STA304 – Group 3, Technical Report

Due Dec. 2, 2024

Yusuf Emre Kenaroglu, 1007318751
Diptanshu Jitendra Chaudhari, 1008331048
Leqa Al Tamsi, 1007505709
Ana Elisa Lopez-Miranda, 1008819879
Raswanth Chandrasekharan, 1009047773

Disclaimer:

We would like to acknowledge that we have made use of LLM tools to help us with the creation of tables in LaTeX.

1 Abstract

Education was one of the many systems COVID-19 disrupted globally. It forced students to adapt to abrupt changes in learning environments. This study investigates how COVID-19 related changes impacted student learning and adaptation, focusing on domestic and international STA304 students at the University of Toronto Mississauga campus. We distributed our survey online by providing a link to it in the course's online discussion platform (Piazza). Analysis reveals that COVID-19 had a significant impact on both groups of students. Surprisingly, it also revealed that both groups shared a common experience, with no significant difference in how the pandemic impacted their academic success, engagement, or skills.

2 Introduction

A rapid increase in health concerns diverged the traditional education systems. These shifts have highlighted differences in how students from diverse backgrounds -specifically as domestic and international students- experienced and adapted to the educational changes. Understanding these impacts is essential to developing more resilient and inclusive academic support structures for future disruptions.
To investigate these effects, our study focuses on students enrolled in STA304 at the University of Toronto, examining how different factors affected their learning

and adaptation during the COVID-19 pandemic. Specifically, we aim to answer the following research questions:

- RQ1: What factors impacted students' education during the COVID-19 period?
  H_o : Domestic and international students experienced the same levels of impact on their academic success, engagement, and adaptability.
  H_a : Domestic and international students experienced different levels of impact on their academic success, engagement, and adaptability.
- RQ2: How do students' valuations of education during the COVID-19 period compare to the present?
  H_o : Students valued their education during COVID-19 the same as they do currently.
  H_a : Students value their education more than they did during COVID-19.

This paper is organized as follows: Section 3 outlines the data collection methodology and sampling design, detailing our use of stratified random sampling to capture perspectives across different student demographics. Section 4 describes the statistical analyses employed to address our research questions. Section 5 discusses our findings, including notable trends in students' perceptions of their education and adaptability. Section 6 addresses the limitations of our study and potential improvements. Section 7 concludes with a summary of findings and directions for future research. Finally, Section 8 includes all the R code we used for this project.

3 Methodology

Between September 24^{th} and October 21^{st}, an online survey meant to understand how COVID-19 has impacted students' learning and adaptation was distributed to students enrolled in STA304 at the University of Toronto Mississauga Campus. The survey included questions that addressed different aspects of participants' lives, including their living conditions, academic engagement, health concerns, and perspectives on their academic experiences both during and post COVID-19. We used stratified random sampling. Since the survey was made available for all students enrolled in the course, each student had an equal chance of filling out our survey. Additionally, we also used a single seed value when carrying out our analysis in R so that the randomization within our sample was consistent in a way that would not create discrepancies. The reason why we used stratified sampling is because we hypothesized that COVID-19 might have impacted domestic and international students differently. Stratified random sampling allowed us to have control over how

many international and domestic students we had in our sample to make sure we were not over or under representing either group.

4 Analysis

For our study, our population size is N = 239 and we plan to collect data using stratified random sampling. To determine a sample size to collect, we'll go with the calculation focusing on the mean population parameter. It is assumed that there is an equal proportion among international and domestic students. Hence, given a bound of error of 0.1125, the sample size calculation is as follows:

$$n = \frac{\sum_{i=1}^{2} \frac{N_i^2 p_i q_i}{a_i}}{N^2 \frac{B^2}{4} + \sum_{i=1}^{L} N_i p_i q_i}$$

Given:
$N = 239$
$N_1 = N_2 = 239 \times 0.5 = 119.5$
$p_1 = q_1 = p_2 = q_2 = 0.5$
$a_1 = \frac{N_1}{N} = \frac{119.5}{239}$, $a_2 = \frac{N_2}{N} = \frac{119.5}{239}$,
$B = 0.1125$

$$\frac{\frac{119.5^2 \cdot 0.25}{119.5/239} + \frac{119.5^2 \cdot 0.25}{119.5/239}}{239^2 \cdot \frac{0.1125^2}{4} + (119.5 \cdot 0.25 + 119.5 \cdot 0.25)} \approx 60$$

Therefore, we sampled 60 out of 89 students for the rest of our analysis.

We found that approximately 54% (33) of participants reported that they were a domestic student whereas approximately 46% (27) of participants reported that they were an international student.

To answer RQ1, We performed simple linear regression, comparison of proportions, and multiple linear regression. Starting with simple regression:
Assumptions we must verify for the 2 simple linear regression tests:

- Linearity: The relationship between X and Y in the below tests is linear.
- Independence of errors: X is independent from the residuals.
- Normality of errors: the residuals must be approximately normally distributed.

