Creative Thinking and LSC101

Statisticians Xhoana Laska, Kaida Lou, Lucy Yin
Master of Statistics Students
NC State University

Researcher Dr. Erica Kosal
Director of Life Sciences First Year (LSFY) Program
NC State University

INTRODUCTION

Launched in 2014, the Life Sciences First Year (LSFY) Program at NC State was created to help life sciences students find the right fit for a degree program. Featuring Biochemistry, Biological Sciences, Genetics, Microbiology, Nutrition Science, Plant Biology and Zoology, the LSFY Program's goal is to challenge students to think critically about what they are interested in studying and creatively about how they can apply their studies to their personal and professional goals.[1] In 2021-2022, 600 students joined the LSFY program in the fall, and 50 students joined in the spring. These 650 students were all required to take the LSC101: Critical & Creative Thinking in the Life Sciences course.

With the LSC101 course, critical thinking was already a measurable component of the class, the researcher would like to measure and assess creative thinking, which can be done using the TTCT (Torrance Test for Creative Thinking) assessment. The TTCT has been used for decades and is verified and reliable. Students enrolled in LSC101 were required to take the TTCT assessment at the beginning and end of the semester. Students must also take the in-house generated and Institutional Review Board (IRB) approved survey and quiz on creative thinking at the beginning and end of the semester. The survey portion was intended to gather students' demographic information and ranked opinions on creativity-related statements, while the quiz portion contained factual questions for students to answer. The same quiz was intended to be given at the beginning and end of the semester, but unfortunately, quiz questions asked at the beginning and end of the semester were not identical.

By using the before-and-after-semester results from the TTCT assessment and in-house survey, the client wanted to investigate the following ideas:

1. Determine if creative thinking increased as a result of taking this LSC101 course.
2. Examine whether patterns exist in creative thinking with demographics (eg. race, gender, hometown, intended major).

In the end, we will provide our client with a summary report of the data analysis. The client will then use the summary report to determine what actions, if any, are needed to help meet the goal of improving critical and creative thinking of students enrolled in LSC101. Since this is the first study of its kind, the client also hopes to publish the data and findings in teaching journals, present them at conferences, and share them with the Department Head of Biological Sciences.

DATA

All data processing and statistical analysis were performed using Python.

[1] "Welcome to the Life Sciences First Year Program", *Life Sciences First Year Program*, North Carolina State University, https://departments.sciences.ncsu.edu/lsfy/

Data Collection

The Fall 2021 LSC101 course was taught by 3 instructors: Dr. Erica Kosal (researcher in this study), Dr. Kenny (Hung-Chieh) Kuo, and Dr. Jason Flores. Twelve different sections of this course were offered. Professor Erica Kosal and Dr. Kenny (Hung-Chieh) Kuo conducted this research, therefore, only students from their sections participated in this study. Data were collected in 4 forms, pre and post-semester TTCT scores, pre and post-semester survey results, pre and post-semester quiz results, and high school GPA information collected from NC State's Undergraduate Admissions.

The TTCT assessment is typically given to K-12 students, all college students are grouped into grade 13. The verbal component of the TTCT assessment consisted of 6 activities and 2 versions of the same test were given to students at the beginning and end of the semester. The 6 subtests prompted students to reference a drawing and answer questions related to "asking", "guessing causes", "guessing consequences", "product improvement", "unusual uses", and "just suppose".[2] The subtests were timed with the goal of writing as many answers as permitted by time. Scoring was done by individuals specifically trained and capable of scoring the assessments with a very high degree of reliability. A score was produced corresponding to each of the 3 dimensions: fluency, flexibility, and originality. Fluency is one of the critical scores since other scores are dependent upon a student giving relevant responses, so flexibility and originality both stem from fluency. To interpret the results produced by professional scorers, we referenced the "Interpreting Test Results" booklet. Both Grade-based and Age-based norms were provided for the TTCT. Grade-based norms typically serve as the primary source for score interpretation, with Age-based norms available for some specialized uses, for our analysis we're working with Age-based norm data. Raw scores for all 3 dimensions were also converted to standard scores with a mean of 100 and a standard deviation of 20. An average standard score was also provided and can be used as a general measure of creative potential.

The researcher provided the following TTCT scores for us to use: fluency SA (age-based standard score for fluency), flexibility SA (age-based standard score for flexibility), originality SA (age-based standard score for originality), average SA (age-based standard score for average), fluency RS (grade-based raw score for fluency), flexibility RS (grade-based raw score for flexibility), and originality RS (grade-based raw score for originality). We have matching scores from the beginning and end of the semester. After reviewing the data, 275 students from the beginning of the semester and 265 students at the end of the semester completed the TTCT assessment, 238 out of 275 at the beginning and 228 out of 265 at the end consented for their data to be used in our study.

