# Mixed Effects Regression Analysis on Evaluations of Freshman Statistics from "General Education" Program

Olivia Wang
ziyanw2@andrew.cmu.edu

28 November 2021

**Abstract**

Dietrich College at Carnegie Mellon University is interested in student work performed in Freshman Statistics from the new "General Education" program. In this paper, we address four key research questions related to evaluations of Freshman Statistics from "General Education" program. The data for this study come from a recent experiment conducted by Dietrich College with 91 artifacts rated by three raters on seven rubrics were randomly sampled from a Fall and Spring section of Freshman Statistics courses for the 2019 calendar year, which is sourced from Junker (2021). To evaluate research questions, we employ exploratory data analysis on summary statistics and barplots; calculate intraclass correlation (ICC) and percent exact agreement; develop multiple mixed-effects models, and investigate other interesting relationship between variable Sex and ratings. The results suggest that the distribution of ratings for each rubric is pretty much unique, and the distribution of ratings given by each rater is also distinguishable; whether raters generally agree on their scores or not depends on different rubrics. While the final mixed-effects model suggests that Rater, Semester, and Rubric are three factors related to the ratings; there are some interesting interactions between factors Rater, Rubric, and Artifact; factor Sex has no significant effect in predicting ratings. This conclusion, however, has some limitations of our small sample size and model shortcomings. Ideally, we would need additional samples to perform more comprehensive analyses.

## 1. Introduction

General Education is very important in college, because it can provide students with the foundation need to become highly intelligent in chosen field of study, and in life after college. To emphasize the importance of general education for undergraduates, Dietrich College at Carnegie Mellon University is in the process of implementing a new "General Education" program for undergraduates. This program specifies a set of courses and experiences that all undergraduates must take, and in order to determine whether the new program is successful, the college hopes to rate student work performed in each of the "General Education" courses each year. Recently, the college has been experimenting with student evaluations of Freshman Statistics, using raters from across the college. In particular, the associate dean of Dietrich College is interested to learn more about the results of the recent experiment which are displayed in four key research questions:

- **Distribution of Ratings:** Is the distribution of ratings for each rubric 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?
- **Agreement among Raters:** 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?

- **Mixed Effects Regression Analysis:** More generally, how are various factors in this experiment (Rater, Semester, Sex, Repeated, Rubric) related to the ratings? Do the factors interact in any interesting ways?
- **Interesting Topics:** Is there anything else interesting to say about this data?

## 2. Data

The data for this study come from a recent experiment of the new "General Education" program with rating work in Freshman Statistics conducted by Dietrich College at Carnegie Mellon University. In the experiment, 91 project papers referred to as "artifacts" were randomly sampled from a Fall and Spring section of Freshman Statistics courses for the 2019 calendar year. To evaluate these artifacts, three raters from three different departments were asked to rate these artifacts on seven rubrics. For all the rubrics, the rating scale is the same with values defined as integers ranging from 1 to 4. The reader should refer to Junker (2021) for detail descriptions about rubrics for rating Freshman Statistics projects and rating scale used for all rubrics.

| 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 1: Variable Definitions for the experimental data from Dietrich College.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | SD |
|---|---|---|---|---|---|---|---|
| Research Question | 1 | 2 | 2 | 2.35 | 3 | 4 | 0.59 |
| Critique Design | 1 | 1 | 2 | 1.85 | 2 | 4 | 0.83 |
| Initial EDA | 1 | 2 | 2 | 2.44 | 3 | 4 | 0.70 |
| Select Method | 1 | 2 | 2 | 2.05 | 2 | 3 | 0.48 |
| Interpret Result | 1 | 2 | 3 | 2.48 | 3 | 4 | 0.61 |
| Visual Organization | 1 | 2 | 2 | 2.41 | 3 | 4 | 0.68 |
| Text Organization | 1 | 2 | 3 | 2.60 | 3 | 4 | 0.70 |

Table 2: Summary Statistics for ratings of each rubric based on whole set of 91 artifacts.

| | N | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | SD |
|---|---|---|---|---|---|---|---|---|
| Rater 1 | 272 | 1 | 2 | 2 | 2.35 | 3 | 4 | 0.70 |
| Rater 2 | 272 | 1 | 2 | 2 | 2.43 | 3 | 4 | 0.70 |
| Rater 3 | 266 | 1 | 2 | 2 | 2.15 | 3 | 4 | 0.69 |

Table 3: Summary Statistics for ratings given by each rater based on whole set of 91 artifacts.

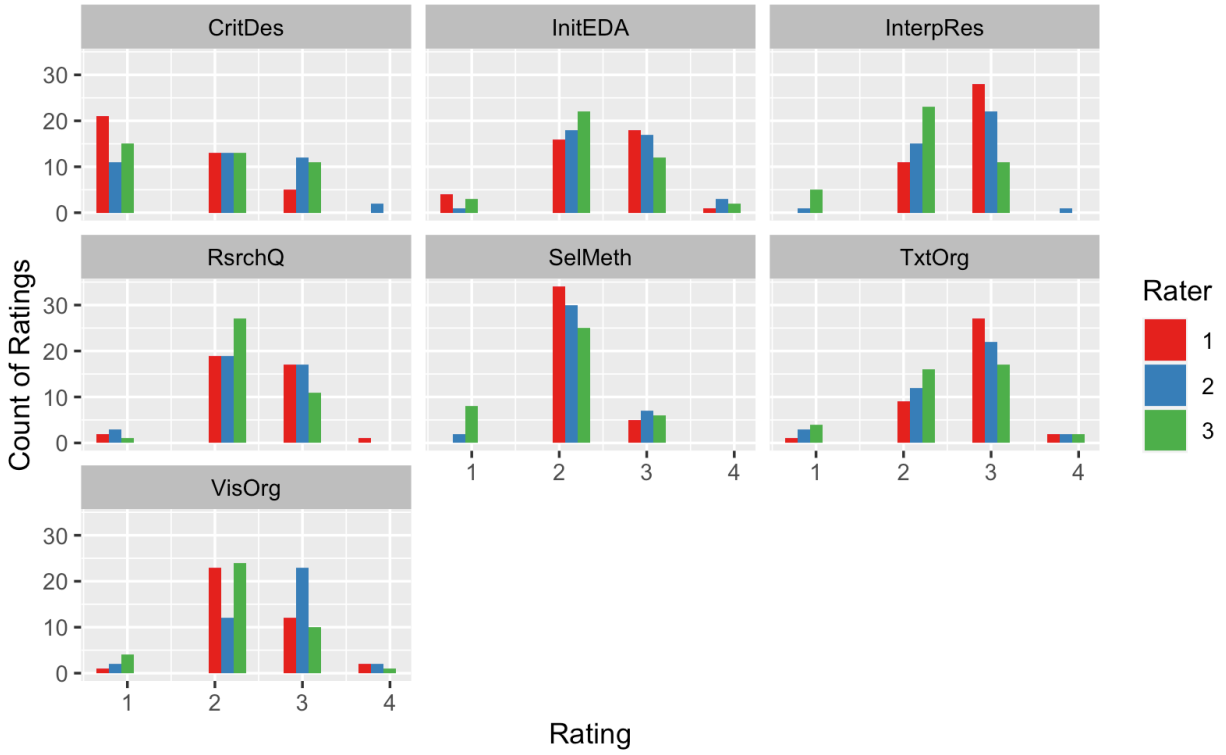



Figure 1: Barplots of ratings for each rubric given by each rater for whole set of 91 artifacts.

In order to fairly rate student work performed in Freshman Statistics courses from General Education program, the raters did not know which class or which students produced the artifacts that they rated. Based on total of 91 artifacts, thirteen of the artifacts were rated by all three raters. The remaining 78 artifacts were rated by only one rater. In addition, the data are given in the two files ratings.csv and tall.csv, which both contain the variables available for analysis are defined in Table 1. However, the data has been organized in two different formats for our analyses. The file ratings.csv contains data organized exactly as in Table 1. The file tall.csv contains the same data, but organized so that each row contains just one rating, in the column labelled Rating, and the rubric for that rating is listed in the column labelled Rubric.

There are total of 117 observations in the file ratings.csv, with three observations having missing values based on variables Sex, CritDes, and VisOrg. These observations have been dropped from the data set (Appendix 1, Part B, p. 6) to simplify the modeling process. Besides, there are total of 819 observations in the file tall.csv, with nine observations having missing values based on variables Sex and Rating. Accordingly, these observations have also been dropped from the data set (Appendix 1, Part C, p. 7).

In Figure 1, we show the distribution of ratings for each rubric given by each rater for full dataset. Across most of the rubrics, more artifacts were rated at middle scores such as 2 or 3 by each rater. Critique Design is the only rubric, on which most of the artifacts were rated at score 1 by Rater 1 and Rater 3. Hence, rubric Critique Design has the lowest mean ratings among all the rubrics as shown in Table 2. Select Method is the only rubric on which no artifact was rated at score 4. Across all the raters, the distributions of ratings given by Rater 1 and Rater 2 on each rubric are relatively similar. Table 3 shows the summary statistics for ratings given by each rater based on whole set of 91 artifacts, we can see that the average ratings given by Rater 3 is lower than other raters. The distribution of the ratings for each rubric given by each rater for 13 artifacts seen by all three raters is very similar to the whole data set, which can be found in Appendix 1, Part A, p. 4.

# 3. Methods

As mentioned before, the associate dean of Dietrich College is interested to learn more about the four key research questions, we outlined the methods for each research question below.

## 3.1 Distribution of Ratings

The first research question considered on the distributions of ratings for each rubric and each rater, we mainly focused on exploratory data analysis on summary statistics and barplots. Specifically, we first set up an assumption to define low/high rating. Then, we examined one-dimensional summary statistics and barplots of ratings for each rubric to compare the qualitative variations among distribution of ratings that each rubric may follow. Similarly, we analyzed one-dimensional summary statistics and barplots of ratings given by each rater to investigate the qualitative diaparity among distribution of ratings given by each rater. We conducted the above analyses on both full dataset for whole set of 91 artifacts and subset of the data for just the 13 artifacts seen by all three raters to make conclusions.

## 3.2 Agreement among Raters

The second research question aimed to measure the agreement among three raters and determine which raters might be contributing to disagreement. The analyses consist of two parts:

### Measure of agreement among raters

To measure the agreement among raters' ratings for each artifact, we calculated the Intraclass Correlation (ICC), which describes the reliability of ratings or measurements for specific clusters. To calculate it, we treated each artifact as a cluster of three ratings given by three raters and fitted seven random-intercept models with Artifact as the grouping variable, one for each rubric. Then, we calculated the seven ICC's, which can be very helpful to determine whether the raters are generally in agreement (high correlation among the raters) or not (low correlation among the raters) on each rubric. The analysis performed on both full dataset and subset of the data for just the 13 artifacts seen by all three raters.

### Find raters contributing to disagreement

To find which raters might be contributing to disagreement across all raters, we calculated the percent exact agreement, which measures the proportion of times on which two raters gave the same rating for each artifact on a specific rubric. To calculate it, we made a 2-way table of counts for the ratings of each pair of raters, on each rubric. Considering there are three pairs of raters, each rubric will get three tables, we got a total of 21 tables. For each table, the percentage of observations on the main diagonal is the percent exact agreement between the two raters. As a result, higher percent exact agreement for three pairs of raters on a specific rubric means all raters agreed on their ratings on a specific rubric. On the contrary, lower percent exact agreement for only one pair of raters on a specific rubric indicates one of the raters might be contributing to disagreement. The analysis performed only on subset of the data for just the 13 artifacts seen by all three raters.

## 3.3 Mixed Effects Regression Analysis

The third research question focused on how various factors in this experiment are related to the ratings, and whether they interacted in any interesting ways. We conducted analyses in two parts:

### Fit seven mixed-effects models for each rubric

We fitted seven separate mixed-effects models for each rubric with Artifact as the grouping variable. The mixed-effects model for each rubric using Artifact as the random grouping variable is a good way to account for the shared differences of ratings on each rubric across the artifacts. To illustrate, the differences of ratings for each artifact should be similar with other artifacts on each rubric.

We started to add fixed effects for all the variables Rater, Semester, Sex and Repeated to each of the seven intercept-only models, one for each rubric. We performed automated backward elimination of fixed effects using BIC as our selection criterion to determine which fixed effects should be included in models. Then, we validated our result with ANOVA test based on AIC, BIC and likelihood ratio test (LRT) to examine the significance of added fixed effects. After that, we started to add corresponding fixed-effect interactions based on added fixed effects. We repeated ANOVA test to examine the significance of added interaction terms. Lastly, we added random effects that are also presented as fixed effects to the seven rubric specific models. We repeated ANOVA test to determine which random effects should be included in models based on statistical significance. As a result, seven final mixed-effects models for each rubric were evaluated based on summary regression statistics. We performed the above analysis on both full dataset and subset of the data for just the 13 artifacts seen by all three raters.

### Fit final mixed-effects model

Considering that the first approach doesn't let us directly examine interactions with Rubric, since each model considers only one rubric at a time. We fitted a final mixed-effects model with Artifact as the grouping variable to directly investigate integrations between Rubric and other factors. We started to add fixed effects for all of the variables Rater, Semester, Sex, Repeated and Rubric to the intercept-only model with random intercept for Rubric. The following modeling process was the same as described in the first approach. As a result, one final mixed-effects model was evaluated based on summary regression statistics to interpret how various factors in this experiment are related to the ratings, and whether they interacted in any interesting ways. We performed the above analysis only on full dataset for whole set of 91 artifacts.

## 3.4 Interesting Topics

Lastly, we illustrated on other things we could say about our analysis, that will be of interest to the associate dean and something that is interesting and useful to the college. We focused on further exploratory data analysis on summary statistics and barplots to examine the distribution of ratings based on variable Sex to investigate if there is any other interesting relationship between variable Sex and ratings. Then, we assessed the qualitative observations from the exploratory data analysis to determine if they aligned with the final mixed-effects model got from Section 4.3, Part 2. If not, possible interpretation would be provided. The analysis performed on both full dataset for whole set of 91 artifacts and subset of the data for just the 13 artifacts seen by all three raters to make conclusions.

# 4. Results

## 4.1 Distribution of Ratings

First, we set up an assumption to define low/high rating. Here, among all rubrics, we defined low rating as artifact was rated less than or equal to 2; high rating as artifact was rated above 2. As shown in Table 2, the summary statistics of ratings for each rubric based on full dataset, the distribution of ratings for each rubric is pretty much distinguishable from the other rubrics. Specifically, rubric Text Organization tends to get higher ratings because it gets the highest values from all numerical summary of ratings shown in Table 2 among all the rubrics. In addition, rubric Critique Design tends to get lower ratings since it has the lowest mean and first quartile of ratings among all the rubrics. Similarly rubric Select Method also tends to get lower ratings because it has the same value of median and third quartile of ratings with the lowest standard deviation of ratings among other rubrics, which indicates that nearly 75% of the artifacts were rated less than or equal to score 2. The analysis of summary statistics of ratings for each rubric based on 13 artifacts seen by all three raters is very similar to the whole data set, which can be found in Appendix 2, Part C, p. 16.

Then, we made visual comparison for barplots of ratings for each rubric based on all artifacts. Figure 2 shows the similar patterns as the summary statistics of ratings for each rubric based on full dataset. We clearly found that there are total of 71 high ratings on rubric Text Organization which is the highest number

of high ratings among all the rubrics. Critique Design is the only rubric, on which most of the artifacts were rated at score 1. Select Method is the only rubric on which extremely majority of the artifacts were rated at score 2 and no artifact was rated at score 4. However, considering that low rating as artifact was rated less than or equal to 2, there are total of 99 low ratings on rubric Select Method which is the highest number of low ratings among all the rubrics. However, the barplots of ratings for each rubric based on 13 artifacts seen by all three raters in Appendix 2, Part A, p. 10 show very similar patterns as the whole data set, except that the number of low ratings on rubric Select Method and Critique Design are the same which is the highest one among all the rubrics.

Hence, we concluded that based on full dataset, the distribution of ratings for each rubric is pretty much distinguishable from the other rubrics. Text Organization tends to get especially high ratings, and rubric Select Method tends to get especially low ratings.

As shown in Table 3, the summary statistics for ratings given by each rater based on full dataset, all of the values from numerical summary of ratings given by three raters are very similar. It's worthy notching that there are 272 ratings given by each of Rater 1 and Rater 2, and there are 266 ratings given by Rater 3 based on whole set of 91 artifacts. The average ratings given by Rater 3 is lower than other raters. Then, we made visual comparison of the barplots of ratings given by each rater based on full dataset from Figure 3, we clearly found that the distributions of ratings given by Rater 1 and Rater 2 on each rubric are relatively similar. So, there is no significant evidence from distributions of ratings given by Rater 1 and Rater 2 that we could conclude which rater tends to give especially high ratings. However, Rater 3 tends to give especially low ratings since the total number of low ratings given by Rater 3 is 190 which is the highest number of low ratings given by all raters. The barplots of ratings given by each rater based on subset of the data in Appendix 2, Part B, p. 13 show some slightly differences from the barplots based on full dataset. For instance, Rater 3 becomes the only rater without giving score 4, and Rater 2 tends to give more ratings at score 2 than score 3.

Hence, we concluded that based on full dataset, the distribution of ratings given by each rater is distinguishable from the other raters. Rater 3 tends to give especially low ratings, there is no rater tends to give especially high ratings.

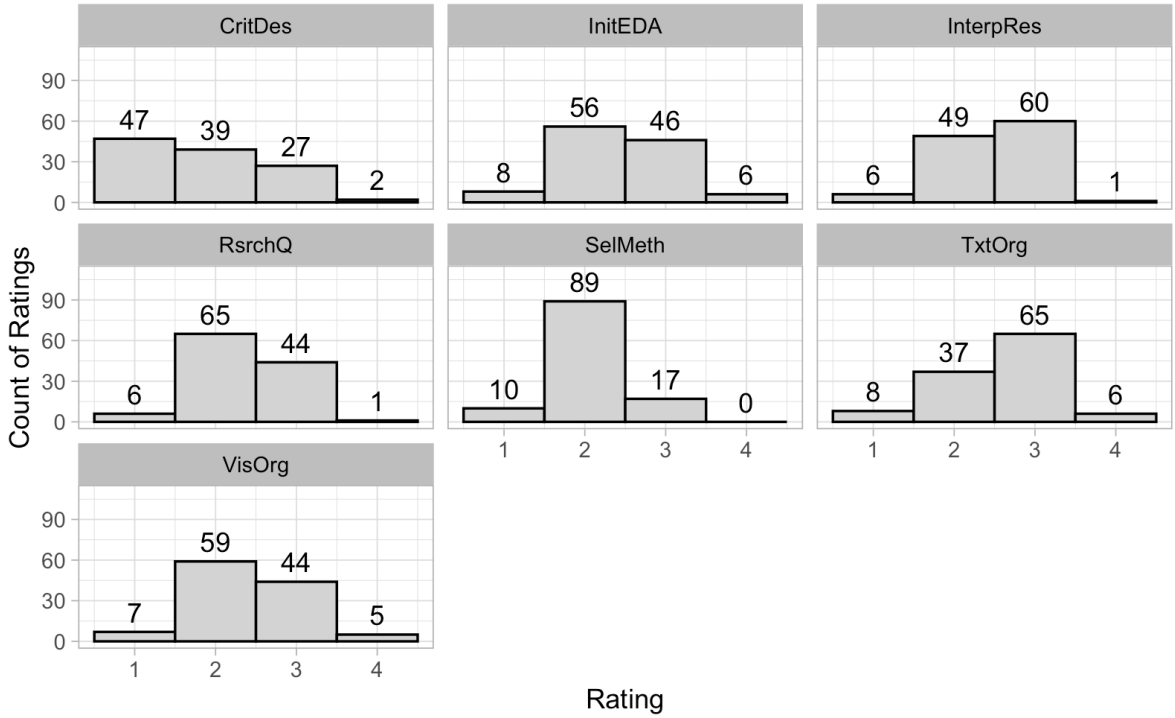



Figure 2: Barplots of ratings for each rubric based on whole set of 91 artifacts.

