Mixed Effects Regression Analysis on Students' Work Ratings in Dietrich College at CMU

Sifeng Li
sifengl@andrew.cmu.edu
7 December 2021

Abstract

We address the question of identifying factors that are related to ratings on student artifacts in order to determine the success of the new Dietrich College General Education program. We examine data collected by Junker (2021), on 91 student artifacts seen by three raters from three different departments. We perform barplots, ICCs, and Percent Exact Agreement (PEA) to examine differences in ratings related to rubrics and raters. We use mixed effect models to determine which factors in the experiment related to ratings. We find that rater, semester, rubric, and the interaction of rater and rubric as fixed effects and rubric and rater as random effects, grouped by artifacts to be our final model. The Dean's office is responsible for understanding how the model works and how to evaluate the success of the new General Education program by investigating how these factors impact ratings.

1 Introduction

Dietrich College at Carnegie Mellon University is now implementing a new "General Education" program for undergraduate students. In order to find out whether the GE course is successful, the college uses raters from across the college to rate 91 artifacts on seven rubrics. With the common understanding that different raters can have subjective opinions on each artifact, we want to further investigate how the distribution of ratings differs from each rubric or each rater and how various factors in the experiment related to the ratings.

In addition to answering the main question posed above, we will address the following questions:

- 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.
- 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?
- More generally, how are the various factors including rater, semester, sex, repeated, and rubric in this experiment related to the ratings? Do the factors interact in any interesting ways?
- Is there anything else interesting to say about this data?

2 Data

The data for this study come from Dietrich College, Carnegie Mellon University, collected by Junker (2021). The dataset contains 91 project papers, referred to as "artifacts", were randomly

chosen from a Fall and Spring section of Freshman Statistics. Three raters from three different departments were asked to rate these artifacts. Specifically, 13 of the 91 artifacts were rated by all three raters and the other 78 of the 91 artifacts were rated by only one rater.

In all, 91 artifacts are presented in the dataset available to us, and the rating rubric and rating scale are presented as following:

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

Table 1: Rubrics for Rating Freshman Statistics Projects

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

Table 2: Rating Scale Used for All Rubrics

In Table 3, we show the summary statistics for variables in the data file called ratings.

Variables	Minimum	Median	Mean	Maximum	S.D.
Sample	1	60	59.890	118	34.092
RsrchQ	1	2	2.350	4	0.592
CritDes	1	2	1.871	4	0.840
InitEDA	1	2	2.436	4	0.700
SelMeth	1	2	2.068	4	0.486
InterpRes	1	3	2.487	4	0.610

VisOrg	1	2	2.414	4	0.673
TxtOrg	1	3	2.598	4	0.696

Table 3: Summary Statistics for Variables of Ratings Dataset

Next, we show the summary statistics for variables in the data file called ratings, but we focus on only 13 ratings that were rated by all three raters in table 4.

Variables	Minimum	Median	Mean	Maximum	S.D.
Sample	1	52	54.28	110	34.330
RsrchQ	1	2	2.282	3	0.560
CritDes	1	2	1.718	3	0.724
InitEDA	1	2	2.385	3	0.544
SelMeth	1	2	2.051	3	0.510
InterpRes	1	3	2.513	4	0.601
VisOrg	1	2	2.282	3	0.605
TxtOrg	1	3	2.667	4	0.621

Table 4: Summary Statistics for Variables of 13 Artifacts in Ratings Dataset

We want to point out that there is missing data in the entire dataset. Specifically, there is a missing value of variable Sex, a missing value of variable CritDes, and a missing value of variable VisOrg. We consider not to worry about the missing value issue with reasons that the two observations with missing data will be eliminated when we use Rating as the response variable to perform models and there are no missing values in the subset including only artifacts that were rated by all three raters.

3 Methods

We will address the methods used for each research question defined in the Introduction section.

3.1 Researching on Distributions of Ratings

First, we make visual observations, specifically barplot, on the 13 artifacts that have been seen by all 3 raters. Then, we calculate the percentage table for the distribution of ratings on each rubric. This analysis can tell us how the overall ratings perform on different rubrics in order to

help us understand whether there are any extreme high/low ratings. variables work in combination to affect the average income per person. Then, we filter the dataset to contain scores given by different raters, and make barplots to illustrate the distribution given by each rater. Detailed R analyses can be found in Technical Appendix pp.7-pp.24.

