
CGPA of GA	0.7125	0.4197883	(2.779702, 3.380298)

Table 6: Summary of CGPA for Each Group

Assumption	Checking method	p-value
Equal Variance	Bartlett test	0.9701
Normality	Shapiro-Wilk test	0.001027

Table 7: Assumption Results

4.5 RQ 4:

To be able to answer to RQ4, we used the Kruskal-Wallis test, as well as RQ2 and RQ3, to determine whether there is a mean difference in time spent on study (1 to 24 hours per day) based on the preferred learning method. Since the normality assumption fails, the ANOVA test is not applicable.

The Kuskal-Wallis test works because only 2 assumptions can be satisfied (the results of the assumption tests are shown in the appendix section at the end of the document):

- Independence of Observations: Since the data were collected independently from STA304 students, the data should be independent.
- Homogeneity of Variance: This condition is satisfied through the Bartlett test.

Category	Min	1st quantile	Median	Mean	3rd quantile	Max
Time Spent on Study of TLL	1.000	2.000	4.000	4.588	5.000	20.000
Time Spent on Study of OTP	1.00	2.00	3.00	3.75	5.00	12.00
Time Spent on Study of GA	1.00	3.25	4.00	5.40	4.75	20.00

Table 8: Summary of Time Spent on Study for Each Group

Category	IQR	Standard deviation	95% Confidence Interval
Time Spent on Study of TLL	3	4.105559	(3.155738, 6.020733)
Time Spent on Study of OTP	3	2.67083	(2.326815, 5.173185)
Time Spent on Study of GA	1.5	5.378971	(1.552116, 9.247884)

Table 9: Summary of Time Spent on Study for Each Group

Assumption	Checking method	p-value
Equal Variance	Bartlett test	0.06048
Normality	Shapiro-Wilk test	8.583e-10

Table 10: Assumption Results

After checking the necessary assumptions, we can conduct the Kruskal-Wallis test and end up with the following results: Kruskal-Wallis chi-squared = 0.74067, df = 2, and p-value = 0.6905. This indicates that we fail to reject the null hypothesis that there is no mean difference in the mean time spent for the preferred learning method chosen by the student.