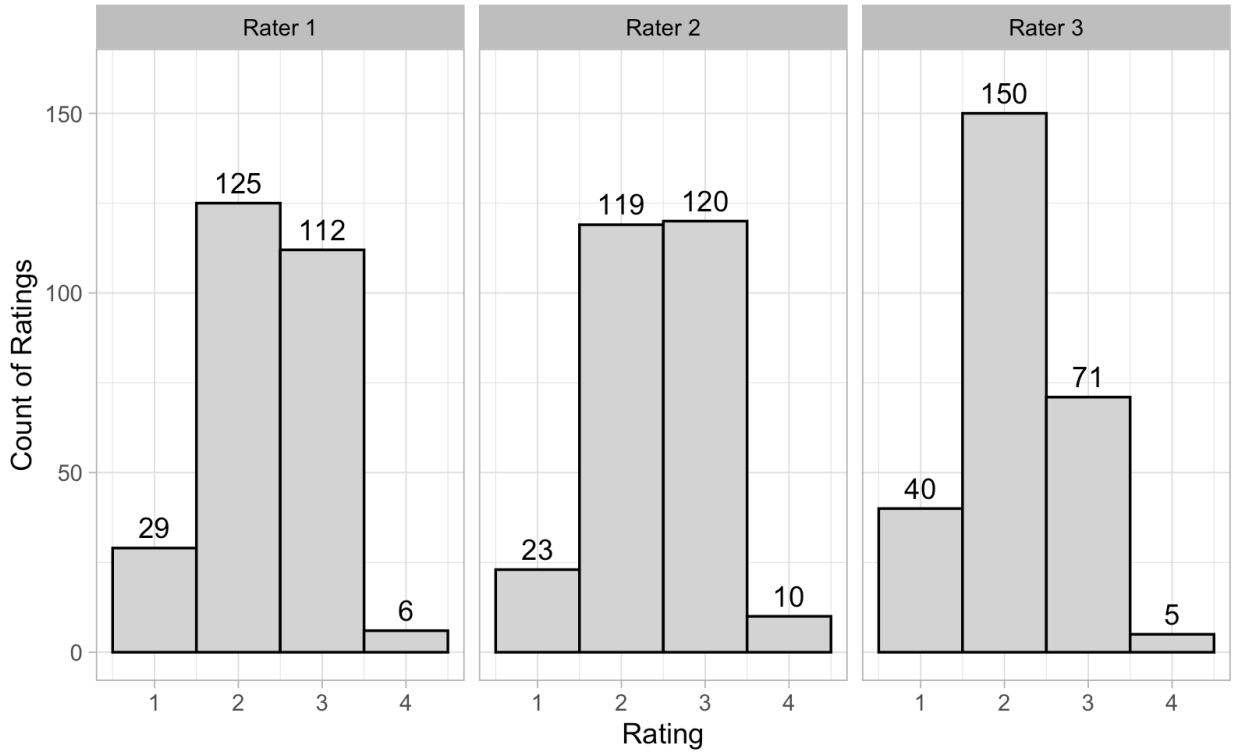



Figure 3: Barplots of ratings given by each rater based on whole set of 91 artifacts.

## 4.2 Agreement among Raters

As mentioned in the Methods Section 3.2, the second research question aimed to measure the agreement among three raters and determine which raters might be contributing to disagreement. The analyses consist of two parts:

| Rubric | ICC Subset | ICC Full |
|---|---|---|
| Research Question | 0.19 | 0.21 |
| Critique Design | 0.57 | 0.67 |
| Initial EDA | 0.49 | 0.69 |
| Select Method | 0.52 | 0.46 |
| Interpret Results | 0.23 | 0.22 |
| Visual Organization | 0.59 | 0.66 |
| Text Organization | 0.14 | 0.19 |

Table 4: Intraclass correlation (ICC) for ratings on each rubric based on full dataset and subset of the data.

### Measure of agreement among raters

The first part is to calculate the Intraclass Correlation (ICC), which helps to measure the agreement among raters' ratings for each artifact. Table 4 presents the output of the seven intraclass correlations for seven

rubrics based on full dataset and subset of the data. In general, ICC between 0.50 and 0.75 is defined as moderate reliability and ICC below 0.50 is defined as poor reliability. Hence, we concluded that raters generally agree more on their ratings on rubrics Initial EDA, Critique Design, Select Method and Visual Organization based on full dataset. However, raters generally disagree on their ratings on the remaining rubrics, especially on rubric Text Organization. Note that these conclusions are the same for raters' ratings on each rubric based on subset of the data, except those raters become not consistent with their ratings on rubric Select Method.

| Rubric | Rater 1 & Rater 2 | Rater 1 & Rater 3 | Rater 2 & Rater 3 |
|---|---|---|---|
| Research Question | 0.38 | 0.77 | 0.54 |
| Critique Design | 0.54 | 0.62 | 0.69 |
| Initial EDA | 0.69 | 0.54 | 0.85 |
| Select Method | 0.92 | 0.62 | 0.69 |
| Interpret Results | 0.62 | 0.54 | 0.62 |
| Visual Organization | 0.54 | 0.77 | 0.77 |
| Text Organization | 0.69 | 0.62 | 0.54 |

Table 5: Percent exact agreement between two raters on each rubric based on full dataset.

### Find raters contributing to disagreement

The second part is to find which raters might be contributing to disagreement across all raters, we calculated total of 21 percent exact agreements between each pair of raters on each rubric based on full dataset. The results are shown in Table 5. Obviously, percent exact agreements are close between each pair of raters on rubrics Critique Design, Interpret Results, Visual Organization and Text Organization, which means that raters are generally in agreement with their ratings on these rubrics. We mainly focused on rubrics Research Question, Initial EDA, and Select Method, since on which there are comparable differences among percent exact agreements between each pair of raters. In particular, we can conclude the followings:

- **Research Question:** Rater 2 might be contributing to disagreement among raters' ratings on rubric Research Question. Because Rater 1 and Rater 2 had very low agreement with their ratings on rubric Research Question; Rater 1 and Rater 3 had higher agreement with their ratings on rubric Research Question.
- **Initial EDA:** Rater 1 might be contributing to disagreement among raters' ratings on rubric Initial EDA. Because Rater 1 and Rater 2 or Rater 3 had moderate agreement with their ratings on rubric Initial EDA; Rater 2 and Rater 3 had very high agreement with their ratings on rubric Initial EDA.
- **Select Method:** Rater 3 might be contributing to disagreement among raters' ratings on rubric Select Method. Because Rater 3 and Rater 1or Rater 2 had moderate agreement with their ratings on rubric Select Method; Rater 1 and Rater 2 had extremely high agreement with their ratings on rubric Select Method.

In conclusion, whether raters generally agree on their scores or not depends on different rubrics. Each rater has a specific rubric on which they disagree on their ratings with other two raters.

## 4.3 Mixed Effects Regression Analysis

As mentioned in Methods Section 3.3, the third research question focused on how various factors in this experiment are related to the ratings, and whether they interacted in any interesting ways. We conducted analyses in two parts:

### Fit seven mixed-effects models for each rubric

In the first part, we fitted seven separate mixed-effects models for each rubric with Artifact as the grouping variable on both full dataset and subset of the data for just the 13 artifacts seen by all three raters.

First of all, we analyzed on the subset of the data, we added significant fixed effects for all the variables Rater, Semester, Sex and Repeated to each of the seven intercept-only models, one for each rubric. After automated backward elimination of fixed effects using BIC as our selection criterion, we only included fixed effect Rater to each of the seven intercept-only models (Appendix 4, Part A, p. 19). However, from the output of each ANOVA test for each rubric, AIC and BIC disagreed, the likelihood ratio test (LRT) was in favor of the original intercept-only model (Appendix 4, Part A, p. 20). Hence, the seven intercept-only models were adequate here, we didn't move to check for any fixed-effect interactions and random effects.

Secondly, we analyzed on full dataset, we repeated the same process as mentioned before. After automated backward elimination of fixed effects using BIC as our selection criterion, we included fixed effect Rater to each of the seven intercept-only models. However, we also included fixed effect Semester to the intercept-only model for rubric Select Method. From the output of each ANOVA test for each of the rubrics in Initial EDA, Research Question, and Text Organization, we ended with the intercept-only model. Similarly, based on the output of each ANOVA test for each of the rubrics in Critique Design, Interpret Results, and Visual organization, we got fixed effect Rater should be added to the original intercept-only model based on statistical significance. For rubric Select Method, the output of ANOVA test shows that all of AIC, BIC, and the likelihood ratio test (LRT) preferred with added variable Semester; AIC and BIC disagreed, the likelihood ratio test (LRT) was in favor of added fixed effect Rater, hence we decided to add both fixed effects Rater and Semester to the intercept-only model for rubric Select Method. Details of these analyses in R can be found in Appendix 4, Part B, p. 20-21.

After fitting the fixed effects, Select Method was the only rubric with model that included more than one fixed effect. Therefore, we added the fixed-effect interaction term between variable Rater and Semester. We repeated ANOVA test again, the output shows that all of AIC, BIC, and the likelihood ratio test (LRT) preferred with the fixed-effect interaction term was not statistically significant to the model for rubric Select Method. Finally, we repeated ANOVA test for each of the four models for rubrics Critique Design, Interpret Results, Visual Organization, and Select Method with significant fixed effects added to see if their corresponding random effects were also significant to the model. As a result, the outputs show that, none of the random effects could be fitted and the summary regression statistics for seven final mixed-effects models for each rubric based on full dataset is shown in Table 6. Details of these analyses in R can be found in Appendix 4, Part C, p. 21-29.

| | Select Method | Initial EDA | Research Question | Text Organization | Critique Design | Interpret Results | Visual Organization |
|---|---|---|---|---|---|---|---|
| $\hat{\beta}_0$: (Intercept) | - | 2.44 | 2.35 | 2.59 | - | - | - |
| $\hat{\beta}_1$: SemesterS19 | -0.36 | - | - | - | - | - | - |
| $\hat{\beta}_2$: Rater1 | 2.25 | - | - | - | 1.69 | 2.70 | 2.38 |
| $\hat{\beta}_3$: Rater2 | 2.23 | - | - | - | 2.11 | 2.59 | 2.65 |
| $\hat{\beta}_4$: Rater3 | 2.03 | - | - | - | 1.89 | 2.14 | 2.28 |
| $\hat{\sigma}^2$: (Std.Dev) | 0.11 | 0.17 | 0.28 | 0.40 | 0.25 | 0.25 | 0.15 |
| $\hat{\tau}^2$: (Std.Dev) | 0.10 | 0.37 | 0.07 | 0.09 | 0.44 | 0.06 | 0.29 |

Table 6: Summary regression statistics for seven final mixed-effects models based on full dataset.

Based on Table 6, the estimated coefficients were provided for seven final mixed-effects models based on full dataset, we summarized the interpretations of seven final mixed-effects models as followings:

**Select Method**

The final mixed-effects model for rubric Select Method, shown below as:

$$\text{Rating} = \beta_1 \times \text{SemesterS19} + \beta_2 \times \text{Rater1} + \beta_3 \times \text{Rater2} + \beta_4 \times \text{Rater3} + (1 \mid \text{Artifact}) - 1 + \varepsilon$$

- **(1 | Artifact):** This is a random intercept term grouped by Artifact, which measures the random effect deviation from the overall mean rating for each artifact. Considering the same rater in the same semester (Semester Spring 19 or Semester Fall 19), different artifacts of the total 91 artifacts tend to get different mean ratings on rubric Select Method.
- **Rater:** Fixed effect Rater has a statistically significant effect on the ratings for rubric Select Method. Considering the same artifact in the same semester, Rater 3 tends to give the lowest rating on rubric Select Method, followed by Rater 2 tends to give 0.20 higher rating than Rater 3 and Rater 1 tends to give 0.02 higher rating than Rater 2.
- **Semester:** Fixed effect Semester has a statistically significant effect on the ratings for rubric Select Method. Considering the same artifact rated by the same rater, the artifact received 0.36 lower rating in Semester Spring 19 than in Semester Fall 19 on rubric Select Method.

**Initial EDA & Research Question & Text Organization**

The final mixed-effects model for rubrics Initial EDA, Research Question and Text Organization is the same, shown below as:

$$\text{Rating} = \beta_0 + (1 \mid \text{Artifact}) + \varepsilon$$

- **(1 | Artifact):** Different artifacts of the total 91 artifacts tend to get different mean ratings on rubrics Initial EDA, Research Question and Text Organization.

**Critique Design & Interpret Results & Visual Organization**

The final mixed-effects model for rubrics Critique Design, Interpret Results, and Visual Organization is the same, shown below as:

$$\text{Rating} = \beta_2 \times \text{Rater1} + \beta_3 \times \text{Rater2} + \beta_4 \times \text{Rater3} + (1 \mid \text{Artifact}) - 1 + \varepsilon$$

- **(1 | Artifact):** Considering the same rater, different artifacts of the total 91 artifacts tend to get different mean ratings on rubrics Critique Design, Interpret Results, and Visual Organization.
- **Rater:** Considering the same artifact in the same semester, 1) For rubric Critique Design, Rater 1 tends to give the lowest rating, followed by Rater 3 tends to give 0.20 higher rating than Rater 1 and Rater 2 tends to give 0.22 higher rating than Rater 3; 2) For rubric Interpret Results, Rater 3 tends to give the lowest rating, followed by Rater 2 tends to give 0.45 higher rating than Rater 3 and Rater 1 tends to give 0.11 higher rating than Rater 2; 3) For rubric Visual Organization, Rater 3 tends to give the lowest rating, followed by Rater 1 tends to give 0.10 higher rating than Rater 3 and Rater 2 tends to give 0.27 higher ratings than Rater 1.

**Conclusion**

In conclusion, from the seven final mixed-effects models for each rubric based on full dataset, we found that fixed effect Rater has a statistically significant effect on the ratings for rubrics Select Method, Critical Design, Interpret Results, and Visual Organization; while fixed effect Semester only has a significant effect on the ratings for rubric Select Method. The negative estimated coefficient for fixed effect Semester indicates that artifacts generally received lower ratings in Semester Spring 19 than in Semester Fall 19 on

rubric Select Method. For ratings on rubrics Initial EDA, Research Question and Text Organization, which are not related to any factors in this experiment. Overall, Rater, Semester, and Rubric are three factors which are related to the ratings.

## Fit final mixed-effects model

In the second part, we fitted a final mixed-effects model with Artifact as the grouping variable on full dataset to directly investigate integrations between Rubric and other factors.

We started to add fixed effects for all the variables for all of the variables Rater, Semester, Sex, Repeated and Rubric to the intercept-only model with random intercept for Rubric. After automated backward elimination of fixed effects using BIC as our selection criterion, we successfully included three fixed effects Rater, Semester and Rubric to the intercept-only model based on statistical significance. Next, we considered about the possible combinations of fixed-effect interactions with significant fixed effects. We repeated ANOVA test again, the output shows that only fixed-effect interaction term between variables Rater and Rubric was statistically significant to the model. Finally, we added three random effects Rater, Semester and interaction term between variables Rater and Rubric which are also presented as fixed effects to the model. The ANOVA test shows that only random effect Rater should be included in model based on statistical significance. Details of these analyses in R can be found in Appendix 4, Part D, p. 30-32. The final mixed-effects model we got, as shown in Table 7 with summary regression statistics:

$$\text{Rating} = \beta_0 + \beta_1 \times \text{SemesterS19} + \beta_2 \times \text{Rater1} + \beta_3 \times \text{Rater2} + \beta_4 \times \text{Rater3} + \beta_5 \times \text{Rater:Rubric} + (0 + \text{Rubric} \mid \text{Artifact}) + (0 + \text{Rater} \mid \text{Artifact}) + \varepsilon$$

| | Estimate | Std. Error | t value |
|---|---|---|---|
| $\hat{\beta}_0$: (Intercept) | 1.76 | 0.11 | 15.41 |
| $\hat{\beta}_1$: SemesterS19 | -0.16 | 0.08 | -2.08 |
| $\hat{\beta}_2$: Rater2 | 0.37 | 0.14 | 2.63 |
| $\hat{\beta}_3$: Rater3 | 0.20 | 0.13 | 1.51 |
| $\hat{\beta}_4$: RubricInitEDA | 0.74 | 0.13 | 5.69 |
| $\hat{\beta}_5$: RubricInterpRes | 0.99 | 0.13 | 7.76 |
| $\hat{\beta}_6$: RubricRsrchQ | 0.73 | 0.12 | 6.16 |
| $\hat{\beta}_7$: RubricSelMeth | 0.41 | 0.13 | 3.29 |
| $\hat{\beta}_8$: RubricTxtOrg | 1.02 | 0.13 | 7.81 |
| $\hat{\beta}_9$: RubricVisOrg | 0.65 | 0.13 | 4.90 |
| $\hat{\beta}_{10}$: Rater2:RubricInitEDA | -0.30 | 0.17 | -1.92 |
| $\hat{\beta}_{11}$: Rater2:RubricInterpRes | -0.51 | 0.15 | -3.34 |
| $\hat{\beta}_{12}$: Rater2:RubricRsrchQ | -0.49 | 0.15 | -3.31 |
| $\hat{\beta}_{13}$: Rater2:RubricSelMeth | -0.39 | 0.15 | -2.57 |
| $\hat{\beta}_{14}$: Rater2:RubricTxtOrg | -0.55 | 0.16 | -3.52 |
| $\hat{\beta}_{15}$: Rater2:RubricVisOrg | -0.11 | 0.16 | -0.66 |
| $\hat{\beta}_{16}$: Rater3:RubricInitEDA | -0.30 | 0.16 | -1.89 |
| $\hat{\beta}_{17}$: Rater3:RubricInterpRes | -0.72 | 0.15 | -4.65 |
| $\hat{\beta}_{18}$: Rater3:RubricRschQ | -0.32 | 0.15 | -2.19 |
| $\hat{\beta}_{19}$: Rater3:RubricSelMeth | -0.39 | 0.15 | -2.59 |
| $\hat{\beta}_{20}$: Rater3:RubricTxtOrg | -0.45 | 0.16 | -2.83 |
| $\hat{\beta}_{21}$: Rater3:RubricVisOrg | -0.28 | 0.16 | -1.73 |

Table 7: Summary regression statistics for final mixed-effects model based on full dataset.

Based on Table 7, the estimated coefficients were provided for final mixed-effects model based on full dataset, we can interpret the model as followings:

- **Semester:** Fixed effect Semester has a statistically significant effect on the ratings. Considering the same artifact rated on the same rubric by the same rater, the artifact received 0.16 lower rating in Semester Spring 19 than in Semester Fall 19.
- **Rubric:** Fixed effect Rubric has a statistically significant effect on the ratings.
    - **Critique Design:** Considering the same artifact in the same semester, Rater 1 tends to give the lowest rating on the rubric Critical Design, followed by Rater 3 tends to give 0.20 higher rating than Rater 1 and Rater 2 tends to give 0.17 higher rating than Rater 3.
    - **Initial EDA:** Considering the same artifact in the same semester, Rater 3 tends to give the lowest ratings on the rubric Initial EDA, followed by Rater 1 tends to give 0.10 higher rating than Rater 3 and Rater 2 tends to give 0.07 higher rating than Rater 1.
    - **Interpret Results:** Considering the same artifact in the same semester, Rater 3 tends to give the lowest ratings on the rubric Interpret Results, followed by Rater 2 tends to give 0.37 higher rating than Rater 3 and Rater 1 tends to give 0.16 higher rating than Rater 2.
    - **Research Question:** Considering the same artifact in the same semester, Rater 3 tends to give the lowest ratings on the rubric Research Question, followed by Rater 2 tends to give 0.01 higher rating than Rater 3 and Rater 1 tends to give 0.12 higher rating than Rater 2.
    - **Select Method:** Considering the same artifact in the same semester, Rater 3 tends to give the lowest ratings on the rubric Select Method, followed by Rater 2 tends to give 0.02 higher rating than Rater 3 and Rater 1 tends to give 0.02 higher rating than Rater 2.
    - **Text Organization:** Considering the same artifact in the same semester, Rater 3 tends to give the lowest ratings on the rubric Text Organization, followed by Rater 2 tends to give 0.06 higher rating than Rater 3 and Rater 1 tends to give 0.18 higher rating than Rater 2.
    - **Visual Organization:** Considering the same artifact in the same semester, Rater 3 tends to give the lowest ratings on the rubric Visual Organization, followed by Rater 1 tends to give 0.08 higher rating than Rater 3 and Rater 2 tends to give 0.26 higher rating than Rater 1.
- **Rater:** Fixed effect Rater has a statistically significant effect on the ratings. Considering all of the artifacts in the same semester, Rater 3 tends to give especially low ratings, Rater 1 tends to give higher ratings.
- **(0 + Rubric | Artifact) + Rubric:** There is an interaction between Rubric and Artifact. There are different average scores on each rubric, but the rubric averages also vary a bit from one Artifact to the next, by a small random effect that depends on Artifact. In all of this, the fact that Rubric scores depend on Artifact is what we might expect since the artifacts aren't all of equal quality on each rubric, and so we should expect the average scores on each Rubric to vary from one Artifact to the next.
- **(0 + Rater | Artifact) + Rater:** There is an interaction between Rater and Artifact. Each Rater's rating on each Artifact differs from what we would expect (from the fixed effects alone) by a small random effect that depends on the Artifact. This interaction suggests that the Raters are not interpreting the evidence in the artifacts in the same way.
- **Rubric + Rater + Rater:Rubric:** There is an interaction between Rater and Rubric. Each Rater uses each Rubric in a way that is not like, or even parallel to, other rater's Rubric usage. This interaction suggests that the Raters are not all interpreting the Rubrics in the same way.