3.2 Researching on Whether Raters Agree on the Score

First, we make visual observations, specifically barplot, on the scores given by different raters. Then, we calculate ICC on each rubric as a measure of rater agreement. We perform the calculation by making Rating to be the response variable, and the random intercept is grouped by Artifact. The ICC is calculated by dividing the random effect variance by the sum of random effect variance and the residual variance. With the value of ICC, we can directly identify whether raters generally agree more on the rating. Furthermore, in order to investigate which raters might be contributing to disagreement, we make a 2-way table of counts for the ratings of each pair of raters on each rubric to illustrate which rater agrees with which rater on each rubric. Detailed R analyses can be found in Technical Appendix pp.24-pp.48.

3.3 Researching on how Various Factors Related to the Ratings

First, we add fixed effects to the seven rubric-specific models using the dataset from the 13 common artifacts that all three raters have seen. We use backwards elimination to yield a model, then we use likelihood ratio test for each rubric to see if there is any difference on the estimates of raters by comparing this model to the intercept-only model. Detailed R analyses can be found in Technical Appendix pp.48-pp.54.

Second, we add fixed effects to the seven rubric-specific models using data without missing ratings. Similar to the first part, we use backwards elimination to yield a model, then we use likelihood ratio test for each rubric to see if there is any difference on the estimates of raters by comparing this model to the intercept-only model. For the rubrics that adding fixed effects improves the fit of the model, we check the t-statistics of the fixed effects to make sure they make sense and try adding interactions and new random effects. Detailed R analyses can be found in Technical Appendix pp.54-pp.59.

Finally, we try adding fixed effects, interactions, and new random effects to the “combined” model with Rubric being the random effect grouped by Artifact. We start with the intercept-only model, then try adding fixed effects using backward elimination. Then, we try adding interactions based on the fixed effects that we think are important, and we decide whether the term is significant based on the AIC, BIC, and p-values from likelihood ratio tests. After having any fixed effects and interactions included in the model, we consider to add new random effects and again, we decide whether the term is significant based on the AIC, BIC, and p-values from likelihood ratio tests. Detailed R analyses can be found in Technical Appendix pp.59-pp.71.

3.4 Interesting Things on the Dataset

This part contains research on interesting facts based on the semester by drawing barplots and making percentage tables for Fall semester and Spring semester respectively. This will help us

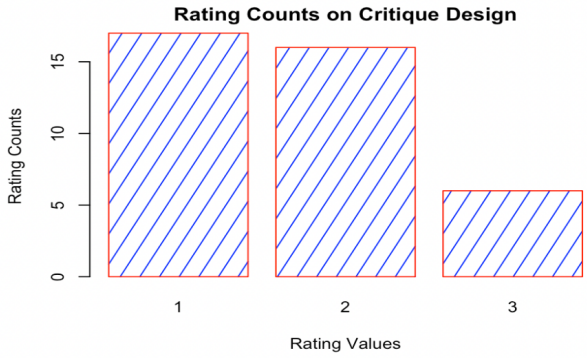
compare how raters perform on giving scores for different rubrics. Detailed R analyses can be found in Technical Appendix pp.71-pp.78.

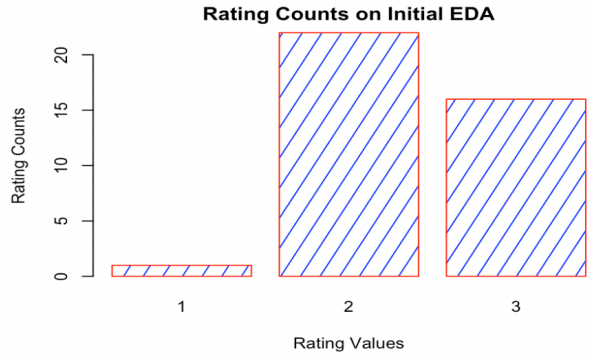
For this paper, all analyses were carried out in R and RStudio (RStudio Team, 2020).

