Tel: (814)
863-0281
Fax: (814)
863-7114

Statistical Consulting Center
Department of Statistics
Eberly College of Science

The Pennsylvania State
University
323 Thomas Building
University Park, Pa 16802

From: *Claire Kelling, Nikolas Siapoutis, Xiufan Yu, Aniruddha Rajendra Rao,* Graduate Student Consultants

To: Yu Yan

Re: Data Analysis Report

RESEARCH FIELD: College of Education

PROJECT TITLE: Gaming and Students Engagement

1.0 - PROJECT DESCRIPTION

Yu Yan is interested to see whether gamification has a positive effect on student engagement in learning English. The study was conducted by Yu Yan on freshman students in a second tier chinese school over a period of six weeks. The students who took the English course had the option to use the English assessment system where they were assigned to either the control group (no gamification) or treatment group (gamification). She recruited 700 students from four instructors that she knew prior to the survey. Around 500 of the students used the survey, and about 300 of them use it more than once. She used this final sample size of people who logged in more than once, because she is not interested in people who just glanced at the website, but did not come back. Yu randomly assigned the students to either the treatment or control group using STATA. She balanced the gender and the total number in the two groups, although her sample is more female (⅔ female, ⅓ male). The use of this system, in the treatment or control group, is not compulsory, but people usually used the system for 10 minutes per week. The system is used for the first six weeks of the course, at the end of which they fill out a survey. She also mentioned that she has the final exam scores from the students at the end of the course, but it is unclear at this point how this will be incorporated into the analysis. Yu has taken STAT 501 and has knowledge of basic statistical concepts.

1.1 - RESEARCH QUESTIONS

There are two main parts to her Research Questions:

1) Database part: Is there any difference between the values of total time spent on the system (or total number of assessments or average time or average assessment) between the control and treatment group?
2) Survey part: Are students of the treatment group (gamification) more satisfied or more motivated than those that are not? How is their achievement/goal orientation different across groups (what are they motivated by)?

1.2 - STATISTICAL QUESTIONS

Yu is interested in creating t-tests/ANOVA to answer her two research questions. She also has a few statistical questions that we will address in our report briefly. These questions are as follows:
Directly related to research question:

- How does she deal with statistical outliers?
- How does she measure engagement? Which dependent variable is the most representative of this?
- There are some violation of assumptions of the t-test and ANOVA. How does she deal with this?
- There are more female students in the sample. Is this okay? (address including female as a covariate)

Other questions, for future projects:

- Also, she is looking to learn about effect size and power analysis.

1.3 - VARIABLES OF INTEREST

First research question:

Yu has created many of the variables that are the most important in our analysis herself, using the raw data from the database. We are interested in the total number of assessments (1 min tests) that students have taken over the full time period of the study and the total time (in seconds) that they took on all of the tests. We are also interested in the variable “m30_n”, which is a variable that Yu generated in order to more accurately capture the number of visits. If the student logs back in after 30 minutes, she counts this as another visit, or another count in her “m30_n” variable. This is necessary because a student may log in and immediately close out of the browser, or the browser may fail and they need to reload it, so she does not think that these two visits, which are seconds apart, should count as separate. We are also interested in “stay_30m”, which is another variable that Yu generated. This variable is the total time that a visitor is taking assessments divided by “m30_n”. This is a measure of the average amount of time that they stayed for each visit. Lastly, we will create our own variable to capture the average number of assessments taken per visit.

In summary, the four variables that we are considering in the first part of our analysis, to measure engagement are as follows:

- total_time
- total_assm
- stay_30m = total visits/ m30_n (a measure of the average time spent per visit)
- avg_assm = total assessments/ m30_n (a measure of the average # of tests per visit)

Second research question:

In her second research question, Yu is interested in motivation and satisfaction. We will consider this separately from the first question. In the first question, Yu is using data from the database but for this second question, she is relying on students to respond to a survey that is not compulsory. As we may expect, some students did not respond to the survey. Therefore, the dataset that we will use in this part of the analysis is representative of a subset of the students that are captured in the full study.

There are two parts to this analysis. First, she is interested to see if students in the treatment group are more satisfied than those that are not. To answer this question, we have 11 questions in the survey that measure satisfaction. First, we will need to match the survey dataset to the database portion of the data. Then, we can either take these satisfaction responses individually, or perhaps aggregate them, and see if there is a difference in satisfaction between the treatment and control group.

The next part that Yu is interested in is motivation. Originally, the research question seeks to see if students of the treatment group are more motivated than those who are not in the treatment group. However, due to the randomization of the assignment of the treatment, this is not an appropriate way to study motivation in our opinion. We say this because if the randomization is effective, a motivated person is equally likely to be in both the treatment and control group. In terms of data, we have four aggregated measures of motivation that were motivated by literature and provided by Yu. We would like to control for these factors in our another analysis. In other words, if we control for motivation, does this affect how students engage with the system. Is there an interaction between motivation and the treatment? The four aggregated measures that were provided to us are as follows:

- mastApp: mastery approach, where students want to learn more in class
- mastAvo: mastery avoidance, where students are afraid of not learning from class
- perfApp: performance approach, where students want to do well on tests
- perfAvo: performance avoidance, where students are afraid of not doing well

These four measures are scaled from each student’s response to 29 survey questions in a way that has been used by numerous past studies in education, according to Yu.