- More troubling are the interaction between Rater and Rubric and the interaction between Rater and Artifact. These interactions suggest that perhaps the raters should be trained more, to make the raters' ratings more similar to each other.

### Conclusion

In conclusion, from the final mixed-effects model based on full dataset, we found that Rater, Semester, and Rubric are three factors in this experiment which are related to the ratings. Besides, there are some interesting interactions between factors Rater, Rubric, and Artifact, which indicate that raters are not all interpreting the rubrics and the evidence in the artifacts in the same way.

## 4.4 Interesting Topics

In this section, we considered about other things we could say about our analysis, that will be of interest to the associate dean and something that is interesting and useful to the college. As mentioned in the Methods Section 3.4, we mainly focused on further exploratory data analysis on summary statistics and barplots to examine the distribution of ratings based on variable Sex to investigate if there is any other interesting relationship between variable Sex and ratings.

Note that there are total of 434 ratings from females and 364 ratings from males based on whole set of 91 artifacts. In Figure 4, the barplots of ratings by sex based on full dataset shows that the distributions of ratings for both females and males are relatively similar. Specifically, both distributions follow the normal distribution roughly centered at ratings of 2 and 3, which indicates that for both females and males, most of the artifacts were rated at middle scores such as 2 or 3. The summary statistics for ratings by sex based on whole set of 91 artifacts shows the similar patterns as the barplots, which can be found in Appendix 5, p. 37.

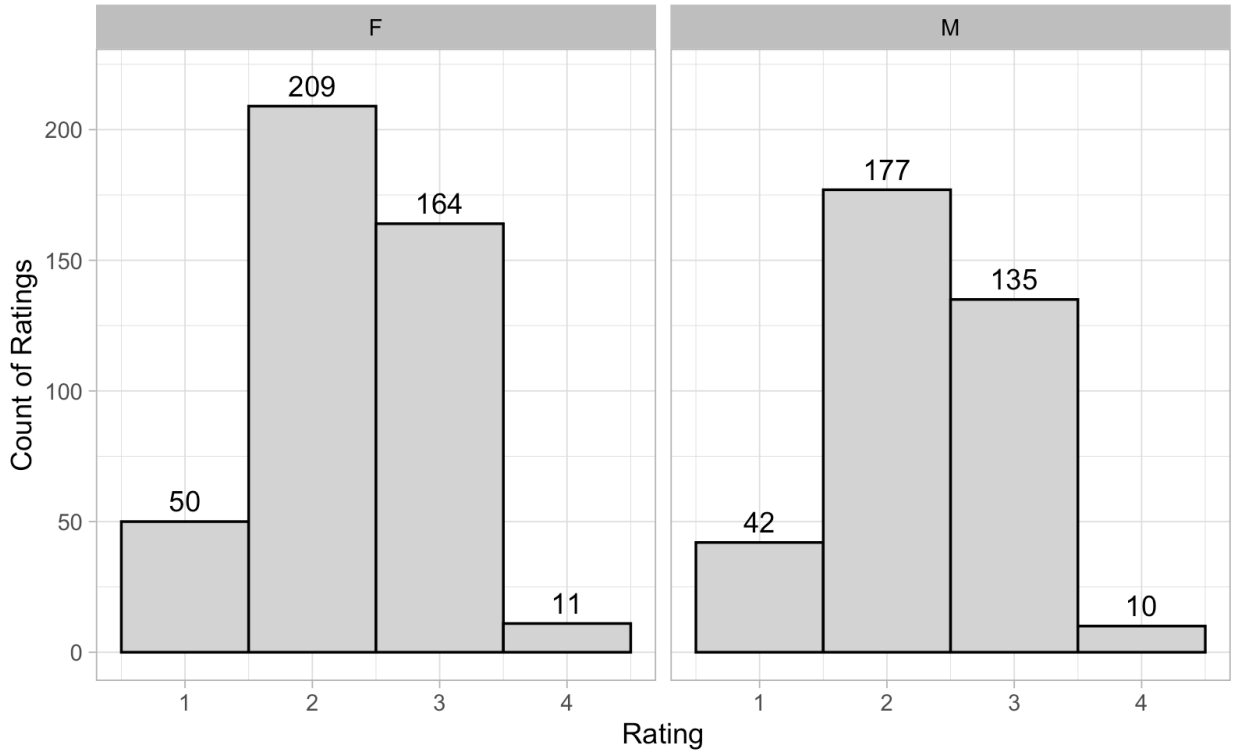



Figure 4: Barplots of ratings by sex based on whole set of 91 artifacts.

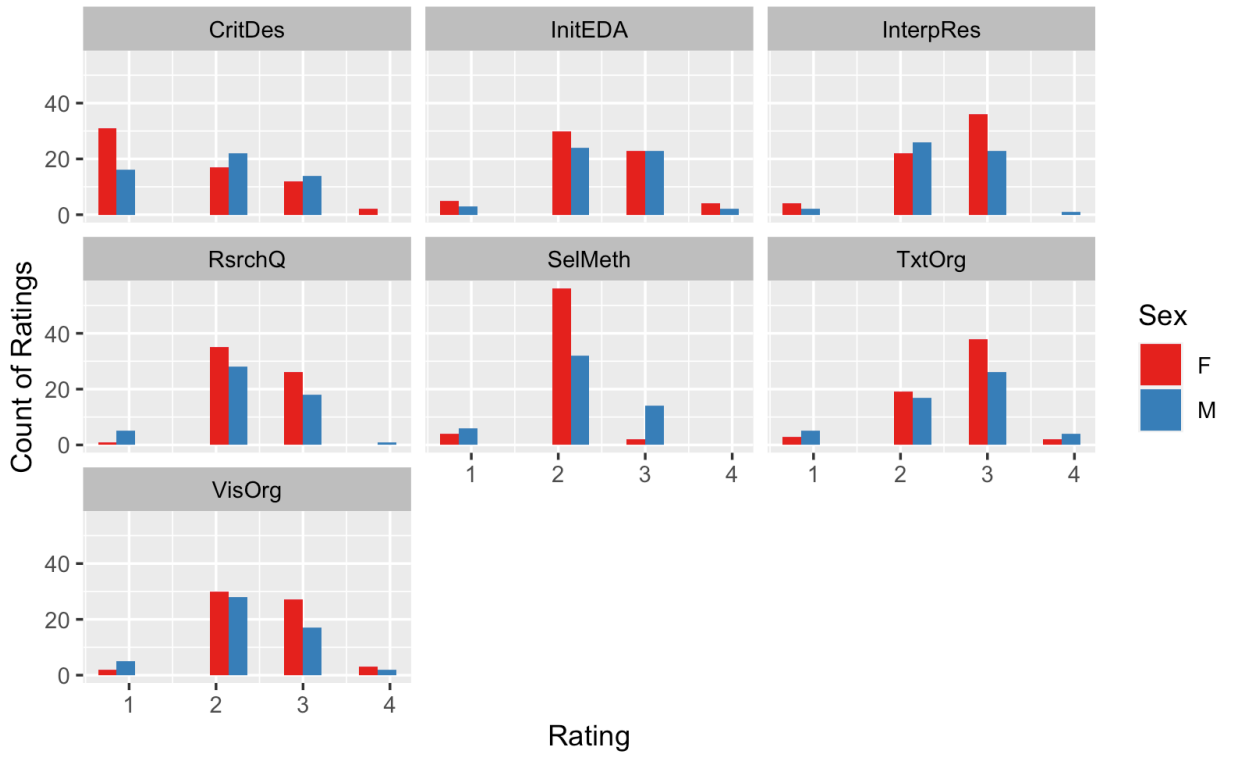



Figure 5: Barplots of ratings for each rubric by sex based on whole set of 91 artifacts.

When compared the barplots of ratings for each rubric by sex based on full dataset shown in Figure 5, we found that the distributions of ratings for both females and males on each rubric are also relatively similar, but there are some slight differences in ratings for each rubric by sex based on whole set of 91 artifacts. To illustrate, across most of the rubrics, it is obviously to see that both females and males received more ratings at middle scores such as 2 or 3. Critique Design is the only rubric, on which females got most of ratings at score 1, and males received most of ratings at score 1 or 2. In addition, Select Method is the only rubric, on which females got extremely most of the ratings at score 2, and none of the females and males received ratings at score 4. We also examined the barplots of ratings for each rubric by sex based on 13 artifacts seen by all three raters in Appendix 5, p. 40, which display very similar patterns as the whole data set, except that none of the females received ratings at score 4 on all the rubrics.

In conclusion, there is no noticeable difference in the distributions of ratings for both females and males on each rubric. We examined the results based on full dataset for whole set of 91 artifacts and subset of the data for 13 artifacts seen by all three raters. The results are aligned with the final mixed-effects model got from Section 4.3, Part 2, since Sex was not included as a significant factor in our final model.

## 5. Discussion

In this report, we performed mixed effects regression analysis on experimental data from a recent experiment of the new "General Education" program with rating work in Freshman Statistics conducted by Dietrich College at Carnegie Mellon University. Specifically, we addressed four key research questions. The first research question considered on the distributions of ratings for each rubric and each rater, we mainly focused on exploratory data analysis on summary statistics and barplots. We concluded that based on full dataset, the distribution of ratings for each rubric is pretty much unique, and the distribution of ratings given by each rater is also distinguishable. From the second research question, we calculated the Intraclass Correlation (ICC) and percent exact agreement to measure the agreement among three raters and determine which raters might be contributing to disagreement. As a result, we found that whether raters

generally agree on their scores or not depends on different rubrics, each rater has a specific rubric on which they disagree on their ratings with other two raters. We fitted seven separate mixed-effects models for each rubric and a final mixed-effects model to solve the third research question. As a result, we found that Rater, Semester, and Rubric are three factors in this experiment which are related to the ratings. Besides, there are some interesting interactions between factors Rater, Rubric, and Artifact. Finally, we focused on further exploratory data analysis on summary statistics and barplots to investigate if there is any other interesting relationship between variable Sex and ratings in the fourth research question. The results shows that factor Sex has no significant effect in predicting ratings.

Based on our analyses, we presented to Dean's Office in Dietrich College at Carnegie Mellon University with following interesting findings: First, raters didn't all interpret the rubrics and the evidence in the artifacts in the same way. Hence students who received lower ratings from Rater 3 would get higher ratings from other two raters. Second, there is a difference in the ratings between Semester Spring 19 and Semester Fall 19. Specifically, considering all other factors are the same, we estimated that students from Spring section of Freshman Statistics would generally perform slightly worse than students from Fall section. As a note, there is no noticeable difference in work performance between females and males in Freshman Statistics.

All in all, we provided Dean's Office in Dietrich College at Carnegie Mellon University with following suggestions: First, there should be more standardized trainings on the raters to make the raters' ratings more similar to each other. Second, there should be more standardized assessments on courses to ensure students from both Fall and Spring section of Freshman Statistics have the same learning experiences. For example, possible standardized assessments to balance the difficulty levels for both Fall and Spring section of Freshman Statistics.

There are some strengths in our work. For example, we applied multiple techniques such as exploratory data analysis on summary statistics and barplots; intraclass correlation (ICC); percent exact agreement; and multiple mixed-effects models to present comprehensive analyses. Besides, we conducted model selection focused on automated backward elimination and ANOVA test based on AIC, BIC and likelihood ratio test (LRT), which allowed for an accurate investigation on the relationships between various factors in this experiment with ratings.

There are some limitations in our work. As mentioned before, there is a difference in the ratings between Semester Spring 19 and Semester Fall 19, while we didn't consider the possible causes of the difference into the final mixed-effects model. One cause could be students have taken some other preliminary courses in Semester Spring which could help to improve their work performance in Freshman Statistics in Semester Fall. The other cause could be there are much more students from statistics related majors take Freshman Statistics in Semester Fall than Semester Spring. Besides, there are only 91 artifacts were sampled in this experiment, which are not sufficient to test all possible random effects. Also, based on the interpretation for our final mixed-effect model in Section 4.3, we found that there could be other factors which had significant effects in predicting ratings without being considered in the experiment.

Overall, further improvements on our work including that repeat this experiment with a larger number of artifacts randomly sampled from a Fall and Spring section of Freshman Statistics; and conduct further analyses on other factors which might have significant effects in predicting ratings without being considered in the previous experiment.

## References

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 13, 2021 from https://canvas.cmu.edu/courses/25337/files/folder/Project02/.

# Technical Appendix

Olivia Wang

## Contents



## Appendix 1. Initial Data Import & Exploration

### Part A

#### Initial Look at the Data

```
## read the data in wide and tall formats...
ratings <- read.csv("ratings.csv",header=T)
tall <-  read.csv("tall.csv",header=T)
summary(ratings)
```

```
##        X            Rater       Sample          Overlap     Semester
##  Min.   :  1   Min.   :1   Min.   :  1.00   Min.   : 1   Length:117
##  1st Qu.: 30   1st Qu.:1   1st Qu.: 31.00   1st Qu.: 4   Class :character
##  Median : 59   Median :2   Median : 60.00   Median : 7   Mode  :character
##  Mean   : 59   Mean   :2   Mean   : 59.89   Mean   : 7
##  3rd Qu.: 88   3rd Qu.:3   3rd Qu.: 89.00   3rd Qu.:10
##  Max.   :117   Max.   :3   Max.   :118.00   Max.   :13
##                                             NA's   :78
##      Sex                RsrchQ        CritDes         InitEDA
##  Length:117         Min.   :1.00   Min.   :1.000   Min.   :1.000
##  Class :character   1st Qu.:2.00   1st Qu.:1.000   1st Qu.:2.000
##  Mode  :character   Median :2.00   Median :2.000   Median :2.000
##                     Mean   :2.35   Mean   :1.871   Mean   :2.436
##                     3rd Qu.:3.00   3rd Qu.:3.000   3rd Qu.:3.000
##                     Max.   :4.00   Max.   :4.000   Max.   :4.000
##                                    NA's   :1
##     SelMeth        InterpRes         VisOrg          TxtOrg
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :3.000   Median :2.000   Median :3.000
##  Mean   :2.068   Mean   :2.487   Mean   :2.414   Mean   :2.598
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :3.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
##                                  NA's   :1
##    Artifact            Repeated
##  Length:117         Min.   :0.0000
##  Class :character   1st Qu.:0.0000
##  Mode  :character   Median :0.0000
##                     Mean   :0.3333
##                     3rd Qu.:1.0000
##                     Max.   :1.0000
##
```

```
summary(tall)
```

```
##        X             Rater     Artifact            Repeated
##  Min.   :  1.0   Min.   :1   Length:819         Min.   :0.0000
##  1st Qu.:205.5   1st Qu.:1   Class :character   1st Qu.:0.0000
##  Median :410.0   Median :2   Mode  :character   Median :0.0000
##  Mean   :410.0   Mean   :2                      Mean   :0.3333
##  3rd Qu.:614.5   3rd Qu.:3                      3rd Qu.:1.0000
##  Max.   :819.0   Max.   :3                      Max.   :1.0000
##
##    Semester             Sex               Rubric             Rating
##  Length:819         Length:819         Length:819         Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:2.000
##  Mode  :character   Mode  :character   Mode  :character   Median :2.000
##                                                           Mean   :2.318
##                                                           3rd Qu.:3.000
##                                                           Max.   :4.000
##                                                           NA's   :2
```

```
## extract the reduced data set with the 13 artifacts that all 3 raters saw...
ratings.13 <- ratings[grep("O",ratings$Artifact),]
tall.13 <-  tall[grep("O",tall$Artifact),]
```

```
ratings$Rater = as.factor(ratings$Rater)
ratings %>%
  pivot_longer(
    cols = RsrchQ:TxtOrg, names_to = "rubric", values_to = "rating") %>%
    ggplot(aes(x = rating, fill = Rater)) +
    geom_histogram(bins = 8, position = "dodge") +
    facet_wrap(~ rubric) +
    theme(strip.background =element_rect(fill = "grey")) +
    theme(strip.text = element_text(colour = 'black')) +
    scale_fill_brewer(palette="Set1") +
    ylab('Count of Ratings') +
    xlab('Rating')
```

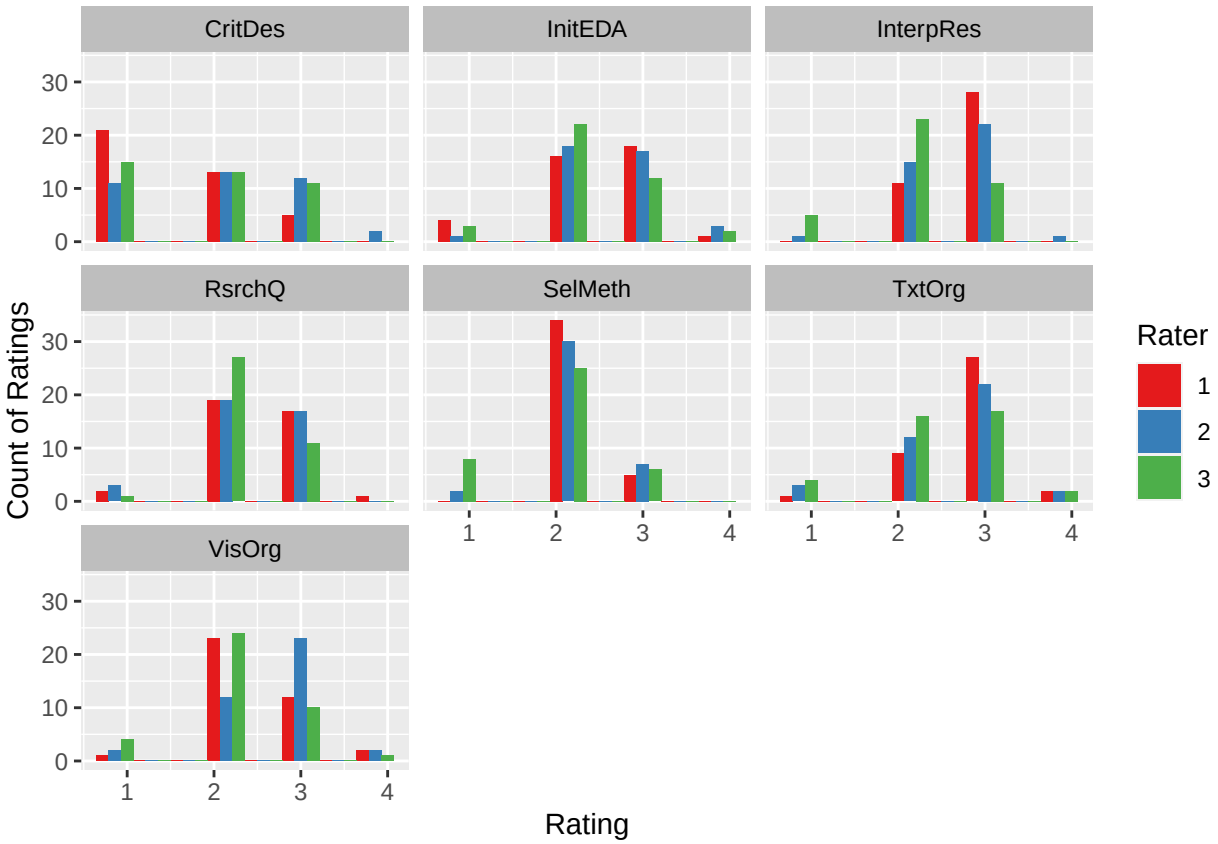



Figure 1: Barplots of ratings for each rubric given by each rater for whole set of 91 artifacts

```
ratings.13$Rater = as.factor(ratings.13$Rater)
ratings.13 %>%
  pivot_longer(
    cols = RsrchQ:TxtOrg, names_to = "rubric", values_to = "rating") %>%
```

```
    ggplot(aes(x = rating, fill = Rater)) +
    geom_histogram(bins = 8, position = "dodge") +
    facet_wrap(~ rubric) +
    theme(strip.background =element_rect(fill = "grey")) +
    theme(strip.text = element_text(colour = 'black')) +
    scale_fill_brewer(palette="Set1") +
    ylab('Count of Ratings') +
    xlab('Rating')
```