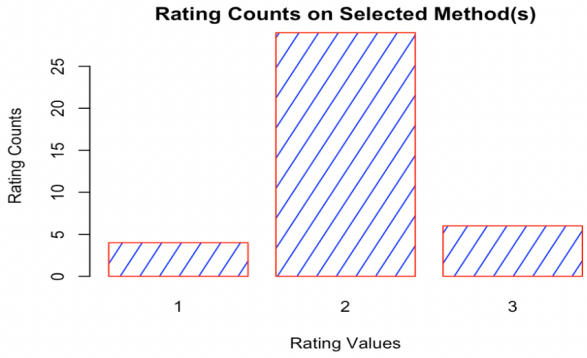
4 Results

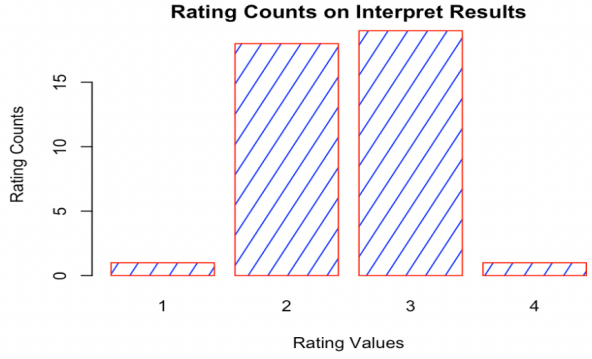
4.1 Researching on Distributions of Ratings

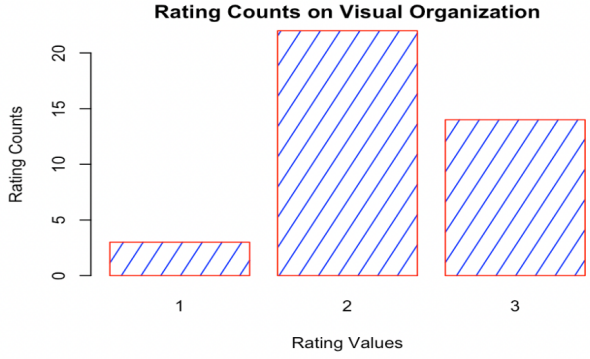
First, we do analysis on drawing barplots and calculating percentage tables to a sub-dataset with only 13 artifacts seen by all 3 raters.











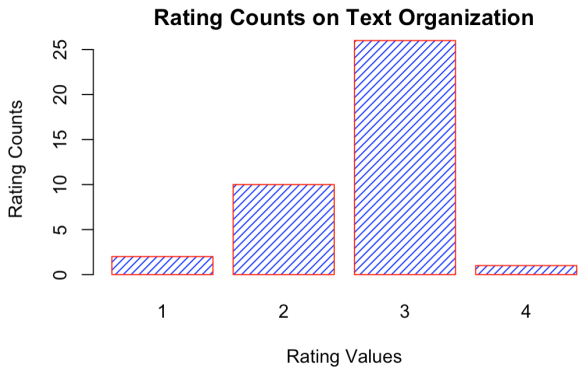
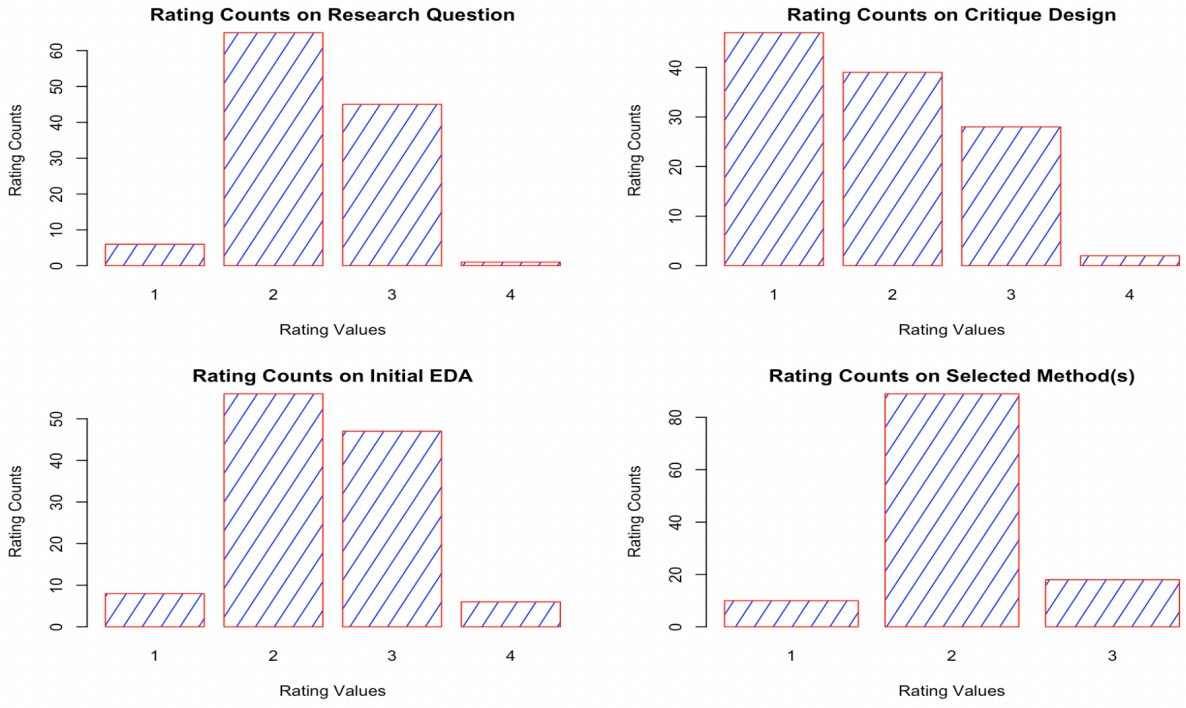


