Evaluating the "General Education" Program for Dietrich College, Carnegie Mellon University

Yanlin Li, yanlinli@andrew.cmu.edu

Abstract

In this project, we investigated several factors that may influence the ratings in the General Education (Gen Ed) program in Dietrich College at Carnegie Mellon University. The data we used contains the rating information of the 91 project papers (referred to as "artifacts"), which were randomly sampled from a Fall and Spring section of the Gen Ed program in Freshman Statistics. The methods we used include exploratory data analysis, intraclass correlation (ICC) analysis, linear mixed model selected by forward and backward fitting, Bayesian Information Criterion (BIC), and Analysis of Variance (ANOVA) tables. We figured out the distribution of ratings for each rubric, rater, semester, and sex, analyzed the agreement rates within the raters, and developed a mixed-effects model to see how factors including raters, semesters, sex, and rubrics are related to the ratings. Finally, we concluded that students did poorly on Critique Design, rating process is fair, ratings are different across semesters and there are no gender bias in this program. Our analysis is still limited by the small data set, unrelated predictors, and lack of other models, which should be tackled during next step.

Key words: Linear mixed model, Mixed Effects Regression Analysis, intraclass correlation, General Education, Carnegie Mellon University

Introduction

Dietrich College at Carnegie Mellon University is in the process of implementing a new program called "General Education" (abbreviation: Gen Ed) for undergraduates. This program includes a set of courses that are mandatory for all undergraduate students to take. Recently, the college is doing an experiment about the program in Freshman Statistics. The students' performance in the program is evaluated on several rubrics by the ratings made by the raters across the college. The raters do not know the information of students, including the students' names, the class they are from, and all the other personal information. The deans office Dietrich College would like to use the experiment data to figure out the factors that can influence the ratings in this program. To be more specific, they are interested in the following four questions below:

1. Is the distribution of ratings for each rubrics pretty much indistinguishable from the other rubrics, or are there rubrics that tend to get especially high or low ratings? Is the distribution of ratings given by each rater pretty much indistinguishable from the other raters, or are there raters that tend to give especially high or low ratings?
2. For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees with the others? Or do they all disagree?
3. More generally, how are the various factors in this experiment (Rater, Semester, Sex, Repeated, Rubric) related to the ratings? Do the factors interact in any interesting ways?
4. Is the distribution of ratings for different semesters and sex groups different from each other?

Data

The data is about the rating information of the 91 project papers (referred to as "artifacts"), which were randomly sampled from a Fall and Spring section of the Gen Ed program in Freshman Statistics. Three raters from different departments were assigned to do the ratings based on 7 rubrics. Only 13 of the 91 artifacts were graded by all the three raters. The rest 78 artifacts were rated by only one rater. Details about the rubrics and rating scale are in the tables (Table 1 & 2) below (Note: These are not the rubrics and rating scale used by instructors or TA's in Freshman Statistics. They are only approved to be used in this experiment.):

Table 1: Rubrics for rating Freshman Statistics projects

Short Name	Full Name	Description
RsrchQ	Research Question	Given a scenario, the student generates, critiques or evaluates a relevant empirical research question.
CritDes	Critique Design	Given an empirical research question, the student critiques or evaluates to what extent a study design convincingly answer that question.
InitEDA	Initial EDA	Given a data set, the student appropriately describes the data and provides initial Exploratory Data Analysis.
SelMeth	Select Method(s)	Given a data set and a research question, the student selects appropriate method(s) to analyze the data.
InterpRes	Interpret Results	The student appropriately interprets the results of the selected method(s).
VisOrg	Visual Organization	The student communicates in an organized, coherent and effective fashion with visual elements (charts, graphs, tables, etc.).
TxtOrg	Text Organization	The student communicates in an organized, coherent and effective fashion with text elements (words, sentences, paragraphs, section and subsection titles, etc.).

Table 2: Rating scale used for all rubrics

Rating	Meaning
1	Student does not generate any relevant evidence.
2	Student generates evidence with significant flaws.
3	Student generates competent evidence; no flaws, or only minor ones.
4	Student generates outstanding evidence; comprehensive and sophisticated.

This table below (Table 3 in Page 4) contains all the variables that are included in the data file. Variables that are not expected to be useful for analysis are shown in parentheses.

Here is a summary table of ratings for different rubrics. (Table 4)

Table 4: Summary table: Rubrics

Variable	Median	Mean	SD
RsrchQ	2	2.35	0.59
CritDes	2	1.87	0.84
InitEDA	2	2.44	0.70
SelMeth	2	2.07	0.49
InterpRes	3	2.49	0.61

Variable	Median	Mean	SD
VisOrg	2	2.41	0.67
TxtOrg	3	2.60	0.70

Here is the counts for other categorical variables. (Table 5 in Page 4)

One thing to highlight is that there are some missing data in this data set. There are two missing rating data. The first one comes from spring semester and rubric Critique Design. The second one is from fall semester and rubric Visual Organization. Considering that rating is our outcome variable, we chose to drop the two data. Besides, there is one artifact with missing sex data (expressed in "–" in the summary table above, count 7 means the seven rubric items with rating). Because we did not have a convincing justification to code it as female or male, we also dropped the data for all the analysis involving it. However, the 13 artifacts seen by all three raters do not have missing values, so we do not need to worry about that for analysis using this subset of data.

Methods

Question 1

Is the distribution of ratings for each rubrics pretty much indistinguishable from the other rubrics, or are there rubrics that tend to get especially high or low ratings? Is the distribution of ratings given by each rater pretty much indistinguishable from the other raters, or are there raters that tend to give especially high or low ratings?

I first used the full set of data to do the analysis below.

To begin with, I made a summary table containing the counts of different ratings (1, 2, 3, and 4) for each rubric. The distribution of ratings for each rubric was also evaluated in the summary table with their mean and standard deviation. Besides, seven bar charts of ratings regarding different rubrics were plotted for comparison. We use these visualizations and tables to compare the ratings for different rubrics. (See Technical Appendix 1 Rubrics Page 14)

Similarly, I made a similar summary table for different raters (rater 1, 2, and 3), which contains the counts of different ratings, mean and standard deviation. Bar charts were also plotted in the same way as what I did for the rubrics. We expect the distribution of ratings for each raters are similar. Any abnormal patterns from the visualizations and tables were reported. (See Technical Appendix 1 Raters Page 16)

Then, I switched to the subset of 13 artifacts seen by all three raters to do the same set of analysis. I compared the results made by the subset with the results for full data to see whether the 13 artifacts are representative of the whole set of 91 artifacts. (See Technical Appendix 1 Distribution of subset of artifacts seen by all three raters Page 14)

Question 2

For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees with the others? Or do they all disagree?

To address this question, I will use the subset of 13 artifacts seen by all three raters. There are two main steps.

First, intraclass correlation (ICC) was used to evaluate the agreement among the raters. ICC is the common correlation among a certain group of values. If the raters generally agree with each other, we expect the the ICC's between the ratings for each artifact is high, and the ICC's within all the ratings made by each rater is low. That means, when one rater's rating goes up from one artifact to the next, we expect the other

Table 3: Variables in the data (Variables that are not expected to be useful for analysis are shown in parentheses)

Variable Name	Values	Description
(X)	1, 2, 3,...	Row number in the data set
Rater	1, 2 or 3	Which of the three raters gave a rating
(Sample)	1, 2, 3,...	Sample number
(Overlap)	1, 2, ..., 13	Unique identifier for artifact seen by all 3 raters
Semester	Fall or Spring	Which semester the artifact came from
Sex	M or F	Sex or gender of student who created the artifact
RsrchQ	1, 2, 3 or 4	Rating on Research Question
CritDes	1, 2, 3 or 4	Rating on Critique Design
InitEDA	1, 2, 3 or 4	Rating on Initial EDA
SelMeth	1, 2, 3 or 4	Rating on Select Method(s)
InterpRes	1, 2, 3 or 4	Rating on Interpret Results
VisOrg	1, 2, 3 or 4	Rating on Visual Organization
TxtOrg	1, 2, 3 or 4	Rating on Text Organization
Artifact	(text labels)	Unique identifier for each artifact
Repeated	0 or 1	1 = this is one of the 13 artifacts seen by all 3 raters

Table 5: Summary table: Categorical variables

Repeated	count	Semester	count	Sex	count	Rater	count
0	546	F19	581	–	7	1	273
1	273	S19	238	F	448	2	273
				M	364	3	273

raters' ratings to go up as well. The ratings should not be predictable by raters. (See Technical Appendix 2 Overall agreement Page 19)

Second, pairwise rating agreement of the three raters were evaluated under the seven rubrics. The purpose of this step is to identify any rater that disagrees with the others, and to ensure that the raters are grading in the same way. A two-way table was made for each pair of raters (rater 1 & rater 2, or rater 2 & rater 3, or rater 1 & rater 3) for each rubric. The percentage of artifacts that the two raters gave the same ratings for that rubric is the agreement rate. We investigate the rates and see whether there is any rater whose agreement rate with both other raters are low. If that happens, we claim that the rater is disagree with the other raters with the ratings regarding that rubric. (See Technical Appendix 2 Agreement for different raters Page 20)

Question 3

More generally, how are the various factors in this experiment (Rater, Semester, Sex, Repeated, Rubric) related to the ratings? Do the factors interact in any interesting ways?

To address this question, we can fit a different model for each rubric, but it is hard to directly examine interactions with Rubric, since each model considers only one Rubric at a time. So we would like to put Rubric into the model and put random slope on it. In order to find the best model, there are three steps: selecting random effects, selecting fixed effects, and finding interactions.

To select random effects, we used the methods of BIC and forward-fitting. The candidates of the random

effects groups are Rater and Artifact. We think both of them may influence the intercept, and the slopes for all the factors except Repeated. (See Technical Appendix 3 Model selection, Random effects Page 21)

After we got the model with random effects, fixed effects were selected by an iterative algorithm we designed. We started from a model with all the possible fixed effects. The algorithm can automatically delete factors from the model until the Bayesian Information Criterion (BIC) value reaches a minimum, or no factors can be deleted according to Analysis of Variance (ANOVA) table. (For the full algorithm and pseudo-codes, See Appendix 3 Model Selection, Fixed Effects Page 21)

Next, we started to work on interaction terms. We fitted nine models with possible interaction terms including pairwise interaction of fixed effects, and the interaction of existed fixed effects and other variables. The BIC values of these models were calculated and compared with the model without interactions. If any interaction term was justified to be added into the model, the same approach would be performed for more interaction terms. (See Appendix 3 Model Selection, Interaction Page 21)

Finally, we got the final optimal model. This model was then used to do the ICC analysis on artifacts, which is similar to what we did in question 2. We assessed the level of how raters agree with each other and compared the results to the results we got in question 2. (See Appendix 3 Model Selection, ICC Page 25)

Question 4

Is the distribution of ratings for different semesters and sex groups different from each other?

In this part, we investigated whether the rating is different for different semester and sex. Data with missing values were deleted. We used plots, summary table and one-way ANOVA tests for our analysis.

First, bar charts of ratings for different semesters and sex groups were plotted. We examined the plots for any differences and similarities. (See Technical Appendix 4 Plots Page 25) Second, we used summary tables with counts, mean, and standard deviation to compare the distributions of rating in different semester and sex categories. (See Technical Appendix 4 Summary tables Page 25) Third, we focused on the means of ratings in different categories, and hypothesized that the means for different semester and sex groups are the same. One-way Analysis of Variance (ANOVA) tests were performed to test this hypothesis and p-values were reported. (See Technical Appendix 4 One-way ANOVA test for means Page 27)

Results

Question 1

Is the distribution of ratings for each rubrics pretty much indistinguishable from the other rubrics, or are there rubrics that tend to get especially high or low ratings? Is the distribution of ratings given by each rater pretty much indistinguishable from the other raters, or are there raters that tend to give especially high or low ratings?