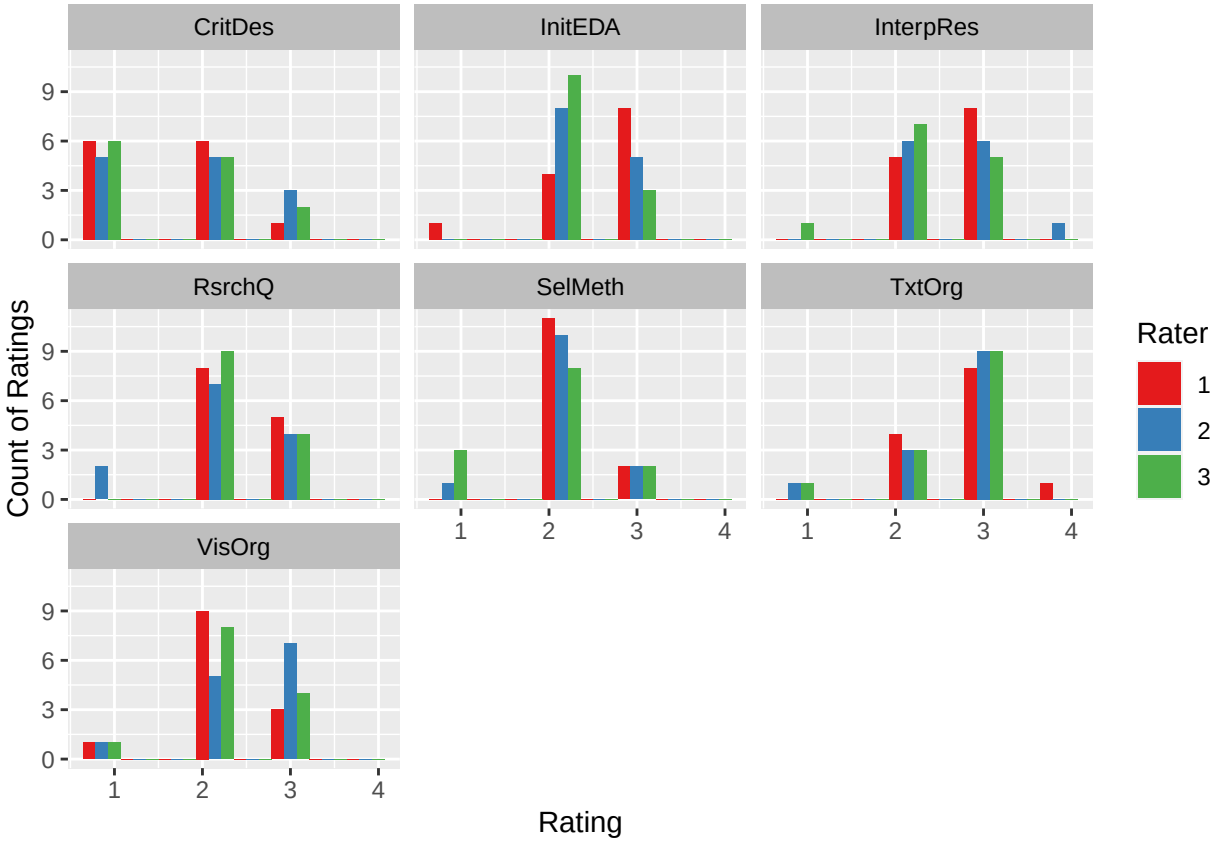



Figure 2: Barplots of ratings for each rubric given by each rater for 13 artifacts seen by all three raters

## Part B

### ratings.csv

We can check to see how many unique values each variable has (this is especially relevant for the categorical variables). From the table below, we found there is a strange thing that Sex variable (Sex or gender of student who created the artifact) has three unique values.

```
apply(ratings,2,function(x) {length(unique(x))}) %>%
  kbl(booktabs=T,col.names="unique values",caption=" ") %>%
  kable_classic(full_width=F)
```

Table 1:

| | unique values |
|---|---|
| X | 117 |
| Rater | 3 |
| Sample | 117 |
| Overlap | 14 |
| Semester | 2 |
| Sex | 3 |
| RsrchQ | 4 |
| CritDes | 5 |
| InitEDA | 4 |
| SelMeth | 3 |
| InterpRes | 4 |
| VisOrg | 5 |
| TxtOrg | 4 |
| Artifact | 91 |
| Repeated | 2 |

We can check for NA's directly:

```
apply(ratings,2,function(x) any(is.na(x)))
```

```
##         X     Rater    Sample   Overlap  Semester       Sex    RsrchQ   CritDes
##     FALSE     FALSE     FALSE      TRUE     FALSE     FALSE     FALSE      TRUE
##   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg  Artifact  Repeated
##     FALSE     FALSE     FALSE      TRUE     FALSE     FALSE     FALSE
```

```
## find the row number (X) of each missing value of three variables (CritDes, VisOrg, Sex)
which(is.na(ratings$CritDes))
```

```
## [1] 44
```

```
which(is.na(ratings$VisOrg))
```

```
## [1] 99
```

```
ratings[ratings$Sex=="--",]
```

```
##   X Rater Sample Overlap Semester Sex RsrchQ CritDes InitEDA SelMeth InterpRes
## 5 5     3      5      NA     Fall  --      3       3       3       3         3
##   VisOrg TxtOrg Artifact Repeated
## 5      3      3        5        0
```

```
ratings[c(44,99),] ## just to check that these are the rows with missing ratings...
```

```
##     X Rater Sample Overlap Semester Sex RsrchQ CritDes InitEDA SelMeth
```

```
## 44 44     2     45      NA   Spring   F      2      NA       2       2
## 99 99     1    100      NA     Fall   F      2       3       2       3
##    InterpRes VisOrg TxtOrg Artifact Repeated
## 44         2      2      3       45        0
## 99         3     NA      2      100        0
```

```
ratings.nonmissing <- ratings[-c(44,99),]  ## now delete them...
ratings.nonmissing[ratings.nonmissing$Sex=="--",]  ## check which rows will be eliminated
```

```
##   X Rater Sample Overlap Semester Sex RsrchQ CritDes InitEDA SelMeth InterpRes
## 5 5     3      5      NA     Fall  --      3       3       3       3         3
##   VisOrg TxtOrg Artifact Repeated
## 5      3      3        5        0
```

```
ratings.nonmissing <- ratings.nonmissing[ratings.nonmissing$Sex!="--",] ## eliminate them
```

- There do appear to be any missing values in the data! As we can see there are two variables (CritDes, and VisOrg) have NA's (In general we might also check for old-fashioned missing value codes like "9", "99", "98", etc., but there's no evidence of that in Table 5 (look at the Min and Max values - no "9's", "99's", etc.))
- Note that none of the missing values occur in the smaller 13-rubric data set. So we don't have to worry about missing data at all in analyses that just involve this smaller data set.
- Specifically the row number (X) of one missing value for CritDes is 44, the row number (X) of one missing value for VisOrg is 99, there is total of 1 missing values for Sex. I am going to eliminate by hand the two observations with missing data. Besides, I can't think of a good justification for imputing the "Sex" of the student who didn't report this to either M or F. So I will eliminate that person from the data set also. Hence I got a new data set ratings.nonmissing.

## Part C

**tall.csv**

We can check for NA's directly:

```
apply(tall,2,function(x) any(is.na(x)))
```

```
##        X    Rater Artifact Repeated Semester      Sex   Rubric   Rating
##    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE     TRUE
```

```
## find the row number (X) of each missing value of variables (Rating, Sex)
which(is.na(tall$Rating))
```

```
## [1] 161 684
```

```
## note that in the "ratings" data frame, the missing "Sex"
## value is "--" while in the "tall" data frame it is ""
## (a string of length 0).

## make the "tall" be consistent with the "ratings" coding.
tall$Sex[nchar(tall$Sex)==0] <- "--"
tall[apply(tall,1,function(x){any(is.na(x))}),]
```

```
##       X Rater Artifact Repeated Semester Sex  Rubric Rating
## 161 161     2       45        0      S19   F CritDes     NA
## 684 684     1      100        0      F19   F  VisOrg     NA
```

```
Rubric.names <- sort(unique(tall$Rubric))
tall[c(161,684),] ## just to check that these are the rows with missing ratings...
```

```
##       X Rater Artifact Repeated Semester Sex  Rubric Rating
## 161 161     2       45        0      S19   F CritDes     NA
## 684 684     1      100        0      F19   F  VisOrg     NA
```

```
tall.nonmissing <- tall[-c(161,684),]  ## now delete them...
tall.nonmissing[tall.nonmissing$Sex=="--",]  ## check which rows will be eliminated
```

```
##       X Rater Artifact Repeated Semester Sex    Rubric Rating
## 5     5     3        5        0      F19  --    RsrchQ      3
## 122 122     3        5        0      F19  --   CritDes      3
## 239 239     3        5        0      F19  --   InitEDA      3
## 356 356     3        5        0      F19  --   SelMeth      3
## 473 473     3        5        0      F19  -- InterpRes      3
## 590 590     3        5        0      F19  --    VisOrg      3
## 707 707     3        5        0      F19  --    TxtOrg      3
```

```
tall.nonmissing <- tall.nonmissing[tall.nonmissing$Sex!="--",] ## eliminate them
```

- There do appear to be any missing values in the data! As we can see there are two variables (Sex and Rating) have NA's. (In general we might also check for old-fashioned missing value codes like "9", "99", "98", etc., but there's no evidence of that in Table 5 (look at the Min and Max values - no "9's", "99's", etc.))
- Second, in any modeling that we do, the "Rating" is the outcome variable, so R will just drop the two observations with missing Rating values. This will mean that the "full" data sets may be different for models that involve different rubrics: For models involving five of the rubrics we will get all the data from all the raters, but for models involving CritDes we would be missing a rating from Rater 2, and for models involving VisOrg we would be missing a rating from Rater 1. We need to be vigilant about when these differences actually occur, since they could undermine some model comparisons (different data sets).
- Note that none of the missing values occur in the smaller 13-rubric data set. So we don't have to worry about missing data at all in analyses that just involve this smaller data set.
- Specifically the row number (X) of two missing values for Rating are 161 and 684, there are total of 7 missing values for Sex. We will also have to be careful of the missing "Sex" value (currently coded as "–". Considering the Research Question #3 in Appendix 4. Now the missing ratings become important. We want to use the same data set for every model fit and model comparison. So I am going to eliminate by hand the two observations with missing data, and only do fitting and comparison on this "slightly" reduced data set. Besides, I can't think of a good justification for imputing the "Sex" of the student who didn't report this to either M or F, and leaving it as "–" makes the models harder to interpret. So I will eliminate that person from the data set also. Hence I got a new data set tall.nonmissing.

# Appendix 2. Research Question #1

## Part A

**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?**

First of all, let us use ratings.csv to do the analysis. So let's make a table with the usual one-dimensional summary statistics based on the subset of the ratings.csv data set with only seven rubrics of all artifacts named as ratings_rubrics.

```
ratings_rubrics <- ratings.nonmissing[,c(7:13)]
apply(ratings_rubrics,2,function(x) c(summary(x),SD=sd(x))) %>% as.data.frame %>% t() %>%
  round(digits=2) %>% kbl(booktabs=T,caption="Summary Statistics for ratings of each rubric based on wh
```

Table 2: Summary Statistics for ratings of each rubric based on whole set of 91 artifacts

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | SD |
|---|---|---|---|---|---|---|---|
| RsrchQ | 1 | 2 | 2 | 2.35 | 3 | 4 | 0.59 |
| CritDes | 1 | 1 | 2 | 1.85 | 2 | 4 | 0.83 |
| InitEDA | 1 | 2 | 2 | 2.44 | 3 | 4 | 0.70 |
| SelMeth | 1 | 2 | 2 | 2.05 | 2 | 3 | 0.48 |
| InterpRes | 1 | 2 | 3 | 2.48 | 3 | 4 | 0.61 |
| VisOrg | 1 | 2 | 2 | 2.41 | 3 | 4 | 0.68 |
| TxtOrg | 1 | 2 | 3 | 2.60 | 3 | 4 | 0.70 |

Secondly, let's consider tall.csv So let's make a table based on the subset of the tall.csv data set with only Rater and Rating of all artifacts named as tall_rubrics.

```
tall_rubrics <- tall.nonmissing[,c(2,8)]
```

Then let's make 3 subsets of ratings given by three raters named as rater1, rater2 and rater3.

```
rater1 = tall_rubrics[which(tall_rubrics$Rater==1),]
rater2 = tall_rubrics[which(tall_rubrics$Rater==2),]
rater3 = tall_rubrics[which(tall_rubrics$Rater==3),]
```

Next, let's make a table with the usual one-dimensional summary statistics based on the subset of ratings given by three raters named as rater1, rater2 and rater3.

```
rater1$Rating <- as.numeric(rater1$Rating)
rater2$Rating <- as.numeric(rater2$Rating)
rater3$Rating <- as.numeric(rater3$Rating)

rater_rating <-
  cbind(c(N=length(rater1$Rating),summary(rater1$Rating),SD=sd(rater1$Rating)),
        c(N=length(rater2$Rating),summary(rater2$Rating),SD=sd(rater2$Rating)),
        c(N=length(rater3$Rating),summary(rater3$Rating),SD=sd(rater3$Rating))) %>%
  as.data.frame()
```

```
colnames(rater_rating) = c('Rater1', 'Rater2', 'Rater3')

rater_rating %>% t() %>%
  round(digits=2) %>% kbl(booktabs=T,caption="Summary Statistics for ratings given by each rater based
```

Table 3: Summary Statistics for ratings given by each rater based on whole set of 91 artifacts

| | N | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | SD |
|---|---|---|---|---|---|---|---|---|
| Rater1 | 272 | 1 | 2 | 2 | 2.35 | 3 | 4 | 0.70 |
| Rater2 | 272 | 1 | 2 | 2 | 2.43 | 3 | 4 | 0.70 |
| Rater3 | 266 | 1 | 2 | 2 | 2.15 | 3 | 4 | 0.69 |

Here are some ideas to compare distributions across Rubrics.

```
## take care that all ratings run from 1 to 4,
## whether or not rater used all categories...
tall$Rating <- factor(tall$Rating,levels=1:4)
for (i in unique(tall$Rubric)) {
  ratings[,i] <- factor(ratings[,i],levels=1:4)
}
```

```
## Barplots for the reduced data set
ggplot(tall.13, aes(y = Rating)) +
  geom_histogram(position = 'dodge', binwidth = 1,color = "black", fill = "lightgrey") +
  xlab('Count of Ratings') +
  scale_x_continuous(limits = c(0, 40)) +
  facet_wrap(~as.factor(Rubric)) +
  stat_bin(binwidth=1, geom="text", aes(label=..count..), vjust=-0.5) +
  coord_flip() +
  theme_light() +
  theme(strip.background =element_rect(fill = "grey")) +
  theme(strip.text = element_text(colour ='black'))
```

```
## Barplots for full data set using tall.nonmissing data set
ggplot(tall.nonmissing, aes(y = Rating)) +
  geom_histogram(position = 'dodge', binwidth = 1, color = "black", fill = "lightgrey") +
  xlab('Count of Ratings') +
  scale_x_continuous(limits = c(0, 110)) +
  facet_wrap(~as.factor(Rubric)) +
  stat_bin(binwidth=1, geom="text", aes(label=..count..), vjust=-0.5) +
  coord_flip() +
  theme_light() +
  theme(strip.background =element_rect(fill = "grey"))+
  theme(strip.text = element_text(colour = 'black'))
```

**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?**

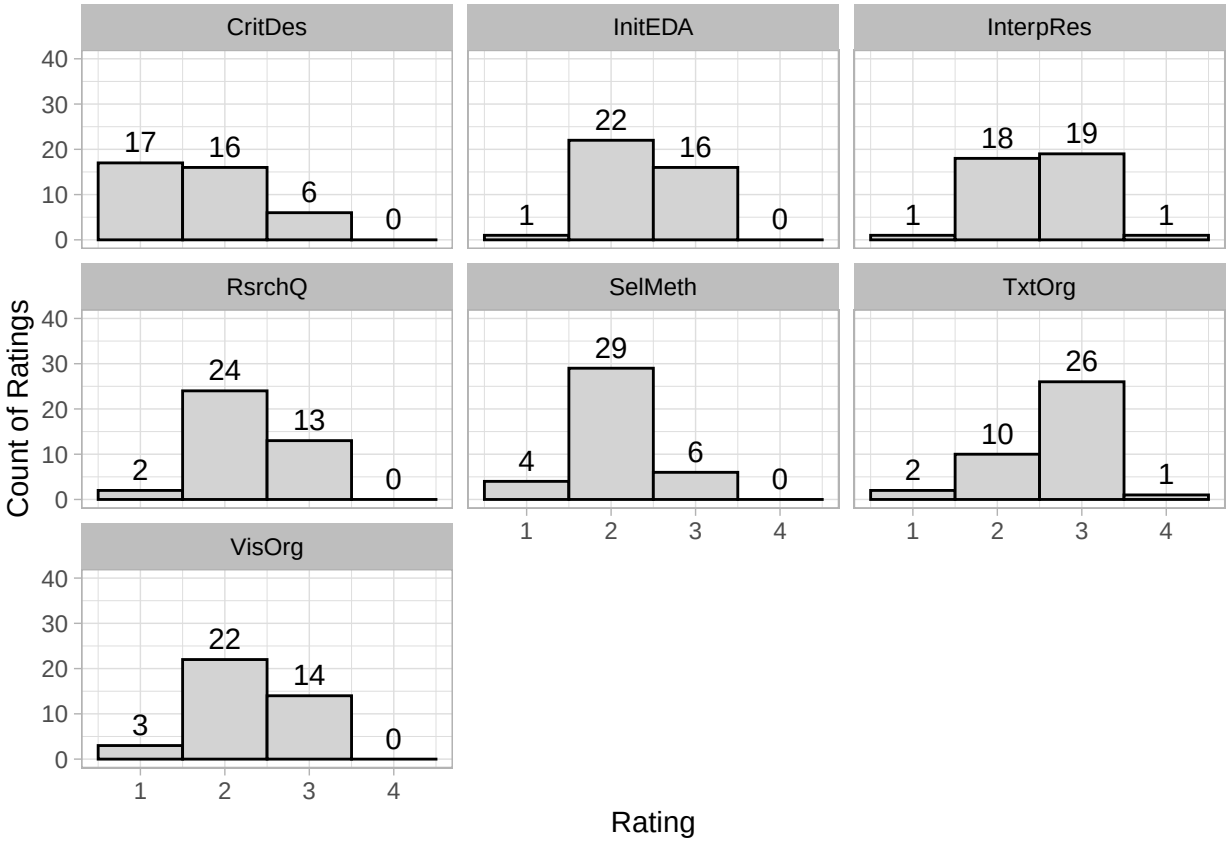



Figure 3: Barplots of ratings for each rubric based on 13 artifacts seen by all three raters