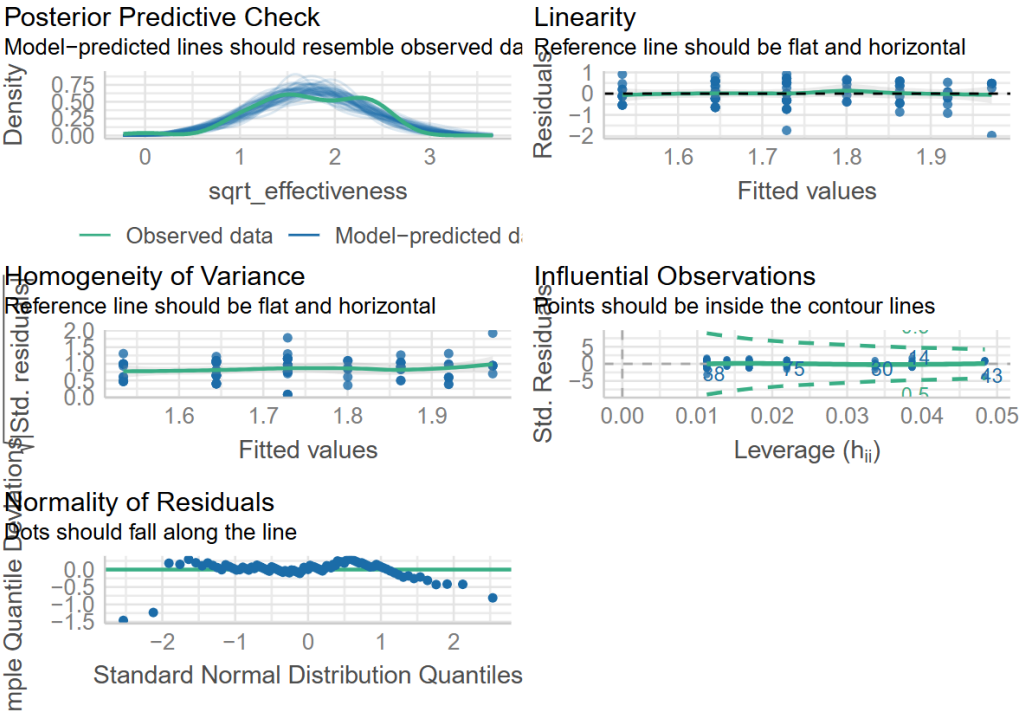
- Homogeneity of variances among errors: the variance of the residuals are the same for all values of X.

The first simple linear regression test we performed was to see if there was a connection between effectiveness of online classes vs. student engagement with the teaching team.

Here, the dependent variable (Y) is the effectiveness of online classes (measured on a scale from 1 to 7). The independent variable (X) is the student engagement with the teaching team (measured on a scale from 1 to 7). We looked at whether or not there was a linear relationship between how much a student engaged with the teaching team and how effective they thought the online education was. We transform the dependent variable and independent variable using sqrt() function so that all the assumptions are verified. Without it, the plots would suggest that the assumptions are not verified. We concluded the following regression formula:

$$Y_{sqrt(eff)} = 1.2673 + 0.2665 \times X_{sqrt(eng)}$$
$$\text{p-value: } 0.028$$

Assumption verification for the first simple linear regression test:



- We verify “Linearity” and “independence of errors” assumptions because the reference lines are flat and horizontal, as shown in the “Linearity” plot.
- We verify “Homogeneity of Variance amongst errors” assumption as the reference line is flat and horizontal, as shown in the “Homogeneity of Variance” plot.
- We verify “Normality of errors” assumption as dots follow along the horizontal line, in ”Normality of Residuals” plot.