The survey consisted of demographic questions and ranked opinions questions and was given to all students at the beginning and end of the semester. Demographic questions asked students for their name, ID number, intended major, zip code, hometown, gender, and race. Ranked opinion questions asked students to rank, on a scale from 1 to 5, creativity-related statements such as "it is important for scientists to be creative", "I consider myself to be creative", etc. Since our research focuses on creative thinking and demographic data, the

[2] https://www.testingmom.com/tests/torrance-test/

ranked opinions portion of the survey was not used in our analysis. A total of 276 students at the beginning and 264 students at the end completed the survey.

The quiz contained factual questions and was graded by the professors. Types of questions on the quiz ranged from short answer, multiple choice, yes/no/maybe, and true/false selections. Because not all questions were asked at the beginning and end of the semester, we did not include the quiz in our analysis.

Students' weighted and unweighted high school GPAs were also provided. Unweighted GPA ranged from 2.974 to 4.481, while weighted GPA ranged from 3.421 to 5.143.

All the data mentioned above were provided to us via Excel and Google Sheets documents, we then uploaded these data into Python for processing, cleaning, and exploratory data analysis (EDA).

Data Processing

Five Excel worksheets were provided to us containing information about all the research subjects involved in this study: pre-semester survey data collected on students for both professors, post-semester survey data collected on students of Dr. Kuo, post-semester survey data collected on students of Dr. Kosal, high school GPA data, and the TTCT score data. In order to create a single clean dataset with all the necessary information, data in each worksheet was first separately cleaned with the goal of achieving consistency in column names, data types, standardized capitalization, etc. We started with the deletion of duplicated rows. Then we identified unique values of every column and fixed misalignments such as: removing extra spaces in first and last names, proper capitalization, etc. Each student ID number had a unique deidentified number assigned. In instances where unusual student ID numbers were observed, such as having greater or lesser than 9 digits, we used the student ID and unique deidentified numbers across different worksheets to double-check and correct these inconsistencies. We also found a double submission of survey responses for the same student, in which case we kept only the later submission. The last step of data cleaning was to remove irrelevant data such as email addresses, timestamps of the survey/quiz submission, and zip codes.

After cleaning the data separately, we found mismatching of data while trying to merge all information together by student. This was most notable when students' gender, race, hometown type, and intended majors did not match from pre and post-semester survey results. Because gender, race, and hometown information were also asked in the TTCT assessment, we utilized the majority rule when assigning a unified value to students with mismatching information. For example, if a student answered "town" for hometown type in the pre-semester survey, answered "suburb" in the post-semester survey, but chose "town" again in the TTCT assessment, "town" was ultimately assigned as this student's hometown type because it was chosen 2 times out of 3. To merge together the TTCT assessment, demographic data, and high school GPA, the unique deidentified number and student ID were used as the linking variables across the different worksheets.

During data processing we found that, although 276 students took the survey at the beginning of the semester and 264 took it at the end, only 257 students took the survey at both times. Out of these 257 students, high school GPA data were only available for 225 students of them. For the TTCT assessment, although 275 students took it at the beginning of the semester and 265 students took it at the end, only 219 complete and consented results were available for us to use. Merging available TTCT results, demographic information, and high school GPA data gave us a total of 198 complete records to work with. These complete records contained all necessary information for our variables of interests: gender, race, hometown type, intended major, weighted high school GPA, and pre and post-semester scores of fluency SA, flexibility SA, originality SA, and average SA. Records with missing values were dropped because the number of subjects with all the necessary information was considered sufficient to make inferences about the population of 600 students that enrolled in the LSC101 class.

METHODS

As directed by our client, we explored the first research idea of determining whether students' creative thinking increased as a result of taking the LSC101 course. Since the TTCT score was the main measurement of creative thinking and we have matched pairs of data from each student, we used the paired sample t-test to explore this research idea.

With the paired sample t-test, the null hypothesis is that the 2 population means are equal, meaning the mean TTCT score from the beginning of the semester should be the same as the mean TTCT score from the end of the semester. Our alternative hypothesis is that the mean TTCT score from the end of the semester is greater than the mean TTCT score from the beginning of the semester.

H_0: $\mu_b = \mu_e$ $\quad\quad$ H_1: $\mu_b < \mu_e$

To calculate the test statistic *t*, we will use the formula $t = \frac{\bar{x}_{diff}}{\frac{s_{diff}}{\sqrt{n}}}$ where $\bar{x}_{diff}$ is the sample mean of the differences, s_{diff} is the sample standard deviation of the differences, and *n* is the sample size which is the number of pairs. If the p-value calculated using the test statistic *t* with *n-1* degrees of freedom is less than our chosen significance level of 5% ($\alpha = 0.05$), then we reject the null hypothesis that there's no difference in mean TTCT score from before and after the semester[3].

There are 3 main assumptions to the paired sample t-test:

1. Observations should be independent of each other.
2. The differences between the matched pairs should be approximately normally distributed.
3. The differences between the matched pairs should not contain extreme outliers.