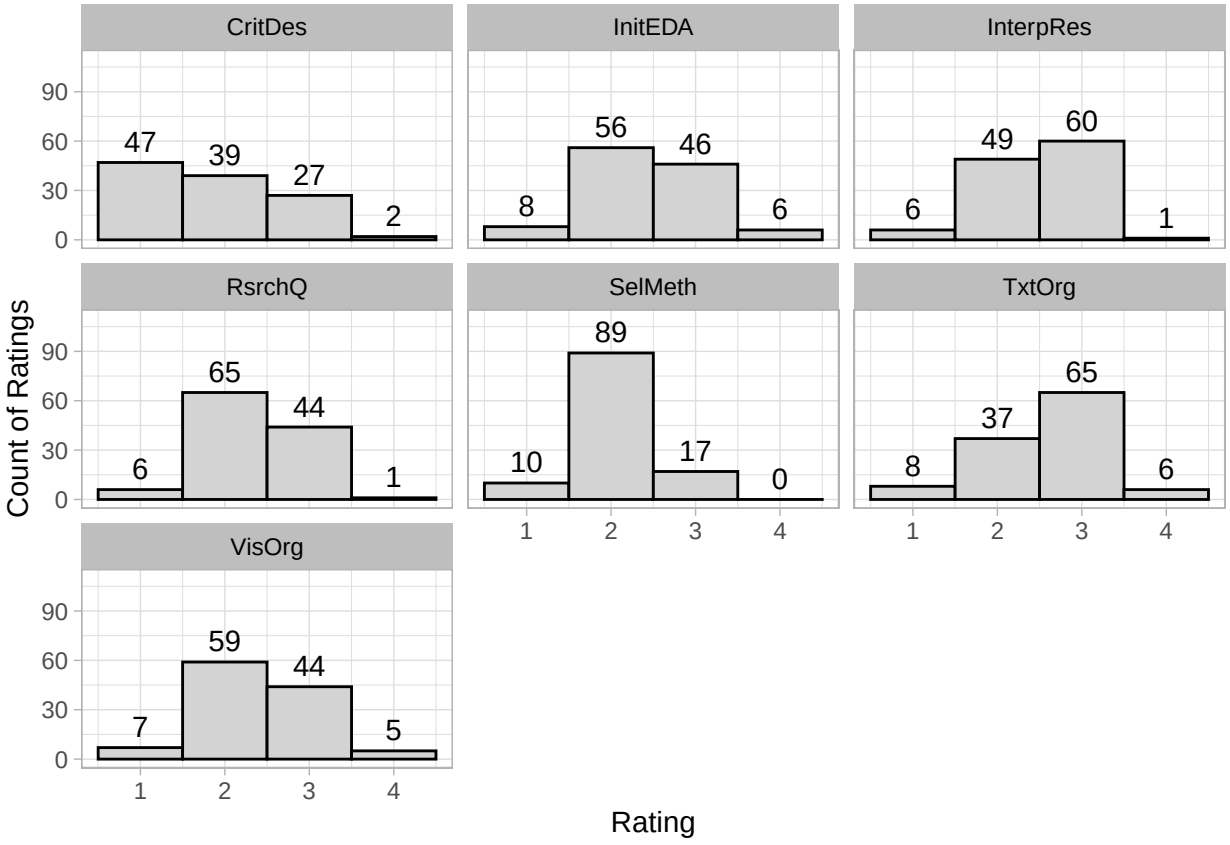



Figure 4: Barplots of ratings for each rubric based on whole set of 91 artifacts

- In order to answer this question, we made an assumption first, among all rubrics, we defined low rating as artifact was rated less than or equal to 2; high rating as artifact was rated above 2.
- Based on the numerical summaries table and the barplots of ratings for seven rubrics for full data set, we can get the followings:
- Rubric SelMeth (Rating on Select Method(s)) tends to get especially low ratings.
- If we take a look at the distribution of ratings on rubric SelMeth, the total number of low ratings on rubric SelMeth is 99 which is the highest one among all the rubrics for rating Freshman Statistics projects.
- The Max Value of ratings on rubric SelMeth is 3 which is the lowest Max Value of all rubrics; besides, both of the Median Value and 3rd Quartile Value of ratings on rubric SelMeth are 2 which indicates that nearly 75% of the artifacts were rated less than or equal to score 2. Considering that the Mean Value of ratings on rubric SelMeth is 2.05 which is close to the Median Value and 3rd Quartile Value of ratings on rubric SelMeth, and the Standard Deviation of ratings on rubric SelMeth is 0.48 which is the lowest one among all rubrics, which means most of the artifacts were rated between 2 and 3 on rubric SelMeth.
- Although, from the numerical summaries table, the Mean Value of CritDes (Rating on Critique Design) is 1.85 which is the lowest Mean Value among all rubrics, and also from the the distribution of rubric CritDes, the number of the artifacts rated at grade 1 is the highest among all rubrics. But considering that we assumed low rating as artifact was rated at grade less than or equal to 2, the number of artifacts were rated as low ratings of rubric CritDes are nearly 86 which is less than rubric SelMeth, so we still continue with rubric SelMeth tends to get especially low ratings.
- Rubric TxtOrg (Rating on Text Organization) tends to get especially high ratings.
- The Max Value of ratings on rubric TxtOrg is 4 which is the highest one of all rubrics; besides, the Mean Value of ratings on rubric TxtOrg is 2.60 which is the highest Mean Value among all the rubrics. Also both of the Median Value and 3rd Quartile Value of ratings on rubric TxtOrg are 3, all of both are also the highest ones among all the rubrics.
- If we take a look at the distribution of ratings on rubric TxtOrg, the number of artifacts rated at grade 1 and 4 are very close. the total number of high ratings for rubric TxtOrg is 71 which is the highest one among all the rubrics for rating Freshman Statistics projects.

## Part B

**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?**

And here are some idea to compare distributions across Raters.

```
## Needed to make the title of each facet more human-readable...
rater.name <- function(x) { paste("Rater",x) }

## Barplots for reduced data...
tall.13$Rating = as.numeric(tall.13$Rating)
ggplot(tall.13, aes(y = Rating)) +
  geom_histogram(position = 'dodge', binwidth = 1, color = "black", fill = "lightgrey") +
  xlab('Count of Ratings') +
  scale_x_continuous(limits = c(0, 60)) +
  facet_wrap( ~ Rater, labeller=labeller(Rater=rater.name)) +
  stat_bin(binwidth=1, geom="text", aes(label=..count..), vjust=-0.5) +
```

```
  coord_flip() +
  theme_light() +
  theme(strip.background =element_rect(fill = "grey")) +
  theme(strip.text = element_text(colour = 'black'))
```

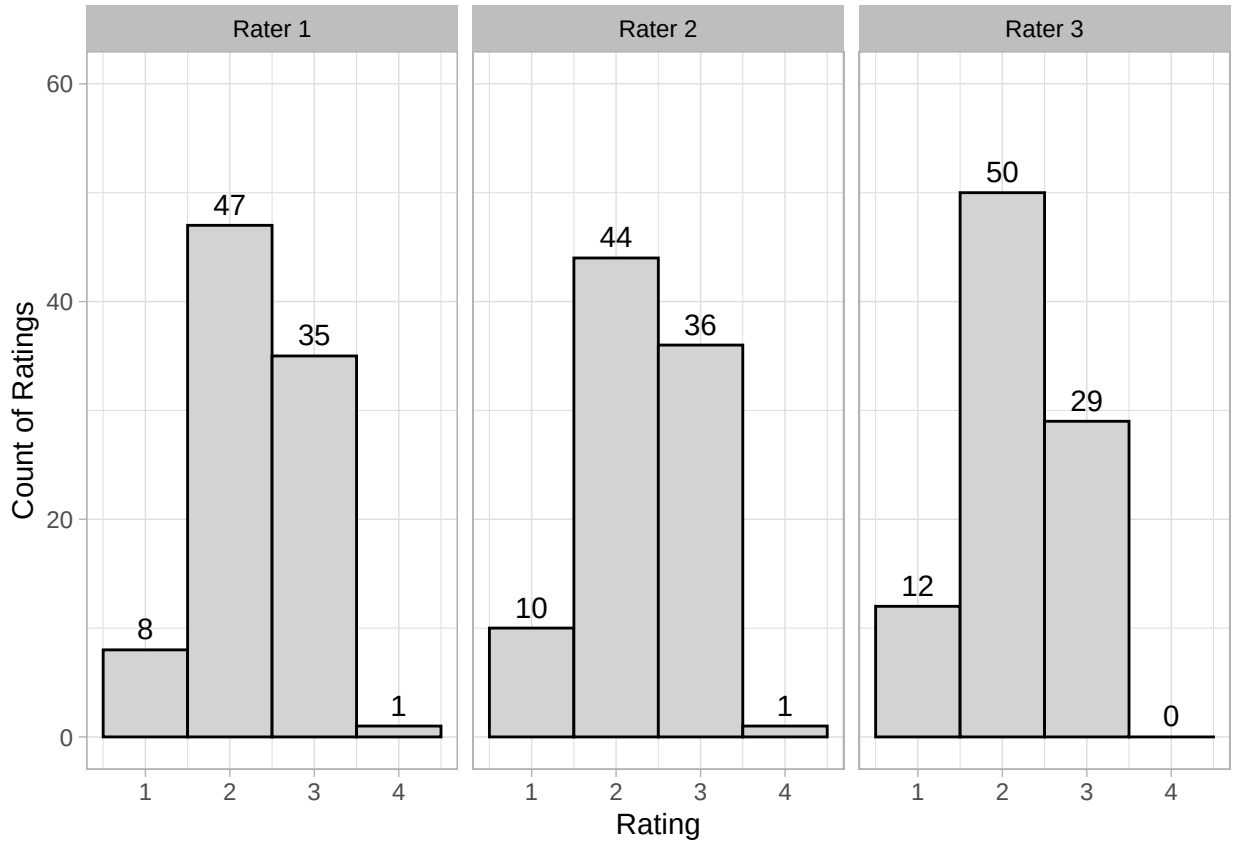



Figure 5: Barplots of ratings given by each rater based on 13 artifacts seen by all three raters

```
## Barplots for full data...
ggplot(tall.nonmissing, aes(y = Rating)) +
  geom_histogram(position = 'dodge', binwidth = 1, color = "black", fill = "lightgrey") +
  xlab('Count of Ratings') +
  scale_x_continuous(limits = c(0, 160)) +
  facet_wrap( ~ Rater, labeller=labeller(Rater=rater.name)) +
  stat_bin(binwidth=1, geom="text", aes(label=..count..), vjust=-0.5) +
  coord_flip() +
  theme_light() +
  theme(strip.background =element_rect(fill = "grey"))+
  theme(strip.text = element_text(colour = 'black'))
```

**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?**

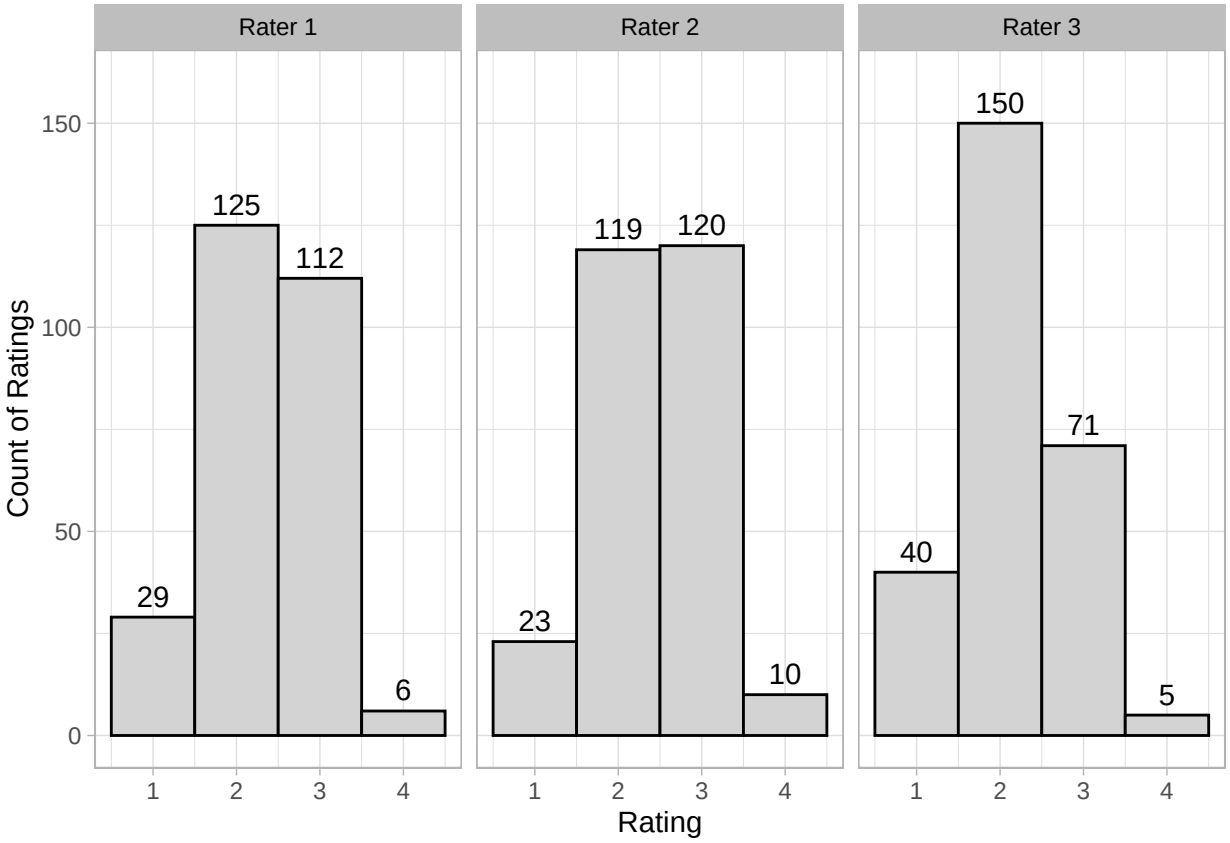



Figure 6: Barplots of ratings given by each rater based on whole set of 91 artifacts

- In order to answer this question, we made an assumption first, among all rubrics, we defined low rating as artifact was rated less than or equal to 2; high rating as artifact was rated above 2.
- Based on the numerical summaries table and the barplots of ratings given by each rater for full data set, we can get the followings:
- From the numerical summaries table, all of the values (Max, Min, Median, Mean, 1st Quartile, 3rd Quartile, Standard Deviation, etc.) of ratings given by three raters are very similar.
- Rater3 tends to give especially low ratings.
- If we take a look at the distribution of ratings given by rater3, we can see the total number of low ratings given by rater3 is 190 which is the highest one among all the ratings given by each rater. Besides from table rater1 (272 objects), rater2 (272 objects) and rater3 (266 objects), all of the raters have rated similar number of artifacts, however, there are nearly 150 artifacts rated at grade 2 from rater3, this number is significantly higher than number of artifacts rated at grade 2 by other two raters, also the number of artifacts rated at grade 1 by rater3 are also higher than number of artifacts rated at grade 1 by rater1 and rater2.
- Rater2 tends to give higher ratings.
- If we take a look at the distribution of ratings given by rater2, we can see the total number of high ratings given by rater2 is 130 which is the highest one among all the ratings given by each rater. Besides from table rater1 (272 objects), rater2 (272 objects) and rater3 (266 objects), all of the raters have rated similar number of artifacts, also we can see the distribution of ratings given by rater2 and rater1 are very similar, both of them are tend to give high ratings. However, we can clearly see the number of artifacts rated at both grade 3 and 4 by rater2 are both higher than by rater1.
- However, there is no significant evidence from distributions of ratings given by rater1 and rater2 that we can conclude which rater tends to give especially high ratings. Because both of distributions of ratings given by rater1 and rater2 on each rubric are very similar.

## Part C

### Compare and determine whether these thirteen artifacts are representative of the whole set of 91 artifacts?

We want to compare and determine whether these thirteen artifacts are representative of the whole set of 91 artifacts.

Let's make a table with the usual one-dimensional summary statistics based on the subset of the ratings.13 data with only seven rubrics of all artifacts named as ratings13_rubrics.

```
ratings13_rubrics <- ratings.13[,c(7:13)]
ratings13_rubrics$RsrchQ = as.numeric(ratings13_rubrics$RsrchQ)
ratings13_rubrics$CritDes = as.numeric(ratings13_rubrics$CritDes)
ratings13_rubrics$InitEDA = as.numeric(ratings13_rubrics$InitEDA)
ratings13_rubrics$SelMeth = as.numeric(ratings13_rubrics$SelMeth)
ratings13_rubrics$InterpRes = as.numeric(ratings13_rubrics$InterpRes)
ratings13_rubrics$VisOrg = as.numeric(ratings13_rubrics$VisOrg)
ratings13_rubrics$TxtOrg = as.numeric(ratings13_rubrics$TxtOrg)
apply(ratings13_rubrics,2,function(x) c(summary(x),SD=sd(x))) %>% as.data.frame %>% t() %>%
  round(digits=2) %>% kbl(booktabs=T,caption="Summary Statistics for ratings of each rubric for just th
```

Table 4: Summary Statistics for ratings of each rubric for just the 13 artifacts seen by all three raters

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | SD |
|---|---|---|---|---|---|---|---|
| RsrchQ | 1 | 2 | 2 | 2.28 | 3 | 3 | 0.56 |
| CritDes | 1 | 1 | 2 | 1.72 | 2 | 3 | 0.72 |
| InitEDA | 1 | 2 | 2 | 2.38 | 3 | 3 | 0.54 |
| SelMeth | 1 | 2 | 2 | 2.05 | 2 | 3 | 0.51 |
| InterpRes | 1 | 2 | 3 | 2.51 | 3 | 4 | 0.60 |
| VisOrg | 1 | 2 | 2 | 2.28 | 3 | 3 | 0.60 |
| TxtOrg | 1 | 2 | 3 | 2.67 | 3 | 4 | 0.62 |

**Compare and determine whether these thirteen artifacts are representative of the whole set of 91 artifacts?**

- Yes. For the following reasons:
- Based on the numerical summaries table and the barplots of ratings from Appendix 2, Part A & B for the subset of the data for just the 13 artifacts seen by all three raters and full data set, we can get the followings:
- Compared one-dimensional summary statistics based on the subset of the data for just the 13 artifacts seen by all three raters with full data set from Appendix 2, Part A & B, we can see both of the Median Values and Min Values of ratings on all rubrics are the same for both ratings data and ratings.13 data; Standard Deviations, 1st Quartiles, 3rd Quartiles of ratings of all rubrics are very similar. Except for most of the Max Values of ratings on all rubrics are 4 in ratings data, however most of the Max Values of all ratings is 3 in ratings.13 data.
- Compared univariate distributions of ratings based on the subset of the data for just the 13 artifacts seen by all three raters with full data set from Appendix 2, Part A & B, we can see the univariate distributions of ratings have very similar trends, except for ratings on some rubrics have different number of unique values for each of situation mentioned in Part A & B.
- Hence, we approved that these thirteen artifacts are representative of the whole set of 91 artifacts.