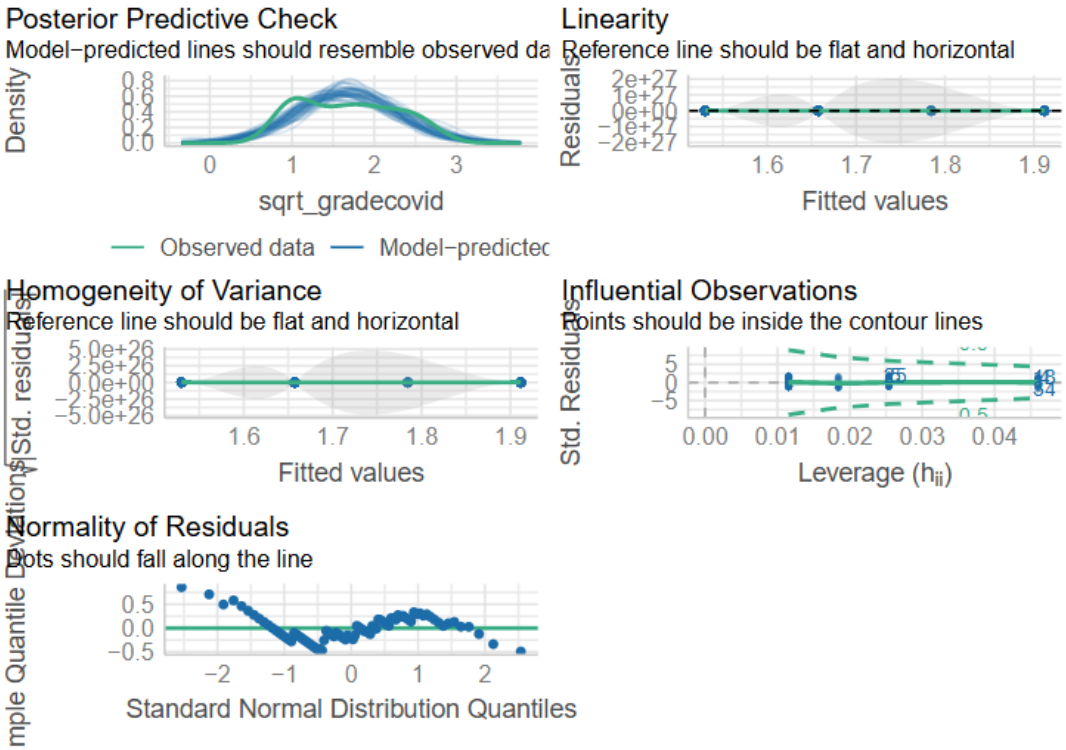
The second simple linear regression test we performed was to see if there was a connection between living arrangements of students and their CGPA during COVID-19.

Here the dependent variable(Y) is the CGPA (measured on a scale from 1 (2.0) to 7 (4.0)). The independent variable(X) is their living arrangement (measured on a scale from 1 (“Owning, living alone”) to 4(“Renting, living with multiple people”)).

We assign 4 to ”Rf = Rental home with 4 or fewer people” and then 1 to ”Owned home with more than 4 people (4)” because we believe that people in rental home with 4 or fewer people likely have more financial responsibility and impacted schedules than people in other ranks which results in lesser marks.
We looked at whether or not there was a linear relationship between the student’s living arrangements and their cumulative grade point averages during COVID-19. We also apply the sqrt() to the dependent variable Y. Without the transformation, the plots would suggest that the assumptions are not verified. We concluded the following regression formula:

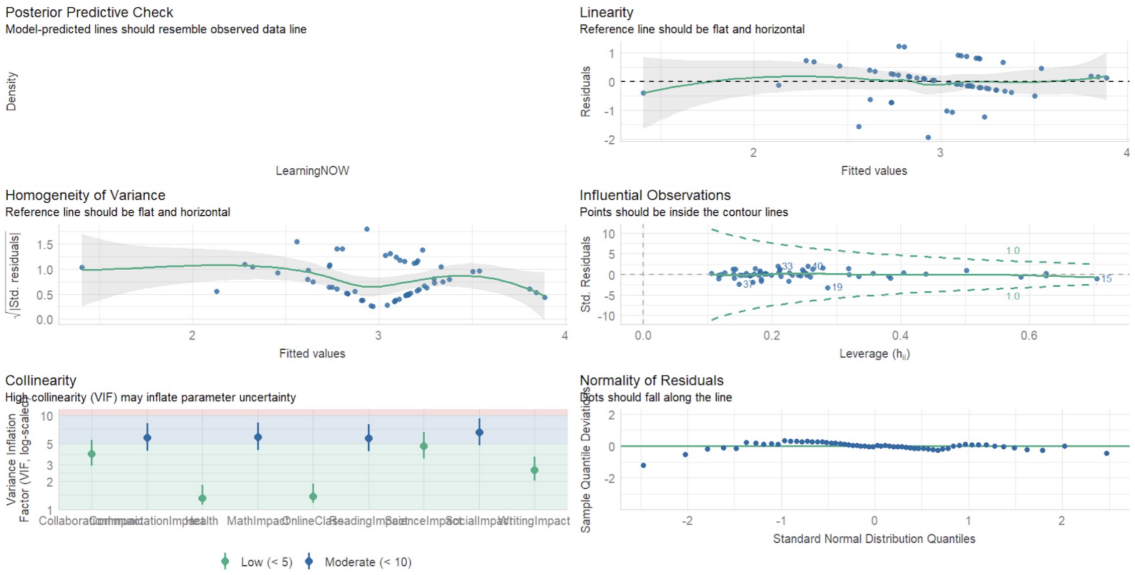
$$Y_{sqrt(GPA)} = 1.40286 + 0.12721 \times X_{living}$$
p-value: 0.0292

Assumption verification for the second simple linear regression test:



- We verify "Linearity" and "independence of errors" assumptions because the reference lines are flat and horizontal, as shown in the "Linearity" plot.
- We verify "Homogeneity of Variance amongst errors" assumption as the reference line is flat and horizontal, as shown in the "Homogeneity of Variance" plot.
- We verify "Normality of errors" assumption as dots follow along the horizontal line, in "Normality of Residuals" plot.