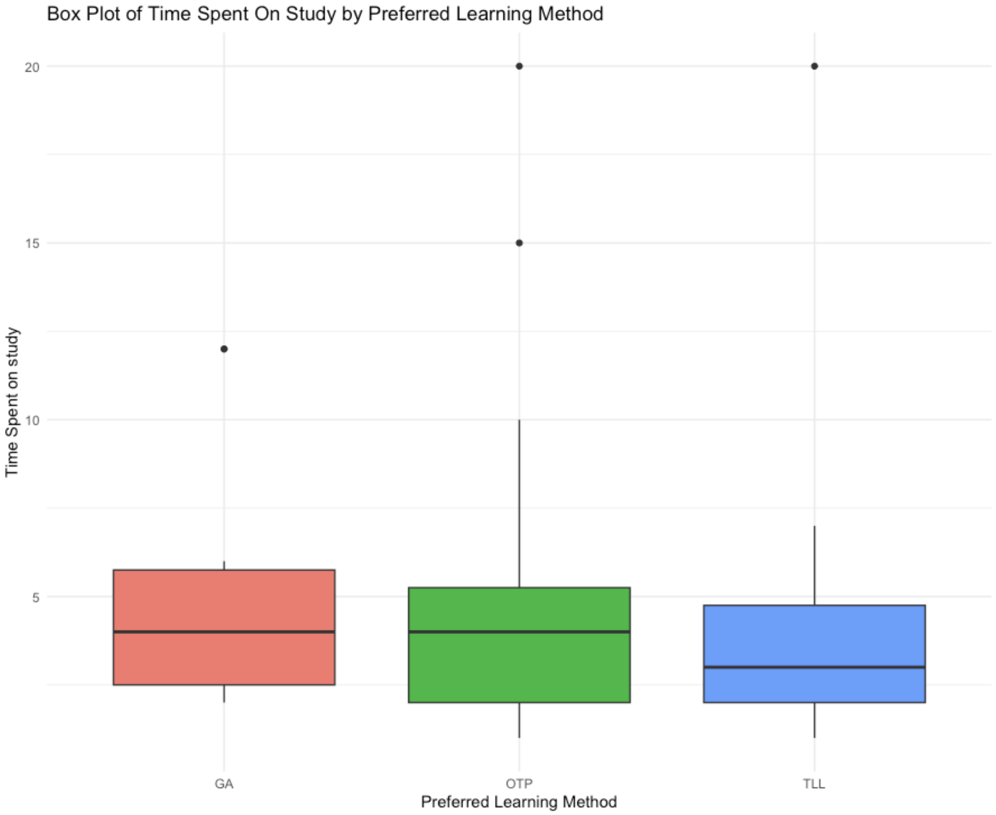

Figure 3: Boxplot for different study methods/resources based on time spent studying per day

5 Discussion / Results

Firstly, the test result to the RQ1A does make sense to us where we believed that there was no relationship between the preferred learning method and their gender. Therefore, we fail to reject the null hypothesis with the p-value 0.5777 greater than the significant level of 0.05.

Because of the chi-square test result to the RQ1B, there's no relationship between the usage of different learning methods and whether they speak English as their first language or not. This is because the p-value is greater than 0.05, and there is no significant evidence in the data to show the dependency. This result surprised us because we initially thought that non-English speakers would feel more comfortable using generative AI to help their learning compared to other methods, due to the plain and short English sentences it offers.

In contrast to RQ1, the result of RQ2 is the same as our expectation. The result of the Kruskal-Wallis test indicates that at least one different method has a different preference level from students since the p-value is less than 0.05 at a significant level. Looking at the confidence interval and box-plot of each group, traditional learning method such as textbook, lecture slides, and lectures are significantly less preferred than online platforems and generative AI.

Similar to the RQ2, we also conducted a Kuskal-Wallis test for the RQ3. The result suggests no statistical significance in the mean difference in CGPA achievement with different learning methods since the p-value is bigger than 0.05.

At last, the result of the ANOVA test of the RQ4 has a p-value greater than 0.05, which means that we didn't find the significant mean difference in the time spent studying with different learning methods.

6 Limitations

Since the sampled population comes from the students of STA304, with a small population size and who predominantly are also in the statistics major/specialist, they might have similar experiences in their academic journey. Thus, this research may not have captured any big differences among the surveyed students.

Furthermore, in this survey, we assumed equal proportions of English native speakers and non-native speakers (0.5 each). However, the raw data we collected showed different proportions, with the number of English non-native speakers greater than that of native speakers.

Thus, future study needs to broaden the scope of the sample to include students who are not enrolled in STA304, so it can capture traits that statistics major/specialist students may not possess. Also, future research could use the actual proportions of native and non-native speakers observed in the collected data.

7 Conclusion

We explored the relationship between students' demographic background such as gender, native language, and preferences for a study resource/learning method. The analysis indicates that traditional learning methods are less preferred than online learning platforms (YouTube, Coursera, etc.) and generative AI (ChatGPT, Gemini, etc.). However, we did not observe a significant association between students' preference for learning methods and their academic performance and study hours. Also, English as a first language was not associated with students' choice of learning methods.

Given these results, we have a new assumption: the grades in specific subjects can be different based on students' preferences for learning methods, even though we did not find the significant difference in CGPA. Future research could test this by collecting the grades of individual subjects (e.g., STA, MAT, ENG, PSY).