# Appendix 3. Research Question #2

```
## useful preliminaries
Rubric.names <- sort(unique(tall$Rubric))

## First we examine the 13 "common" artifacts that all 3 raters saw...
ICC.vec <- NULL
for (i in Rubric.names) {
  tmp <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=tall.13[tall.13$Rubric==i,])
  sig2 <- summary(tmp)$sigma^2
  tau2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
  ICC <- tau2 / (tau2 + sig2)
  ICC.vec <- c(ICC.vec,ICC)
}
names(ICC.vec) <- Rubric.names
agreement.results <- cbind(ICC.common=ICC.vec,"       a12"=0,a23=0,a13=0)
```

```
agreement.tables <- as.list(rep(NA,7))
names(agreement.tables) <- Rubric.names
for (i in Rubric.names) {
  r12 <- data.frame(r1=factor(ratings.13[ratings.13$Rater==1,i],levels=1:4),
                    r2=factor(ratings.13[ratings.13$Rater==2,i],levels=1:4),
                    a1=ratings.13[ratings.13$Rater==1,"Artifact"],
                    a2=ratings.13[ratings.13$Rater==2,"Artifact"])
  if(any(r12[,3]!=r12[,4])) { stop(paste("Rater 1-2 Artifact mismatch on rubric",i)) }
  a12 <- mean(r12[,1]==r12[,2])
  r12 <- table(r12[,1:2])  ## print this to see how much agreement there is among raters 1-2
  r23 <- data.frame(r2=factor(ratings.13[ratings.13$Rater==2,i],levels=1:4),
                    r3=factor(ratings.13[ratings.13$Rater==3,i],levels=1:4),
                    a2=ratings.13[ratings.13$Rater==2,"Artifact"],
                    a3=ratings.13[ratings.13$Rater==3,"Artifact"])
  if(any(r23[,3]!=r23[,4])) { stop(paste("Rater 2-3 Artifact mismatch on rubric",i)) }
  a23 <- mean(r23[,1]==r23[,2])
  r23 <- table(r23[,1:2])  ## print this to see how much agreement there is among raters 2-3
  r13 <- data.frame(r1=factor(ratings.13[ratings.13$Rater==1,i],levels=1:4),
                    r3=factor(ratings.13[ratings.13$Rater==3,i],levels=1:4),
                    a1=ratings.13[ratings.13$Rater==1,"Artifact"],
                    a3=ratings.13[ratings.13$Rater==3,"Artifact"])
  if(any(r13[,3]!=r13[,4])) { stop(paste("Rater 1-3 Artifact mismatch on rubric",i)) }
  a13 <- mean(r13[,1]==r13[,2])
  r13 <- table(r13[,1:2])  ## print this to see how much agreement there is among raters 1-3
  agreement.results[i,2:4] <- c(a12,a23,a13)
  agreement.tables[[i]] <- list(r12,r23,r13)
}
round(agreement.results,2)
```

```
##           ICC.common        a12  a23  a13
## CritDes         0.57       0.54 0.69 0.62
## InitEDA         0.49       0.69 0.85 0.54
## InterpRes       0.23       0.62 0.62 0.54
## RsrchQ          0.19       0.38 0.54 0.77
## SelMeth         0.52       0.92 0.69 0.62
## TxtOrg          0.14       0.69 0.54 0.62
## VisOrg          0.59       0.54 0.77 0.77
```

```
if (F) { print(agreement.tables) }

## Now add in ICC's calculated from all the data...
ICC.vec <- NULL
for (i in Rubric.names) {

  tmp <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=tall.nonmissing[tall.nonmissing$Rubric==i,])
  sig2 <- summary(tmp)$sigma^2
  tau2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
  ICC <- tau2 / (tau2 + sig2)
  ICC.vec <- c(ICC.vec,ICC)
}
names(ICC.vec) <- Rubric.names
agreement.results <- cbind(ICC.alldata=ICC.vec,agreement.results)
round(agreement.results,2)
```

```
##           ICC.alldata ICC.common        a12  a23  a13
## CritDes          0.67       0.57       0.54 0.69 0.62
## InitEDA          0.69       0.49       0.69 0.85 0.54
## InterpRes        0.22       0.23       0.62 0.62 0.54
## RsrchQ           0.21       0.19       0.38 0.54 0.77
## SelMeth          0.46       0.52       0.92 0.69 0.62
## TxtOrg           0.19       0.14       0.69 0.54 0.62
## VisOrg           0.66       0.59       0.54 0.77 0.77
```

**For each rubric, do the raters generally agree on their scores?**

- Based on the rules of thumb for interpreting ICC, we would conclude that an ICC of 0.782 indicates that the rubrics can be rated with "good" reliability by different raters.
- The output of the values of seven ICC's, we usually define ICC between 0.50 and 0.75 as moderate reliability. In this case, we can conclude that raters generally are consistent with one another in how they rate on rubrics CritDes, SelMeth, InitEDA and VisOrg. However, we usually define ICC below 0.50 as poor reliability. In this case, we can see that raters generally are not consistent with one another in how they rate on rubrics RsrchQ, InterpRes and TxtOrg.

**Is there one rater who disagrees with the others? Or do they all disagree?**

- Obviously, percent exact agreements are close between each pair of raters on rubrics CritDes, InterpRes, VisOrg and TxtOrg, which means that raters are generally in agreement with their ratings on these rubrics. We mainly focused on rubrics RsrchQ, InitEDA, and SelMeth, since on which there are comparable differences among percent exact agreements between each pair of raters.
- RsrchQ: Rater 2 might be contributing to disagreement among raters' ratings on rubric RsrchQ. Because Rater 1 and Rater 2 had very low agreement with their ratings on rubric RsrchQ; Rater 1 and Rater 3 had higher agreement with their ratings on rubric Research RsrchQ.
- InitEDA: Rater 1 might be contributing to disagreement among raters' ratings on rubric InitEDA. Because Rater 1 and Rater 2 or Rater 3 had moderate agreement with their ratings on rubric InitEDA; Rater 2 and Rater 3 had very high agreement with their ratings on rubric InitEDA.
- SelMeth: Rater 3 might be contributing to disagreement among raters' ratings on rubric SelMeth. Because Rater 3 and Rater 1or Rater 2 had moderate agreement with their ratings on rubric SelMeth; Rater 1 and Rater 2 had extremely high agreement with their ratings on rubric SelMeth.

**You can re-do the ICC calculations on the full data set (but not the percent exact agreement calculations—why not?). Do the seven ICC's for the full data set agree with the seven ICC's for the subset corresponding to the 13 artifacts that all three raters saw?**

- We do not need to re-do the percent exact agreement calculations because percent exact agreement requires that both raters rated the same artifact, this metric was only calculated for the subset of the data rather than the full data set.
- The seven ICC's for the full data set agree with the seven ICC's for the subset corresponding to the 13 artifacts that all three raters saw, because all of the seven ICC's for the full data set are close to the seven ICC's for the subset, the maximum difference is around 0.1.

# Appendix 4. Research Question #3

## Part A

**Adding fixed effects to the seven rubric-specific models using just the data from the 13 common artifacts that all three raters saw.**

First, we try to add fixed effects to our seven rubric-specific models... In principle it will matter whether we use only the data reduced to the 13 common artifacts, or the full data set.

I will start with the reduced data - tall.13 (so of course I can't check "repeated" on this reduced data—since all the repeated = 1 on this reduced data).

```
library(LMERConvenienceFunctions)
library(LMERConvenienceFunctions, warn.conflicts=F, quietly=T)
library(RLRsim)

## So my starting model for experimenting was
tmp <- lmer(as.numeric(Rating) ~ -1 + as.factor(Rater) +
              Semester + Sex + (1|Artifact),
            data=tall.13[tall.13$Rubric=="RsrchQ",],REML=FALSE)

## So a typical function call would be
tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)

## Anyway, backwards elimination yields a model
## with raters only:
formula(tmp.back_elim)

## The estimates for raters don't look that different from each other,
## so we can test to see if they are different by comparing with the
## intercept-only model
tmp.int_only <- update(tmp.back_elim, . ~ . + 1 - as.factor(Rater))
anova(tmp.int_only,tmp.back_elim)

## Again the models are nested so I really only need to look at the p-value
## from the likelihood ratio chi-squared test.
anova(tmp.int_only,tmp.back_elim)$"Pr(>Chisq)"[2]

## it looks like the intercept-only model is adequate here (the p-value
## is much greater than 0.05 or any other common significance level).
## Note: since no main effects were retained, there's really no reason to
## check for interactions.

Rubric.names <- sort(unique(tall$Rubric))
model.formula.13 <- as.list(rep(NA,7))
names(model.formula.13) <- Rubric.names
for (i in Rubric.names) {
  ## fit each base model
  rubric.data <- tall.13[tall.13$Rubric==i,]
  tmp <- lmer(as.numeric(Rating) ~ -1 + as.factor(Rater) +
              Semester + Sex + (1|Artifact),
            data=rubric.data,REML=FALSE)
  ## do backwards elimination
```

```
  tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)
  ## check to see if the raters are significantly different from one another
  tmp.single_intercept <- update(tmp.back_elim, . ~ . + 1 - as.factor(Rater))
  pval <- anova(tmp.single_intercept,tmp.back_elim)$"Pr(>Chisq)"[2]
  ## choose the best model
  if (pval<=0.05) {
    tmp_final <- tmp.back_elim
  } else {
    tmp_final <- tmp.single_intercept
  }
  ## and add to list...
  model.formula.13[[i]] <- formula(tmp_final)
}
```

```
## see what "final models" we got...
model.formula.13
```

```
## $CritDes
## as.numeric(Rating) ~ (1 | Artifact)
##
## $InitEDA
## as.numeric(Rating) ~ (1 | Artifact)
##
## $InterpRes
## as.numeric(Rating) ~ (1 | Artifact)
##
## $RsrchQ
## as.numeric(Rating) ~ (1 | Artifact)
##
## $SelMeth
## as.numeric(Rating) ~ (1 | Artifact)
##
## $TxtOrg
## as.numeric(Rating) ~ (1 | Artifact)
##
## $VisOrg
## as.numeric(Rating) ~ (1 | Artifact)
```

So, it looks like we don't need to add any fixed effects or interactions to the models for each rubric, using only the data reduced to the 13 common rubrics.

## Part B

**Adding fixed effects to the seven rubric-specific models using all the data.**

Now let's try with the full data... We use tall.nonmissing data set to conduct modeling analysis.

```
model.formula.alldata <- as.list(rep(NA,7))
names(model.formula.alldata) <- Rubric.names

for (i in Rubric.names) {
  ## fit each base model
```

```
  rubric.data <- tall.nonmissing[tall.nonmissing$Rubric==i,]
  tmp <- lmer(as.numeric(Rating) ~ -1 + as.factor(Rater) +
              Semester + Sex + (1|Artifact),
            data=rubric.data,REML=FALSE)
  ## do backwards elimination
  tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)
  ## check to see if the raters are significantly different from one another
  tmp.single_intercept <- update(tmp.back_elim, . ~ . + 1 - as.factor(Rater))
  pval <- anova(tmp.single_intercept,tmp.back_elim)$"Pr(>Chisq)"[2]
  ## choose the best model
  if (pval<=0.05) {
    tmp_final <- tmp.back_elim
  } else {
    tmp_final <- tmp.single_intercept
  }
  ## and add to list...
  model.formula.alldata[[i]] <- formula(tmp_final)
}
```

```
## see what "final models" we got...
model.formula.alldata
```

```
## $CritDes
## as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##
## $InitEDA
## as.numeric(Rating) ~ (1 | Artifact)
##
## $InterpRes
## as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##
## $RsrchQ
## as.numeric(Rating) ~ (1 | Artifact)
##
## $SelMeth
## as.numeric(Rating) ~ as.factor(Rater) + Semester + (1 | Artifact) -
##     1
##
## $TxtOrg
## as.numeric(Rating) ~ (1 | Artifact)
##
## $VisOrg
## as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
```

## Part C

### Trying interactions and new random effects for the seven rubric specific models using all the data.

Now we see there are some differences among the models: For `InitEDA`, `RsrchQ` and `TxtOrg`, the models are just the simple random-intercept models. For the other four, the models are a little more complex. We should examine each of these 4 models to see (a) if the fixed effects make sense to us; and (2) if there are any interactions or additional random effects to consider.

```r
## refit the model and check on the t-statistics -- do all the variables matter?
fla <- formula(model.formula.alldata[["SelMeth"]])
tmp <- lmer(fla,data=tall.nonmissing[tall.nonmissing$Rubric=="SelMeth",])
round(summary(tmp)$coef,2)  ## fixed effects and their t-values
## apparently they do.

## now check to make sure we really need "Rater" as a factor...
tmp.single_intercept <- update(tmp, . ~ . + 1 - as.factor(Rater))
anova(tmp.single_intercept,tmp)
## looks like we do, so we keep "tmp" as our best model so far...

## now let's check for fixed-effect interactions... Since only Rater and Semester
## are involved, we only need to examine Rater*Semester
tmp.fixed_interactions <- update(tmp, . ~ . + as.factor(Rater)*Semester - Semester)
anova(tmp,tmp.fixed_interactions)
## Looks like the fixed-effect interactions are not needed; again we keep
## "tmp" as our best model so far...

## Finally we check for random effects, we should try
## (Rater|Artifact) and (Semester|Artifact).
## isn't even possible, so no testing is needed.
```

```r
## So this is our final model for SelMeth:
summary(tmp)
```

**SelMeth**

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + Semester + (1 | Artifact) -
##     1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "SelMeth", ]
##
## REML criterion at convergence: 143.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0480 -0.3923 -0.0551  0.2674  2.5827
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.08973  0.2996
##  Residual             0.10842  0.3293
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1  2.25037    0.07503  29.992
## as.factor(Rater)2  2.22653    0.07424  29.991
## as.factor(Rater)3  2.03316    0.07521  27.033
## SemesterS19       -0.35860    0.09796  -3.661
```

```
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2 a.(R)3
## as.fctr(R)2  0.285
## as.fctr(R)3  0.287  0.280
## SemesterS19 -0.413 -0.391 -0.394
```

```
formula(tmp)
```

```
## as.numeric(Rating) ~ as.factor(Rater) + Semester + (1 | Artifact) -
##     1
```

The final Rubric specific model for rubric SelMeth: $as.numeric(Rating)\ as.factor(Rater) + Semester + (1|Artifact) - 1$.

- $(1|Artifact)$: This is a random intercept term based on Artifact, which measures the random effect deviation from the overall mean rating for each artifact. Considering the same rater in the same semester (Semester Spring 19 or Semester Fall 19), different artifacts of the total 91 artifacts tend to get different mean ratings on rubric Select Method.
- Rater: Fixed effect Rater has a statistically significant effect on the ratings for rubric Select Method. Considering the same artifact in the same semester, Rater 3 tends to give the lowest rating on rubric Select Method, followed by Rater 2 tends to give 0.20 higher rating than Rater 3 and Rater1 tends to give 0.02 higher rating than Rater 2.
- Semester: Fixed effect Semester has a statistically significant effect on the ratings for rubric Select Method. Considering the same artifact rated by the same rater, the artifact received 0.36 lower rating in Semester Spring 19 than in Semester Fall 19 on rubric Select Method.

```
## refit the model and check on the t-statistics -- do all the variables matter?
fla <- formula(model.formula.alldata[["InitEDA"]])
tmp <- lmer(fla,data=tall.nonmissing[tall.nonmissing$Rubric=="InitEDA",])
round(summary(tmp)$coef,2)  ## fixed effects and their t-values
## looks like we do, so we keep "tmp" as our best model so far...

## now let's check for fixed-effect interactions... Since we do not have
## any fixed effects are involved, so the fixed-effect interactions are not needed.
## Finally we check for random effects.  We should only add random effects that
## are also present as fixed effects.  This means, for this model isn't even possible,
## so no testing is needed.
```

```
## So this is our final model for InitEDA:
summary(tmp)
```

**InitEDA**

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (1 | Artifact)
```

```
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "InitEDA", ]
##
## REML criterion at convergence: 239
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8889 -0.3391 -0.1427  0.4276  1.6035
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3651   0.6042
##  Residual             0.1655   0.4068
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.44226    0.07537    32.4
```

```
formula(tmp)
```

```
## as.numeric(Rating) ~ (1 | Artifact)
```

The final Rubric specific model for rubric InitEDA: $as.numeric(Rating)$ $(1|Artifact)$.

```
## refit the model and check on the t-statistics -- do all the variables matter?
fla <- formula(model.formula.alldata[["RsrchQ"]])
tmp <- lmer(fla,data=tall.nonmissing[tall.nonmissing$Rubric=="RsrchQ",])
round(summary(tmp)$coef,2)  ## fixed effects and their t-values
## looks like we do, so we keep "tmp" as our best model so far...

## now let's check for fixed-effect interactions... Since we do not have
## any fixed effects are involved, so the fixed-effect interactions are not needed.

## Finally we check for random effects.  We should only add random effects that
## are also present as fixed effects.  This means, for this model isn't even possible,
## so no testing is needed.
```

```
## So this is our final model for RsrchQ:
summary(tmp)
```

**RsrchQ**

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (1 | Artifact)
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "RsrchQ", ]
##
## REML criterion at convergence: 209.1
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2694 -0.5285 -0.3736  0.9743  2.4770
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.07276  0.2697
##  Residual             0.27825  0.5275
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.35169    0.05794   40.59
```

```
formula(tmp)
```

```
## as.numeric(Rating) ~ (1 | Artifact)
```

The final Rubric specific model for rubric RsrchQ: $as.numeric(Rating)$ $(1|Artifact)$.

```
## refit the model and check on the t-statistics -- do all the variables matter?
fla <- formula(model.formula.alldata[["TxtOrg"]])
tmp <- lmer(fla,data=tall.nonmissing[tall.nonmissing$Rubric=="TxtOrg",])
round(summary(tmp)$coef,2)  ## fixed effects and their t-values

## looks like we do, so we keep "tmp" as our best model so far...

## now let's check for fixed-effect interactions... Since we do not have
## any fixed effects are involved, so the fixed-effect interactions are not needed.

## Finally we check for random effects.  We should only add random effects that
## are also present as fixed effects.  This means, for this model isn't even possible,
## so no testing is needed.
```

```
## So this is our final model for TxtOrg:
summary(tmp)
```

**TxtOrg**

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (1 | Artifact)
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "TxtOrg", ]
##
## REML criterion at convergence: 247.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3557 -0.7550  0.3834  0.5302  2.4132
```

```
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.09371  0.3061
##  Residual             0.39573  0.6291
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.58745    0.06821   37.93
```

```
formula(tmp)
```

```
## as.numeric(Rating) ~ (1 | Artifact)
```

The final Rubric specific model for rubric TxtOrg: $as.numeric(Rating)$ $(1|Artifact)$.

- $(1|Artifact)$: Different artifacts of the total 91 artifacts tend to get different mean ratings on rubrics Initial EDA, Research Question and Text Organization.

```
## refit the model and check on the t-statistics -- do all the variables matter?
fla <- formula(model.formula.alldata[["CritDes"]])
tmp <- lmer(fla,data=tall.nonmissing[tall.nonmissing$Rubric=="CritDes",])
round(summary(tmp)$coef,2)  ## fixed effects and their t-values
## apparently they do.

## now check to make sure we really need "Rater" as a factor...
tmp.single_intercept <- update(tmp, . ~ . + 1 - as.factor(Rater))
anova(tmp.single_intercept,tmp)
## looks like we do, so we keep "tmp" as our best model so far...

## now let's check for fixed-effect interactions... Since only Rater
## is involved, so the fixed-effect interactions are not needed.

## Finally we check for random effects, we should try
## (as.factor(Rater)|Artifact)
## isn't even possible, so no testing is needed.
```

```
## So this is our final model for CritDes:
summary(tmp)
```

**CritDes**

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "CritDes", ]
##
```

```
## REML criterion at convergence: 271
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.55495 -0.50027 -0.08228  0.64663  1.60935
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.4349   0.6595
##  Residual             0.2473   0.4972
## Number of obs: 115, groups:  Artifact, 89
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1   1.6863     0.1207   13.98
## as.factor(Rater)2   2.1129     0.1219   17.34
## as.factor(Rater)3   1.8908     0.1219   15.51
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2
## as.fctr(R)2 0.244
## as.fctr(R)3 0.244  0.246
```

```
formula(tmp)
```

```
## as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
```

The final Rubric specific model for rubric CritDes: $as.numeric(Rating)\ as.factor(Rater) + (1|Artifact) - 1$.