[3] https://www.statology.org/paired-samples-t-test/

We used fluency SA (age-based standard score for fluency), flexibility SA (age-based standard score for flexibility), originality SA (age-based standard score for originality), average SA (age-based standard score for average) and calculated differences between the pre and post-semester scores for each of these dimensions. We made the assumption that each student's performance on the TTCT assessment was independent of other students. And by visually examining the histograms of the score differences we saw the paired differences are approximately normally distributed although a slight left skew was observed (Figure 1). Since the t-test is fairly robust to non-normality, we concluded this normality assumption was met.

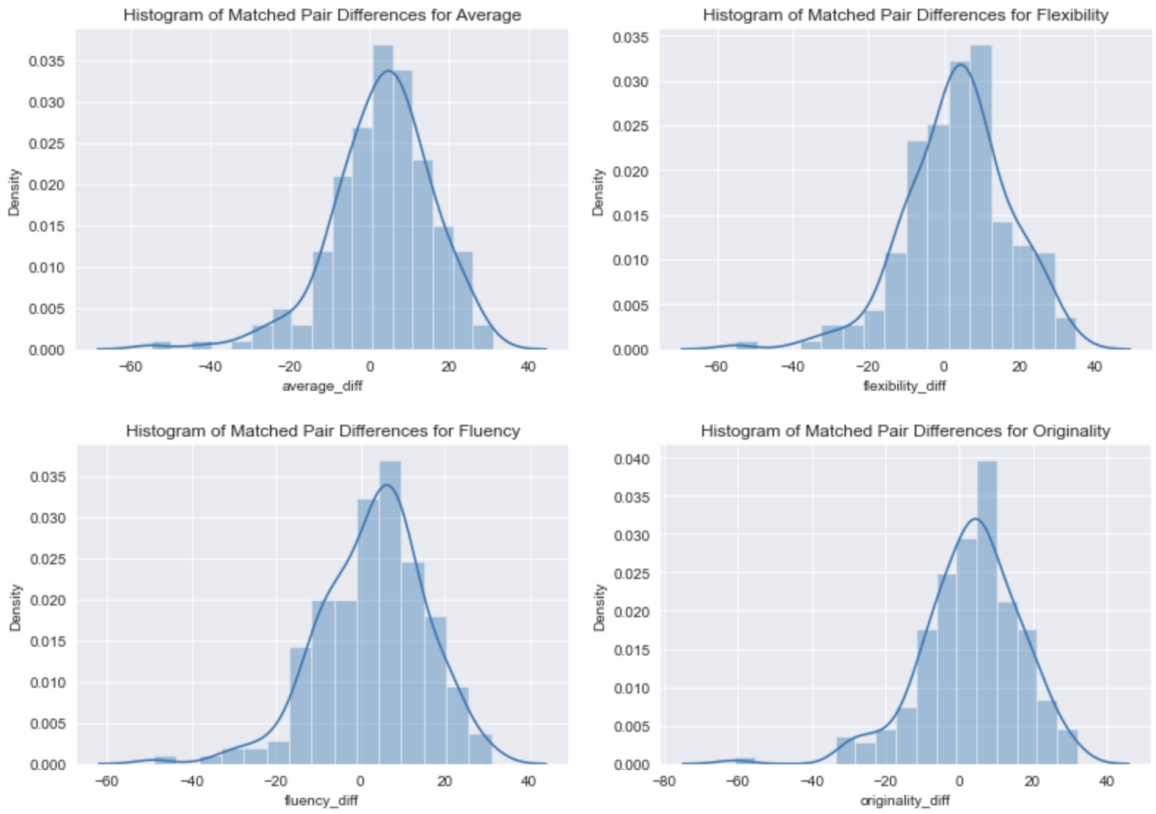

Figure 1. Histograms of Average Difference, Flexibility Difference, Fluency Difference, and Originality Difference

We used boxplots to explore whether extreme observations were observed. Based on a visual examination of the box plots, there were some observations that could be identified as outliers (Figure 2). We used Tukey's Box Plot Method[4] and identified observation 76 as the probable outlier for average difference and originality difference. No probable outliers were found for flexibility difference and fluency difference. A probable outlier is defined as an outlier located outside the outer fence, which is 3 interquartile ranges (IQR) below Q1 and 3 IQRs above Q3. Following Tukey, only the probable outliers are treated. As a result, we plan to run the paired sample t-test twice, once with observation 76 omitted and once with it included and compare any difference in output.