Rubrics

Here is a summary table for different rubrics with the number of different grades and the grade means and standard deviation. (Table 6 in Page 7)

We can see that the mean grade for Text Organization is the highest, and the one for Critique Design is the lowest. Critique Design also has the highest standard deviation and the highest number of rating 1.

Here are the bar charts for the seven rubrics. (Figure 1 in Page 6) The graphs are similar except the one for Critique Design. Only the distribution of Critique Design skews to the right, and all the other are highest with ratings 2 or 3.

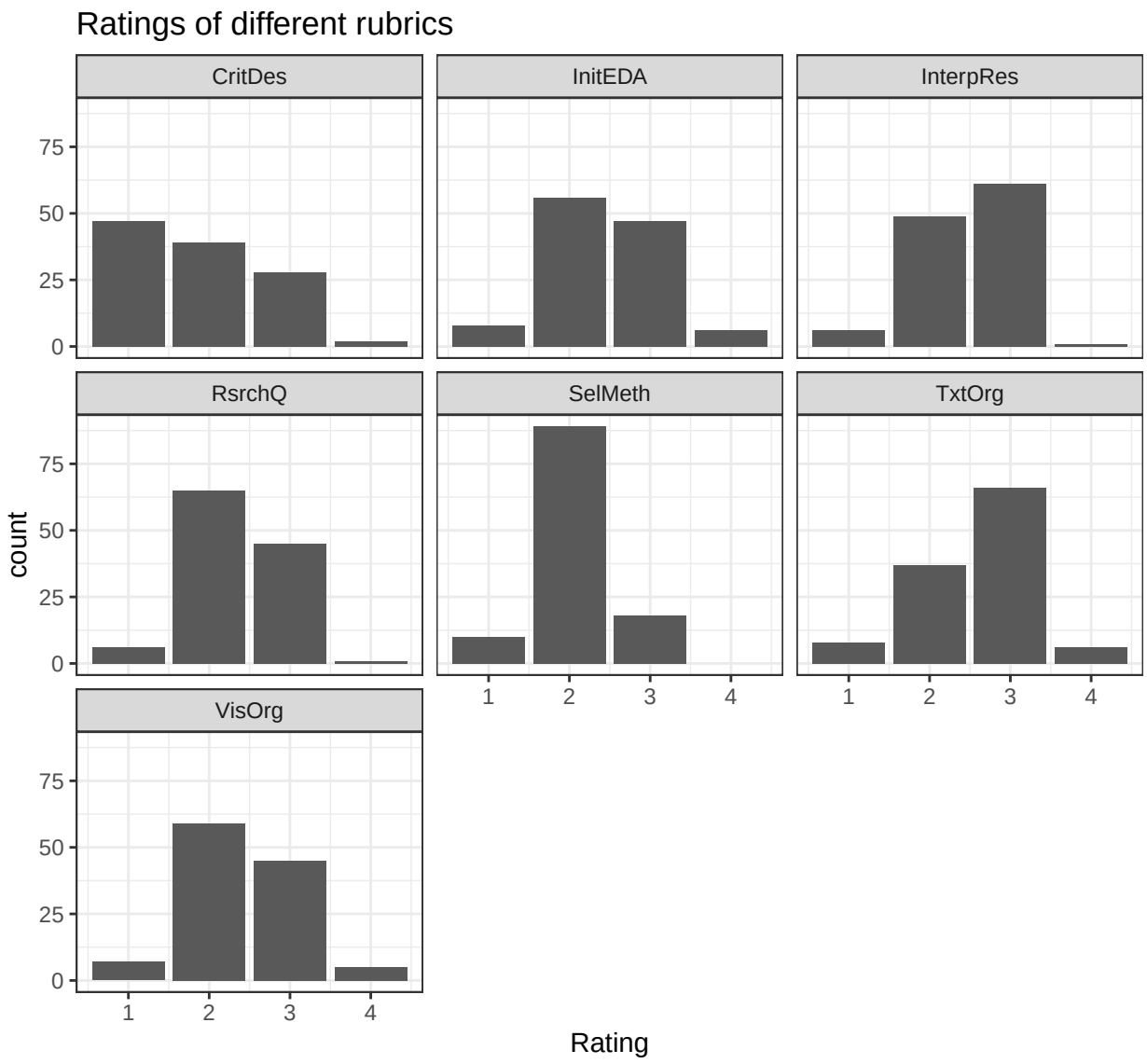

Figure 1: Ratings of different rubrics

Table 6: Summary table: Rubics

	Count				Distribution	
Rubrics	Rating 1	Rating 2	Rating 3	Rating 4	Mean	SD
RsrchQ	6	65	45	1	2.35	0.59
CritDes	47	39	28	2	1.87	0.84
InitEDA	8	56	47	6	2.44	0.70
SelMeth	10	89	18	0	2.07	0.49
InterpRes	6	49	61	1	2.49	0.61
VisOrg	7	59	45	5	2.41	0.67
TxtOrg	8	37	66	6	2.60	0.70

Raters

Here is a summary table for ratings produced by the three raters. (Table 7 in Page 7)

Table 7: Summary table: Raters

	Count				Distribution	
Raters	Rating 1	Rating 2	Rating 3	Rating 4	Mean	SD
Rater 1	40	150	78	5	2.18	0.69
Rater 2	23	119	120	10	2.43	0.70
Rater 3	29	125	112	6	2.35	0.70

We can see that the ratings have similar distributions. Among the three raters, Rater 1 is the most likely to give low ratings among the three raters. The mean ratings made by Rater 1 is also lowest. Rater 2 is the most likely to give high ratings and the mean ratings made by Rater 2 is also highest. The standard deviation for the three raters are similar.

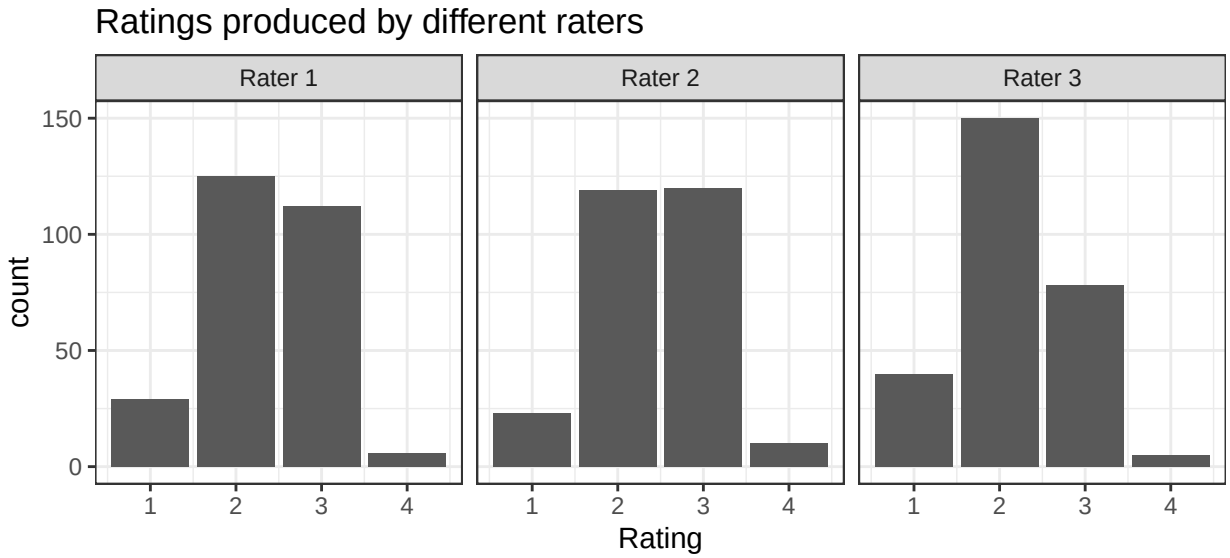

Figure 2: Ratings produced by different raters

The same pattern can be observed from the three bar charts above. (Figure 2)

Then we did the same analysis for the subset of 13 artifacts seen by all three raters. The results are similar. (See Technical Appendix 1 Distribution of subset of artifacts seen by all three raters Page 14, for details)

Thus, we conclude that these thirteen artifacts are representative of the whole set of 91 artifacts.

Question 2

For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees with the others? Or do they all disagree?

The table below (Table 8) shows the intraclass correlation (ICC) values and the agreement rates for artifacts and raters. The last column is the rater that tend to disagree with the others in each rubric.

Table 8: Pairwise agreement rate for different rubrics

	ICC		Agreement rates			
Rubrics	Rater	Artifacts	Rater 1 & 2	Rater 2 & 3	Rater 1 & 3	Disagree Rater
RsrchQ	0.00	0.19	0.38	0.54	0.77	Rater 2
CritDes	0.00	0.57	0.54	0.69	0.62	None
InitEDA	0.00	0.49	0.69	0.85	0.54	Rater 1
SelMeth	0.00	0.52	0.92	0.69	0.62	Rater 3
InterpRes	0.01	0.23	0.62	0.62	0.54	None
VisOrg	0.00	0.59	0.54	0.77	0.77	None
TxtOrg	0.00	0.14	0.69	0.54	0.62	None

Overall agreement

We can see that the ICC values for raters are all very small (< 0.05). The ICC value means correlation between ratings on any two different artifacts by the same rater. The low value indicates that there is not much correlation between artifacts, grouped by rater. Thus, one cannot predict the rating on one artifact from the rating on another. The ICC value is the highest for rating on Interpret Results.

The ICC values for artifacts are all high (> 0.1), which means that the link between the ratings from different raters for certain artifacts are high. The consistency of ratings can be high for different raters, and these raters generally agree on their scores. The consistency are especially high for rating on Visual Organization, Critique Design, and Select Methods.

This part of the analysis can be conducted by the full data. (For the full result table, see Technical Appendix 2, Full data, Page 20) The ICC values produced by full data are higher for raters, and similar for artifacts.

Agreement for different raters

The columns about Agreement Rates in the table show the percentage of ratings that each pair of raters agree with each other (i.e. giving the same rating). The last column is the rater that tend to disagree with the others in each rubric. Here, "disagree with others" means that the agreement rates of that rater with both other raters are significantly lower than the agreement rate between the other two raters. "None" in this column means that the rate of agreement for each pairs are similar. We can see that all the raters only disagree with the others in one rubric item. None of the raters seems to consistently disagree with other raters in all rubrics. Besides, we can see from the table that the agreement rate between rater 1 and rater 2 on Research Question is especially low.

Question 3

More generally, how are the various factors in this experiment (Rater, Semester, Sex, Repeated, Rubric) related to the ratings? Do the factors interact in any interesting ways?

As is described in Methods section, the random effects of artifact on rubric and rater were first selected by forward-fitting and Bayesian Information Criterion (BIC). (See details in Appendix 3 Model Selection, Random effects Page 21) Fixed effects of rubric, rater, and repeated were left in the model after backward-fitting of fixed effects. (Algorithm can be found in Appendix 3 Model Selection, Fixed Effects Page 21) All the possible interaction terms were added into the model, and tested by BIC. After the test, we found that the model with no interaction got the lowest BIC value, so we choose not to add any interaction term into the model. (See details in Appendix 3 Model Selection, Interaction Page 21)

Finally, we got our optimal model:

Rating ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact) + Rubric + Rater + Repeated

The expression above can be a bit confusing to you. Actually, it just means that the model contains:

1. Fixed effects of Rubric, Rater, and Repeated (An estimate for all the artifacts)
2. Random slope of Artifact on Rubric and Rater (Different adjustments to slope estimates for different artifacts)

This table below shows the fixed effects in the model. (Table 9) The column estimate is the esmated coefficients of the fixed effects.