Figure 1: Barplots of all Seven-Specific Rubrics on 13 Artifacts Dataset

The frequency tables (including percentage of rating given each rubric) on page 9-11 of the Technical Appendix suggests that: for 13 Artifacts Dataset, the distribution of ratings for each rubric is pretty much not indistinguishable from the other rubrics except for the rating on critique design and the rating on text organization.

- For rubrics other than rating on critique design and the rating on text organization, we can observe that raters give score 2 most frequently on artifacts.
- For the rating of critique design, we can observe that raters give score 1 most frequently on artifacts.
- For the rating on text organization, we can observe that raters give score 3 most frequently on artifacts.

Then, we do analysis on drawing barplots and calculating percentage tables to the entire dataset.



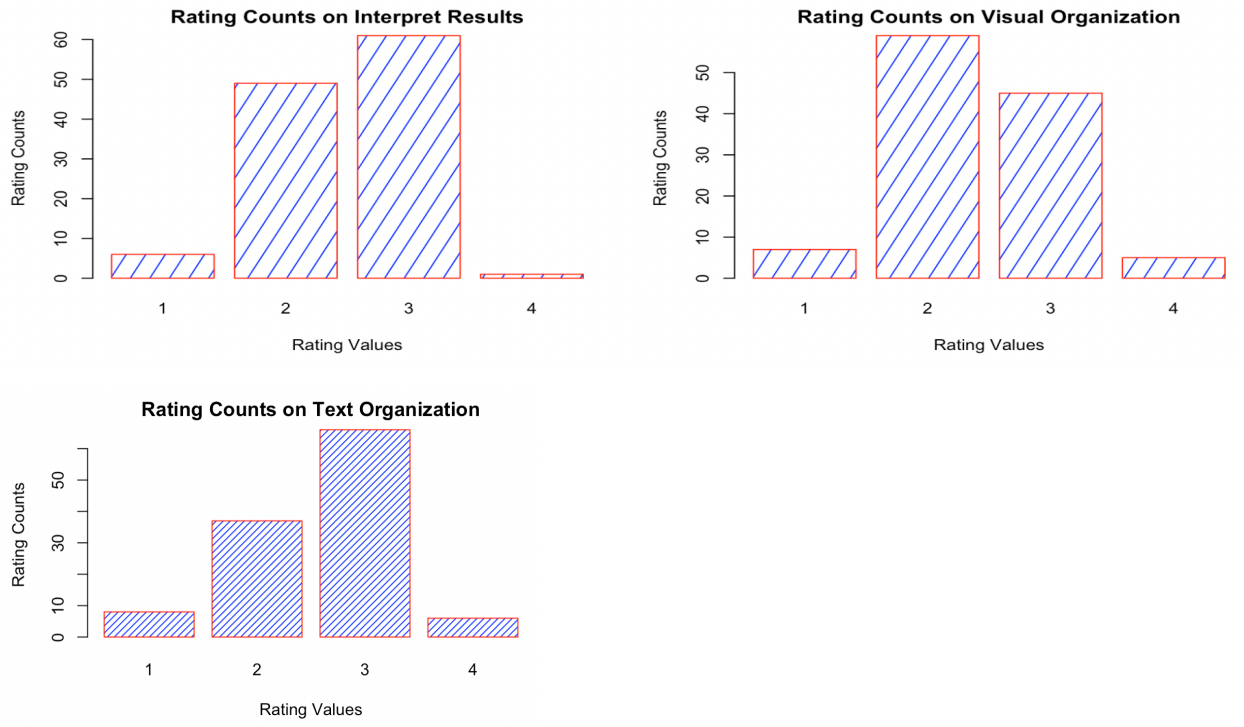


Figure 2: Barplots of all Seven-Specific Rubrics on Full Dataset

Comparing the full data with the subset containing 13 artifacts, we believe that the 13 artifacts are representative of the whole set of 91 artifacts. Also, we can observe that the distribution of these ratings in the subset of the data are comparable to those in the full dataset.

- For the distribution of rating on Interpret Results and Text Organizations, they are more indistinguishable from each other.
- It is obvious that Rating on Critique Design tends to get especially low ratings. We believe that the 13 artifacts are representative of the whole set of 91 artifacts.

As for illustrating the ratings given by each rater, we filter 3 datasets containing each rater's rating on seven rubrics and perform barplots on three sub-datasets, respectively.

	Rater 1	Rater 2	Rater 3
Rating Score 1	8	10	12
Rating Score 2	47	44	50
Rating Score 3	35	36	29
Rating Score 4	1	1	0

Table 5: Counts for Each Rating Score Given by Each Rater

From the above table, we compare the distribution of 3 raters rating on different artifacts, we can observe that the distribution of these ratings given by each rater is pretty much indistinguishable from the other users. All three of them give Rating Score 2 most frequently and Rating Score 4 least frequently. No rater tends to give especially high or low ratings.

4.2 Researching on Whether Raters Agree on the Score

In researching this question, we focus on the sub dataset containing only 13 artifacts seen by all 3 raters.

First, we measure the agreement among different raters by calculating the intraclass correlation (ICC) and fit seven random-intercept models as one model for each rubric.