We conducted multiple linear regression on this data set using co-variates that we thought would be most helpful for our predictions. We included the effects COVID-19 had on the following: Health, online classes, impacts on math, reading, writing, communication, social skills, collaboration skills. By understanding the overall impact during COVID-19 to learning skills, we can interpret the learning level now.



We assumed the following for multiple linear regression:

Linearity: the relationship between the explanatory variables and the response variable is linear.

Independence of errors: the response variable is independent of the residuals.

Normality of errors: the residuals must be approximately normally distributed.

Homogeneity of variances amongst errors: the variance of the residuals are the same for all explanatory variables.

No multi-collinearity: the independent variables are not too highly correlated with each other.

- Linearity:
  We can see that the data points do not form a horizontal bound around the 0 line. The green line serves as our trend line for the points that hovers around zero but curves dramatically in areas of scarce data points, like below two on the fitted values. We can see a clear pattern in the residuals, it is decreasing in between the fitted values of 2.5 and 3.5. However, using the performance package we can see that there are no major errors and that our confidence bounds still capture the trend line and zero line. Though it is not perfect, we can still use our linearity assumption.

- Homogeneity of Variance:

Our criteria for homogeneity of variance is that the trend line, green line, is horizontal and the points are randomly bounced around and form a band across the horizontal trend line. However, in this case we see our line curves heavily from 2.5 fitted values onwards. We can see a clear U-shape in the data beneath the trend line. Similarly the performance package has not indicated any drastic errors and our confidence bounds still lay along our desired line, so we can still use our homogeneity of variance assumption.

- Influential Observation:
  All of the points hover around 0 which mean they have very little leverage, or impact to our data. We can also see that all of our data points are in between the contour lines, which means that we don't have any extremely impactful observations. This is good because that means that no one data point, or cluster of them are skewing our results.

- Collinearity:
  The VIF plot takes each variable and plots its VIF value from 1-10. It is a general rule of thumb that the VIF being less than 10 is workable. The VIF is the variable inflation factor is exactly that, due to the collinearity between the variables we see the variance itself is inflated. We see that all of our VIF are below 10 and therefore do not require any major corrections. We are able to say our independence assumption holds.

- Normality of Residuals:
  Here our ideal circumstance is that all the points are along the green trend line. We can see that the points follow along the green trend line and deviate slightly the farther we go down the quantiles. Majority of our points are in the inner quantiles and therefore we can strongly say that the normality assumption holds for our data set.

Graphing the results, we can see that they do look relatively normal and roughly follow a linear relationship. Our outliers also do seem to fall into the necessary bounds needed which means are model is performing decently.
We want to examine the differences in the impact of COVID-19 on education between international and domestic students. We focus on three key factors: Academic Success, Engagement and Learning. A two-sample z-test was conducted to determine if

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	2.97635	0.51267	5.806	6.06e-07 ***
Health	-0.03110	0.05498	-0.566	0.57452
OnlineClass	0.02724	0.05986	0.455	0.65133
MathImpact	0.35483	0.12100	2.932	0.00527 **
ScienceImpact	-0.35147	0.11342	-2.786	0.00778 **
ReadingImpact2	0.06288	0.39662	0.159	0.87474
ReadingImpact3	-0.25836	0.38031	-0.679	0.50040
ReadingImpact4	-0.76730	0.43487	-1.764	0.08445 .
ReadingImpact5	-0.50968	0.42366	-1.201	0.23600
ReadingImpact6	-0.28740	0.46118	-0.623	0.53631
ReadingImpact7	-1.46208	0.77516	-1.886	0.06574 .
WritingImpact	0.06749	0.07826	0.862	0.39308
CommunicationImpact	0.05073	0.13070	0.389	0.69884
SocialImpact	0.09275	0.13521	0.686	0.49624
CollaborationImpact	-0.15819	0.10107	-1.565	0.12455

Table 1: Multiple Linear Regression Results for LearningNOW on Various Impact Factors

there are statistically significant differences in the proportions of students impacted by those factors between international and domestic students.

Let p_1 be the proportion of international students impacted and p_2 be the proportion of domestic students impacted. Here, n_1 denotes the sample size of international students and n_2 denotes the sample size of domestic students.

Assumptions:

- The responses were randomly sampled from international and domestic student groups, and each student's response is considered independent of others.
- Each factor under investigation has binary outcomes ("impacted" or "not impacted") which satisfies the binomial distribution assumption. We analyzed the Likert scale data, with options ranging from "Strongly Disagree", "Disagree", "Slightly Disagree", "I'm not sure", "Slightly Agree", "Agree" and "Strongly Agree". The "Strongly Disagree" response was denoted as 1 and "Strongly Agree" as 7. For Academic Success and Engagement, responses of 1-3 on the questionnaire were considered an impact, while responses of 5-7 indicated no

impact. For Learning, responses of 1-3 indicated no impact, while responses of 5-7 indicate an impact. Responses of 4 were excluded.