[4] https://towardsdatascience.com/detecting-and-treating-outliers-in-python-part-1-4ece5098b755

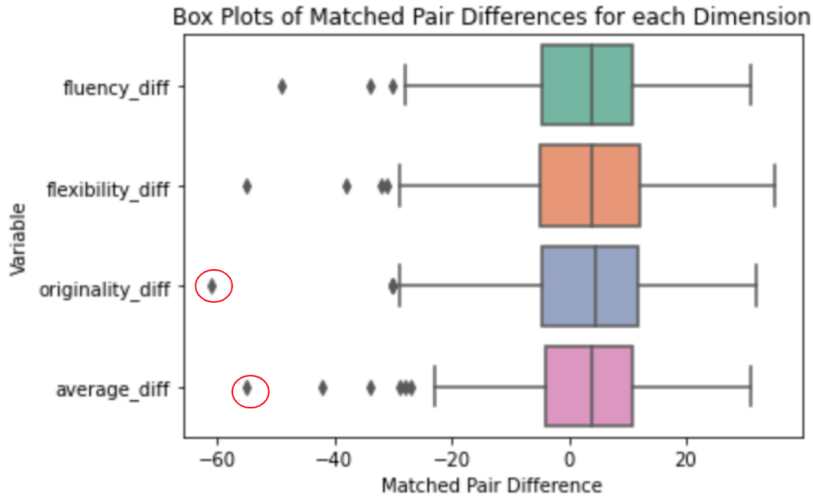

Figure 2. Box Plot of Differences.

To explore the second research idea of finding patterns in demographics, we just used the average SA difference as the response variable as the average standard score can be used as a general measure of creative potential. We first utilized the Random Forest Model and its variable importance plots to identify which demographic variables were more important in predicting the matched pairs' average differences. Although the Random Forest Model did not provide useful predictions, possibly due to our small sample size, we were able to identify gender and race as relatively more important in predicting average differences.

In order to investigate the relationship between creative thinking and demographic information, we narrowed our data down to average standard TTCT scores, intended major, hometown type, gender, race, and weighted high school GPA. We approached this research idea in 3 ways:

1. Using the average TTCT scores as the response, we analyzed each level of the predictor variables to find any increase in post-semester TTCT scores compared to the pre-semester TTCT scores. For example, did the students from hometown type of city improve their TTCT scores after taking the LSC101 class? What about students from the hometown type of suburbs? Etc. We used the paired sample t-test to tackle this idea.
2. After identifying which levels of the predictor variables had a corresponding increase in average difference, we tested if the average difference in TTCT scores was different among those specific levels. For example, do the different hometown types have a varying average difference in TTCT scores? For predictors with 2 levels, we use the two-sample t-test, for predictors with more than 2 levels we use the one-way ANOVA. If a difference was found, we determine which levels are different from the others by applying the Tukey-Kramer method which works for groups of varying sizes.
3. While controlling for race and gender, we tested if other factors or interaction of those factors affected the average difference in TTCT score. We used the multi-way ANOVA method to answer this. The procedure involved testing the significance of the overall model, testing for interaction effects, and in case of no interaction effects, interpreting the significant main effects.

In addition to the paired sample t-test introduced for the first research idea, we incorporated 2 sample t-tests, one-way ANOVA, multi-way ANOVA, and Tukey-Kramer procedures to investigate the second research idea.

The 2 sample t-test is used to test whether or not the means of the 2 populations are equal. The null hypothesis is that the 2 population means are equal, the alternative hypothesis is the two population means are not equal.

H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$

To calculate the test statistic *t*, we will use the formula $t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ where $\bar{x}_1$ and $\bar{x}_2$ are sample means, n_1 and n_2 are sample sizes, and $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$ where s_1^2 and s_2^2 are sample variances. If the p-value calculated using the test statistic *t* with $(n_1 + n_2 - 1)$ degrees of freedom is less than our chosen significance level of 5% ($\alpha = 0.05$), then we can reject the null hypothesis[5].

There are 4 main assumptions to the 2 sample t-test:

1. The data should be approximately normally distributed.
2. The 2 samples should have approximately the same variance, otherwise use Welch's t-test.
3. The data in both samples were obtained using a random sampling method.
4. The observations in one sample should be independent of the observations in the other sample.

The one-way ANOVA is used to compare the means of 3 or more independent groups to determine if there is a difference between the corresponding population means. The null hypothesis is that all population means are equal, and the alternative hypothesis is that at least 1 population mean is different from the rest.

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$ H_1: $at\ least\ 1\ \mu\ differs$

One-way ANOVA can be performed in Python. If the p-value calculated is less than our chosen significance level of 5% ($\alpha = 0.05$), then we reject the null hypothesis and conclude that at least one of the population means is different from the others[6].

There are 4 main assumptions to the one-way ANOVA test:

1. Population should be normally distributed.
2. Population variances should be equal.
3. Samples were obtained using a random sampling method.
4. The observations in one sample should be independent of the observations in the other sample.