Rubric	ICC Score
Research Question	0.19
Critique Design	0.57
Initial EDA	0.49
Select Method(s)	0.52
Interpret Results	0.23
Visual Organization	0.59
Text Organization	0.14

Table 6: ICC Score for Each Rubric

From the above table, we can notice that the ICC scores reflect the correlation between any two rater's ratings on the same artifact. We would expect the correlation to be higher if the raters are consistent with one another in how they rate, i.e. raters agree more when their correlations are higher. With the above explanation, we can conclude that:

- For Research Question, the ICC value of 0.19 indicates that these raters do not agree much on the rating.
- For Critique Design, the ICC value of 0.57 indicates that these raters do not agree much on the rating.
- For Initial EDA, the ICC value of 0.49 indicates that these raters do not agree much on the rating.
- For Select Method(s), the ICC value of 0.52 indicates that these raters do not agree much on the rating.
- For Interpret Results, the ICC value of 0.23 indicates that these raters do not agree much on the rating.
- For Visual Organization, the ICC value of 0.59 indicates that these raters do agree on the rating.

- For Text Organization, the ICC value of 0.14 indicates that these raters do not agree much on the rating.

The ICC's can help us determine whether the raters are generally in agreement on each rubric, but they cannot tell us which raters might be contributing to disagreement. Then, we perform a 2-way table of counts for the ratings of each pair of raters on each rubric and calculate the Percent Exact Agreement (PEA) to identify which rater agrees with which rater on each rubric.

	PEA between Rater 1 and Rater 2	PEA between Rater 2 and Rater 3	PEA between Rater 1 and Rater 3
Research Question	0.39	0.54	0.77
Critique Design	0.54	0.69	0.62
Initial EDA	0.69	0.85	0.54
Select Method(s)	0.92	0.69	0.62
Interpret Results	0.62	0.62	0.54
Visual Organization	0.54	0.77	0.77
Text Organization	0.69	0.54	0.62

Table 7: Percent Exact Agreement (PEA) between Every 2 Raters for Each Rubric

From the above table, we can notice that the Percent Exact Agreement reflects the correlation between any two rater's ratings on the same artifact more specifically. We would expect the coefficient to be higher if those two raters are consistent with one another in how they rate, i.e. two raters agree more when the coefficient is higher. With the above explanation, we can conclude that:

- For Research Question, only rater 1 and rater 3 agree on the rating; for rater 1 and rater 2 as well as rater 2 and rater 3, they do not agree much on the rating.
- For Critique Design, none of the 3 ratings groups agree much on the rating.
- For Initial EDA, only rater 2 and rater 3 agree on the rating; for rater 1 and rater 2 as well as rater 1 and rater 3, they do not agree much on the rating.
- For Select Method(s), only rater 1 and rater 2 agree on the rating; for rater 2 and rater 3 as well as rater 1 and rater 3, they do not agree much on the rating.
- For Interpret Results, none of the 3 ratings groups agree much on the rating.
- For Visual Organization, rater 2 and rater 3 as well as rater 1 and rater 3 agree on the rating; for rater 1 and rater 2, they do not agree much on the rating.
- For Text Organization, none of the 3 ratings groups agree much on the rating.

4.3 Researching on how Various Factors Related to the Ratings

For the seven rubric-specific models using just the data from the 13 common artifacts that all three raters saw, we find that adding fixed effects does not improve the fit of any of the seven models (Details can be found on Technical Appendix pp.53-pp.54). With the fact that we do not find any fixed effects are significant, we decide not to add any interaction terms or new random effects further, using only the data reduced to the 13 common rubrics.

For the seven rubric specific models using the entire dataset, we find that adding fixed effects perform some different results. For InitEDA, RsrchQ, and TxtOrg, adding fixed effects does not improve the fit of those models, i.e. those models are just simple random-intercept models. However, for CritDes, InterpRes, and VisOrg, adding Rater and removing the intercept improves the fit of those models; for SelMeth, adding Semester and removing the intercept improves the fit of the model. Therefore, we think that for rubrics CritDes, InterpRes, and VisOrg, Rater is related to Ratings; but for only one rubric SelMeth, Semester is related to Ratings (Details can be found on Technical appendix pp.58-pp.59).

Next, based on the results of likelihood ratio test and t-values, we find that for rubric CritDes, InterpRes, and VisOrg, including Rater in the model is important. Since Rater is the only fixed effect included in those models, there is no need to try fixed effects interactions. Moreover, since there are more random effects than number of observations in the dataset, the model with the random intercept of Rater grouped by Artifact cannot be fit, so we decide not to include any new random intercepts into the model. Therefore, for rubric CritDes, InterpRes, and VisOrg, the final model includes Rater as a fixed effect, but no additional fixed interactions or random effects included (Details can be found on Technical appendix pp.59-pp.64).

In addition to the above findings, based on the results of likelihood ratio test and t-values, we find that for rubric SelMeth, including Semester in the model is important. Since there are more random effects than number of observations in the dataset, the model with the random intercept of Rater grouped by Artifact and the model with the random intercept of Semester grouped by Artifact cannot be fit, so we decide not to include any new random intercepts into the model. Therefore, for rubric SelMeth, the final model includes Rater and Semester as the fixed effects, but no additional fixed interactions or random effects included (Details can be found on Technical appendix pp.59-pp.64).

We summarize the coefficients of the fixed effects for the final seven rubric-specific models as Table 8:

	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
(Intercept)	2.35	-	2.44	-	-	-	2.59
Rater1	-	1.69	-	2.25	2.70	2.38	-
Rater 2	-	2.11	-	2.23	2.59	2.65	-
Rater 3	-	1.89	-	2.03	2.14	2.28	-
Semester S19	-	-	-	-0.36	-	-	-

Table 8: Fixed Effects Coefficients for the Seven Rubric-Specific Models

From the above table, we can interpret the result as:

- For the rubric RsrchQ, the overall mean rating is 2.35; for the rubric InitEDA, the overall mean rating is 2.44; for the rubric TxtOrg, the overall mean rating is 2.59.
- Compared to the fall semester, the ratings on rubric SelMeth are 0.36 units lower on average.
- For the rubric CritDes, Rater 2 gives the highest rating on average among three raters and Rater 1 gives the lowest rating on average among three raters. Ratings given by Rater 2 are 0.42 units higher than ratings given by Rater 1 on average.
- For the rubric SelMeth, Rater 1 and Rater 2 give approximately the same ratings on average and are around 0.20 units higher than rating given by Rater 3 on average.
- For the rubric InterpRes, Rater 1 gives the highest rating on average among three raters and Rater 3 gives the lowest rating on average among three raters. Ratings given by Rater 1 are 0.56 units higher than ratings given by Rater 3 on average.
- For the rubric VisOrg, Rater 2 gives the highest rating on average among three raters and Rater 3 gives the lowest rating on average among three raters. Ratings given by Rater 2 are 0.37 units higher than ratings given by Rater 3 on average.