- The condition $n_1\ p_1$, $n_1(1-p_1)$, n_2, p_2, $n_2(1-p_2) \geq 5$ is met for all factors, as shown in the tables below.

Factor 1: Academic Success assesses whether students cared about their academic success during COVID-19. If students showed low signs of caring for their academic success, they will be considered impacted by COVID-19.

Group	**Sample Size**	**Proportion Impacted**	**np**	**n(1-p)**
International	41	59%	24	17
Domestic	48	48%	23	25

Table 2: Sample Size and Proportion Impacted by Group (Third Dataset)

Factor 2: Engagement assesses whether students engaged with the teaching team during COVID-19. Students who showed low levels of engagement were considered impacted by COVID-19.

Group	**Sample Size**	**Proportion Impacted**	**np**	**n(1-p)**
International	41	66%	27	14
Domestic	48	63%	30	18

Factor 3: Learning assesses whether students developed new learning strategies because of COVID-19. Students who developed new learning strategies were considered impacted by COVID-19.

Group	**Sample Size (n)**	**Proportion Impacted (p)**	**np**	**n(1-p)**
International	41	59%	24	17
Domestic	48	48%	23	25

Below are the sample data and test results for each factor:

Factor	**z-test**	**p-value**	**Significant?**
Academic	0.17108	0.6791	No
Engagement	0.011461	0.9147	No
Learning	0.61994	0.4311	No

To answer RQ2, We performed comparison of means Additionally, we also calculated Cronbach's Alpha to determine internal consistency to questions dealing with skill

impairment due to COVID-19 as well as to further highlight potential differences in domestic and international student experiences.

The Cronbach's Alpha analysis for responses related to skill impacts caused by COVID-19 is as follows:

Domestic Students: 0.8869
International Students: 0.8730

For comparison of means, we assumed independence, randomness and normality. Since our sample size is $60 > 30$, we were able to do so by the central limit theorem. We also assumed that the variances for the two strata are equal.

The sampled strata were combined into one. For each stratum, the Levene test was performed to check for equal variances. The degrees of freedom for each test was 58. Each test passed and thus the variance for the two independent groups are equal.

Category	F-Value	p-value
Health	1.9029	0.173
Online Class	1.159	0.2861
MathImpact	1.7295	0.1936
ScienceImpact	0.5646	0.4555
ReadingImpact	0.0174	0.8956
WritingImpact	0.1454	0.7044
CommunicationImpact	0.5423	0.4645
SocialImpact	1.1023	0.2981
CollaborationImpact	1.391	0.2431
AcademicCOVID	0.0216	0.8837
AcademicNOW	0.0035	0.9528
EngagementCOVID	1.6133	0.2091
EngagementNOW	1.7727	0.1883
LearningCOVID	0.1981	0.6579
LearningNOW	0.8986	0.3471

Table 3: F-Test Results and p-values for Various Categories

Category	T-Test	p-value
Health	-1.7239	0.09004
OnlineClass	-1.3767	0.1739
MathImpact	-0.3191	0.7508
ScienceImpact	0.0493	0.9608
ReadingImpact	-2.1232	0.0380
WritingImpact	-0.2323	0.8171
CommunicationImpact	0.0414	0.9671
SocialImpact	0.4558	0.6502
CollaborationImpact	0.7290	0.4689
AcademicCOVID	1.1971	0.2362
AcademicNOW	0.0595	0.9528
EngagementCOVID	0.6798	0.4993
EngagementNOW	-1.8016	0.0768
LearningCOVID	-1.8016	0.0768
LearningNOW	-0.8987	0.3725

Table 4: T-Test Results and p-values for Various Categories

The only spot where international and domestic students had a statistically significant difference was on the impact COVID-19 had on their reading skills.

5 Discussion/Results

For the first simple linear regression analysis:
Since the p-value of 0.028 is less than 0.05, this model is significant and engagement in classes is deemed to be useful in explaining effectiveness of online classes.
Notes about meaning of test results,

- Pr($> |t|$): Since Pr($> |t|$) is significantly less than 0.05 we can say that there is a significant connection between engagement and effectiveness.
- Multiple R-squared: 5.427% of variation in effectiveness can be explained by engagement in online classes.

Therefore, we can conclude that a student's engagement with the teaching team is a factor to determine whether they find the class effective.

For the second simple linear regression analysis:
Since the p-value of 0.0292 is less than 0.05, this model is significant and engagement in classes is deemed to be useful in explaining effectiveness of online classes.

Notes about meaning of test results,

- Pr($> |t|$): Since Pr($> |t|$) is significantly less than 0.05 we can say that there is a significant connection between engagement and effectiveness.
- Multiple R-squared: 5.348% of variation in grades can be explained by living arrangements of students during COVID-19 classes.