```
## refit the model and check on the t-statistics -- do all the variables matter?
fla <- formula(model.formula.alldata[["InterpRes"]])
tmp <- lmer(fla,data=tall.nonmissing[tall.nonmissing$Rubric=="InterpRes",])
round(summary(tmp)$coef,2)  ## fixed effects and their t-values
## apparently they do.

## now check to make sure we really need "Rater" as a factor...
tmp.single_intercept <- update(tmp, . ~ . + 1 - as.factor(Rater))
anova(tmp.single_intercept,tmp)
## looks like we do, so we keep "tmp" as our best model so far...

## now let's check for fixed-effect interactions... Since only Rater
## is involved, so the fixed-effect interactions are not needed.

## Finally we check for random effects, we should try
## (as.factor(Rater)|Artifact)
## isn't even possible, so no testing is needed.
```

```
## So this is our final model for InterpRes:
summary(tmp)
```

**InterpRes**

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "InterpRes", ]
##
## REML criterion at convergence: 199.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5317 -0.7627  0.2635  0.6614  2.6535
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.06224  0.2495
##  Residual             0.25250  0.5025
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1  2.70421    0.08912   30.34
## as.factor(Rater)2  2.58574    0.08912   29.01
## as.factor(Rater)3  2.13918    0.09027   23.70
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2
## as.fctr(R)2 0.061
## as.fctr(R)3 0.062  0.062
```

```
formula(tmp)
```

```
## as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
```

The final Rubric specific model for rubric InterpRes: $as.numeric(Rating)\ as.factor(Rater) + (1|Artifact) - 1$.

```
## refit the model and check on the t-statistics -- do all the variables matter?
fla <- formula(model.formula.alldata[["VisOrg"]])
tmp <- lmer(fla,data=tall.nonmissing[tall.nonmissing$Rubric=="VisOrg",])
round(summary(tmp)$coef,2)  ## fixed effects and their t-values
## apparently they do.

## now check to make sure we really need "Rater" as a factor...
tmp.single_intercept <- update(tmp, . ~ . + 1 - as.factor(Rater))
anova(tmp.single_intercept,tmp)
## looks like we do, so we keep "tmp" as our best model so far...

## now let's check for fixed-effect interactions... Since only Rater
## is involved, so the fixed-effect interactions are not needed.
```

```
## Finally we check for random effects, we should try
## (as.factor(Rater)|Artifact)
## isn't even possible, so no testing is needed.
```

```
## So this is our final model for InterpRes:
summary(tmp)
```

**VisOrg**

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "VisOrg", ]
##
## REML criterion at convergence: 219.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5004 -0.3365 -0.2483  0.3841  1.8552
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.2907   0.5392
##  Residual             0.1467   0.3830
## Number of obs: 115, groups:  Artifact, 89
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1  2.37794    0.09658   24.62
## as.factor(Rater)2  2.64891    0.09564   27.70
## as.factor(Rater)3  2.28355    0.09658   23.64
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2
## as.fctr(R)2 0.263
## as.fctr(R)3 0.265  0.263
```

```
formula(tmp)
```

```
## as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
```

The final Rubric specific model for rubric VisOrg: $as.numeric(Rating)\ as.factor(Rater) + (1|Artifact) - 1$.

- $(1|Artifact)$: Considering the same rater: different artifacts of the total 91 artifacts tend to get different mean ratings on rubrics Critique Design, Interpret Results, and Visual Organization.
- Rater: Considering the same artifact in the same semester, 1) For rubric Critique Design, Rater 1 tends to give the lowest rating, followed by Rater 3 tends to give 0.20 higher rating than Rater 1 and Rater 2 tends to give 0.22 higher rating than Rater 3; 2) For rubric Interpret Results, Rater 3 tends to give the lowest rating, followed by Rater 2 tends to give 0.45 higher rating than Rater 3 and Rater 1 tends to give 0.11 higher rating than Rater 2; 3) For rubric Visual Organization, Rater 3 tends to give the lowest rating, followed by Rater 1 tends to give 0.10 higher rating than Rater 3 and Rater 2 tends to give 0.27 higher ratings than Rater 1.

## Part D

**Trying to add fixed effects, interactions, and new random effects to the "combined" model Rating ~ 1 + (0 + Rubric|Artifact), using all the data.**

Now we try something similar with the mixed-effects model suggested on p. 4 of the project assignment sheet.

```
## Start with the "combined" intercept-only model...
comb.0 <- lmer(as.numeric(Rating) ~ 1 + (0 + Rubric | Artifact),
               data=tall.nonmissing)
summary(comb.0)
display(comb.0)

## Although the random effects are highly correlated, we can still proceed with
## our variable selection...
## Try adding fixed effects with no interactions...
comb.full <- update(comb.0, . ~ . + as.factor(Rater) + Semester +
                      Sex + Repeated + Rubric)
summary(comb.full)
formula(comb.full)
comb.back_elim <- fitLMER.fnc(comb.full, log.file.name = FALSE)
summary(comb.back_elim)
formula(comb.back_elim)

## The final model fit is a boundary fit again, but we will proceed to try
## interactions
comb.inter <- update(comb.back_elim, . ~ . + as.factor(Rater)*Semester*Rubric)
ss <- getME(comb.inter,c("theta","fixef"))
comb.inter.u<- update(comb.inter,start=ss,
             control=lmerControl(optimizer="bobyqa",
                                 optCtrl=list(maxfun=2e5)))
comb.inter_elim <- fitLMER.fnc(comb.inter.u, log.file.name = FALSE)
anova(comb.back_elim,comb.inter_elim,comb.inter.u)
## the models are nested so we can use AIC, BIC or likelihod ratio (deviance)
## tests... AIC and the LRT agree on comb.inter_elim; BIC likes the simpler
## comb.back_elim.
```

```
formula(comb.inter_elim)
```

```
## as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) +
##     Semester + Rubric + as.factor(Rater):Rubric
```

```
summary(comb.inter_elim)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) +
##     Semester + Rubric + as.factor(Rater):Rubric
##    Data: tall.nonmissing
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))
##
## REML criterion at convergence: 1419.6
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9280 -0.5122 -0.0447  0.4827  3.5854
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.50348  0.7096
##           RubricInitEDA   0.35480  0.5956   0.44
##           RubricInterpRes 0.15192  0.3898   0.35 0.82
##           RubricRsrchQ    0.17953  0.4237   0.63 0.44 0.72
##           RubricSelMeth   0.06727  0.2594   0.42 0.60 0.74 0.36
##           RubricTxtOrg    0.26069  0.5106   0.42 0.64 0.67 0.55 0.64
##           RubricVisOrg    0.25491  0.5049   0.34 0.71 0.68 0.51 0.38 0.77
##  Residual                 0.18519  0.4303
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##                                  Estimate Std. Error t value
## (Intercept)                       1.75945    0.11785  14.929
## as.factor(Rater)2                 0.36537    0.13296   2.748
## as.factor(Rater)3                 0.21421    0.13297   1.611
## SemesterS19                      -0.17780    0.08228  -2.161
## RubricInitEDA                     0.74625    0.13676   5.457
## RubricInterpRes                   1.01453    0.13479   7.527
## RubricRsrchQ                      0.74926    0.12419   6.033
## RubricSelMeth                     0.42672    0.13040   3.272
## RubricTxtOrg                      1.04967    0.13551   7.746
## RubricVisOrg                      0.68354    0.13947   4.901
## as.factor(Rater)2:RubricInitEDA  -0.30843    0.17249  -1.788
## as.factor(Rater)3:RubricInitEDA  -0.29522    0.17282  -1.708
## as.factor(Rater)2:RubricInterpRes -0.53674    0.17008  -3.156
## as.factor(Rater)3:RubricInterpRes -0.75247    0.17049  -4.414
## as.factor(Rater)2:RubricRsrchQ   -0.50157    0.16151  -3.106
## as.factor(Rater)3:RubricRsrchQ   -0.37068    0.16179  -2.291
## as.factor(Rater)2:RubricSelMeth  -0.39602    0.16467  -2.405
## as.factor(Rater)3:RubricSelMeth  -0.41324    0.16504  -2.504
## as.factor(Rater)2:RubricTxtOrg   -0.58380    0.17141  -3.406
## as.factor(Rater)3:RubricTxtOrg   -0.48649    0.17177  -2.832
## as.factor(Rater)2:RubricVisOrg   -0.14444    0.17442  -0.828
## as.factor(Rater)3:RubricVisOrg   -0.33380    0.17481  -1.910

##
## Correlation matrix not shown by default, as p = 22 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it

## optimizer (bobyqa) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
## Finally, we consider adding random effects to what seems like the
## best model so far, comb.inter_elim...
```

```
## The fixed-effects terms we have to work with are:
## as.factor(Rater)
## Semester
## as.factor(Rater):Rubric

m0 <- comb.inter_elim
mA <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) +
             (0 + as.factor(Rater) | Artifact) + as.factor(Rater) +
             Semester + Rubric + as.factor(Rater):Rubric, data=tall.nonmissing)
anova(m0,mA)
## AIC and BIC both like including (0 + as.factor(Rater) | Artifact) in the model

m0 <- comb.inter_elim
mA <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) +
             (0 + Semester | Artifact) + as.factor(Rater) +
             Semester + Rubric + as.factor(Rater):Rubric, data=tall.nonmissing)
anova(m0,mA)
## AIC and BIC do not like (0 + Semester | Artifact) in the model...

m0 <- comb.inter_elim
mA <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) +
             (0 + as.factor(Rater) | Artifact) +
             (0 + as.factor(Rater):Rubric | Artifact) + as.factor(Rater) +
             Semester + Rubric + as.factor(Rater):Rubric, data=tall.nonmissing)
```

```
## Error: number of observations (=810) <= number of random effects (=1890) for term (0 + as.factor(Rat
```

```
## anova(m0,mA)     -- Not needed!
## There are not enough observations to fit mA here, so we need not do any
## formal model comparison...
```

```
## So, to summarize, the "final" model appears to be
comb.final <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) +
             (0 + as.factor(Rater) | Artifact) + as.factor(Rater) +
             Semester + Rubric + as.factor(Rater):Rubric, data=tall.nonmissing)
```

**Do you find that any of these fixed effects have a significant effect in predicting ratings? Are there any other random effects that you can justify adding to these models?**

- In conclusion, from the final mixed-effects model based on full dataset, we found that Rater, Semester, and Rubric are three factors in this experiment which are related to the ratings. Besides, there are some interesting interactions between factors Rater, Rubric, and Artifact, which indicate that raters are not interpreting the rubrics and the evidence in the artifacts in the same way.

```
formula(comb.final)
```

```
## as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) |
##     Artifact) + as.factor(Rater) + Semester + Rubric + as.factor(Rater):Rubric
```

```
summary(comb.final)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) |
##     Artifact) + as.factor(Rater) + Semester + Rubric + as.factor(Rater):Rubric
##    Data: tall.nonmissing
##
## REML criterion at convergence: 1370.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.06443 -0.46911 -0.02987  0.45353  2.74012
##
## Random effects:
##  Groups     Name               Variance Std.Dev. Corr
##  Artifact   RubricCritDes      0.49628  0.7045
##             RubricInitEDA      0.31787  0.5638    0.32
##             RubricInterpRes    0.10204  0.3194    0.14  0.67
##             RubricRsrchQ       0.17898  0.4231    0.50  0.19  0.54
##             RubricSelMeth      0.03823  0.1955    0.14  0.23  0.38 -0.24
##             RubricTxtOrg       0.25027  0.5003    0.27  0.44  0.36  0.31  0.21
##             RubricVisOrg       0.23237  0.4821    0.17  0.50  0.45  0.28 -0.16
##  Artifact.1 as.factor(Rater)1 0.01282  0.1132
##             as.factor(Rater)2 0.11176  0.3343   -0.49
##             as.factor(Rater)3 0.09412  0.3068    0.33  0.66
##  Residual                     0.13469  0.3670
##
##
##
##
##
##
##
##   0.54
##
##
##
##
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                      1.75754    0.11403  15.413
## as.factor(Rater)2                0.36607    0.13918   2.630
## as.factor(Rater)3                0.19593    0.12966   1.511
## SemesterS19                     -0.15917    0.07647  -2.081
## RubricInitEDA                    0.73952    0.12996   5.690
## RubricInterpRes                  0.99152    0.12770   7.764
## RubricRsrchQ                     0.72620    0.11792   6.158
## RubricSelMeth                    0.41071    0.12469   3.294
## RubricTxtOrg                     1.01579    0.12999   7.814
## RubricVisOrg                     0.65424    0.13353   4.900
```

```
## as.factor(Rater)2:RubricInitEDA   -0.29984    0.15609  -1.921
## as.factor(Rater)3:RubricInitEDA   -0.29478    0.15635  -1.885
## as.factor(Rater)2:RubricInterpRes -0.51323    0.15348  -3.344
## as.factor(Rater)3:RubricInterpRes -0.71484    0.15364  -4.653
## as.factor(Rater)2:RubricRsrchQ    -0.48743    0.14721  -3.311
## as.factor(Rater)3:RubricRsrchQ    -0.32241    0.14726  -2.189
## as.factor(Rater)2:RubricSelMeth   -0.38642    0.15030  -2.571
## as.factor(Rater)3:RubricSelMeth   -0.38720    0.14961  -2.588
## as.factor(Rater)2:RubricTxtOrg    -0.55106    0.15646  -3.522
## as.factor(Rater)3:RubricTxtOrg    -0.44490    0.15673  -2.839
## as.factor(Rater)2:RubricVisOrg    -0.10488    0.15861  -0.661
## as.factor(Rater)3:RubricVisOrg    -0.27519    0.15885  -1.732
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00493103 (tol = 0.002, component 1)
```

```
display(comb.final)
```

```
## lmer(formula = as.numeric(Rating) ~ (0 + Rubric | Artifact) +
##     (0 + as.factor(Rater) | Artifact) + as.factor(Rater) + Semester +
##     Rubric + as.factor(Rater):Rubric, data = tall.nonmissing)
##                                 coef.est coef.se
## (Intercept)                      1.76     0.11
## as.factor(Rater)2                0.37     0.14
## as.factor(Rater)3                0.20     0.13
## SemesterS19                     -0.16     0.08
## RubricInitEDA                    0.74     0.13
## RubricInterpRes                  0.99     0.13
## RubricRsrchQ                     0.73     0.12
## RubricSelMeth                    0.41     0.12
## RubricTxtOrg                     1.02     0.13
## RubricVisOrg                     0.65     0.13
## as.factor(Rater)2:RubricInitEDA -0.30     0.16
## as.factor(Rater)3:RubricInitEDA -0.29     0.16
## as.factor(Rater)2:RubricInterpRes -0.51   0.15
## as.factor(Rater)3:RubricInterpRes -0.71   0.15
## as.factor(Rater)2:RubricRsrchQ  -0.49     0.15
## as.factor(Rater)3:RubricRsrchQ  -0.32     0.15
## as.factor(Rater)2:RubricSelMeth -0.39     0.15
## as.factor(Rater)3:RubricSelMeth -0.39     0.15
## as.factor(Rater)2:RubricTxtOrg  -0.55     0.16
## as.factor(Rater)3:RubricTxtOrg  -0.44     0.16
## as.factor(Rater)2:RubricVisOrg  -0.10     0.16
## as.factor(Rater)3:RubricVisOrg  -0.28     0.16
##
## Error terms:
##  Groups     Name              Std.Dev. Corr
##  Artifact   RubricCritDes     0.70
##             RubricInitEDA     0.56      0.32
##             RubricInterpRes   0.32      0.14  0.67
##             RubricRsrchQ      0.42      0.50  0.19  0.54
##             RubricSelMeth     0.20      0.14  0.23  0.38 -0.24
##             RubricTxtOrg      0.50      0.27  0.44  0.36  0.31  0.21
##             RubricVisOrg      0.48      0.17  0.50  0.45  0.28 -0.16  0.54
##  Artifact.1 as.factor(Rater)1 0.11
```

```
##             as.factor(Rater)2 0.33     -0.49
##             as.factor(Rater)3 0.31      0.33  0.66
##  Residual                     0.37
## ---
## number of obs: 810, groups: Artifact, 90
## AIC = 1484.6, DIC = 1233.2
## deviance = 1301.9
```

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

- In conclusion, we accept comb.final as our final model, we can interpret the pieces as follows:

**Semester**

- Fixed effect Semester has a statistically significant effect on the ratings. Considering the same artifact rated on the same rubric by the same rater, the artifact received 0.16 lower rating in Semester Spring 19 than in Semester Fall 19.

**Rubric**

- Fixed effect Rubric has a statistically significant effect on the ratings.
- Critique Design: Considering the same artifact in the same semester, Rater 1 tends to give the lowest rating on the rubric Critical Design, followed by Rater 3 tends to give 0.196 higher rating than Rater 1 and Rater 2 tends to give 0.17 higher rating than Rater 3.
- Initial EDA: Considering the same artifact in the same semester, Rater 3 tends to give the lowest ratings on the rubric Initial EDA, followed by Rater 1 tends to give 0.10 higher rating than Rater 3 and Rater 2 tends to give 0.07 higher rating than Rater 1.
- Interpret Results: Considering the same artifact in the same semester, Rater 3 tends to give the lowest ratings on the rubric Interpret Results, followed by Rater 2 tends to give 0.37 higher rating than Rater 3 and Rater 1 tends to give 0.16 higher rating than Rater 2.
- Research Question: Considering the same artifact in the same semester, Rater 3 tends to give the lowest ratings on the rubric Research Question, followed by Rater 2 tends to give 0.01 higher rating than Rater 3 and Rater 1 tends to give 0.12 higher rating than Rater 2.
- Select Method: Considering the same artifact in the same semester, Rater 3 tends to give the lowest ratings on the rubric Select Method, followed by Rater 2 tends to give 0.02 higher rating than Rater 3 and Rater 1 tends to give 0.02 higher rating than Rater 2.
- Text Organization: Considering the same artifact in the same semester, Rater 3 tends to give the lowest ratings on the rubric Text Organization, followed by Rater 2 tends to give 0.06 higher rating than Rater 3 and Rater 1 tends to give 0.18 higher rating than Rater 2.
- Visual Organization: Considering the same artifact in the same semester, Rater 3 tends to give the lowest ratings on the rubric Visual Organization, followed by Rater 1 tends to give 0.08 higher rating than Rater 3 and Rater 2 tends to give 0.26 higher rating than Rater 1.

**Rater**

- Fixed effect Rater has a statistically significant effect on the ratings. Considering all of the artifacts in the same semester, Rater 3 tends to give especially low ratings, Rater 1 tends to give higher ratings.

$(0 + Rubric|Artifact) + Rubric$

- There is an interaction between Rubric and Artifact. There are different average scores on each rubric, but the rubric averages also vary a bit from one Artifact to the next, by a small random effect that depends on Artifact. In all of this, the fact that Rubric scores depend on Artifact is what we might expect since the artifacts aren't all of equal quality on each rubric, and so we should expect the average scores on each Rubric to vary from one Artifact to the next.

$(0 + as.factor(Rater)|Artifact) + as.factor(Rater)$

- There is an interaction between Rater and Artifact. Each Rater's rating on each Artifact differs from what we would expect (from the fixed effects alone) by a small random effect that depends on the Artifact. This interaction suggests that the Raters are not interpreting the evidence in the artifacts in the same way.

$Rubric + as.factor(Rater) + as.factor(Rater) : Rubric$

- There is an interaction between Rater and Rubric. Each Rater uses each Rubric in a way that is not like, or even parallel to, other rater's Rubric usage. This interaction suggests that the Raters are not all interpreting the Rubrics in the same way.
- More troubling are the interaction between Rater and Rubric and the interaction between Rater and Artifact. These interactions suggest that perhaps the raters should be trained more, to make the raters' ratings more similar to each other.

# Appendix 5. Research Question #4

First, let's make a table with the usual one-dimensional summary statistics for ratings by sex based on whole set of 91 artifacts.