8 Appendix

```
########################################################################
# Author: ---                                              #
# Citation:                                                            #
# 1. https://chatgpt.com/share/6730aeb8-5a64-8006-a17b-5a670c72e635    #
# 2. STA304 Tutorial 1 slides.                                         #
########################################################################
# Import library
library(ggplot2)

# Setting the working directory.
setwd("--Working directory--")

# The following code reads the sampling data from the CSV file.
data = read.csv("Cleaned_sampling_data(final_data).csv")


######################## RQ1 ########################
# There are two questions we can ask to conduct the research for RQ1
# Q1.1 Is there any association between preferred learning methods and gender?
# Q1.2 Is there any association between the preferred learning method and first language
↪  in
# English?

##### RQ1.1 Gender dependence #####
split_data_gender = split(data, data$gender)

# First, we split the data based on gender.

##### Male Section #####
Male = split_data_gender[["M"]]

Male_preferred_learning_method = split(Male, Male$preferred_learning_method)

# Next, we split the male data further based on the learning method.

## TLL
Male_TLL = Male_preferred_learning_method[["TLL"]]
# Count the number of males who prefer the TLL learning method.
number_Male_TLL = length(Male_TLL$gender)


## GA
```

```
Male_GA = Male_preferred_learning_method[["GA"]]
# Count the number of males who prefer the GA learning method.
number_Male_GA = length(Male_GA$gender)

## OTP
Male_OTP = Male_preferred_learning_method[["OTP"]]
# Count the number of males who prefer the OTP learning method.
number_Male_OTP = length(Male_OTP$gender)

##### Female Section #####
Female = split_data_gender[["F"]]

# Next, we further split the female data based on the learning method.
Female_preferred_learning_method = split(Female, Female$preferred_learning_method)

## TLL
Female_TLL = Female_preferred_learning_method[["TLL"]]
# Count the number of females who prefer the TLL learning method.
number_Female_TLL = length(Female_TLL$gender)

## GA
Female_GA = Female_preferred_learning_method[["GA"]]
# Count the number of females who prefer the GA learning method.
number_Female_GA = length(Female_GA$gender)

## OTP
Female_OTP = Female_preferred_learning_method[["OTP"]]
# Count the number of females who prefer the OTP learning method.
number_Female_OTP = length(Female_OTP$gender)

# Set up the data frame for the chi-square test.
RQ1_Gender = data.frame(

  Learning_Method = c(rep("TLL", number_Male_TLL), rep("GA", number_Male_GA), rep("OTP",
  ↪  number_Male_OTP),rep("TLL", number_Female_TLL), rep("GA", number_Female_GA),
  ↪  rep("OTP", number_Female_OTP)),
  Gender = c(rep("M", number_Male_TLL + number_Male_GA + number_Male_OTP), rep("F",
  ↪  number_Female_TLL + number_Female_GA + number_Female_OTP))

)

# Create a contingency table for the chi-square test.
contingency_table_RQ1_1 = table(RQ1_Gender$Gender, RQ1_Gender$Learning_Method)

# Perform the independence test.
```

```r
RQ1_1_test = chisq.test(contingency_table_RQ1_1)

# Check the assumption that each cell value should be greater than 5.
RQ1_1_test$expected

# If the assumption does not hold, the following results should be ignored.
RQ1_1_test

# Unfortunately, the assumption failed. Therefore, we conduct Fisher's test
# instead of the chi-square test.
fisher.test(contingency_table_RQ1_1)

##### RQ1.2 English Speaker dependence #####

# First, we split the data based on whether their language is English or not.
split_data_english_first_language = split(data, data$english_first_language)

##### English Speaker Section #####
English_speaker = split_data_english_first_language[["Y"]]

# Next, we further split the data based on the learning method.
English_speaker_preferred_learning_method = split(English_speaker,
↪  English_speaker$preferred_learning_method)