[5] https://www.statology.org/two-sample-t-test/
[6] https://www.statology.org/one-way-anova/

If the p-value from the ANOVA is less than the significance level of 5% ($\alpha = 0.05$), we reject the null hypothesis and conclude that we have sufficient evidence to say that at least one of the means of the groups is different from the others. To find out exactly which groups are different we conduct the Tukey-Kramer test[7], which compares the mean between each pairwise combination of groups, for groups of different sizes.

The multi-way ANOVA is used to estimate how the mean of a quantitative variable changes according to the levels of multiple categorical variables, with the goal of finding how multiple independent variables, in combination, affect the dependent variable[8].

There are 4 main assumptions to the multi-way ANOVA test:

1. The response variable is approximately normally distributed for each group.
2. The variances for each group should be roughly equal.
3. Samples were obtained using a random sampling method.
4. The observations in each group are independent of each other.

To satisfy test assumptions, we first regrouped levels of predictor variables (Table 1). For the intended major, since the majority of the students identified with the biology - ba/bs majors, we grouped biology (ba) and biology (bs) into biology, which then had a total of 115 students; we grouped the other intended majors into “others” which combined together biochemistry, genetics, microbiology, nutrition, plant biology, zoology and major outside of science - this new “others” major then had 83 students. For the hometown type, we combined rural area and village into “others” with 35 students, while leaving other levels - city with 52 students, town with 62 students, and suburbs with 49 students as they were. For gender, 167 students identified as female, we then grouped 27 male students, 2 non-binary students, and 2 students who preferred not to answer into an “others” category now with 31 students. For race, leaving the Caucasian; White as it is with 141 students, we combined Asian, African-American; Black; African, Hispanic; Latinx, Indian, Multiracial, Native Hawaiian or Other Pacific Islander into “others” which then had 57 students. Finally, for weighted high school GPA, we converted this continuous variable into a categorical variable with 3 levels. Weighted high school GPA ranged from 3.421 to 5.143, looking for natural break points in the spread while keeping the count of students in mind, the 3 levels consisted of lower ranging from 3.421 to 4.299, medium ranging from 4.3 to 4.499, and higher ranging from 4.5 to 5.143.

[7] https://www.geeksforgeeks.org/tukey-kramer-test-for-post-hoc-analysis/
[8] https://www.statology.org/two-way-anova/

New Intended Major		New Gender		New Race		New Hometown		New HS Weighted GPA	
Level	Count	Level	Count	Level	Count	Level	Count	Level	Count
Biology	115	Female	167	Caucasian; White	141	Town	62	Lower	62
Others	83	Others	31	Others	57	City	52	Medium	69
						Suburb	49	Higher	67
						Others	35		

Table 1. New variable level grouping.

Keeping the statistical practice ethical guidelines in mind, which lists in section D that an ethical statistical practitioner "considers the impact of statistical practice on society, groups, and individuals. Recognizes that statistical practice could adversely affect groups or the public perception of groups, including marginalized groups. Considers approaches to minimize negative impacts in applications or in framing results in reporting[9]". In our analysis, we did not analyze between levels of gender and race to avoid drawing any conclusions that risk adversely affecting any one or more gender or racial groups. For example, we chose not to test if different gender or racial groups have different mean score differences nor draw conclusions that one group did better than the other group.

RESULTS

Before running actual statistical tests, we decided to first take a look at our data. For our first research idea of examining if creative thinking increased as a result of taking the LSC101 course, we started by producing scatter plots of the TTCT score differences for all four dimensions, each dot representing a student (Figure 3). The red line was drawn at zero score difference. Any dot above the red line is indicative that the student's score improved at the end of the semester, hence a positive score difference; vice versa, any dot below the red line indicates the score decreased at the end of the semester. If a dot landed on the red line, the student's score stayed the same. Visually, we saw that the majority of students improved their TTCT scores after taking the LSC101 class.

[9] Ethical Guidelines for Statistical Practice (February 2022). Section D, number 6.