Therefore, we can conclude that a student's living arrangements during COVID-19 is a key factor to determine how many marks they will achieve.

For the multiple linear regression analysis:
The coefficients seem to show ,That for every one unit increase in health concern that learning now decreased. This is also true for Science, Reading. For Math, Social skills and communication Learning seems to increase per unit increase.

We will carefully observe the coefficients and see what effect the changes have on our response, actively used learning strategies acquired by students. Beginning with health impacts on studies, we can see that it has a coefficient of -0.0311. This implies that as health concerns increased, current learning decreased. Thus, showing that health stress can negatively impact learning even after the fact.

Online class has a coefficient 0.02724, this means that for every increase in online classes effectiveness for the student, their learning now improved. This means that to an extent, the better that online classes were, the better students were able to learn after the pandemic as well.

For the next variable, impact on math it is speaking on if online learning impacted a students math skills and by how much. We can see that it has the coefficient of 0.35483 showing that the more negative impact COVID-19 had on one's math skills the better their learning actually has gotten.

Impact in the science skills seemed to have a negative impact on learning by -0.31597. This means the more impact COVID-19 learning had on science skills negatively impacts learning now as well.

COVID-19's negative impacts on reading skills can be seen via the higher values having more negative impact on acquiring learning strategies. We can see though for participants who disagreed with COVID-19 negatively impacting their reading skills reported a positive impact of 0.06288 on acquiring learning strategies, which means low or no negative impacts on reading allowed for students to acquire new learning strategies. However, noticing that as the agreement of COVID-19's negative impacts on reading skills result in greater negative impacts to acquiring learning strategies. Observing the negative coefficient for those who highly agreed with COVID-19 negatively impacting their reading skills (-1.46208), we can clearly see how the effects of COVID-19 influencing other academic parts of students.

Impacts to writing, communication and social skills do not seem to display the same characteristics as COVID-19's negative impacts on reading skills.

Interestingly we can see that negative impacts on collaboration skills also negatively impacted acquiring learning strategies (-0.15819), which shows that the greater the negative impact of pandemic learning on collaboration skills, the more adversely a student's ability to acquire learning strategies was affected.

For comparing proportions analysis:
Based on the results of the two-sample z-tests for proportions, there was no statistically significant difference between international and domestic students in the areas of academic success, engagement with the teaching team or the development of new learning strategies during COVID-19. For all three factors the p-value exceeded $\alpha = 0.05$ which means that they impacted international and domestic students the same way.

For Cronbach Alpha analysis:

Responses regarding skill impacts due to COVID-19 domestic and international students had fairly high internal consistency. The actual values themselves however show that neither group experienced significantly different impacts to their skills due to COVID-19.

For comparing means analysis:
The only category that raised a notable p-value was reading impact. The p-value for reading impact was 0.0380 which is lower than $\alpha = 0.05$. When p-values are below the alpha level, they raise a statistically significant result. Therefore, there is a statistically significant difference in the impact felt by international students and domestic students in the reading category. All other results gave a result that was higher than $\alpha = 0.05$ and are thus not statistically significant. Therefore, there are no statistically significant differences between international and domestic students for all other categories.

6 Limitations

We initially had three research questions. Our third research question was aimed at exploring how students adjusted to the changes in circumstances during COVID-19. However, we had to remove it because our survey did not provide the data we originally expected. Specifically, we asked participants whether they developed new learning strategies due to COVID-19 and whether they still use those strategies today using Likert scale answers. We realized that the way we structured this question was not serving our needs, so we decided to focus on the other research questions instead. Additionally, the survey itself was complex. We had to get creative with our questions so that we could keep it under the maximum number of questions, following previous steps' guidelines. We think having done so reduced our response rate despite our efforts to incentivize participants to complete it. To mitigate the issues we had with our survey, we could add open-ended questions that ask participants to explain the steps they implemented, if any, as they transitioned from an in-person education practice to an online one. By doing so, we would have sufficient information on how students actually adjust themselves for continued education during COVID-19.

7 Conclusion

We have seen that there is a linear relationship between students engaging with the teaching staff and how effective they thought classes were (RQ1/RQ2). We have also seen that there is a linear relationship between living conditions of students and their academic performance(RQ1). Additionally, we have seen that health concerns negatively impact how students acquire new skills during their education(RQ1). We were unable to identify a significant difference between domestic and international students when it comes to how much COVID-19 impacted them in terms of academic success(RQ1). This, however, does not mean that either international or domestic students were not affected by COVID-19. These findings are useful for institutions and educators to understand how students cope during unprecedented times so that pandemics like COVID-19 can be handled more rigorously in the future.