Table 9: Fixed effects in the final model

Terms	estimate	std error	t-statistic	p-value
(Intercept)	1.97	0.10	20.54	1.00
RubricInitEDA	0.54	0.09	5.73	1.00
RubricInterpRes	0.59	0.10	5.84	1.00
RubricRsrchQ	0.46	0.09	5.32	1.00
RubricSelMeth	0.16	0.09	1.74	0.96
RubricTxtOrg	0.69	0.10	6.88	1.00
RubricVisOrg	0.53	0.10	5.41	1.00
Rater2	0.00	0.08	-0.01	0.50
Rater3	-0.19	0.07	-2.81	0.00
Repeated	-0.08	0.08	-1.01	0.16

We can start from interpreting the fixed effects.

From the summary, we can see that almost all the fixed effects are insignificant (p value higher than 0.05). We include them just because they work well in improving the model (lower BIC values). Despite from the insignificance, we still have one significant finding and two insignificant (not convinced by the data) findings.

1. Significant: The slope estimate of rater 3 is negative, so artifacts graded by rater 3 are generally lower in scores. This slope can be influenced by random effects, so this inference can be different for each artifact.
2. Insignificant: All the rubric terms have positive slope estimate, so they have higher average ratings than Critique Design. This inference can be different for each artifact for the same reason.
3. Insignificant: The slope estimate for repeated (= 1) is negative, so we expect a slightly lower score for the artifacts seen by all the three raters

Table 10 shows the first six rows of the random effects (six artifacts). For the full random effects in the model, see table in Appendix 3 Final model, Page 23-24.

Given the random effects, we now have all the estimates for our final model. So here is how we interpret the whole model:

Table 10: Random effects of Artifact in final model

	Random slopes on Rubrics							Random slopes on Raters		
	CritDes	InitEDA	InterpRes	RsrchQ	SelMeth	TxtOrg	VisOrg	Rater1	Rater2	Rater3
100	0.63	-0.23	-0.07	-0.12	0.24	-0.41	-0.32	0.04	-0.05	0.05
101	-0.76	0.31	-0.17	-0.74	0.04	0.21	0.33	-0.06	0.07	-0.06
102	-0.93	-0.32	-0.13	-0.65	0.06	-1.01	-0.43	-0.08	0.09	-0.09
103	-0.04	0.30	0.16	-0.23	0.29	0.17	-0.25	0.05	-0.05	0.05
104	-0.77	0.27	0.19	-0.21	0.09	0.16	-0.15	-0.02	0.02	-0.02
105	-0.74	-0.47	-0.26	-0.35	-0.08	-0.46	-0.39	-0.06	0.07	-0.07

1. Interpreting intercept: We do not have any random intercept, so for all the artifacts, the intercept should be 1.97.

2. Interpreting slopes: Factor Repeated does not have a random intercept, so the adjustment should always be -0.08 for all the artifacts seen by all three raters, and 0 for others. Slopes for rubrics and raters contain both fixed effects and random effects. For each artifact, the random effect on each rubric and rater terms are different. The slope for each term for each artifact is the addition of its fixed effect and random effect. We take Artifact 100 as an example. The fixed effect of rubric Initial EDA is 0.54 for all the artifacts. The random effect of artifact 100 on the slope of Initial EDA is -0.23. So for this artifact, the slope for Initial EDA is 0.54 - 0.23 = 0.31. Similar calculations can be done for all the artifacts and factors.

Finally, we assessed the ICC values using this model. The table below shows these values. (Table 11)

Table 11: ICC analysis for Artifacts under final model

	ICC for Artifacts	
Rubrics	Final model	Previous
CritDes	0.83	0.19
InitEDA	0.73	0.57
InterpRes	0.49	0.49
RsrchQ	0.58	0.52
SelMeth	0.27	0.23
TxtOrg	0.69	0.59
VisOrg	0.67	0.14

We can see that considering all the other possible factors, the correlations of ratings for each rubric within artifacts become higher. This means that the raters actually agree more on their ratings.

Question 4

Is the distribution of ratings for different semesters and sex groups different from each other?

As is mentioned in Methods sections, We uses plots, summary table and one-way ANOVA tests for this part.

Plots

From the bar charts above (Figure 3), we can see that:

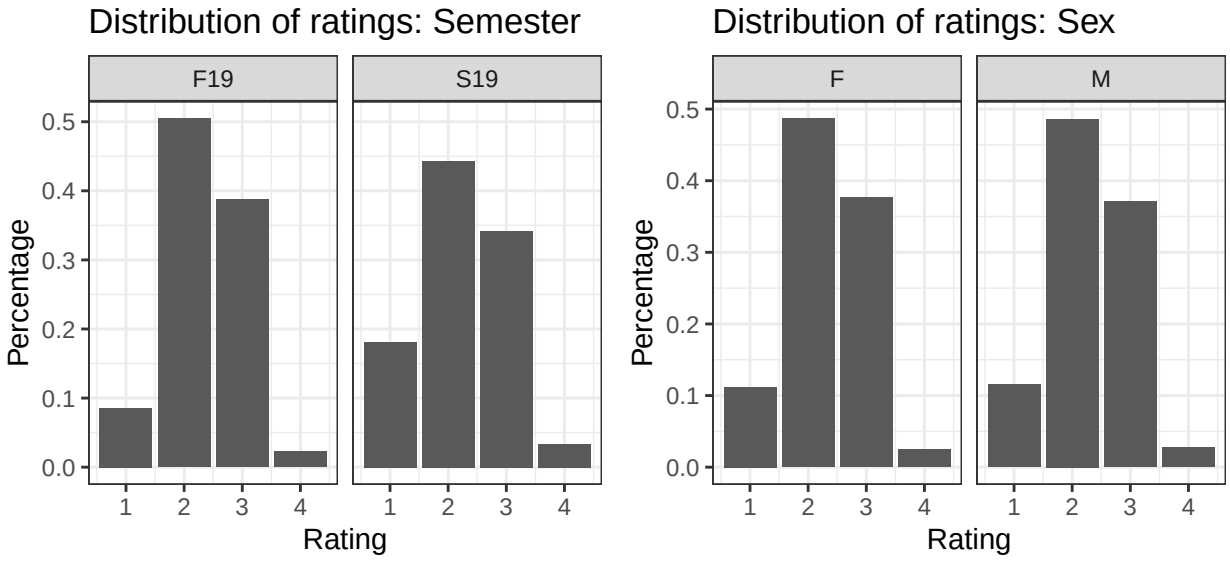

Figure 3: Distribution of ratings: Semester and Sex

1. For both semesters, 2 are the most common rating, followed by 3, 1, and 4. However, the percentage of artifacts (and rubrics) with ratings 2 and 3 is lower in spring semester than fall semester. The situation is opposite for ratings 1 and 4.
2. The distributions of ratings are similar for different sex of authors.

Summary tables

Table 12: Summary tables for semester and sex

Semester	count	mean	sd
F19	573	2.347295	0.6662287
S19	237	2.227848	0.7803091

Sex	count	mean	sd
F	446	2.313901	0.7000061
M	364	2.310440	0.7079251

From the summary tables above (Table 12), we can see that:

1. There are more samples for fall semester than spring semester. The mean rating for fall is higher than that in spring, and the standard deviation for fall is lower than spring.
2. There are more samples for female authors than male authors. The mean and standard deviation of ratings are similar across sex groups.

One-way ANOVA test for means

In this part, we hypothesize that the mean ratings are the same across different semesters and sex groups.

From the outputs above (Table 13), we can see that:

1. When we hypothesize that the mean ratings are the same for the two semesters, the p-value we got from one-way ANOVA test is $0.03 < 0.05$. This means we should reject the null hypothesis, and conclude that the mean ratings are different across semesters. To be more specific, students got significantly higher ratings in fall semester than spring semester.

Table 13: ANOVA tables for semester and sex

term	df	sumsq	meansq	statistic	p.value
Semester	1	2.39	2.39	4.86	0.03
Residuals	808	397.58	0.49	NA	NA
term	df	sumsq	meansq	statistic	p.value
Sex	1	0.00	0.0	0	0.94
Residuals	808	399.97	0.5	NA	NA

2. When we hypothesize that the mean ratings are the same for different sex groups, the p-value we got from one-way ANOVA test is $0.94 > 0.05$. This means we should accept the null hypothesis, and conclude that the mean ratings are the same across sex groups.

Discussion

The main purpose of this project is to give a detailed picture of the Gen Ed program to the dean's office of Dietrich College at Carnegie Mellon University. The analysis we have done are closely related to the performance of different students in the program, grading fairness, rubric design, and rating consistency across semesters. The main take away from this project can be summarized into four bullet points:

1. Many students in the program did poorly on Critique Design. The mean rating of this rubric is significantly lower than average, and the ratings are highly differentiated among students. Thus, the department should offer some students additional help in doing Critique Design.
2. The rating process is approximately fair, and the raters generally agree with each other. None of the three raters consistently disagree with the others. Instead, the raters' tendentiousness in grading is different for different artifacts. This demonstrates the objectiveness of raters when rating artifacts.
3. The distribution of ratings are different in the two semesters. The mean rating is significantly higher in fall semester than in spring semester, and the students' ratings are more similar in fall semester. It is understandable that there may be difference in the course contents across semesters, but actions should be taken to standardize the difficulty of this program.
4. There is no gender bias in this program. Male and female students did equally well in this program, which is good for a Gen Ed program.

Next, we will answer the four research questions we stated in Introduction, and if necessary, discuss some more findings in detail.

Question 1

Is the distribution of ratings for each rubrics pretty much indistinguishable from the other rubrics, or are there rubrics that tend to get especially high or low ratings? Is the distribution of ratings given by each rater pretty much indistinguishable from the other raters, or are there raters that tend to give especially high or low ratings?

Answer: The distribution of ratings are similar for all rubrics except Critique Design, which tend to get especially low ratings. The distribution of ratings given by each raters are indistinguishable. Ratings given by rater 1 is lowest in average, but the difference is not large.

Question 2

For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees with the others? Or do they all disagree?

Answer: Yes, raters generally agree on their scores, and none of the raters consistently disagree with the others.

Additional discussion: From our analysis, there are three main observations:

1. We cannot predict the rating on one artifact from the rating on another, which was graded by the same rater.
2. The ratings for one artifact are consistent for most rubrics, indicating that the raters generally agree with each other. The consistency is the lowest for Text Organization, which is understandable because raters can have different preference in writing. The rates are especially high for rating on Visual Organization, Critique Design, and Select Methods. These are objective parts of the works.
3. Rater 1 and rater 2 are highly disagree on Research Question. They should reach a consensus on what is a good research question.

Question 3

More generally, how are the various factors in this experiment (Rater, Semester, Sex, Repeated, Rubric) related to the ratings? Do the factors interact in any interesting ways?

Answer: Rater, rubric, and repeated are the factors that can help predict ratings. The relationship between ratings and these factors is different for each artifact. No interaction between factors can be observed from our analysis.

Additional discussion: To be honest, we cannot generate any significant fact that is true for all artifacts from the model. The effect of rubrics and raters on ratings are different for each artifact. The only fixed effect repeated is not significant, so no inference can be generated from its coefficient. This is actually a good result, because the ratings should only be linked to how well the students did in his/her artifact. None of the factors in our data is actually related to a student's well-being. Rater, semester, sex, rubric, and whether being seen by all the raters are all external factors. So, the ultimate inference from this model should be:

The assessment process of Gen Ed program is fair and unbiased.

Question 4

Is the distribution of ratings for different semesters and sex groups different from each other?

Answer: The distributions of ratings for different semesters are different. Compared to spring semester, the mean rating for fall semester is significantly higher, and the standard deviation is lower. The distributions of ratings are similar for different sex groups.

Weaknesses & Next step

1. The data set is small for analysis, especially when there are only 13 artifacts that have been graded by all raters. This is thus hard to draw a unbiased and convincing conclusion on the performance of Gen Ed program. In the future, we can include more data in our analysis for a better judgment.