# TLL
English_speaker_TLL = English_speaker_preferred_learning_method[["TLL"]]
number_English_speaker_TLL = length(English_speaker_TLL$english_first_language)

# GA
English_speaker_GA = English_speaker_preferred_learning_method[["GA"]]
number_English_speaker_GA = length(English_speaker_GA$english_first_language)

# OTP
English_speaker_OTP = English_speaker_preferred_learning_method[["OTP"]]
number_English_speaker_OTP = length(English_speaker_OTP$english_first_language)

##### Non-English Speaker Section #####

Non_English_speaker = split_data_english_first_language[["N"]]
# We then further split the data based on the learning method.
Non_English_speaker_preferred_learning_method = split(Non_English_speaker,
↪  Non_English_speaker$preferred_learning_method)

# TLL
Non_English_speaker_TLL = Non_English_speaker_preferred_learning_method[["TLL"]]
```

```
number_Non_English_speaker_TLL = length(Non_English_speaker_TLL$english_first_language)

# GA
Non_English_speaker_GA = Non_English_speaker_preferred_learning_method[["GA"]]
number_Non_English_speaker_GA = length(Non_English_speaker_GA$english_first_language)

# OTP
Non_English_speaker_OTP = Non_English_speaker_preferred_learning_method[["OTP"]]
number_Non_English_speaker_OTP = length(Non_English_speaker_OTP$english_first_language)

RQ1_first_language = data.frame(

  first_language = c(rep("Y", number_English_speaker_TLL + number_English_speaker_GA +
  ↪  number_English_speaker_OTP), rep("N", number_Non_English_speaker_TLL +
  ↪  number_Non_English_speaker_GA + number_Non_English_speaker_OTP)),
  Learning_Method = c(rep("TLL", number_English_speaker_TLL), rep("GA",
  ↪  number_English_speaker_GA), rep("OTP", number_English_speaker_OTP),
                      rep("TLL", number_Non_English_speaker_TLL), rep("GA",
                      ↪  number_Non_English_speaker_GA), rep("OTP",
                      ↪  number_Non_English_speaker_OTP))
)

# Create the contingency table for the further chi-square test.
contingency_table_RQ1_2 = table(RQ1_first_language$first_language,
↪  RQ1_first_language$Learning_Method)

# Perform the independence test.
RQ1_2_test = chisq.test(contingency_table_RQ1_2)

# Check the assumption that each cell value should be greater than 5.
RQ1_2_test$expected

# If the assumption does not hold, the following results should be ignored.
RQ1_2_test


######################## RQ2 ########################
# To complete RQ2, we first need to extract a row of data based
# on their preferred_learning_method.
split_data_anova = split(data, data$preferred_learning_method)

# Split the data into different preferred learning method groups.
data_TLL =  split_data_anova[["TLL"]]
data_GA =  split_data_anova[["GA"]]
data_OTP =  split_data_anova[["OTP"]]
```

```
# The following code tries to extract the preference level data
# for each learning method.
data_TLL_prefer = data_TLL$textbook_usefulness
data_GA_prefer = data_GA$chatgpt_usefulness
data_OTP_prefer = data_OTP$online_platform_usefulness

# The following code tries to obtain the length of data for each learning method.
TLL_length_RQ2 = length(data_TLL_prefer)
GA_length_RQ2 = length(data_GA_prefer)
OTP_length_RQ2 = length(data_OTP_prefer)

# The following code tries to construct clean data for ANOVA for RQ2.
clean_RQ2_data = data.frame(
  preference_level=c(data_TLL_prefer, data_GA_prefer, data_OTP_prefer),
  Group = c(rep("TLL",TLL_length_RQ2), rep("GA", GA_length_RQ2), rep("OTP",
   ↪  OTP_length_RQ2))
)

# To use ANOVA, it is better to satisfy the assumptions of normality,
# equal variance, and independence.

# Check the assumptions:

# Check for equal variance.
bartlett.test(preference_level˜Group, data = clean_RQ2_data)

# Check for normality.
# Run ANOVA on preference_level from clean_RQ2_data to obtain residuals.
model <- aov(preference_level˜Group, data = clean_RQ2_data)
residuals_RQ2 <- residuals(model)
shapiro.test(residuals_RQ2)

# Since the normality assumption does not hold,
# we need to use the Kruskal-Wallis test.
kruskal.test(preference_level˜Group, data = clean_RQ2_data)