8 Appendix

Click here for the ChatGPT conversation used as part of table creation.

```
library(dplyr)
library(tidyr)
library(ltm)
library(ggplot2)
library(readr)

set.seed(8) # We all used the same seed number so that the
    randomization was consistent across all of our work.

group <- read.csv("STA304_Group_3_cleaned_up.csv")
groupD <- subset(group, Status == "D")
groupI <- subset(group, Status == "I")
groupD_sampled <- groupD[sample(nrow(groupD), 33), ]
groupI_sampled <- groupD[sample(nrow(groupD), 27), ]
groupD_sampled$Group <- "D"
groupI_sampled$Group <- "I"
combined_sample <- rbind(groupD_sampled, groupI_sampled)
install.packages(car)
library(car)
```

```
leveneTest(Health~Group, data = combined_sample)
leveneTest(OnlineClass~Group, data = combined_sample)
leveneTest(MathImpact~Group, data = combined_sample)
leveneTest(ScienceImpact~Group, data = combined_sample)
leveneTest(ReadingImpact~Group, data = combined_sample)
leveneTest(WritingImpact~Group, data = combined_sample)
leveneTest(CommunicationImpact~Group, data = combined_sample
    )
leveneTest(SocialImpact~Group, data = combined_sample)
leveneTest(CollaborationImpact~Group, data = combined_sample
    )
leveneTest(AcademicCOVID~Group, data = combined_sample)
leveneTest(AcademicNOW~Group, data = combined_sample)
leveneTest(EngagementCOVID~Group, data = combined_sample)
leveneTest(EngagementNOW~Group, data = combined_sample)
leveneTest(LearningCOVID~Group, data = combined_sample)
leveneTest(LearningNOW~Group, data = combined_sample)

t.test(groupD_sampled$Health, groupI_sampled$Health, var.
    equal = TRUE)
t.test(groupD_sampled$OnlineClass, groupI_sampled$
    OnlineClass, var.equal = TRUE)
t.test(groupD_sampled$MathImpact, groupI_sampled$MathImpact,
     var.equal = TRUE)
t.test(groupD_sampled$ScienceImpact, groupI_sampled$
    ScienceImpact, var.equal = TRUE)
t.test(groupD_sampled$ReadingImpact, groupI_sampled$
    ReadingImpact, var.equal = TRUE)
t.test(groupD_sampled$WritingImpact, groupI_sampled$
    WritingImpact, var.equal = TRUE)
t.test(groupD_sampled$CommunicationImpact, groupI_sampled$
    CommunicationImpact, var.equal = TRUE)
t.test(groupD_sampled$SocialImpact, groupI_sampled$
    SocialImpact, var.equal = TRUE)
t.test(groupD_sampled$CollaborationImpact, groupI_sampled$
    CollaborationImpact, var.equal = TRUE)
```

```
t.test(groupD_sampled$AcademicCOVID, groupI_sampled$
    AcademicCOVID, var.equal = TRUE)
t.test(groupD_sampled$AcademicNOW, groupI_sampled$
    AcademicNOW, var.equal = TRUE)
t.test(groupD_sampled$EngagementCOVID, groupI_sampled$
    EngagementCOVID, var.equal = TRUE)
t.test(groupD_sampled$EngagementNOW, groupI_sampled$
    EngagementNOW, var.equal = TRUE)
t.test(groupD_sampled$LearningNOW, groupI_sampled$
    LearningNOW, var.equal = TRUE)

data <- read.csv("path_to_folder/STA304_Group_3_cleaned_up.
    csv")
n1 <- sum(data$Status == "I")
n2 <- sum(data$Status == "D")
p1_academics- mean(data$AcademicCOVID[data$Status == "I"] %
    in% c(1, 2))
p2_academics<- mean(data$AcademicCOVID[data$Status == "D"] %
    in% c(1, 2))
test_result <- prop.test(c(p1_academics, p2_academics), c(n1
    , n2))


p1_engagement <- sum(data$EngagementCOVID[data$Status == "I"
    ] %in% c(1, 2, 3))
p2_engagement <- sum(data$EngagementCOVID[data$Status == "D"
    ] %in% c(1, 2, 3))
test_result <- prop.test(c(p1_engagement, p2_engagement), c(
    n1, n2))


p1_learning <- sum(data$LearningCOVID[data$Status == "I"] %
    in% c(3, 4))
p2_learning <- sum(data$LearningCOVID[data$Status == "D"] %
    in% c(3, 4))
test_result <- prop.test(c(p1_learning, p2_learning), c(n1,
    n2))
```

```r
#simple linear regression

# Test 1
dataset <- data.frame(
  engagement = c(