Although the above models allow us to look at the relationship between different factors and Ratings, they do not allow us to examine the interactions with rubric. Therefore, we try modelling a single model, starting with the model that includes Rubric as a random effect grouped by artifact.

Based on the result of likelihood ratio tests and AIC values, we find that the final combined result includes Rater, Semester, Rubric, and the interaction of Rater and Rubric as fixed effects and Rubric and Rater as random effects, grouped by artifacts (Details can be found on Technical appendix pp.64-pp.71).

	Estimate Coefficient	Standard Error
(Intercept)	1.76	0.11
Rater 2	0.37	0.14
Rater 3	0.21	0.13
Semester S19	-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
Rater2:RubricInitEDA	-0.30	0.16
Rater3:RubricInitEDA	-0.29	0.16
Rater2:RubricInterpRes	-0.51	0.15
Rater3:RubricInterpRes	-0.71	0.15
Rater2:RubricRsrchQ	-0.49	0.15
Rater3:RubricRsrchQ	-0.32	0.15
Rater2:RubricSelMeth	-0.39	0.15
Rater3:RubricSelMeth	-0.39	0.15
Rater2:RubricTxtOrg	-0.55	0.16
Rater3:RubricTxtOrg	-0.44	0.16
Rater2:RubricVisOrg	-0.10	0.16
Rater3:RubricVisOrg	-0.28	0.16

Table 9: Fixed Effects Coefficients for the Final Model

From the above table, we can interpret the fixed effect results as:

- Compared to Rater 1, we would expect that the ratings given by Rater 2 are 0.37 units higher on average and the ratings given by Rater 3 are 0.21 units higher on average, with keeping other predictors constant.
- Compared to fall semester, we would expect that the ratings given by spring semester are 0.26 units lower on average, with keeping other predictors constant.
- Compared to rubric CritDes, we would expect that the ratings on rubric InitEDA are 0.74 units higher on average, the ratings on rubric InterpRes are 0.99 units higher on average, the ratings on rubric RsrchQ are 0.73 units higher on average, the ratings on rubric SelMeth are 0.41 units higher on average, the ratings on rubric TxtOrg are 1.02 units higher on average, and the ratings on rubric VisOrg are 0.65 units higher on average, with keeping other predictors constant.

From the above table, we can interpret the interaction term results as:

- Compared to the rating on rubric CritDes given by Rater 1, we would expect that the ratings on rubric InitEDA given by Rater 2 are 0.30 units lower on average and the ratings on rubric InitEDA given by Rater 3 are 0.29 units lower on average.

- Compared to the rating on rubric CritDes given by Rater 1, we would expect that the ratings on rubric InterpRes given by Rater 2 are 0.51 units lower on average and the ratings on rubric InterpRes given by Rater 3 are 0.71 units lower on average.
- Compared to the rating on rubric CritDes given by Rater 1, we would expect that the ratings on rubric RsrchQ given by Rater 2 are 0.49 units lower on average and the ratings on rubric RsrchQ given by Rater 3 are 0.32 units lower on average.
- Compared to the rating on rubric CritDes given by Rater 1, we would expect that the ratings on rubric SelMeth given by Rater 2 are 0.39 units lower on average and the ratings on rubric SelMeth given by Rater 3 are 0.39 units lower on average.
- Compared to the rating on rubric CritDes given by Rater 1, we would expect that the ratings on rubric TxtOrg given by Rater 2 are 0.55 units lower on average and the ratings on rubric TxtOrg given by Rater 3 are 0.45 units lower on average.
- Compared to the rating on rubric CritDes given by Rater 1, we would expect that the ratings on rubric VisOrg given by Rater 2 are 0.10 units lower on average and the ratings on rubric VisOrg given by Rater 3 are 0.28 units lower on average.

4.4 Interesting Things on the Dataset

For this part, we would like to research interesting facts based on fall semester and spring semester. We filter two sub-datasets then draw barplots as well as calculate frequency tables on each rubric for different semesters.