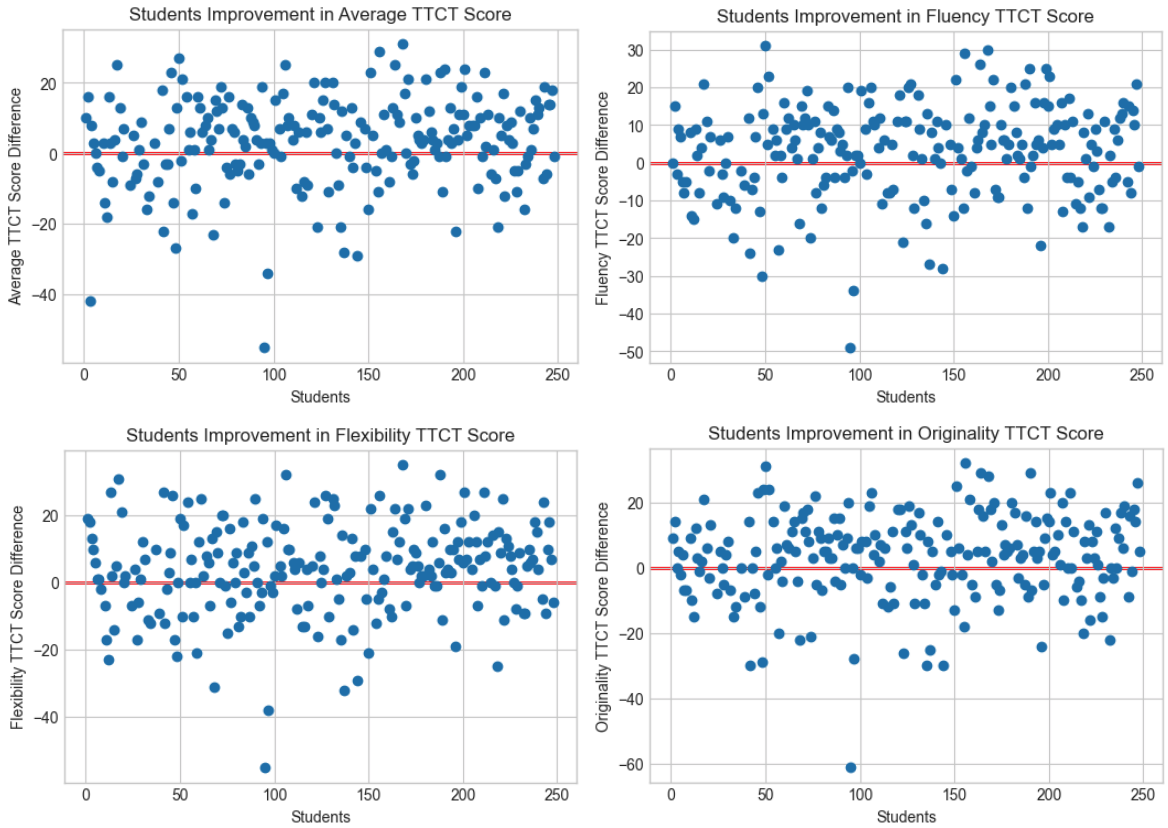

Figure 3. Scatter plots of TTCT score differences for each student.

Depending on the dimension, at the end of the semester, around 61% - 63% of the students performed better on the TTCT assessment, 32% - 35% of the students performed worse, and about 2% - 6% of the students had no change in scores (Table 2). This translated to about 2 out 3 students improving or having no change to their creative thinking based on these TTCT scores.

Score Dimension	% of Score Increase	% of Score Decrease	% of Score No Change
Fluency SA	63.1%	35.4%	1.5%
Flexibility SA	62.1%	32.3%	5.6%
Originality SA	60.6%	33.8%	5.6%
Average	60.6%	35.4%	2.0%

Table 2. Percentage of students with score increase, decrease, and no change.

To see if what we observed in the data has statistical backing, we then ran the paired sample t-tests using each of the 4 dimensions' score differences as the response variable. All p-values were well below our chosen significance level of 5% ($\alpha = 0.05$) (Table 3). Whether we included the probable outlier in the data or not, the p-values were all suggesting that we reject the null hypothesis. We concluded there was a statistically significant difference in the pre and

post-semester mean TTCT scores, which indicates students did improve their creative thinking after taking the LSC101 course.

Dimension	P-Value (including probable outlier obs. 76)	P-Value (excluding probable outlier obs. 76)
Fluency SA	0.0013 (< 0.05)	0.0003 (< 0.05)
Flexibility SA	0.0002 (< 0.05)	0.0000 (< 0.05)
Originality SA	0.0008 (< 0.05)	0.0001 (< 0.05)
Average	0.0010 (< 0.05)	0.0002 (< 0.05)

Table 3. P-values calculated from paired-sample t-tests.

For our second research idea of finding patterns in creative thinking with demographics, we also plotted our data to see if we observe any obvious trends. Looking at the stacked bar plots of score differences for intended major, hometown, and high school GPA, we saw for all levels of these categorical variables, over 50% of students improved their average TTCT score after the course (Figure 4). Looking just at the intended major, a noticeably larger percentage of students intending to study biology improved their TTCT score; similarly, for hometown, a relatively larger percentage of students that grew up in town showed improvements. We also observed a slightly larger percentage of students with medium or higher high school GPAs showed improvements in TTCT scores, although this percentage difference compared with the lower GPA group was not too drastic.