2. This analysis can only focus on the fairness of the rating process and the design of the rubrics because no data was provided about the well-being of students. For next step, more information about the students' performance, such as study hours and GPA should be included to predict ratings.

3. We used only linear mixed model in this project. In the future, when it is actually needed to predict the ratings, we can try other models such as neural network, random forest, and generalized additive model.

Reference

1. Junker, B. W. (2021). Project 02 assignment sheet and data for 36-617: Applied Regression Analysis. Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh PA. Accessed Nov 29, 2021 from https://canvas.cmu.edu/courses/25337/files/folder/Project02

Technical Appendix

1.

Is the distribution of ratings for each rubrics pretty much indistinguishable from the other rubrics, or are there rubrics that tend to get especially high or low ratings? Is the distribution of ratings given by each rater pretty much indistinguishable from the other raters, or are there raters that tend to give especially high or low ratings?

Rubrics

Here is a summary table for different rubrics with the number of different grades and the grade means and standard deviation. (Table 14)

Table 14: Summary table: Rubics

	Count				Distribution	
Rubrics	Rating 1	Rating 2	Rating 3	Rating 4	Mean	SD
RsrchQ	6	65	45	1	2.35	0.59
CritDes	47	39	28	2	1.87	0.84
InitEDA	8	56	47	6	2.44	0.70
SelMeth	10	89	18	0	2.07	0.49
InterpRes	6	49	61	1	2.49	0.61
VisOrg	7	59	45	5	2.41	0.67
TxtOrg	8	37	66	6	2.60	0.70

We can see that the mean grade for Text Organization is the highest, and the one for Critique Design is the lowest. Critique Design also has the highest standard deviation and the highest number of grade 1.

Here are the bar charts for the seven rubrics. (Figure 4 on Page 15) The graphs are similar except the one for Critique Design. Only the distribution of Critique Design skews to the right, and all the other are highest with ratings 2 or 3.

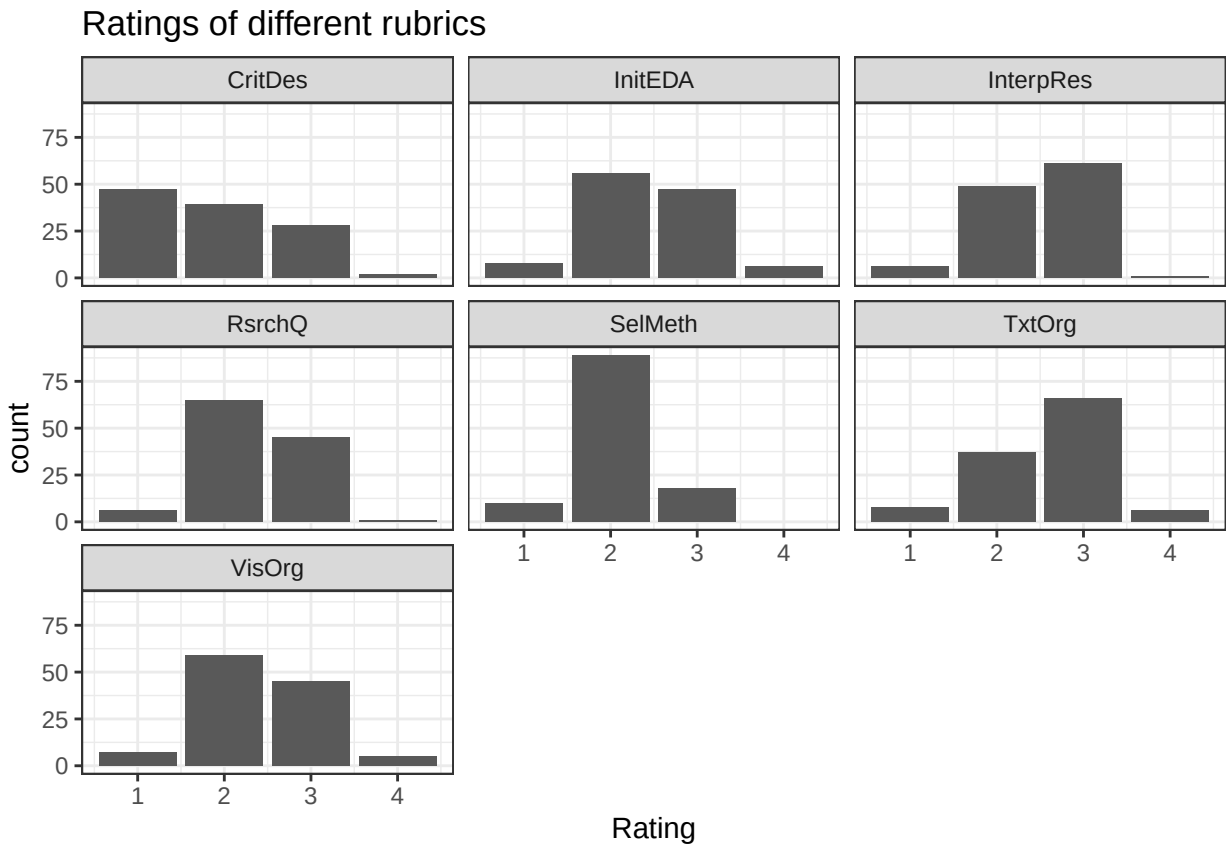

Figure 4: Ratings of different rubrics

Table 15: Summary table: Raters

	Count				Distribution	
Raters	Rating 1	Rating 2	Rating 3	Rating 4	Mean	SD
Rater.1	40	150	78	5	2.18	0.69
Rater.2	23	119	120	10	2.43	0.70
Rater.3	29	125	112	6	2.35	0.70

Raters

Here is a summary table for ratings produced by the three raters. (Table 15)

We can see that the ratings have similar distributions. Among the three raters, Rater 1 is the most likely to give low ratings among the three raters. The mean ratings made by Rater 1 is also lowest. Rater 2 is the most likely to give high ratings and the mean ratings made by Rater 2 is also highest. The standard deviation for the three raters are similar.

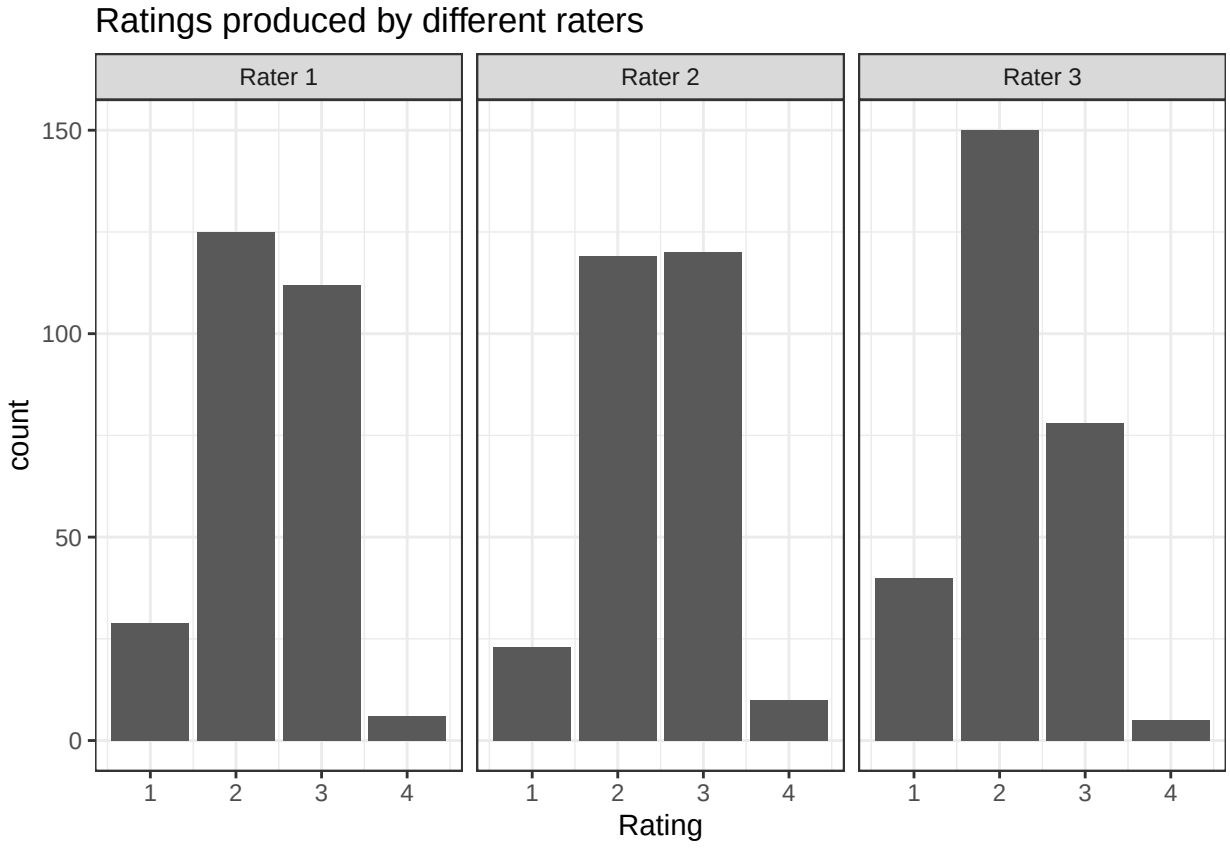

Figure 5: Ratings produced by different raters

The same pattern can be observed from the three bar charts above. (Figure 5)

Distribution of subset of artifacts seen by all three raters

In this part, we will do all the analysis appeared above to evaluate whether a subset of 13 artifacts seen by all three raters are representative of the whole set of 91 artifacts.

Table 16: Summary table: Rubics (subset of 13 artifacts)

	Count				Distribution	
Rubrics	Rating 1	Rating 2	Rating 3	Rating 4	Mean	SD
RsrchQ	2	24	13	0	2.28	0.56
CritDes	17	16	6	0	1.72	0.72
InitEDA	1	22	16	0	2.38	0.54
SelMeth	4	29	6	0	2.05	0.51
InterpRes	1	18	19	1	2.51	0.60
VisOrg	3	22	14	0	2.28	0.60
TxtOrg	2	10	26	1	2.67	0.62

Rubrics We can see from Table 16 in Page 18 that the mean grade for Text Organization is also the highest, and the one for Critique Design is the lowest in the selected data. Critique Design also has the highest standard deviation and the highest number of grade 1.