# Additionally, we can construct the box plot.
ggplot(clean_RQ2_data, aes(x = Group, y = preference_level, fill = Group)) +
  geom_boxplot() +
  labs(title = "Box Plot of cGPA by Preferred Learning Method",
       x = "Preferred Learning Method",y = "cGPA") +
  theme_minimal() + theme(legend.position = "none")

######################## RQ3 ########################
```

```r
# To complete RQ3, we first need to extract a row of data
# based on their preferred_learning_method.
split_data_anova = split(data, data$preferred_learning_method)

# The following code tries to obtain the data corresponding
# to CGPA for each group.
data_TLL_CGPA = data_TLL$cgpa
data_GA_CGPA = data_GA$cgpa
data_OTP_CGPA = data_OTP$cgpa

# The following code tries to count the data in each group.
TLL_length_RQ3 = length(data_TLL_CGPA)
GA_length_RQ3 = length(data_GA_CGPA)
OTP_length_RQ3 = length(data_OTP_CGPA)

clean_RQ3_data = data.frame(
  CGPA=c(data_TLL_CGPA, data_GA_CGPA, data_OTP_CGPA),
  Group = c(rep("TLL",TLL_length_RQ3), rep("GA", GA_length_RQ3), rep("OTP",
  ↪   OTP_length_RQ3))
)


# To use ANOVA, it is better to satisfy the assumptions of normality,
# equal variance,  and independence.

# Check the assumptions:

# Check for equal variance.
bartlett.test(CGPA~Group, data = clean_RQ3_data)

# Check for normality.
# Run ANOVA on preference_level from clean_RQ2_data to obtain residuals.
model <- aov(CGPA~Group, data = clean_RQ3_data)
residuals_RQ3 <- residuals(model)
shapiro.test(residuals_RQ3)

# Since the normality assumption does not hold,
# we need to use the Kruskal-Wallis test.
kruskal.test(CGPA~Group, data = clean_RQ3_data)

ggplot(data,
       aes(x = preferred_learning_method,
           y = cgpa, fill = preferred_learning_method)) +
  geom_boxplot() +
  labs(title = "Box Plot of cGPA by Preferred Learning Method", x = "Preferred Learning
  ↪  Method",y = "cGPA") +
  theme_minimal() + theme(legend.position = "none")

######################## RQ4 ########################
# To complete RQ4, we first need to extract a row of data based
# on their preferred_learning_method.
split_data_anova = split(data, data$preferred_learning_method)

# The following code tries to obtain the data corresponding to
# time spent on study for each group.
data_TLL_time_spent = data_TLL$study_hours_per_day
data_GA_time_spent = data_GA$study_hours_per_day
data_OTP_time_spent = data_OTP$study_hours_per_day

# The following code tries to count the data in each group.
TLL_length_RQ4 = length(data_TLL_time_spent)
GA_length_RQ4 = length(data_GA_time_spent)
OTP_length_RQ4 = length(data_OTP_time_spent)

clean_RQ4_data = data.frame(
  time_spent=c(data_TLL_time_spent, data_GA_time_spent, data_OTP_time_spent),
  Group = c(rep("TLL",TLL_length_RQ4), rep("GA", GA_length_RQ4), rep("OTP",
  ↪  OTP_length_RQ4))
)

# To use ANOVA, it is better to satisfy the assumptions of normality,
# equal variance, and independence.

# Check the assumption:

# Check for equal variance.
bartlett.test(time_spent~Group, data = clean_RQ4_data)

# Check for normality.
shapiro.test(clean_RQ4_data$time_spent)

# Since the normality assumption does not hold,
# we need to use the Kruskal-Wallis test.
kruskal.test(time_spent~Group, data = clean_RQ4_data)

# Additionally, we can construct a box plot.
ggplot(data, aes(x = preferred_learning_method,
                 y = clean_RQ4_data$time_spent,
                 fill = preferred_learning_method)) +
  geom_boxplot() +
  labs(title = "Box Plot of cGPA by Preferred Learning Method", x = "Preferred Learning
  ↪  Method",y = "cGPA") +
  theme_minimal() + theme(legend.position = "none")
```

References

[1] We used ChatGPT on R code to understand how to group the list of data based on the value in the certain column. https://chatgpt.com/share/6730aeb8-5a64-8006-a17b-5a670c72e635

[2] We used ChatGPT to generate the latex template to include R code in latex. https://chatgpt.com/share/6732d20e-4f5c-8007-b25b-b495abaaa165