Additional Analysis:
In addition to to the two research questions that Yu provided, there is also data on the final exam score of the student. We will conduct additional analysis to see if there is an effect of the treatment, as well as other factors, on the final exam score.

Explanatory Variables:
In addition to the variables discussed above, we will control for a couple more variables in our analysis. Specifically, we would like to control for the effect of gender and the teacher and/or english class that they are in. Also, clearly, we have a variable that indicates whether the student is in the treatment group (game) or not. We will match the students in the two datasets through their login id’s.

Variable	**Type**	**Description**
loginID	Discrete	Student login identity
female	Discrete	1= female, 0= male
game	Discrete	1= system with game (treatment group), 0= system without game (control group)

teacher	Discrete	Four different teachers (1= first teacher ,…,4= fourth teacher)
course_total	Discrete	Final exam score in English test (0 - 100)
total_assm	Discrete	Total number of assessments
total_time	Continuous	Total time spent in the system
maxLevel	Discrete	Maximum level of achievement in the test (1-12)
finalLevel	Discrete	Final level of achievement in the test (1-12)
m30_n	Discrete	Total number of visits
stay_30m	Continuous	Average time spent per visit
avg_assm	Continuous	Average number of tests per visit
mastApp	Continuous	Mastery approach (1-5)
mastAvo	Continuous	Mastery avoidance (1-5)
perfApp	Continuous	Performance approach (1-5)
perfAvo	Continuous	Performance avoidance (1-5)

Table 1: Variable type and description

2.0 - EXPLORATORY DATA ANALYSIS (EDA)

Below, we have included some exploratory data analysis for different subsets of the data.

2.1 -EDA for Original Data

We do EDA to get a rough idea about the variables we are interested in and if some of them have any kind of relation. We have 342 students. However, there are 2 students with missing course total. After discussing with Yu, we decided to drop the two students. We now have 340 samples, while 167 are in the gamification group and the rest 173 are in non-gamification (control) group. Since for the first research question we are interested behavior of the variables between the treatment and control group.

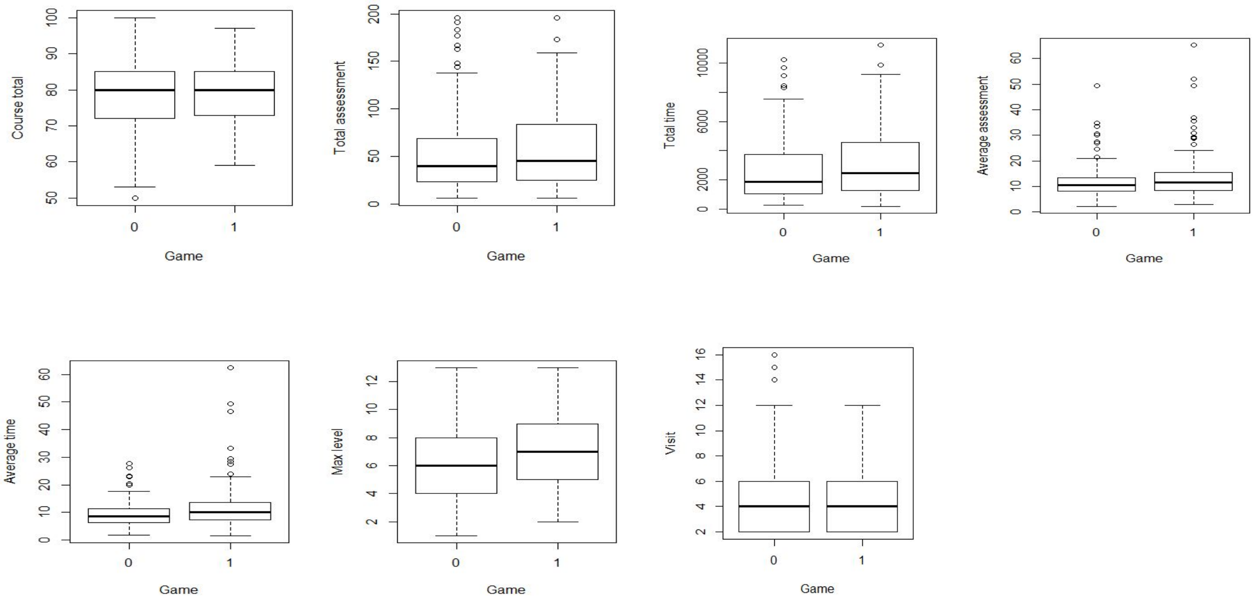

Figure 2.1.1: Boxplots of different variables wrt treatment and control group.

We can see from the boxplot that some of the variables like total time, average time, Max level, average assessment look to be higher for treatment group whereas for other variables it looks the same. Hence in the next section we test if there is any significant statistical difference in the variable of the two groups. Also we can clearly see some outliers. Therefore, we do testing with and without outlier to check if the results are consistent.

Also we are interested in fitting a model of course total. So we check the relation of various variable wrt course total.