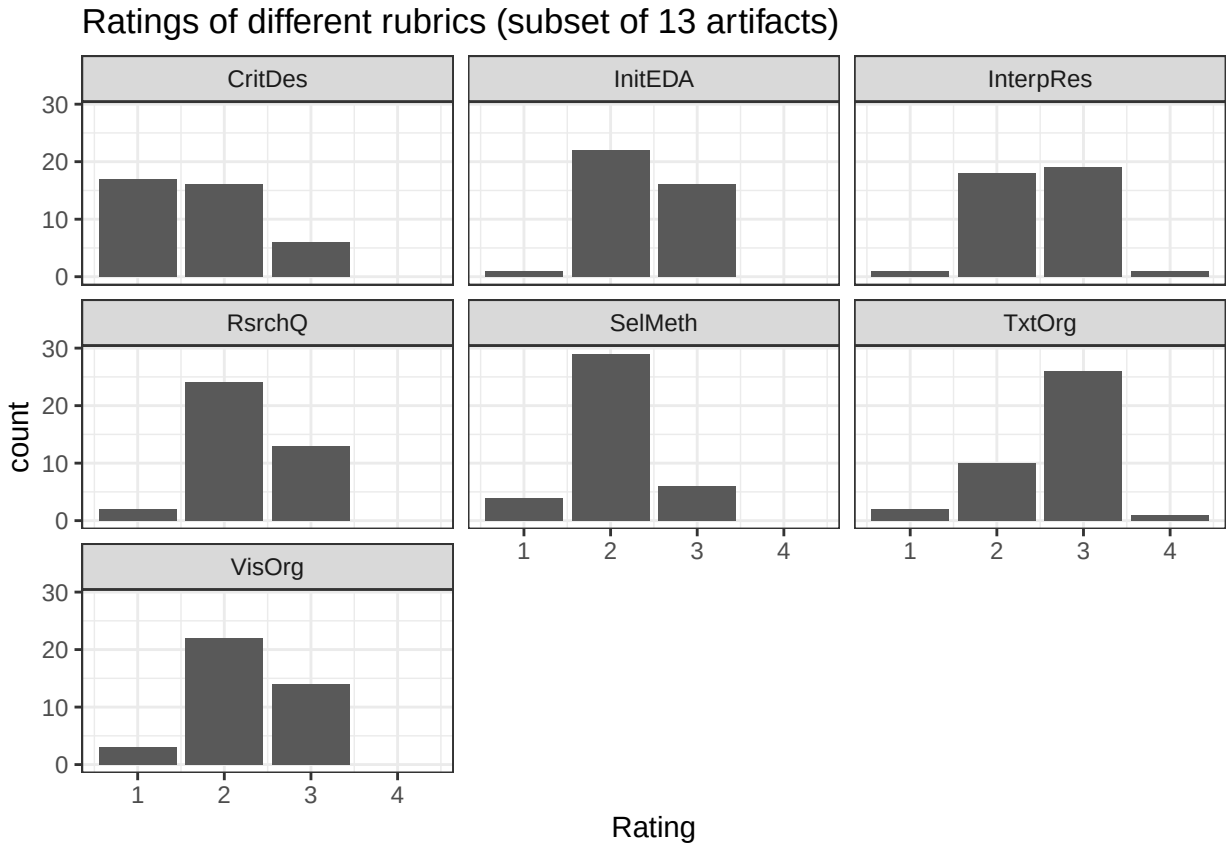

Figure 6: Ratings of different rubrics (subset of 13 artifacts)

We can see the similar patterns from these bar charts (Figure 6 on Page 17) and the bar charts from the full data.

Raters We can see from Table 17 that in the subset of full data, the mean ratings for rater 1 is also the lowest, and that for rater 2 is the highest. The standard deviation for the ratings for the three raters are also similar.

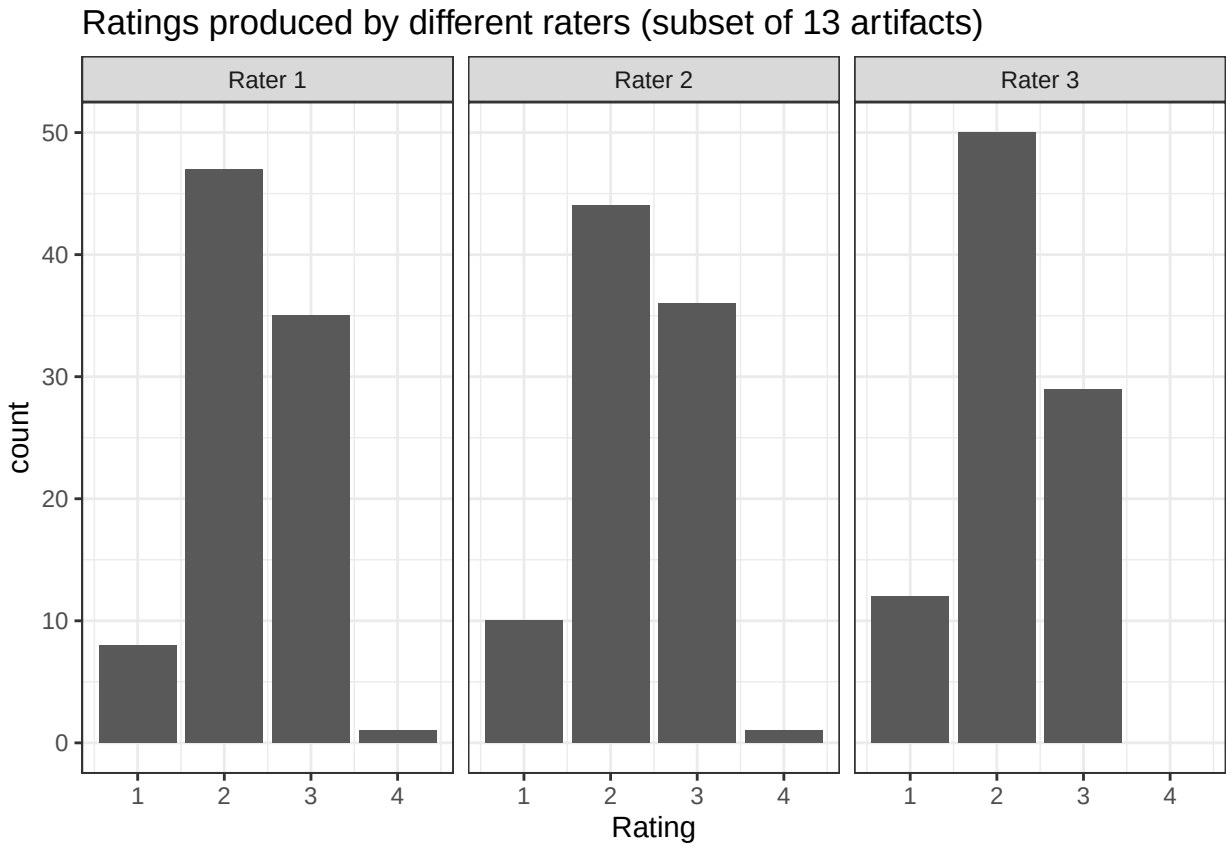

Figure 7: Ratings produced by different raters (subset of 13 artifacts)

Table 17: Summary table: Raters (subset of 13 artifacts)

	Count				Distribution	
Raters	Rating 1	Rating 2	Rating 3	Rating 4	Mean	SD
Rater 1	12	50	29	0	2.19	0.65
Rater 2	10	44	36	1	2.31	0.68
Rater 3	8	47	35	1	2.32	0.65

The patterns in these plots (Figure 7 on Page 19) are approximately the same with the plots produced by full data

Thus, we conclude that these thirteen artifacts are representative of the whole set of 91 artifacts.

2

For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees with the others? Or do they all disagree?

Overall agreement

Table 18: ICC analysis for different rubrics regarding raters and artifacts

Rubrics	ICC of Raters	ICC of Artifacts
RsrchQ	0.0000000	0.1891892
CritDes	0.0000000	0.5725594
InitEDA	0.0033333	0.4929577
SelMeth	0.0000000	0.5212766
InterpRes	0.0108695	0.2295720
VisOrg	0.0000000	0.5924529
TxtOrg	0.0000000	0.1428571

From Table 18, we can see that the ICC values for raters are all very small (< 0.05). The ICC value means correlation between ratings on any two different artifacts by the same rater. The low value indicates that there is not much correlation between artifacts, grouped by rater. Thus, one cannot predict the rating on one artifact from the rating on another. The ICC value is the highest for rating on Interpret Results, which is understandable because the raters can have their preference in case of interpretations.

The ICC values for artifacts are all high (> 0.1), which means that the link between the ratings from different raters for certain artifacts are high. The consistency of ratings can be high for different raters, and these raters generally agree on their scores. The consistency are especially high for rating on Visual Organization, Critique Design, and Select Methods, which are objective parts of the artifacts.

Agreement for different raters

Table 19: Pairwise agreement rate for different rubrics

	Agreement rate			
Rubrics	Rater 1 & 2	Rater 2 & 3	Rater 1 & 3	Disagree Rater
RsrchQ	0.38	0.54	0.77	Rater 2
CritDes	0.54	0.69	0.62	None
InitEDA	0.69	0.85	0.54	Rater 1
SelMeth	0.92	0.69	0.62	Rater 3
InterpRes	0.62	0.62	0.54	None
VisOrg	0.54	0.77	0.77	None
TxtOrg	0.69	0.54	0.62	None

The table above (Table 19) shows the percentage of ratings that each pair of raters agree with each other. The last column is the rater that tend to disagree with the others in each rubric. None in this column means that the rate of agreement for each pairs are similar.

Full data

Here is the ICC analysis done with the full data. (Table 20)

Table 20: ICC analysis for different rubrics regarding raters and artifacts - Full data

Rubrics	ICC of Raters	ICC of Artifacts
RsrchQ	0.0000000	0.2096214
CritDes	0.0780793	0.6730647
InitEDA	0.0026139	0.6867210
SelMeth	0.0199487	0.4719014
InterpRes	0.1988079	0.2200285
VisOrg	0.0792071	0.6607372
TxtOrg	0.0321074	0.1879927

We can see higher ICC values for raters, and similar ICC for artifacts.

We cannot redo the agreement rate part with the full data, because not all the artifacts were graded by all the three raters. If an artifact is not graded by one rater, we cannot evaluate whether the rater agree with the other raters in case of this artifact.

3

More generally, how are the various factors in this experiment (Rater, Semester, Sex, Repeated, Rubric) related to the ratings? Do the factors interact in any interesting ways?

We can fit a different model for each rubric, but it is hard to directly examine interactions with Rubric, since each model considers only one Rubric at a time. So we would like to put Rubric into the model and put random slope of Artifacts on it.

Model selection

To select an appropriate model, we started by forward-selecting the random effects and then backward-selecting fixed effects. The starting model for selection is:

Rating ~ Semester + Rater + Sex + Repeated + Rubric + (0 + Rubric | Artifact)

Random effects The random effects was selected automatically by `fitLMER.fnc` function. This function uses BIC and forward-fitting to select the random effects. The candidates of the random effects groups are Rater and Artifact. We think both of them may influence the intercept, and the slopes for all the factors except Repeated. After selection, the model is one below

Rating ~ Semester + Rater + Sex + Repeated + Rubric + (0 + Rubric | Artifact) + (0 + Rater | Artifact)

Fixed effects Fixed effects was selected by ANOVA table. The algorithm (pseudo-code) is below:

1. Do a while loop. The loop ends when all the fixed effect variables are deleted, or all the p-values for the ANOVA tables are below a threshold of 0.05.
2. For each loop, we fit several models. In each model, one of the fixed effect is deleted. An ANOVA table is created for each model, comparing it with the optimal model generated in the last loop.
3. If all the p-values in the ANOVA tables are below the threshold of 0.05, or the BIC value for the simpler model (the model without variable which has the largest p-value) is larger than the complex model, we end the loop and claim that we do not need to delete variables anymore. The optimal model from the last loop is the final model.
4. If some p-values are greater than 0.05, then we delete the variable with the largest p-value and continue to run the loop.

Rating ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact) + Rubric + Rater + Repeated

The model with the selected fixed effects is printed above. We can see that fixed effects Rubric, Rater, and Repeated have significant effects in predicting ratings. Random effect term `(0 + Rater | Artifact)`, which is a random slope of raters corresponded to Artifact should also be added in the model.

Interaction To explore how interaction terms can improve the model, we started from models with one interaction term. More interactions will be added and tested if any interaction term is better than the original one. We tried nine possible models with one interaction term, which can be categorized into three groups:

1. Model 1 - 3: Pairwise interactions of the three fixed effects
2. Model 4 - 6: Interaction of Sex with each of the three fixed effects
3. Model 7 - 9: Interaction of Semester with each of the three fixed effects

Bayesian Information Criterion (BIC) was used to compare these models.

Table 21: BIC values for all the possible models

Model	BIC
No Interaction	1645.056

Model	BIC
Model 1	1755.440
Model 2	1697.361
Model 3	1729.399
Model 4	1738.357
Model 5	1707.870
Model 6	1699.864
Model 7	1729.329
Model 8	1704.357
Model 9	1696.340

From Table 21, we can see that the model with no interaction performs the best (lowest in BIC). So we choose not to add any interaction term into the model.