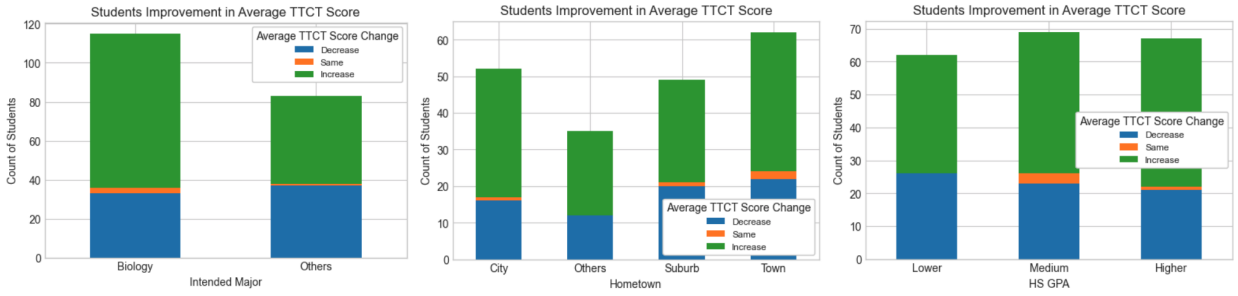

Figure 4. Stacked bar plots of the score differences for each explanatory categorical variable.

To check if our observations have merit, we ran paired sample t-tests on individual demographic variables, with only the average difference as the response and the probable outlier (observation 76) included in the data (Table 4). We saw for the intended major, “Biology” students yielded a significant p-value, meaning we’d reject the null hypothesis that there’s no difference in mean TTCT score from before and after the semester. The same was true for students who chose “Town” as their hometown, the null hypothesis was also rejected in this case. For high school weighted GPA, both the “Medium” and the “Higher” grouped students had significant p-values, therefore for both of these groups, we again rejected that there’s no difference in mean TTCT score. For students intending to study biology, students who grew up

in town, and students whose high school GPA was in the medium or higher groups, the test confirmed they improved their creative thinking after taking the LSC101 course.

Intended Major		**Hometown**		**HS Weighted GPA**	
Level	P-Value	Level	P-Value	Level	P-Value
Biology	0.0001 (< 0.05)	Town	0.0176 (< 0.05)	Lower	0.3044 (> 0.05)
Others	0.6830 (> 0.05)	City	0.1341 (> 0.05)	Medium	0.0362 (< 0.05)
		Suburb	0.2004 (> 0.05)	Higher	0.0170 (< 0.05)
		Others	0.1572 (> 0.05)		

Table 4. P-values calculated from paired-sample t-tests on individual variables.

Since we found 2 significant levels in high school GPA, we proceeded to run the 2-sample t-test to see if significant differences between the 2 levels were observed. To use the 2-sample t-test, the remaining assumption we have yet to validate was whether the variances between the 2 groups were equal. We used Barlett's test on all 3 GPA levels to test for equality of variances and found the p-value to be 0.2747, which is greater than 0.05, indicating we failed to reject the null hypothesis and concluded the 3 levels had equal variances. Running the 2-sample t-test on the "Medium" and "Higher" levels of high school GPA yielded a p-value of 0.7148, which is greater than the significance level of 5%, hence we could not reject the null hypothesis. We concluded the population mean of medium GPA group was not significantly different from the mean of higher GPA group, although the higher GPA group had a smaller p-value.

The final approach to tackle our second research idea was the most complex. While controlling for gender and race, we examined the main effects of all 5 demographic variables as well as interaction effects between intended major, hometown, and high school GPA. We plotted the mean of score difference for all levels of hometown, intended major, and high school GPA to first attempt to visually identify any interactions (Figure 5). In the first plot with hometown and intended major, we saw that irrespective of where students grew up, students intending to study biology had higher improvement in creative thinking. So, visually, there was no interaction effect between hometown and intended major. In the second plot with hometown and high school GPA, since the effect of high school GPA for students growing up in the city was different than those growing up in town, we did observe an interaction effect. We saw that for students with lower high school GPA, the effect on hometown was negative. In the third plot with intended major and high school GPA, we saw that irrespective of the intended major, students with higher high school GPA had higher improvement than the group with medium high school GPA, which in turn had higher improvement than students with lower high school GPA. This suggests no interaction effect existed between the intended major and high school GPA.