    5, 7, 4, 5, 5, 2, 1, 2, 5, 1, 2, 6, 3, 1, 7, 2, 5, 4, 7,
        2,
    3, 1, 1, 2, 3, 2, 2, 5, 6, 6, 2, 1, 2, 1, 6, 5, 2, 3, 2,
        2,
    3, 3, 7, 3, 3, 3, 3, 3, 6, 2, 4, 1, 3, 3, 3, 6, 2, 3, 6,
        2,
    3, 2, 1, 1, 6, 3, 3, 2, 5, 2, 3, 7, 1, 2, 5, 4, 2, 4, 5,
        7,
    3, 2, 2, 1, 4, 3, 3, 5, 4
  ),
  effectiveness = c(
    2, 5, 6, 6, 2, 3, 2, 5, 5, 4, 3, 4, 2, 6, 6, 2, 2, 5, 6,
        3,
    4, 3, 2, 2, 5, 2, 1, 4, 3, 1, 2, 1, 2, 1, 3, 5, 2, 0, 5,
        5,
    3, 5, 0, 2, 2, 1, 6, 2, 6, 6, 2, 2, 1, 3, 2, 3, 1, 6, 2,
        4,
    5, 1, 1, 2, 4, 3, 2, 5, 6, 4, 5, 6, 3, 3, 1, 4, 3, 2, 6,
        6,
    1, 5, 1, 1, 3, 6, 7, 3, 6
  )
)


dataset$sqrt_effectiveness <- sqrt(dataset$effectiveness)
dataset$sqrt_engagement <- sqrt(dataset$engagement)
# the reason why I are using sqrt() is because I want to
    eradicate the
#case where the normality of residuals assumption is not
    fulfilled.
```

```r
# performing linear regression
model <- lm(sqrt_effectiveness ~ sqrt_engagement, data=
    dataset)
summary(model)

#verifying assumptions
install.packages("patchwork")
install.packages("see")
install.packages("performance")
library("performance")
check_model(model)

#test 2

dataset2 <- data.frame(
  livingarrangement = c(
    2, 2, 1, 4, 1, 2, 2, 1, 2, 1, 2, 1, 2, 3, 2, 1, 2, 4, 1,
        4,
    2, 2, 1, 1, 1, 4, 3, 1, 1, 2, 2, 3, 1, 4, 2, 2, 2, 2, 1,
        4,
    2, 2, 2, 2, 4, 2, 4, 4, 1, 1, 3, 1, 4, 2, 1, 2, 2, 2, 1,
        1,
    1, 1, 2, 1, 2, 2, 1, 4, 2, 2, 4, 2, 2, 2, 2, 4, 2, 4, 2,
        4,
    2, 3, 2, 2, 2, 4, 4, 1, 3
  ),
  gradecovid = c(
    6, 1, 2, 7, 7, 1, 1, 4, 4, 1, 4, 1, 2, 1, 1, 1, 2, 7, 2,
        3,
    2, 3, 2, 5, 6, 2, 7, 3, 1, 1, 2, 5, 1, 1, 1, 4, 1, 1, 4,
        3,
    5, 3, 1, 1, 3, 3, 6, 3, 1, 1, 1, 6, 2, 4, 1, 7, 7, 4, 1,
        3,
    3, 1, 3, 2, 1, 6, 1, 6, 5, 3, 4, 1, 6, 1, 7, 4, 1, 5, 6,
        3,
```

```
    2, 3, 5, 5, 6, 4, 3, 5, 2
  )
)

dataset2$sqrt_gradecovid <- sqrt(dataset2$gradecovid)
#fit slr model
#I am using the sqrt() transformation to gradecovid only; If
    I did not do this
#then the plots from check_model() would suggest that the
   assumptions are not verified.

#performing simple linear regression
model <- lm(sqrt_gradecovid ~ livingarrangement,data=
   dataset2)
summary(model)


#verifying assumptions
check_model(model)


##cronbachs alpha

data <- read_csv("STA304_Group_3_cleaned_up.csv")

column_sets <- list(
  impacts = c("MathImpact", "ScienceImpact", "ReadingImpact"
     ,
              "WritingImpact", "CommunicationImpact",
              "SocialImpact", "CollaborationImpact")
)

calculate_alpha <- function(data, columns) {
  data_subset <- data %>%
    select(all_of(columns)) %>%
    drop_na()
  cronbach.alpha(data_subset)$alpha
```

```
}

stratified_sample <- data %>%
  group_by(Status) %>%
  sample_n(size = 30, replace = FALSE) %>%
  ungroup()

alpha_df <- data.frame(Category = character(), Group =
    character(), Alpha = numeric())

for (name in names(column_sets)) {
  columns <- column_sets[[name]]

  alpha_domestic <- calculate_alpha(filter(stratified_sample
      , Status == "D"), columns)
  alpha_international <- calculate_alpha(filter(stratified_
     sample, Status == "I"), columns)

  alpha_df <- rbind(alpha_df,
                    data.frame(Category = name, Group = "
                       Domestic", Alpha = alpha_domestic),
                    data.frame(Category = name, Group = "
                       International", Alpha = alpha_
                       international))
  cat("Cronbach's Alpha for", name, "(Domestic):", round(
     alpha_domestic, 4), "\n")
  cat("Cronbach's Alpha for", name, "(International):",
     round(alpha_international, 4), "\n\n")
}
```