Final model

This table below (Table 22) shows the fixed effects in the model.

Table 22: Fixed effects in the final model

Terms	estimate	std error	t-statistic	p-value
(Intercept)	1.9740090	0.0961070	20.5396981	1.0000000
RubricInitEDA	0.5444742	0.0949677	5.7332547	1.0000000
RubricInterpRes	0.5852104	0.1002030	5.8402475	1.0000000
RubricRsrchQ	0.4596902	0.0864125	5.3197175	0.9999999
RubricSelMeth	0.1624726	0.0932457	1.7424132	0.9592819
RubricTxtOrg	0.6855404	0.0996731	6.8778846	1.0000000
RubricVisOrg	0.5342426	0.0986747	5.4141831	1.0000000
Rater2	-0.0006143	0.0773912	-0.0079375	0.4968334
Rater3	-0.1922401	0.0685028	-2.8063104	0.0025056
Repeated	-0.0828972	0.0824826	-1.0050256	0.1574423

From the summary, we can see that almost all the fixed effects are insignificant. We include them just because they work well in improving the model (lower BIC values). Despite that, there are still some useful information in this table.

1. All the rubrics will result in a higher rating except Critique Design, which means that this is the part that have lower ratings. This effect is not significant, so the coefficients may be a result of a coincidence.
2. Rater 3 is more likely to give lower scores comparing with other raters. This is the only significant fixed effect in the model.
3. Artifacts seen by all three raters can have a lower score. This effect is also not significant. We cannot convince this discovery from the model.

Table 23 shows the random effects in the final model

Table 23: Random effects of Artifact in final model

	Random slopes on Rubrics							Random slopes on Raters		
	CritDes	InitEDA	InterpRes	RsrchQ	SelMeth	TxtOrg	VisOrg	Rater1	Rater2	Rater3
100	0.63	-0.23	-0.07	-0.12	0.24	-0.41	-0.32	0.04	-0.05	0.05
101	-0.76	0.31	-0.17	-0.74	0.04	0.21	0.33	-0.06	0.07	-0.06
102	-0.93	-0.32	-0.13	-0.65	0.06	-1.01	-0.43	-0.08	0.09	-0.09
103	-0.04	0.30	0.16	-0.23	0.29	0.17	-0.25	0.05	-0.05	0.05
104	-0.77	0.27	0.19	-0.21	0.09	0.16	-0.15	-0.02	0.02	-0.02
105	-0.74	-0.47	-0.26	-0.35	-0.08	-0.46	-0.39	-0.06	0.07	-0.07
106	0.02	-1.00	-0.31	0.25	-0.11	0.07	-0.39	-0.02	0.02	-0.02
107	-0.75	-0.43	0.15	0.18	-0.10	-0.42	-0.36	0.00	0.00	0.00
111	-0.74	-0.47	-0.26	-0.35	-0.08	-0.46	-0.39	-0.06	0.07	-0.07
112	-0.68	0.28	0.37	0.18	-0.09	0.27	0.34	0.03	-0.04	0.03
113	-0.91	-0.93	-0.11	-0.16	-0.01	-0.49	-0.44	-0.04	0.05	-0.05
114	-0.77	0.27	0.19	-0.21	0.09	0.16	-0.15	-0.02	0.02	-0.02
115	-0.68	0.28	0.37	0.18	-0.09	0.27	0.34	0.03	-0.04	0.03
116	-0.74	-0.34	-0.18	-0.30	-0.09	0.18	0.22	-0.03	0.04	-0.03
117	-0.85	-0.14	0.15	-0.16	-0.07	0.29	0.76	0.01	-0.02	0.01
118	-0.84	-0.22	0.10	-0.16	0.00	0.20	0.26	-0.01	0.01	-0.01
13	0.57	-0.67	-0.49	0.10	-0.18	-0.68	-0.59	-0.08	-0.37	-0.54
15	0.87	0.53	-0.11	-0.08	-0.04	0.39	0.89	0.02	0.09	0.13
16	0.81	1.13	0.26	0.03	0.18	0.87	0.59	0.03	0.14	0.20
17	0.40	-0.15	-0.04	0.49	-0.24	-0.17	-0.04	-0.03	-0.14	-0.20
21	0.81	0.98	0.32	0.37	0.05	0.82	0.51	0.05	0.22	0.32
22	0.83	0.36	0.18	0.43	0.07	0.21	-0.12	0.03	0.13	0.19
23	-0.16	-0.73	-0.40	0.00	-0.21	-0.67	-0.59	-0.08	-0.39	-0.56
24	0.19	-0.16	-0.16	-0.07	0.07	0.27	-0.12	-0.01	-0.04	-0.05
25	1.28	0.01	-0.44	0.15	-0.15	-0.08	0.04	-0.06	-0.27	-0.39
26	-0.69	-0.44	-0.15	-0.44	0.10	-0.39	-0.45	0.02	0.08	0.12
27	0.25	0.40	-0.04	-0.12	0.08	0.31	-0.04	-0.01	-0.02	-0.03
28	-0.17	-0.57	-0.47	-0.34	-0.08	-0.62	-0.52	-0.10	-0.47	-0.68
32	0.78	0.39	0.18	0.40	-0.01	0.25	0.35	0.04	0.18	0.27
33	0.10	0.18	-0.12	-0.39	0.16	-0.36	-0.36	0.02	0.11	0.16
34	0.61	-0.26	-0.09	-0.13	0.25	-0.45	-0.45	0.04	0.17	0.25
35	-0.57	-0.21	-0.08	-0.23	-0.10	-0.21	0.26	-0.01	-0.02	-0.04
36	0.19	-0.16	-0.16	-0.07	0.07	0.27	-0.12	-0.01	-0.04	-0.05
37	0.92	-0.10	-0.25	0.04	0.10	0.27	-0.12	-0.01	-0.02	-0.04
38	0.16	-0.15	-0.17	-0.13	-0.07	-0.21	0.26	0.00	-0.01	-0.02
39	-0.45	0.16	0.10	0.09	-0.13	-0.26	-0.20	0.00	0.02	0.03
40	0.27	0.38	-0.05	-0.15	0.03	-0.22	-0.12	-0.01	-0.05	-0.07
45	-0.11	-0.25	-0.19	-0.14	0.04	0.21	-0.18	0.03	-0.18	-0.12
46	0.20	0.31	-0.11	-0.20	0.01	-0.28	-0.17	0.03	-0.19	-0.12
47	1.14	-0.21	-0.20	0.55	-0.19	-0.81	-0.64	0.06	-0.41	-0.28
48	0.61	0.67	0.19	0.11	0.24	0.56	-0.25	-0.06	0.37	0.25
49	-0.82	0.23	0.31	0.10	-0.19	0.24	0.74	-0.03	0.19	0.13
53	1.25	0.18	-0.03	0.21	0.13	0.03	0.04	-0.06	0.41	0.27
54	-0.56	0.42	-0.07	-0.61	0.15	0.31	-0.02	0.04	-0.29	-0.19
55	-0.07	-0.33	0.08	0.27	-0.14	-0.38	0.11	-0.01	0.08	0.05

Table 23: Random effects of Artifact in final model *(continued)*

	Random slopes on Rubrics							Random slopes on Raters		
	CritDes	InitEDA	InterpRes	RsrchQ	SelMeth	TxtOrg	VisOrg	Rater1	Rater2	Rater3
56	0.82	-0.17	-0.31	-0.07	-0.05	-0.29	0.20	0.02	-0.11	-0.07
57	-0.67	-0.26	-0.12	-0.25	-0.07	0.25	0.28	0.02	-0.11	-0.08
6	-0.52	-0.25	-0.09	-0.20	-0.02	-0.25	-0.21	-0.02	-0.08	-0.11
61	0.10	0.38	-0.07	-0.17	0.03	0.81	0.45	0.01	-0.04	-0.03
62	1.44	0.97	0.17	0.36	-0.06	0.25	0.83	-0.04	0.27	0.18
63	0.70	0.32	0.13	0.35	-0.03	0.18	0.28	-0.02	0.16	0.10
64	0.64	-0.09	-0.06	0.06	0.08	0.20	0.28	0.00	0.01	0.01
65	0.96	-0.36	-0.47	0.21	-0.20	0.22	0.21	0.01	-0.04	-0.03
66	0.78	0.86	0.24	0.26	-0.07	-0.31	0.29	-0.02	0.14	0.09
67	-0.71	0.16	0.28	0.07	-0.21	-0.86	0.12	-0.01	0.04	0.03
68	0.64	-0.24	0.00	0.40	-0.05	0.14	0.20	-0.02	0.14	0.09
7	-0.46	0.32	0.04	-0.26	-0.01	-0.21	-0.12	-0.01	-0.06	-0.09
72	-0.01	0.39	0.14	-0.13	0.01	-0.28	0.28	0.00	-0.02	-0.01
73	-0.67	-0.86	-0.25	-0.16	-0.01	0.18	-0.27	0.03	-0.21	-0.14
74	-0.92	0.14	0.16	-0.48	0.09	-0.42	0.04	-0.03	0.19	0.13
75	0.04	0.36	0.13	-0.10	0.09	-0.32	-0.18	0.01	-0.10	-0.06
76	0.14	-0.26	-0.23	-0.14	0.00	-0.32	-0.26	0.03	-0.21	-0.14
77	-0.07	-0.17	0.01	-0.08	-0.01	-0.32	0.19	0.01	-0.04	-0.03
78	0.52	0.12	0.06	0.10	0.09	0.03	0.04	-0.06	0.39	0.26
79	-0.21	0.06	0.15	0.00	0.06	0.04	0.04	-0.06	0.37	0.25
8	-0.52	-0.25	-0.09	-0.20	-0.02	-0.25	-0.21	-0.02	-0.08	-0.11
84	-0.01	-0.23	0.15	0.34	-0.07	0.22	0.26	0.03	-0.04	0.03
85	0.89	0.35	0.33	0.81	-0.11	0.30	0.36	0.08	-0.09	0.08
86	-0.01	-0.31	0.11	0.34	0.00	0.14	-0.25	0.01	-0.01	0.01
87	0.02	-1.00	-0.31	0.25	-0.11	0.07	-0.39	-0.02	0.02	-0.02
88	0.80	0.51	0.25	0.41	-0.07	0.37	0.87	0.06	-0.07	0.07
9	-0.51	-0.40	-0.02	0.14	-0.15	-0.30	-0.28	0.00	0.01	0.01
92	-0.75	-0.38	0.18	0.23	-0.04	0.13	-0.26	0.01	-0.01	0.01
93	-0.75	-0.33	0.22	0.27	0.02	0.68	-0.16	0.02	-0.02	0.02
94	0.88	1.00	0.34	0.37	0.01	0.33	0.47	0.05	-0.05	0.05
95	-0.04	0.30	0.16	-0.23	0.29	0.17	-0.25	0.05	-0.05	0.05
96	0.06	0.35	0.29	0.30	-0.05	0.27	0.36	0.04	-0.04	0.04
O1	-0.41	0.49	0.21	0.06	-0.07	-0.14	-0.18	-0.22	0.30	-0.17
O10	-0.38	0.05	0.25	0.09	0.05	0.20	0.06	0.12	-0.31	-0.04
O11	-0.85	-0.42	0.33	0.44	-0.03	0.27	-0.33	0.06	0.07	0.19
O12	-0.28	-0.40	-0.24	-0.25	0.01	-0.28	-0.38	0.07	0.03	0.17
O13	0.12	0.18	-0.26	-0.29	-0.03	0.34	0.21	0.00	0.09	0.08
O2	-0.05	0.39	0.25	0.18	-0.05	0.24	0.39	0.18	-0.96	-0.55
O3	0.36	-0.06	0.07	0.29	-0.01	0.32	0.19	-0.03	0.21	0.13
O4	-0.03	0.11	0.29	-0.15	0.40	0.25	-0.20	-0.08	0.45	0.27
O5	0.95	-0.35	0.03	0.25	0.29	0.03	-0.41	0.04	0.11	0.20
O6	-0.81	-0.68	-0.31	-0.33	-0.08	0.03	-0.13	-0.05	0.15	0.05
O7	0.21	0.27	0.04	0.03	-0.02	0.18	0.27	0.17	0.20	0.56
O8	0.32	-0.23	-0.48	-0.30	0.02	-0.61	-0.92	-0.05	-0.25	-0.36
O9	-0.85	0.51	0.34	-0.23	0.05	0.36	0.51	0.05	0.05	0.15

ICC

The ICC values for the full model is in the table below (Table 24):

Table 24: ICC analysis for Artifacts under full model

Rubrics	ICC.Artifact
CritDes	0.8320207
InitEDA	0.7337102
InterpRes	0.4935779
RsrchQ	0.5768422
SelMeth	0.2736317
TxtOrg	0.6864548
VisOrg	0.6690094

This time, the values of the ICC values are higher for most of the rubrics.