```
ratings.nonmissing %>%
  pivot_longer(
    cols = RsrchQ:TxtOrg, names_to = "rubric", values_to = "rating") %>%
    group_by(Sex) %>%
    dplyr::summarise(
      N = length(rating),
      Min. = min(rating, na.rm = T),
      "1st Qu." = quantile(rating, 0.25, na.rm = T),
      Median = median(rating, na.rm = T),
      Mean = round(mean(rating, na.rm = T), 2),
      "3rd Qu." = quantile(rating, 0.75, na.rm = T),
      Max. = max(rating, na.rm = T),
      SD = round(sd(rating, na.rm = T), 2)
      ) %>%
  kbl(booktabs=T, caption = "Summary statistics for ratings by sex based on whole set of 91 artifacts")
  kable_classic(latex_options = "HOLD_position")
```

Table 5: Summary statistics for ratings by sex based on whole set of 91 artifacts

| Sex | N | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | SD |
|---|---|---|---|---|---|---|---|---|
| F | 434 | 1 | 2 | 2 | 2.31 | 3 | 4 | 0.71 |
| M | 364 | 1 | 2 | 2 | 2.31 | 3 | 4 | 0.71 |

Then let's look at the barplots of ratings for each rubric by sex based on whole set of 91 artifacts.

```
ratings.nonmissing %>%
  pivot_longer(
    cols = RsrchQ:TxtOrg, names_to = "rubric", values_to = "rating") %>%
    ggplot(aes(x = rating, fill = Sex)) +
    geom_histogram(bins = 8, position = "dodge") +
    facet_wrap(~ rubric) +
    theme(strip.background =element_rect(fill = "grey")) +
    theme(strip.text = element_text(colour = 'black')) +
    scale_fill_brewer(palette="Set1") +
    ylab('Count of Ratings') +
    xlab('Rating')
```

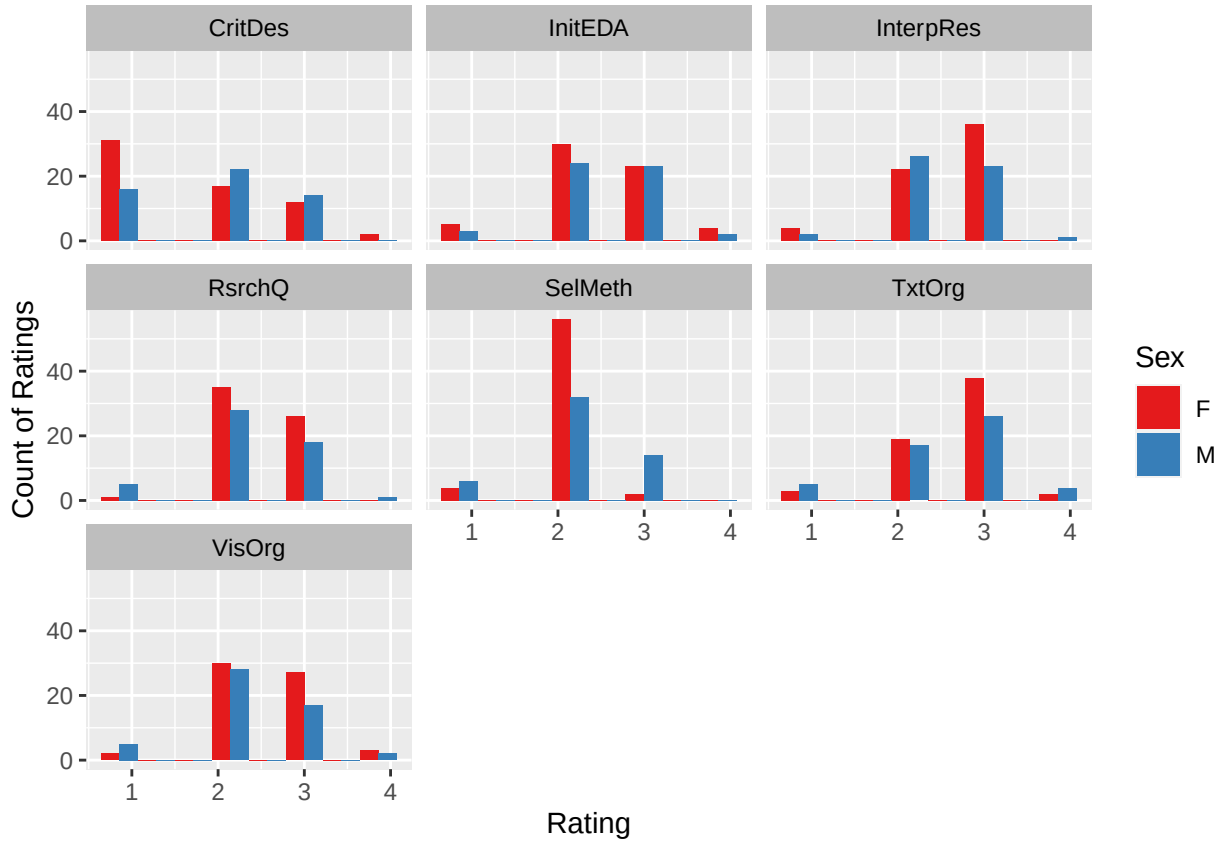
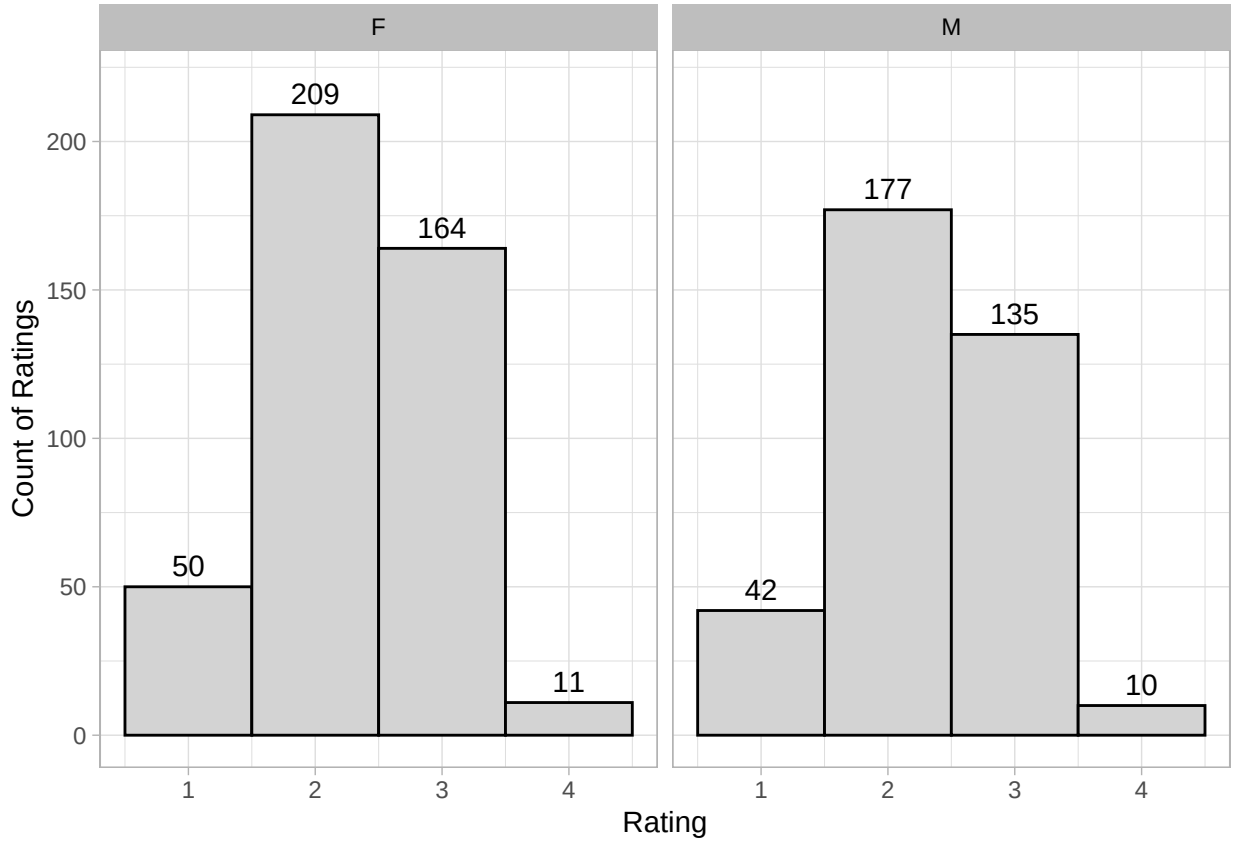



Figure 7: Barplots of ratings for each rubric by sex based on whole set of 91 artifacts

Next, let's look at the barplots of ratings by sex based on whole set of 91 artifacts.

```
ratings.nonmissing %>%
    pivot_longer(
      cols = RsrchQ:TxtOrg, names_to = "rubric", values_to = "rating") %>%
      ggplot(aes(x = rating)) +
      geom_histogram(bins = 8, position = 'dodge',
                     binwidth = 1, color = "black", fill =  "lightgrey") +
      ylim(0,220) +
      facet_wrap(~ Sex) +
      stat_bin(binwidth=1, geom="text", aes(label=..count..), vjust=-0.5) +
      theme_light() +
      theme(strip.background =element_rect(fill = "grey")) +
      theme(strip.text = element_text(colour = 'black')) +
      scale_fill_brewer(palette="Spectral") +
      ylab('Count of Ratings') +
      xlab('Rating')
```




First, let's make a table with the usual one-dimensional summary statistics for ratings by sex based on 13 artifacts seen by all three raters.

```
ratings.13 %>%
  pivot_longer(
    cols = RsrchQ:TxtOrg, names_to = "rubric", values_to = "rating") %>%
    group_by(Sex) %>%
    dplyr::summarise(
      N = length(rating),
      Min. = min(rating, na.rm = T),
```

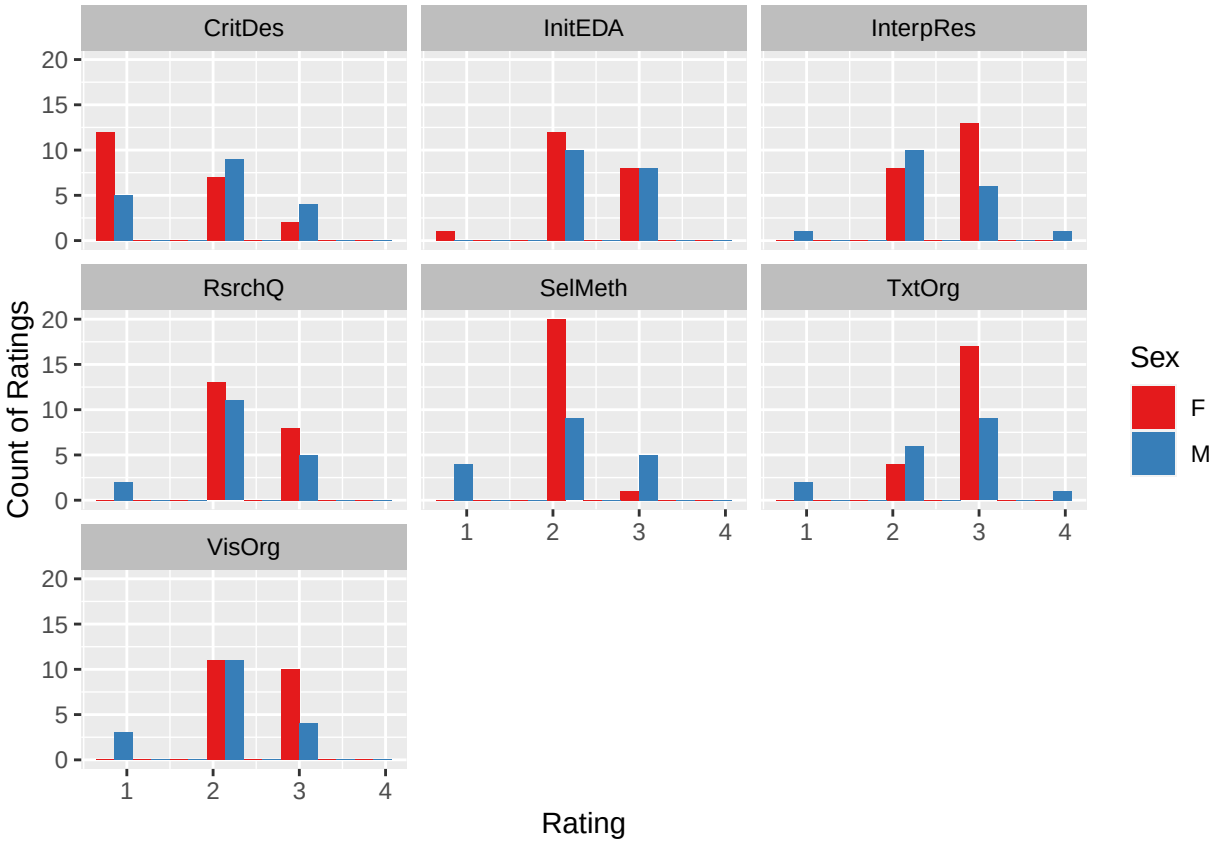
```
    "1st Qu." = quantile(rating, 0.25, na.rm = T),
    Median = median(rating, na.rm = T),
    Mean = round(mean(rating, na.rm = T), 2),
    "3rd Qu." = quantile(rating, 0.75, na.rm = T),
    Max. = max(rating, na.rm = T),
    SD = round(sd(rating, na.rm = T), 2)
    ) %>%
  kbl(booktabs=T, caption = "Summary statistics for ratings by sex based on 13 artifacts seen by all th
  kable_classic(latex_options = "HOLD_position")
```

Table 6: Summary statistics for ratings by sex based on 13 artifacts seen by all three raters

| Sex | N | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | SD |
|---|---|---|---|---|---|---|---|---|
| F | 147 | 1 | 2 | 2 | 2.31 | 3 | 3 | 0.63 |
| M | 126 | 1 | 2 | 2 | 2.22 | 3 | 4 | 0.69 |

Then let's look at the barplots of ratings for each rubric by sex based on 13 artifacts seen by all three raters.

```
ratings.13 %>%
  pivot_longer(
    cols = RsrchQ:TxtOrg, names_to = "rubric", values_to = "rating") %>%
    ggplot(aes(x = rating, fill = Sex)) +
    geom_histogram(bins = 8, position = "dodge") +
    facet_wrap(~ rubric) +
    theme(strip.background =element_rect(fill = "grey")) +
    theme(strip.text = element_text(colour = 'black')) +
    scale_fill_brewer(palette="Set1") +
    ylab('Count of Ratings') +
    xlab('Rating')
```




Next, let's look at the barplots of ratings by sex based on 13 artifacts seen by all three raters.

```
ratings.13 %>%
    pivot_longer(
      cols = RsrchQ:TxtOrg, names_to = "rubric", values_to = "rating") %>%
      ggplot(aes(x = rating)) +
      geom_histogram(bins = 8, position = 'dodge',
                     binwidth = 1, color = "black", fill =  "lightgrey") +
      ylim(0,80) +
      facet_wrap(~ Sex) +
      stat_bin(binwidth=1, geom="text", aes(label=..count..), vjust=-0.5) +
      theme_light() +
      theme(strip.background =element_rect(fill = "grey")) +
      theme(strip.text = element_text(colour = 'black')) +
      scale_fill_brewer(palette="Spectral") +
      ylab('Count of Ratings') +
      xlab('Rating')
```

- Note that: there are total of 434 ratings from females and 364 ratings from males based on whole set of 91 artifacts. The barplots of ratings by sex based on full dataset shows that the distributions of ratings for both females and males are relatively similar. Specifically, both distributions follow the normal distribution roughly centered at ratings of 2 and 3, which indicates that for both females and males, most of the artifacts were rated at middle scores such as 2 or 3. The summary statistics for ratings by sex based on whole set of 91 artifacts shows the similar patterns as the barplots.
- When compared the barplots of ratings for each rubric by sex based on full dataset, we found that the distributions of ratings for both females and males on each rubric are also relatively similar, but

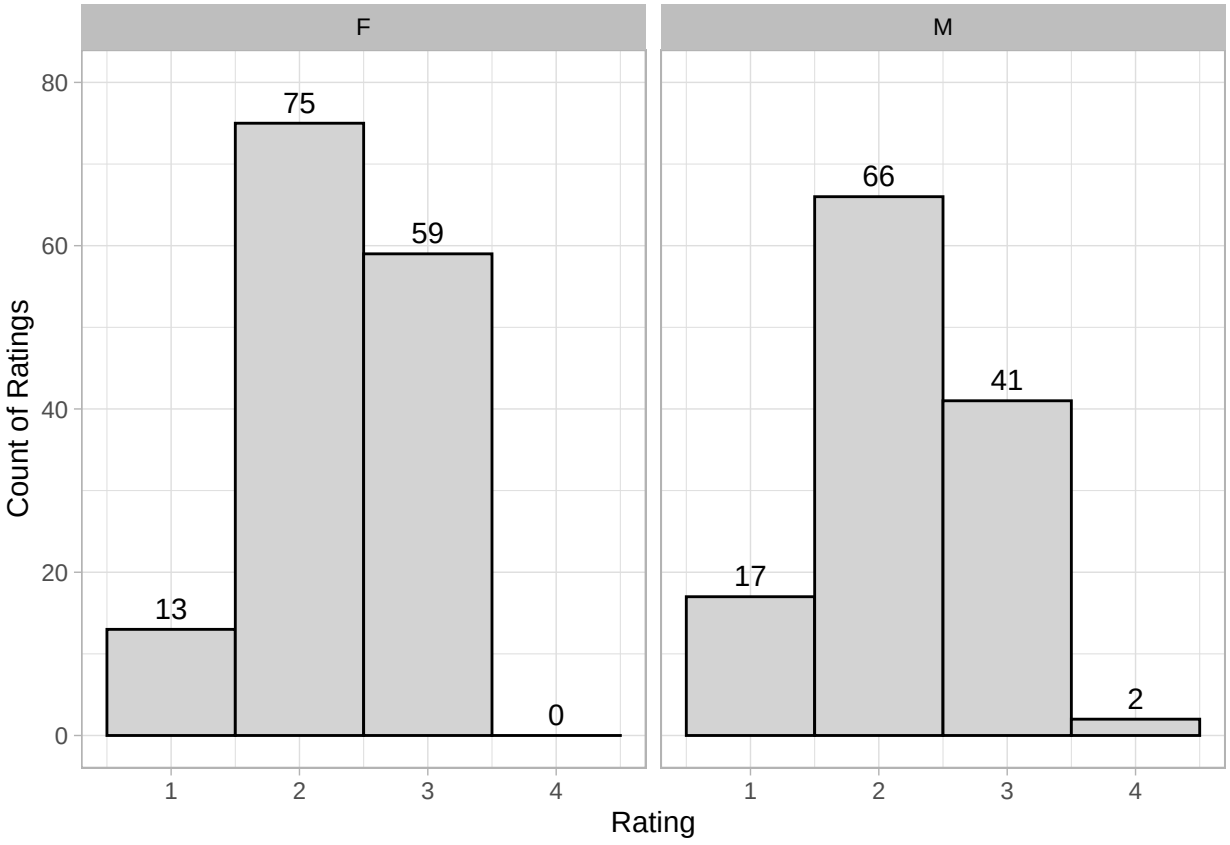



Figure 8: Barplots of ratings by sex based on 13 artifacts seen by all three raters

there are some slight differences in ratings for each rubric by sex based on whole set of 91 artifacts. To illustrate, across most of the rubrics, it is obviously to see that both females and males received more ratings at middle scores such as 2 or 3. Critique Design is the only rubric, on which females got most of ratings at score 1, and males received most of ratings at score 1 or 2. In addition, Select Method is the only rubric, on which females got extremely most of the ratings at score 2, and none of the females and males received ratings at score 4. We also examined the barplots of ratings for each rubric by sex based on 13 artifacts seen by all three raters, which display very similar patterns as the whole data set, except that none of the females received ratings at score 4 on all the rubrics.