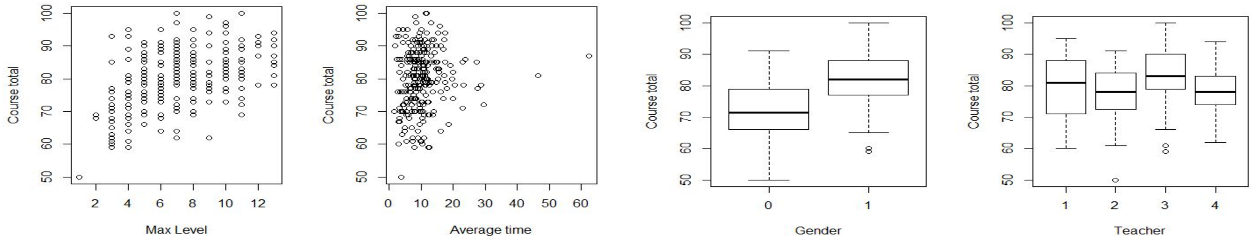

Figure 2.1.2: Histograms and Boxplots for different variables against Course total

Interestingly, from the first two Scatter plots in Figure 2.1.2, it looks like max level and total score are positively related whereas for average time it looks to have some outliers. Seeing the box plot in Figure 2.1.2, we can infer that females in general perform better and there seems to be a difference in the total score for students under different teachers. Using this information we fit a linear model for total scores using these variables and a few variables from the survey data.

2.2 -EDA for Survey Data

We first incorporate the information from original data into the survey data by loginID in order to obtain the game group and other background information, and then we take a look at the combined survey data. There are 278 students who filled out the survey. However, there are 2 students whose information cannot be found in the original dataset, which means we are unable to obtain their group, gender or any other information, so we just ignore these two students. Explicitly, we now obtain a survey data with 276 samples, while 134 are in the gamification group and the rest 142 are in non-gamification group. Furthermore, we notice there are 13 students who only answered the last eleven survey questions related to the evaluation of the system without providing any information regarding their study motivation. Since we regard the analysis on system part and the motivation part as separated, we do not delete these 13 students' information. Another thing we notice is that for the survey question #10, there are a lot of missing values (118 out of 276) . Due to the large amount of missing values, it is not reasonable to delete all the 118 students' information. We will address this question separately in Section 3.

We start with the four aggregated measures that were provided to us which reflects student's motivation. The boxplots in Figure 2.2.1 show slightly differences between the gamification group and non-gamification group, especially in mastAvo and perfAvo, so the groups may not be balanced, despite randomization.

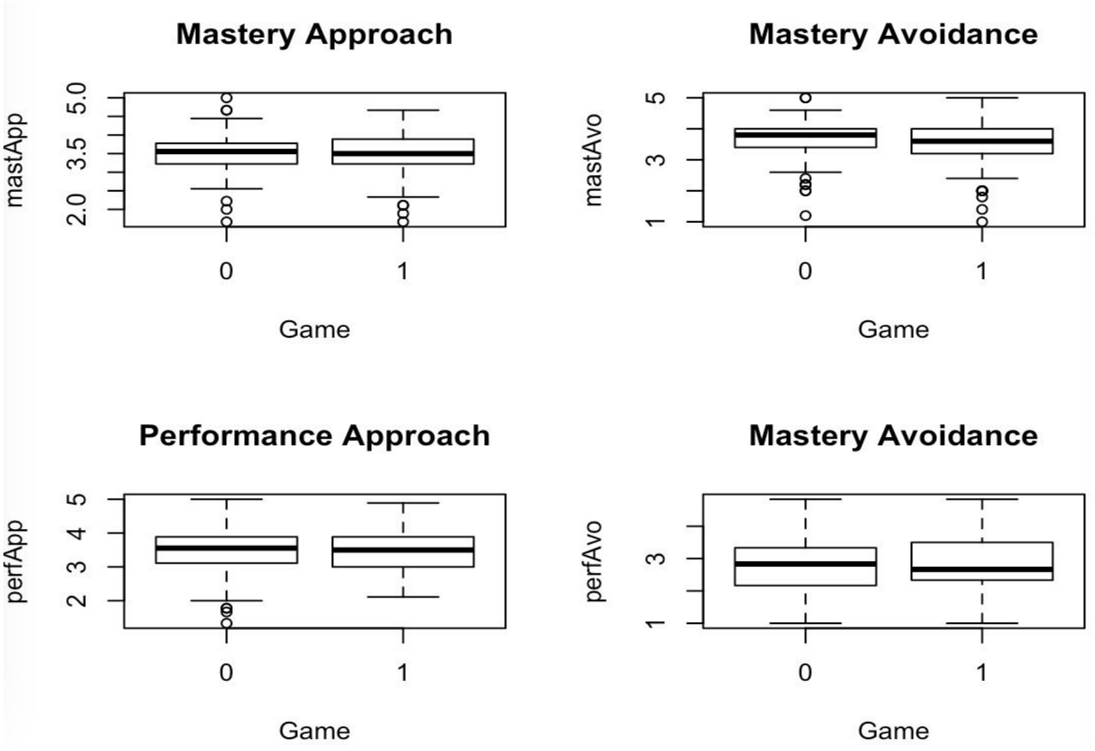

Figure 2.2.1: Boxplots for four aggregated measure variables: mastApp, mastAvo, perfApp, perfAvo

Now we will briefly discuss the 11 survey questions. The plot of their correlation matrix is displayed in Figure 2.2.2.