4

In this part, we will investigate whether the rating is different for different semester and sex. Data with missing values are deleted. We will use plots, summary table and one-way ANOVA tests for our analysis.

Plots

From the bar charts (Figure 8 on Page 26), we can see that:

1. For both semesters, 2 are the most common rating, followed by 3, 1, and 4. However, the percentage of artifacts (and rubrics) with ratings 2 and 3 is lower in S19 than F19. The situation is opposite for ratings 1 and 4.
2. The distributions of ratings are similar for different sex of authors.

Summary tables

Table 25: Summary tables for semester and sex

Semester	count	mean	sd	Sex	count	mean	sd
F19	573	2.347295	0.6662287	F	446	2.313901	0.7000061
S19	237	2.227848	0.7803091	M	364	2.310440	0.7079251

From the summary tables above (Table 25), we can see that:

1. There are more samples for semester F19 than S19. The mean rating for F19 is higher than S19, and the standard deviation for F19 is lower than S19.
2. There are more samples for female authors than male authors. The mean and standard deviation of ratings are similar across sex groups.

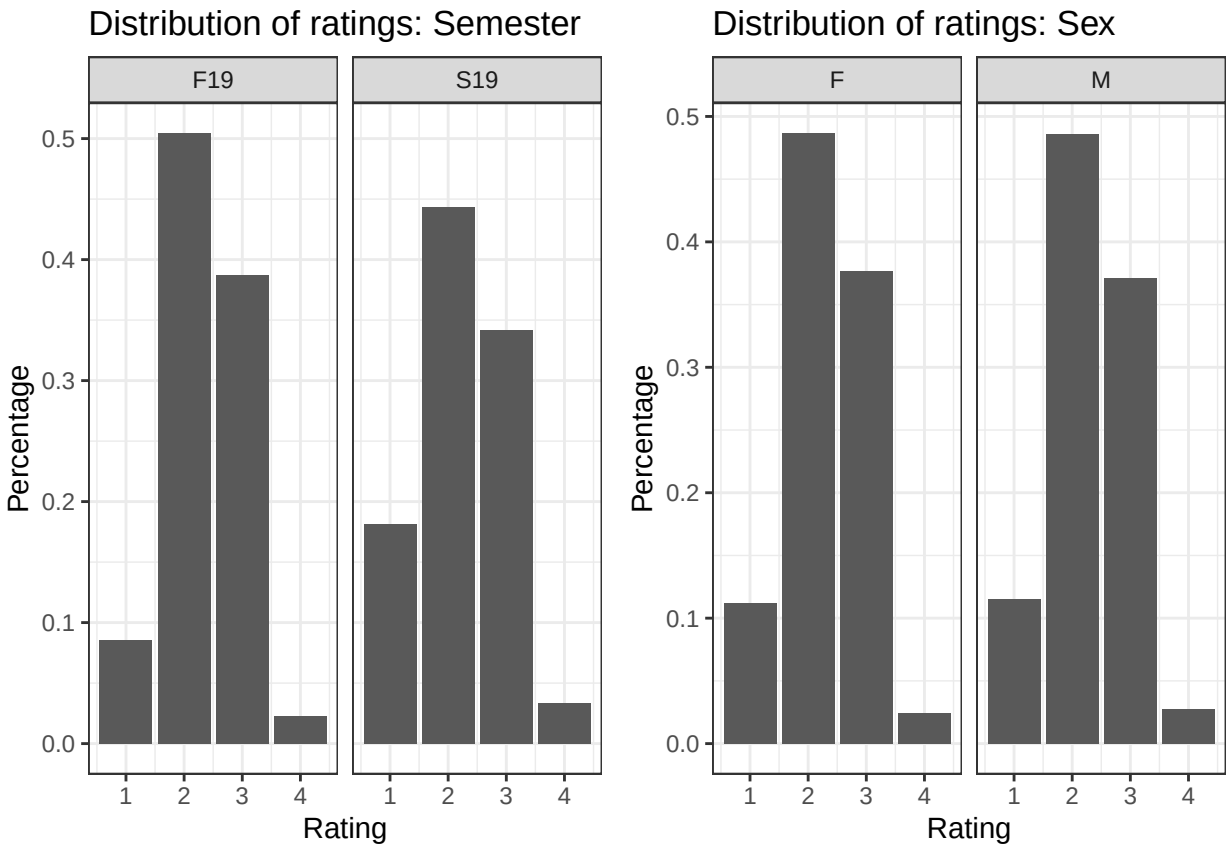

Figure 8: Distribution of ratings: Semester and Sex

Table 26: ANOVA tables for semester and sex

term	df	sumsq	meansq	statistic	p.value
Semester	1	2.39	2.39	4.86	0.03
Residuals	808	397.58	0.49	NA	NA
term	df	sumsq	meansq	statistic	p.value
Sex	1	0.00	0.0	0	0.94
Residuals	808	399.97	0.5	NA	NA

One-way ANOVA test for means

In this part, we hypothesize that the mean ratings are the same across different semesters and sex groups.

From the outputs above (Table 26), we can see that:

1. When we hypothesize that the mean ratings are the same for the two semesters, the p-value we got from one-way ANOVA test is $0.03 < 0.05$. This means we should reject the null hypothesis, and conclude that the mean ratings are different across semesters
2. When we hypothesize that the mean ratings are the same for different sex groups, the p-value we got from one-way ANOVA test is $0.94 > 0.05$. This means we should accept the null hypothesis, and conclude that the mean ratings are the same across sex groups.