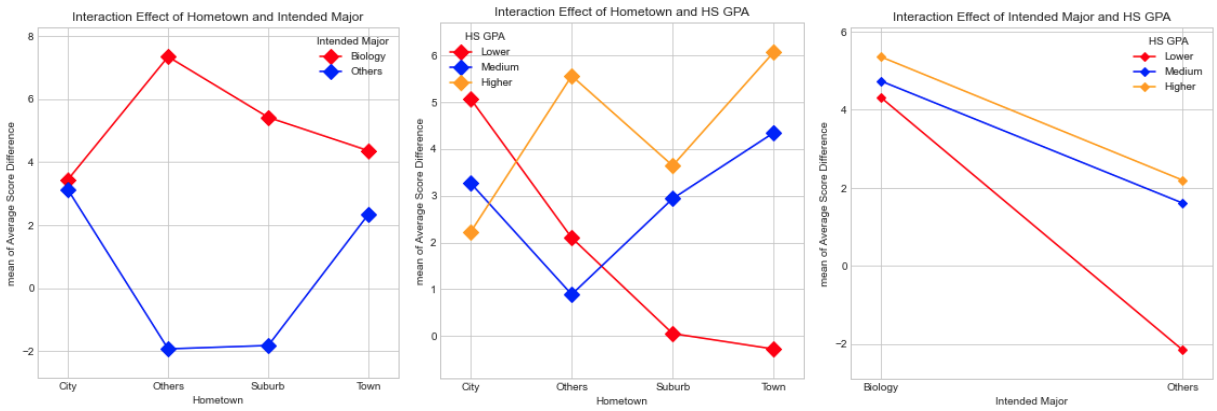

Figure 5. Interaction plots of hometown type, intended major, high school GPA.

We then ran a multi-way ANOVA model. It was crucial to test for interaction terms before assessing the individual effects of factors on the response as if factors do interact, we would not be able to separate their effects. We found that there were no significant interaction effects between the intended major, high school GPA, and hometown, so it was reasonable to look at the main effects. Only the intended major had a significant main effect on the average score difference (Table 5).

5-Way ANOVA	
Predictors	P-Value
Intended Major	0.0215 (< 0.05)
Hometown	0.9473 (> 0.05)
HS Weighted GPA	0.5518 (> 0.05)
Race	0.9411 (> 0.05)
Gender	0.3196 (> 0.05)
Major & Hometown	0.2026 (> 0.05)
Major & HS GPA	0.6585 (> 0.05)
Hometown & HS GPA	0.4463 (> 0.05)
Major & Hometown & HS GPA	0.2570 (> 0.05)

Table 5. P-values calculated from 5-way ANOVA.

Because the intended major's main effect was significant, we used the Tukey-Kramer test to assess the null hypothesis that "Biology" is equal to "Others" in terms of average score differences (Table 6). A p-value of 0.0206 indicates we reject the null hypothesis and conclude that mean differences in the average score were different between biology students and students intending to study other majors. This finding also matched the results from our first

approach where we found the p-value for biology students was significant, but the p-value for other students was not.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Biology	Others	-4.2685	0.0206	-7.8748	-0.6622	True

Table 6. P-values calculated from Tukey's test on intended major.

To verify these findings, we built a 5-way factorial effects model that parametrized the effects of the predictor variables on the response. The model is:

$$Y_{ijklms} = \mu + \alpha_i + \beta_j + \gamma_k + \lambda_l + \upsilon_m + (\gamma\lambda)_{kl} + (\gamma\upsilon)_{km} + (\lambda\upsilon)_{lm} + (\gamma\lambda\upsilon)_{klm} + E_{ijklms}$$

where α_i is the factorial effect for the i-th level of gender, β_j is the factorial effect for the j-th level of race, γ_k is the factorial effect for the k-th level of intended major, λ_l is the factorial effect for the l-th level of hometown, υ_m is the factorial effect for the m-th level of high school GPA, and the other terms are the corresponding interaction terms. When a predictor's effect on the response variable is assessed in a model that contains other predictor variables, that predictor's effect is said to be adjusted for the other predictors. Thus, the gender and race's effects on TTCT score differences were adjusted for the other predictors.

The 5-way factorial effects model was not useful in predicting the response because testing the significance of the overall model produced a p-value of 0.365. Also, R^2 was 0.136 which indicates that the predictors explained only 13.6% of the variability of the response. However, we fitted other 5-way factorial models that included the main effects of the predictors and all the possible combinations of the interaction terms, those models also were not predictive of the response as p-values were very large and small R^2 values suggested low variability explained. Thus, we conclude that some other "lurking" explanatory variables might better explain the variability in the score difference.

DISCUSSION

Based on visual examination of data as well as statistical results, by taking the LSC101 course, students' creative thinking abilities indeed increased, which means the goal of the course was met. If we just dissect intended major, hometown, and high school GPA individually without considering the effects of other variables, we found for each variable at least one level of students increased their creative thinking abilities. Although it's tempting to make statements such as "biology students, students growing up in town, and students with medium or higher weighted high school GPAs are more likely to increase their creative thinking abilities by taking the LSC101 course", in any real-life scenario we also need to consider other factors at play. By doing the multi-way ANOVA we found while controlling for race and gender, only the intended major had a slightly significant effect on the improvement of average TTCT scores, and no

significant interaction patterns in demographics were observed. Since other demographic variables, together, did not appear to have an effect on the change in students' creative thinking, we suspect the presence of unidentified "lurking" variables. Given the right resources, further research can be conducted to explore this possibility further. It is also worth noting that since we regrouped certain levels of explanatory variables to satisify statistical tests' assumptions, the effects of smaller levels, such as hometown type of village or intended major of microbiology, were not directly tested or analyzed.