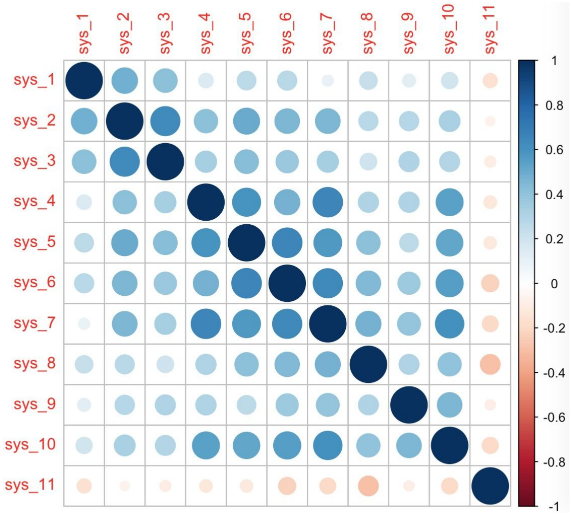

Figure 2.2.2: correlation plot for the eleven survey questions

In this plot, the larger the dot and the darker the color (either blue or red) means that the answers are more correlated for that question. As we expect, the correlation plot on the left of Figure 2.2.2 shows these 11 questions are not independent. Moreover it shows the 11th question are negatively related with all the first ten questions, while the relationship among the first ten questions are all positive. Figure 2.2.3 gives us some first impressions on the differences of people's opinions on the evaluation of the system between two groups. It clearly shows there are obvious differences on the answers for the first question between two groups. And there may be difference on the question 5 and 10, while we cannot see big differences for the remaining questions.

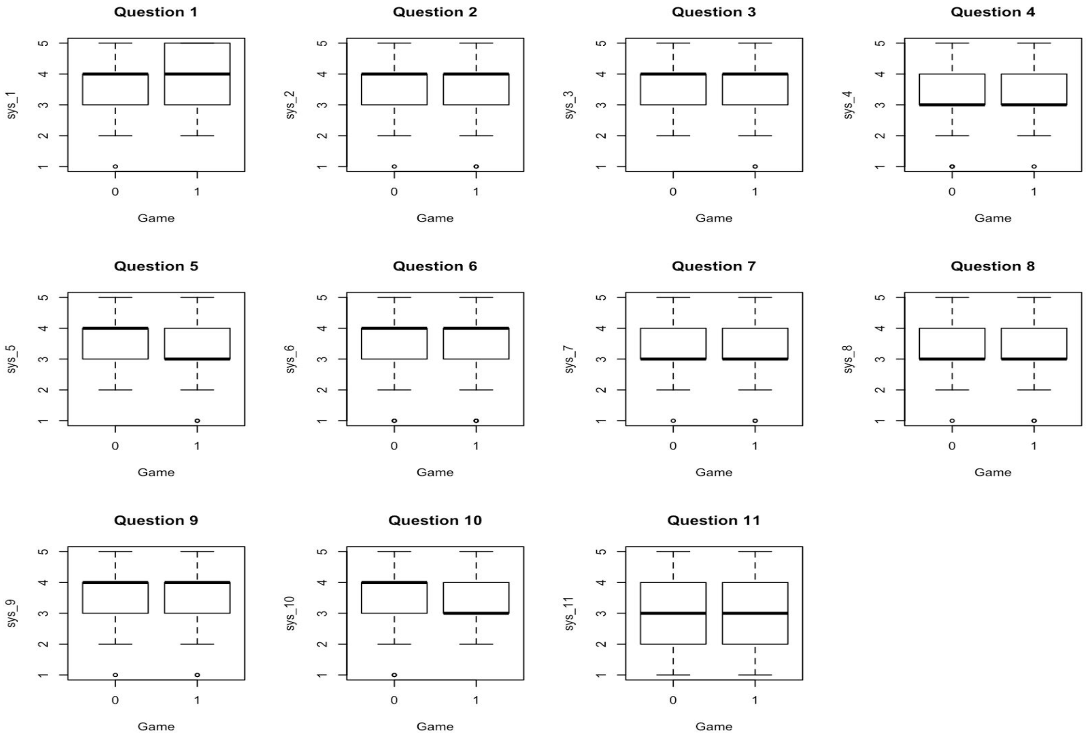

Figure 2.2.3: Boxplots for system questions 1 to 11.

2.3 -Distribution of Original data vs Survey Data vs Missing Data:

We wish to check if all the data follow similar distribution because for some analysis we combine the data and use it but it will be only valid if the data sets are similar.

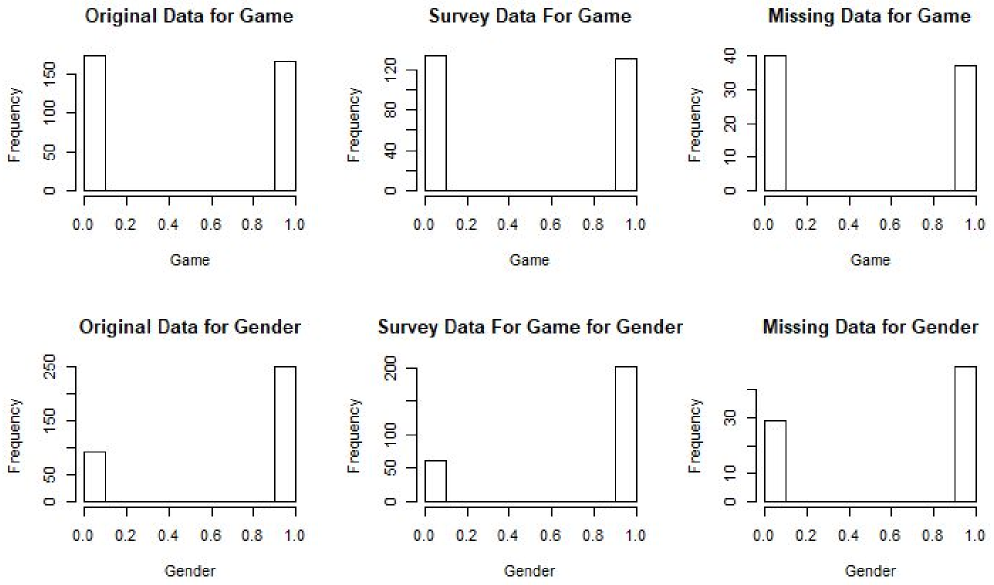

Figure 2.3.1: Comparing distributions of different data subsets, where we see no major concerns