Codes

```
# 1

## Rubrics

smry_rubric = tibble(
  Variable = names(ratings)[7:13],
  Count.1 = rep(NA, 7),
  Count.2 = rep(NA, 7),
  Count.3 = rep(NA, 7),
  Count.4 = rep(NA, 7),
  Mean = rep(NA, 7),
  SD = rep(NA, 7)
)

for (i in 7:13){
  smry_rubric = mutate(smry_rubric,
    Count.1 = replace(Count.1, i-6, sum(na.omit(ratings[[i]] == 1))),
    Count.2 = replace(Count.2, i-6, sum(na.omit(ratings[[i]] == 2))),
    Count.3 = replace(Count.3, i-6, sum(na.omit(ratings[[i]] == 3))),
    Count.4 = replace(Count.4, i-6, sum(na.omit(ratings[[i]] == 4))),
    Mean = replace(Mean, i-6, round(mean(na.omit(ratings[[i]])),2)),
    SD = replace(SD, i-6, round(sd(na.omit(ratings[[i]])),2))
  )
}

kable(smry_rubric, col.names = c("Rubrics", "Rating 1", "Rating 2", "Rating 3",
                                 "Rating 4", "Mean", "SD"),
      caption = "Summary table: Rubics") %>%
  add_header_above(c(" " = 1, "Count" = 4, "Distribution" = 2)) %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))

tall %>% na.omit() %>% ggplot(aes(x = Rating)) +
  facet_wrap( ~ Rubric) +
  geom_bar() +
  theme_bw() +
  labs(title = "Ratings of different rubrics", "pipe")

### Raters


raters = tall %>%
  pivot_wider(names_sep = ".",
              names_from = Rater,
              values_from = Rating)

smry_raters = tibble(
  Variable = c("Rater.1", "Rater.2", "Rater.3"),
  Count.1 = rep(NA, 3),
  Count.2 = rep(NA, 3),
  Count.3 = rep(NA, 3),
  Count.4 = rep(NA, 3),
```

```
  Mean = rep(NA, 3),
  SD = rep(NA, 3)
)

for (i in 7:9){
  smry_raters = mutate(smry_raters,
    Count.1 = replace(Count.1, i-6, sum(na.omit(raters[[i]] == 1))),
    Count.2 = replace(Count.2, i-6, sum(na.omit(raters[[i]] == 2))),
    Count.3 = replace(Count.3, i-6, sum(na.omit(raters[[i]] == 3))),
    Count.4 = replace(Count.4, i-6, sum(na.omit(raters[[i]] == 4))),
    Mean = replace(Mean, i-6, round(mean(na.omit(raters[[i]])),2)),
    SD = replace(SD, i-6, round(sd(na.omit(raters[[i]])),2))
  )
}

kbl(smry_raters, col.names = c("Raters", "Rating 1", "Rating 2", "Rating 3",
                               "Rating 4", "Mean", "SD"),
    caption = "Summary table: Raters") %>%
  add_header_above(c(" " = 1, "Count" = 4, "Distribution" = 2)) %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))

rater.name <- function(x) { paste("Rater",x) }
tall %>% na.omit() %>% ggplot(aes(x = Rating)) +
  facet_wrap( ~ Rater, labeller=labeller(Rater=rater.name)) +
  geom_bar() +
  theme_bw() +
  labs(title = "Ratings produced by different raters")

### Distribution of subset of artifacts seen by all three raters

ratings.13 = ratings %>% filter(Repeated == 1)
raters.13 = raters %>% filter(Repeated == 1)


#### Rubrics


smry_rubric_13 = tibble(
  Variable = names(ratings.13)[7:13],
  Count.1 = rep(NA, 7),
  Count.2 = rep(NA, 7),
  Count.3 = rep(NA, 7),
  Count.4 = rep(NA, 7),
  Mean = rep(NA, 7),
  SD = rep(NA, 7)
)

for (i in 7:13){
  smry_rubric_13 = mutate(smry_rubric_13,
    Count.1 = replace(Count.1, i-6, sum(na.omit(ratings.13[[i]] == 1))),
    Count.2 = replace(Count.2, i-6, sum(na.omit(ratings.13[[i]] == 2))),
    Count.3 = replace(Count.3, i-6, sum(na.omit(ratings.13[[i]] == 3))),
    Count.4 = replace(Count.4, i-6, sum(na.omit(ratings.13[[i]] == 4))),
```

```
    Mean = replace(Mean, i-6, round(mean(na.omit(ratings.13[[i]])),2)),
    SD = replace(SD, i-6, round(sd(na.omit(ratings.13[[i]])),2))
  )
}

kbl(smry_rubric_13, col.names = c("Rubrics", "Rating 1", "Rating 2", "Rating 3",
                                  "Rating 4", "Mean", "SD"),
    caption = "Summary table: Rubics (subset of 13 artifacts)") %>%
  add_header_above(c(" " = 1, "Count" = 4, "Distribution" = 2)) %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))

tall.13 %>% na.omit() %>% ggplot(aes(x = Rating)) +
  facet_wrap( ~ Rubric) +
  geom_bar() +
  theme_bw() +
  labs(title = "Ratings of different rubrics (subset of 13 artifacts)")

#### Raters


smry_raters_13 = tibble(
  Variable = c("Rater 1", "Rater 2", "Rater 3"),
  Count.1 = rep(NA, 3),
  Count.2 = rep(NA, 3),
  Count.3 = rep(NA, 3),
  Count.4 = rep(NA, 3),
  Mean = rep(NA, 3),
  SD = rep(NA, 3)
)

for (i in 7:9){
  smry_raters_13 = mutate(smry_raters_13,
    Count.1 = replace(Count.1, i-6, sum(na.omit(raters.13[[i]] == 1))),
    Count.2 = replace(Count.2, i-6, sum(na.omit(raters.13[[i]] == 2))),
    Count.3 = replace(Count.3, i-6, sum(na.omit(raters.13[[i]] == 3))),
    Count.4 = replace(Count.4, i-6, sum(na.omit(raters.13[[i]] == 4))),
    Mean = replace(Mean, i-6, round(mean(na.omit(raters.13[[i]])),2)),
    SD = replace(SD, i-6, round(sd(na.omit(raters.13[[i]])),2))
  )
}

kbl(smry_raters_13, col.names = c("Raters", "Rating 1", "Rating 2", "Rating 3",
                                  "Rating 4", "Mean", "SD"),
    caption = "Summary table: Raters (subset of 13 artifacts)") %>%
  add_header_above(c(" " = 1, "Count" = 4, "Distribution" = 2)) %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))

tall.13 %>% na.omit() %>% ggplot(aes(x = Rating)) +
  facet_wrap( ~ Rater, labeller=labeller(Rater=rater.name)) +
  geom_bar() +
  theme_bw() +
  labs(title = "Ratings produced by different raters (subset of 13 artifacts)")
```

```
## 2



### Overall agreement


icc.raters = c()
icc.art = c()

for (i in 1:7){
  subset.rubric = common[common$Rubric==rubrics[i],]
  rater.lmer = lmer(Rating ~ 1 + (1|Rater), data=subset.rubric)
  sds = as.data.frame(summary(rater.lmer)$varcor)$vcov
  icc = sds[1]/(sds[1]+sds[2])
  icc.raters = c(icc.raters, icc)
}

for (i in 1:7){
  subset.rubric <- common[common$Rubric==rubrics[i],]
  art.lmer = lmer(Rating ~ 1 + (1|Artifact), data=subset.rubric)
  sds = as.data.frame(summary(art.lmer)$varcor)$vcov
  icc = sds[1]/(sds[1]+sds[2])
  icc.art = c(icc.art, icc)
}

tibble(
  Rubrics = rubrics,
  ICC.Raters = icc.raters,
  ICC.Artifacts = icc.art
) %>%
  kable(col.names = c("Rubrics", "ICC of Raters", "ICC of Artifacts"),
        caption = "ICC analysis for different rubrics regarding raters
        and artifacts", "pipe") %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))


### Agreement for different raters


agree.12 = c()
agree.23 = c()
agree.13 = c()
for (i in 7:13){
  dat_12 = data.frame(r1=repeated[repeated$Rater==1,i],
                      r2=repeated[repeated$Rater==2,i])
  r1 = factor(dat_12$r1,levels=1:4)
  r2 = factor(dat_12$r2,levels=1:4)
  t12 = table(r1,r2)
  prop12 = sum(diag(t12))/sum(t12)
  agree.12 = c(agree.12, prop12)

  dat_23 = data.frame(r2=repeated[repeated$Rater==2,i],
```

```
                      r3=repeated[repeated$Rater==3,i])
  r2 = factor(dat_23$r2,levels=1:4)
  r3 = factor(dat_23$r3,levels=1:4)
  t23 = table(r2,r3)
  prop23 = sum(diag(t23))/sum(t23)
  agree.23 = c(agree.23, prop23)

  dat_13 = data.frame(r1=repeated[repeated$Rater==1,i],
                      r3=repeated[repeated$Rater==3,i])
  r1 = factor(dat_13$r1,levels=1:4)
  r3 = factor(dat_13$r3,levels=1:4)
  t13 = table(r1,r3)
  prop13 = sum(diag(t13))/sum(t13)
  agree.13 = c(agree.13, prop13)
}

tibble(
  Rubrics = rubrics,
  Agreement.12 = round(agree.12,2),
  Agreement.23 = round(agree.23, 2),
  Agreement.13 = round(agree.13, 2),
  Disagree.rater = c("Rater 2","None", "Rater 1", "Rater 3", "None", "None", "None")
) %>%
  kbl(col.names = c("Rubrics", "Rater 1 & 2", "Rater 2 & 3", "Rater 1 & 3",
                    "Disagree Rater"),
      caption = "Pairwise agreement rate for different rubrics") %>%
  add_header_above(c(" " = 1, "Agreement rate" = 3, " " = 1)) %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))

### Full data

icc.raters.full = c()
icc.art.full = c()

for (i in 1:7){
  subset.rubric <- tall[tall$Rubric==rubrics[i],]
  rater.lmer = lmer(Rating ~ 1 + (1|Rater), data=subset.rubric)
  sds = as.data.frame(summary(rater.lmer)$varcor)$vcov
  icc = sds[1]/(sds[1]+sds[2])
  icc.raters.full = c(icc.raters.full, icc)
}

for (i in 1:7){
  subset.rubric <- tall[tall$Rubric==rubrics[i],]
  art.lmer = lmer(Rating ~ 1 + (1|Artifact), data=subset.rubric)
  sds = as.data.frame(summary(art.lmer)$varcor)$vcov
  icc = sds[1]/(sds[1]+sds[2])
  icc.art.full = c(icc.art.full, icc)
}

tibble(
  Rubrics = rubrics,
  ICC.Raters = icc.raters.full,
```

```
  ICC.Artifacts = icc.art.full
) %>%
  kable(col.names = c("Rubrics", "ICC of Raters", "ICC of Artifacts"),
        caption = "ICC analysis for different rubrics regarding
        raters and artifacts - Full data", "pipe") %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))

## 3


### Model selection


lmer.radinit = lmer(Rating ~ Semester + Rater + Sex + Repeated + Rubric
                    + (0 + Rubric | Artifact),
                    data=dat, REML=F)


#### Random effects

lmer.rad = fitLMER.fnc(lmer.radinit,
                       ran.effects = c("(1 | Artifact)", "(1|Rater)",
                                       "(0 + Rubric|Rater)", "(0 + Sex|Rater)",
                                       "(0 + Semester|Rater)",
                                       "(0 + Semester|Artifact)",
                                       "(0 + Rater|Artifact)",
                                       "(0 + Sex|Artifact)"),
                       method = "BIC",
                       set.REML.FALSE = TRUE,
                       log.file.name = FALSE)

#### Fixed effects

variables = c("Semester", "Sex", "Rubric", "Rater", "Repeated")
lmer.last = lmer.rad
bic.last = BIC(lmer.rad)
while (length(variables) > 0){
  p.step = c()
  for (j in 1:length(variables)){
    p.step = c(p.step,
               anova(lmer.last,lmer(as.formula(paste(c("Rating ~
               (0 + Rubric | Artifact) + (0 + Rater | Artifact)",
                                                 variables[-j]), collapse=" + ")),
                           data=dat, REML=F))$Pr[2])
  }
  md  = lmer(as.formula(paste(c("Rating ~ (0 + Rubric | Artifact) +
                                        (0 + Rater | Artifact)",
                                variables[-(p.step==max(p.step))]),
                              collapse="+")),
                           data=dat, REML=F)
  bic = BIC(md)
  if (max(p.step) < 0.05 | bic > bic.last){
    break
```

```
  }
  variables = variables[-(p.step==max(p.step))]
  lmer.last = md
  bic.last = bic
}
lmer.3 = lmer.last


#### Interaction


m1 = lmer(Rating ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact) +
            Rubric * Rater + Repeated, data = dat)
m2 = lmer(Rating ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact) +
            Rubric + Rater * Repeated, data = dat)
m3 = lmer(Rating ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact) +
            Rubric * Repeated + Rater, data = dat)
m4 = lmer(Rating ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact) +
            Rubric * Sex + Rater + Repeated, data = dat)
m5 = lmer(Rating ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact) +
            Rubric + Rater * Sex + Repeated, data = dat)
m6 = lmer(Rating ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact) +
            Rubric + Rater + Repeated * Sex, data = dat)
m7 = lmer(Rating ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact) +
            Rubric * Semester + Rater + Repeated, data = dat)
m8 = lmer(Rating ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact) +
            Rubric + Rater * Semester + Repeated, data = dat)
m9 = lmer(Rating ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact) +
            Rubric + Rater + Repeated * Semester, data = dat)

tibble(
  Model = c("No Interaction", "Model 1", "Model 2", "Model 3",
            "Model 4", "Model 5", "Model 6", "Model 7", "Model 8", "Model 9"),
  BIC = c(BIC(lmer.3), BIC(m1), BIC(m2), BIC(m3), BIC(m4), BIC(m5),
          BIC(m6), BIC(m7), BIC(m8), BIC(m9))
) %>%
  kable(caption = "BIC values for all the possible models", "pipe") %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))

### Final model

tidy(lmer.3) %>%
  filter(effect == "fixed") %>%
  select(term, estimate, std.error, statistic) %>%
  mutate(p.value = pnorm(statistic)) %>%
  kable(caption = "Fixed effects in the final model","pipe",
        col.names = c("Terms", "estimate", "std error", "t-statistic", "p-value")) %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))

ranef(lmer.3)$Artifact %>%
  kbl(caption = "Random effects of Artifact in final model", digits = 2,
      col.names = c("CritDes",  "InitEDA", "InterpRes", "RsrchQ",
                    "SelMeth", "TxtOrg", "VisOrg", "Rater1", "Rater2",'Rater3'),
```

```
      booktabs = TRUE, longtable = TRUE) %>%
  add_header_above(c(" " = 1, "Random slopes on Rubrics" = 7,
                     "Random slopes on Raters" = 3)) %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))

### ICC

rubrics.3 = c("CritDes", "InitEDA", "InterpRes", "RsrchQ", "SelMeth", "TxtOrg", "VisOrg")
sds = as.data.frame(summary(lmer.3)$varcor)$vcov
icc = c()
for (i in 1:7){
  icc = c(icc, sds[i]/(sds[i]+sds[30]))
}

tibble(
  Rubrics = rubrics.3,
  ICC.Artifact = icc
) %>%
  kable(caption = "ICC analysis for Artifacts under full model", "pipe") %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))

## 4


### Plots


p1 = dat %>% ggplot(aes(x = Rating, y = ..prop..)) +
  facet_wrap( ~ Semester) +
  geom_bar() +
  theme_bw() +
  labs(title = "Distribution of ratings: Semester", y = "Percentage")

p2 = dat %>% ggplot(aes(x = Rating, y = ..prop..)) +
  facet_wrap( ~ Sex) +
  geom_bar() +
  theme_bw() +
  labs(title = "Distribution of ratings: Sex", y = "Percentage")

grid.arrange(p1, p2, nrow = 1)



### Summary tables


semester.dat = dat %>%
  dplyr::group_by(Semester) %>%
  dplyr::summarise(
    count = n(),
    mean = mean(Rating),
    sd = sd(Rating)
  )
```

```r
sex.dat = dat %>%
  dplyr::group_by(Sex) %>%
  dplyr::summarise(
    count = n(),
    mean = mean(Rating),
    sd = sd(Rating)
  )

kable(
  list(semester.dat, sex.dat),
  caption = 'Summary tables for semester and sex', "pipe"
) %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))

### One-way ANOVA test for means

semester.aov <- tidy(aov(Rating ~ Semester, data = dat))

sex.aov <- tidy(aov(Rating ~ Sex, data = dat))

kable(
  list(semester.aov, sex.aov),
  caption = 'ANOVA tables for semester and sex', digits = 2, "pipe"
) %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"))
```