The game variable is equally distributed in all the data set whereas for gender Original data and Survey data have similar distribution but it is different for the missing data (we check this using test for Population Proportions, result is in the appendix). These two variables were of main concerns. We can saw the distribution of Course total, teachers, max level which were fairly similar for Original Data and Survey Data, whereas the distribution of visits looked dissimilar (Graphs are available in Appendix).

3.0 –STATISTICAL ANALYSIS
Research Question 1: Is there a difference in engagement between groups?

In order to answer the first research question, we need to make multiple comparisons of the averages between the control and treatment group. Specifically, we want to test the hypothesis:

H_0: Average of total assessments students spend on the system in control group = Average of total assessments spend on the system in treatment group

H_1: Average of total assessments students spend on the system in control group < Average of total assessments spend on the system in treatment group

Similarly for average time.

We decided to apply two different tests for this question and compare our results. Every statistical test has assumptions which must be met prior to the application of it. So before applying our tests, we check its main assumptions.

First, we want to apply two sample t-test. The main assumptions are:

1. The data follow the normal probability distribution.
2. Checking for equal or unequal variance of the two populations.
3. The two samples are independent and both samples are simple random samples from their respective population.

Clearly, the third assumption holds since the samples from the two groups are not related. Since our sample size in each group is quite large (greater than 30), the normality assumption is not a problem. Also, we notice that the variances of the populations are not equal.

Applying the two sample t-test for heterogeneous variance, we can conclude that the treatment group averages are greater than the control group for the average time. On the other hand, we found total assessment did not have significant difference between the groups (at significance level of 0.05)

Variable	**p-value**	**Conclusion between groups**
Total number of assessments	0.1260	Same
Average time	0.0003217	Treatment Higher

Table 3.1: Our table including our test, p-value and some conclusions for the first research question

Secondly, if the normality assumption made in t-test is in doubt (not large enough size), the nonparametric Wilcoxon-Mann-Whitney test is sometimes suggested. Applying this test in our data, we get the same results as the two sample t-test.

Now we check if there are outliers and if these affect our results. In statistics, an outlier is defined as an observation which stands far away from the most of other observations. Often an outlier is present due to the measurements error. Therefore, one of the most important task in data analysis is to identify and (if is necessary) to remove the outliers. There are different methods to detect the outliers, including standard deviation approach and Tukey's method which uses interquartile (IQR=Q3-Q1) range approach. In this project, we use the Tukey's method. According to Tukey's method, outliers are values below Q1-1.5 IQR or above Q3+1.5 IQR where Q1 is the value in the data set that holds 25% of the values below it and Q3 is the value in the data set that holds 25% of the values above it.

Without the outliers,the treatment group averages are greater than the control group for all the variables of interest.

Similarly, t test can be done for variables like total time, average assessment, total number of visits, total course, max level (final level).

Research Question 2: Is there a difference in satisfaction between groups? What role does motivation play?

a) **Satisfaction:**

For Yu's second research question, we first analyze the difference in satisfaction between groups. Correlation plot in EDA part in Section 2.2 shows the eleven survey questions are not independent. Thus our first thought is to do the multivariate two sample t-test to analyze the eleven questions together. Though the data is not normal distributed, out of the same consideration as in the analysis for research question 1, it is reasonable to believe that the large sample size will assure that their asymptotic distribution are good enough for performing our t-test. However, due to the large amount of missing values in question #10, we are unable to combine it with other questions. And we don't think it is reasonable to just delete all the samples containing just one missing values, because in that way we might lose a lot of useful information. What we are going to do is first we do t-test just on question #10. This t-test gives us a result of p-value=0.456. This is actually a pretty high p-value, which implies, regarding the answers of question 10, we have high confidence that there is no differences between two groups. And then we analyze the other ten survey questions(exclude question #10) together. The test we are considering can be described explicitly as following.

H_0 : The means for the ten questions are all the same for the control group and treatment group.
H_1 : The means for the ten questions are NOT all the same for the two groups mentioned before.

Note: In the statement of the hypothesis, what we mean by "the ten questions" is the eleven questions excluding question #10 because of the large amount of missing data(analyzed before).

Multivariate two sample t-test give us a result of p-value=0.036, which implies we have sufficient evidence to reject H_0 and conclude that the means for all ten questions are NOT all the same for the gamification group and non-gamification group. In addition, we also calculate the cronbach's alpha to see how consistent these questions are. It gives us a result of 0.731, which means we are not able to regard these eleven questions as highly internal consistent when using 0.8 as the criterion. Thus in the next step, we consider testing these questions separately.

To explore the detailed differences,Let $\mu_{0,k}$ denote the mean for question k in control group (non-gamification group), $\mu_{1,k}$ denote the mean for question k in treatment group (gamification group). Then for question k=1,...,11, individually, we perform the following three t-tests (whose hypotheses are illustrated in Table 3.2) for each question. And the outcome p-value are summarized in Table 3.3 .

Table 3.2: Alternative Hypotheses for three t-tests against H_0 : $\mu_{0,k} = \mu_{1,k}$

Test 1: t-test (two.sided)	$H_1 : \mu_{0,k} \neq \mu_{1,k}$
Test 2: t-test (less)	$H_1 : \mu_{0,k} < \mu_{1,k}$
Test 3: t-test (greater)	$H_1 : \mu_{0,k} > \mu_{1,k}$

Table 3.3: summary for individually t-test on 11 questions

Question (k)	1	2	3	4	5	6	7	8	9	10	11
mean-difference ($\mu_{0,k} - \mu_{1,k}$)	-0.23	-0.12	-0.08	-0.03	0.14	-0.01	0.08	-0.07	0.15	0.11	-0.13
p-value (two.sided)	0.027	0.254	0.440	0.788	0.177	0.876	0.483	0.450	0.184	0.456	0.250
p-value (less)	0.014	0.127	0.220	0.394	0.911	0.438	0.758	0.225	0.908	0.772	0.125
p-value (greater))	0.986	0.873	0.780	0.606	0.089	0.562	0.242	0.775	0.092	0.228	0.875

Taking alpha-level of 0.05 as our cutoff for statistical significance, based on the results in Table 3.3, only for question 1, we have sufficient evidence to say that there is significant difference between gamification group and non-gamification group, more specifically, for question 1, it is true that students of the treatment group (gamification) are more satisfied than those of the control group (non-gamification). However, for question 2 to 11, we don't have enough evidence to say that. The t-test study and p-value results are actually consistent with the statements we made in EDA part in Section 2.2.

b) Motivation:

Next, we analyze the role of motivation in this problem. The way we chose to do this is by including motivation in our ANOVA when we are testing to see if there is a difference in the engagement of students. In other words, does motivation play a statistically significant role when looking at engagement levels? We have four different kinds of motivation that we have included in the model, which were provided to us in the dataset. These four kinds of motivation include mastery approach and avoidance as well as performance approach and avoidance.

We use ANOVA to see what factors are significant. We examine the following variables: the four motivation kinds, the teacher (categorical variable), gender (indicator), maxLevel, and game (indicator). We perform the analysis for all four possible measures of engagement, as well as with and without outliers, which was calculated again using Tukey's rule, as in the first research question.

The results of this analysis are summarized in the table below. If the independent variable is significant in the effect on the engagement variable, then the p-value is included. For this problem, we use an alpha-level of 0.05 as my cutoff for statistical significance.

Including outliers/Excluding outliers		Engagement (Dependent) Variable	
		Stay_30m (Average Time)	Total Assessment
	mastApp		

Independent Variable	mastAvo		0.003 / <0.001
	perfApp	0.008 / 0.023	
	perfAvo		
	teacher	<0.001 / <0.001	<0.001 / <0.001
	female		
	maxLevel	<0.001 / <0.001	<0.001 / <0.001
	game	0.038 / NA	

Table 3.4: ANOVA fits for various measures of engagement, including significant variables

From this table, we see that there are a couple of variables that are significant across engagement variables. For example, no matter the engagement variable, maxLevel and teacher are significant. In other words, the teacher that the student has and the maximum level that the student reaches during the training has an impact on their level of engagement. Perhaps the teachers are encouraging use of the system in different ways. Also, it seems clear that a student who reaches a higher level in the system would spend more time and complete more tests than those who do not.

In terms of motivation, we also see a couple of measures that are significant across many engagement variables. For example, “performance approach” is significant for both stay_30m as well as the average number of assessments, though only after removing outliers for the second part. Also, mastery avoidance is significant when it comes to total time and total assessment. Gender is significant only for total time when outliers are included and game is only significant for stay_30m when outliers are included. Therefore, we conclude that performance approach and mastery avoidance are perhaps the two most important types of motivations when it comes to engagement with this online system.

Additional Analysis: What is the effect of these factors on educational outcomes, such as the final exam score?

Finally, we complete some additional analysis where we see what independent variables may affect the final exam score. For this part of the analysis, we use a regression model. We use the same variables as above, as well as stay_30m, as our independent variables and we use course_total as our response. In other words, our baseline regression is as follows:

$$\begin{aligned} course\_total \; = \; & \beta_0 + \beta_1 stay\_30m + \beta_2 mastapp + \beta_3 mastavo + \beta_4 perfapp + \beta_5 perfAvo + \beta_6 as.factor(teacher) \\ & + \beta_7 female + \beta_8 \, maxLevel + \beta_9 game + \varepsilon \end{aligned}$$

After we fit this model, we notice that the following variables are significant: . Below, we include a table with the estimates for the coefficients as well as their p-values. Therefore, in order to improve the final exam scores, game does not appear to be a significant variable, but motivation (such as mastery approach) as well as the teacher and the maximum level achieved in the game does appear to be important. Also, gender is important as well. There is positive coefficient, so females generally do better than males. We see that the assumptions are reasonably satisfied, with plots in the appendix.

Variable	Intercept	stay_30m	mastApp	mastAvo	perfApp	perfAvo	teacher2	teacher3	teacher4	female	maxLevel	game
β estimate	61.019	-0.097	2.844	-1.021	0.839	-1.025	-4.011	0.704	-4.234	6.452	1.449	-0.203
p-value	<0.001	0.238	0.005	0.235	0.328	0.108	0.006	0.555	<0.001	<0.001	<0.001	0.821

Table 3.5.1: Regression coefficients and p-values for the model measuring the final exam score with just fixed effects

The plot to assess the assumptions has been included in the appendix. (Figure App1), where we see that the assumptions are reasonably satisfied.

In addition to this basic linear model, we also fit a mixed effects model, where we include teacher as a random effect. We may not be interested in the specific effect of teacher, but we would like to control for it in this way. So, in order to see the effect of gamification, first we fit the basic model with all of our motivation variables, stay_30m, maxLevel, and gender as fixed effects, and with teacher as a random effect. Then we fit the mixed effects model which includes game, as well as all of the other variables above. We conduct a Likelihood Ratio Test (LRT) between the two models to see if the game makes the model a better fit. We see that the p-value for this is 0.86, so adding game into the model does not make the model a better fit. The estimates for this model are quite similar to the estimates above in Table 3.5.1, and they are below in Table 3.5.2.

Variable	Intercept	stay_30m	mastApp	mastAvo	perfApp	perfAvo	female	maxLevel
β estimate	59.073	-0.102	2.838	-0.914	0.779	-1.010	6.521	1.427

Table 3.5.1: Regression coefficients and p-values for the model measuring the final exam score with random effects for the teacher

3.1 - SUPPLEMENTARY MATERIALS

In addition to the data analysis described above, Yu Yan also wanted to learn about a couple more topics. Specifically, she wanted to learn about power analysis and effect size. These are two fundamental and important topics in statistics. In order to focus our efforts on the analysis of her data and since neither of these topics are too relevant to her current research questions, we will mostly proceed by pointing Yu to additional resources to learn more about power analysis and effect size, with brief descriptions.

There are several great resources available through Penn State Online statistics courses. For example, there is a particularly comprehensive set of lessons on sample size, power, and precision of an estimator in STAT 509: Design and Analysis of Clinical Trials. There is also more basic examples of power analysis in the general Resources Page as well as in STAT 500: Applied Statistics. While most of the above lessons tend to focus on power analysis, there are also a few lessons that focus on effect size, such as in STAT 200: Elementary Statistics and have a bit more detail, such as in STAT 502: Analysis of Variance and Design of Experiments.

4.0 - RECOMMENDATIONS

1. Gamification helps in increasing the engagement of students in terms of Total assessment but not in terms of average time.

2. There exists significant difference on the satisfaction between gamification group and non-gamification group. Especially for the survey question 1, we have sufficient evidence to say that students of the gamification group are more satisfied than those of the non-gamification group.
3. The performance approach and mastery avoidance are the two most important types of motivations when it comes to engagement with this online system.

5.0 - CONSIDERATIONS

1. The results after removing outliers are the same when the average mean for treatment group is higher but in the case where the averages between the groups are same, it sometimes change.
2. The distribution of the Original Data and Survey Data is similar but there is variation wrt the Missing Data.
3. Boxplots show slight differences in the four aggregated motivation measures between the gamification group and non-gamification group, especially in mastAvo and perfAvo, which is a little surprising since the background and study motivation of students in each group should be the same because of the randomization.

Claire Kelling, Nikolas Siapoutis, Xiufan Yu, Aniruddha Rajendra Rao

6.0 - APPENDIX

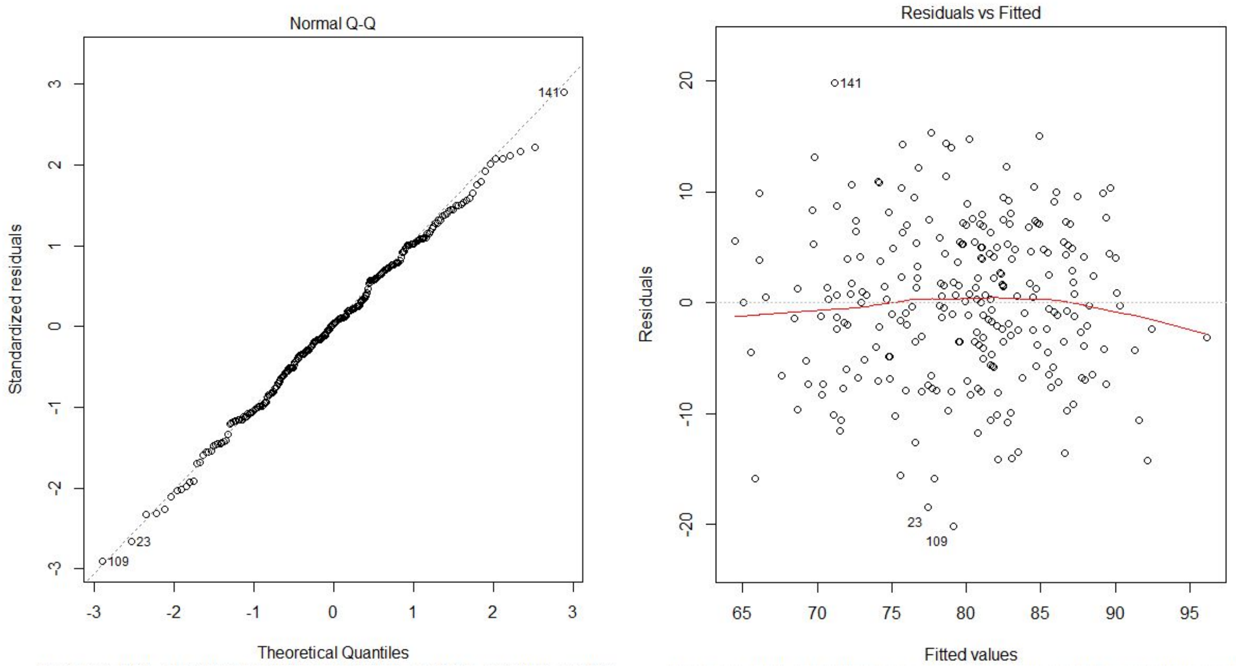

Figure App1: Diagnostic plots of residuals of the regression model, where the assumptions are satisfied.

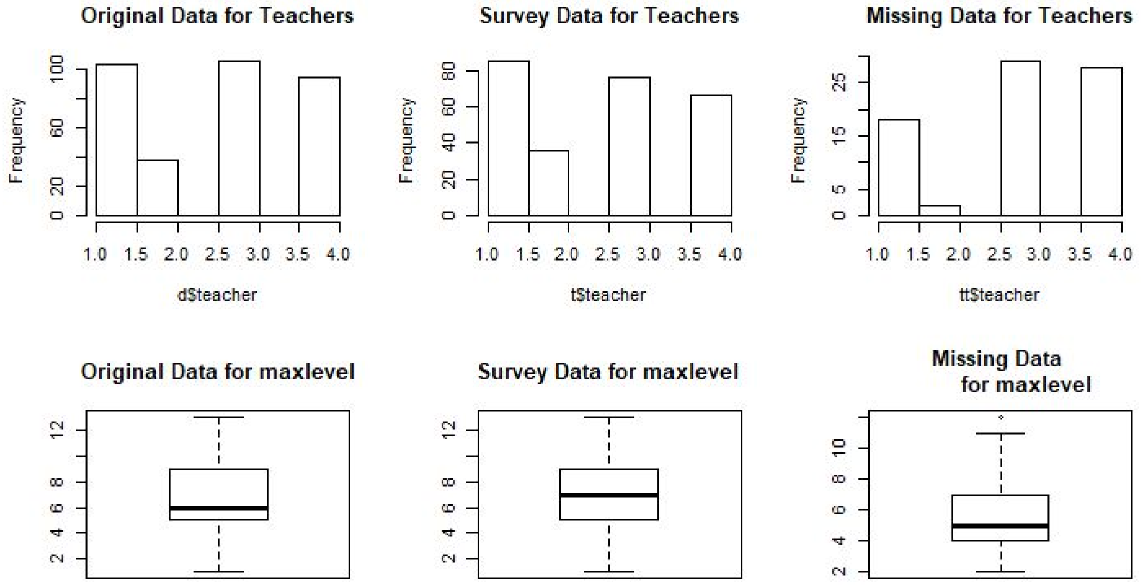

Figure App2: Comparison plots for teachers and maxlevel for different datasets

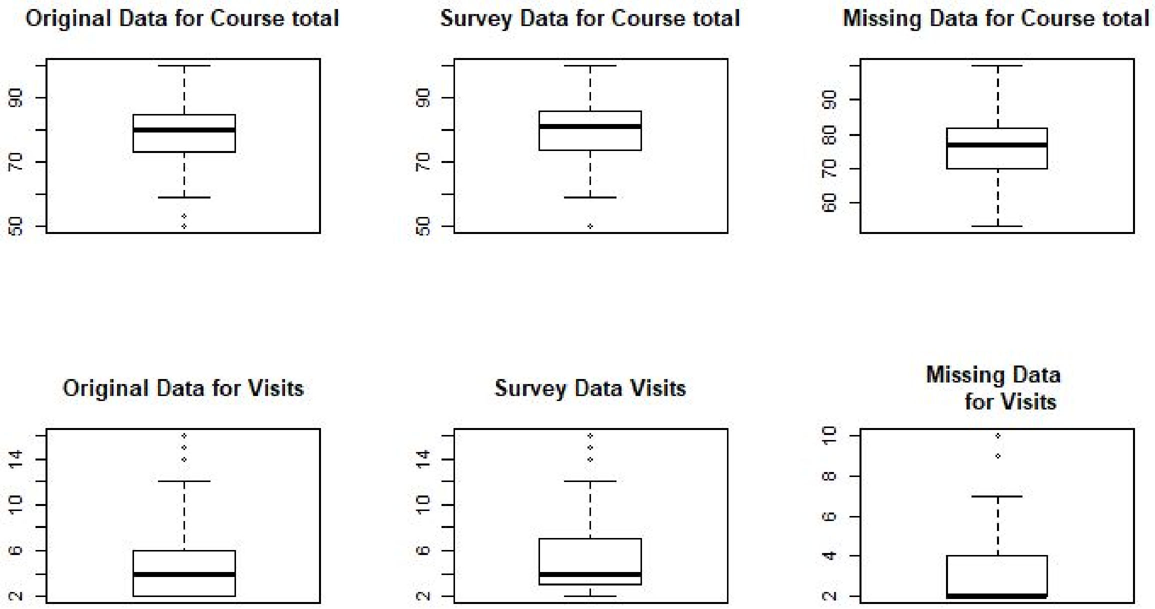

Figure App3:Comparison plots for Course Total and Visits for different datasets