The frequency tables (including percentage of rating given each rubric) on page 75-78 of the Technical Appendix suggests that:

- For rubric Research Question, raters give more score 2 in Fall semester but give more score 3 in Spring semester.
- For rubric Critique Design, raters give approximately the same large amount of score 1 and score 2 in Fall semester but give obviously more score 1 in Spring semester.
- For rubric Initial EDA, raters give approximately the same amount of score 2 and score 3 in Fall semester but give obviously more score 2 in Spring semester.
- For rubric Select Method(s), raters give obviously more score 2 in both Fall and Spring semester.
- For rubric Interpret Results, raters give obviously more score 3 in both Fall and Spring semester.
- For rubric Visual Organization, raters give obviously more score 2 in both Fall and Spring semester.
- For rubric Text Organization, raters give obviously more score 3 in both Fall and Spring semester.

5 Discussion

The study aims to help the Dean's Office at Carnegie Mellon University gain first-hand information on students' performance in each General Education course each year, and thus identify whether the new program is successful. Also, the Dean's Office is able to determine further directions on understanding how to implement a general education course based on the

conclusions from this paper. We recommend that the Dean's Office at CMU open courses for students to practice their critical evaluation skills, hold training sessions for raters to make the ratings consistent, and train students to write standard research papers following the "IMRAD" rule.

5.1 Researching on Distributions of Ratings

In our frequency table, we conclude that the distribution of ratings for each rubric is pretty much not indistinguishable from the other rubrics except for the rating on critique design and the rating on text organization. For rubrics other than rating on critique design and the rating on text organization, we can observe that raters give score 2 most frequently on artifacts.
For the rating of critique design, we can observe that raters give score 1 most frequently on artifacts. For the rating on text organization, we can observe that raters give score 3 most frequently on artifacts.

We can notice that the university puts a great amount of effort into training students to communicate in an organized and effective way through writing academic papers. Moreover, from here, we suggest that the Dean's Office at CMU should consider open courses involving teaching students how to critically evaluate the study design towards answering the research question.

Besides, we conclude that the distribution of these ratings given by each rater is pretty much indistinguishable from the other users. All three of them give Rating Score 2 most frequently and Rating Score 4 least frequently. No rater tends to give especially high or low ratings.

From here, we know that the overall student performance is within the average range without some of them performing exceptionally well or extremely poor.

5.2 Researching on Whether Raters Agree on the Score

In our table with ICC scores for each rubric, we conclude that for rubric Visual Organization, raters generally make agreements on the ratings. However, for those six rubrics other than Visual Organization, they do not agree much on the ratings. As for comparing Percent Exact Agreement (PEA) among every two raters, we conclude that none of the seven-specific rubrics have the case for all three raters agreeing on the ratings.

From there, we suggest that the Dean's Office might consider holding training sessions for teaching assistants on how to grade the student's works by initiating the bottomline and rubrics in order to improve the PEA among every two raters and help raters make the ratings consistent.

5.3 Researching on how Various Factors Related to the Ratings

By observing the seven random-intercept models, we notice that for rubrics CritDes, InterpRes, and VisOrg, Rater is related to Ratings; but for only one rubric SelMeth, Semester is related to Ratings. By observing the final single model, we notice that the model including Rater, Semester, Rubric, and the interaction of Rater and Rubric as fixed effects and Rubric and Rater as random effects, grouped by artifacts can be considered as the best fit.

From there, we conclude that with adding the fixed effects into the model, there might be cases where unfair ratings happened. It is possible since those three raters come from different departments and thus might focus on different scoring criterias. Then, we suggest that the Dean's Office might consider holding group meetings to let all raters have discussions together in order to keep the ratings consistent.

5.4 Interesting Things on the Dataset

From the frequency table for Fall semester and Spring semester, we conclude that students in both semesters do pretty well on Interpret Results and Text Organizations. For students who take this course in Fall semester, they perform better on Critique Design and Initial EDA; however, for students who take this course in Spring semester, they perform better on Research Question.

From there, we suggest that the Dean's Office at CMU might consider training students to write standard research papers based on the "Introduction, Methods, Results, and Discussions" (IMRAD) rule, and make sure students are getting a clear understanding of each part in the research paper.

5.5 Limitations and Future Works

There are some limitations that we would like to discuss regarding our data analysis. The first scope is that we have missing values for variable Sex that may cause the results to be a little bit biased. One possible improvement can be made is to research more on secondary resources and include information on the variable sex then do the analysis again.

In addition to that, the sample size of the data using only 13 artifacts that were rated by all three raters is relatively small, thus it might not be a good representation of the actual results, then there will be problems regarding our analysis. One possible improvement can be made to come up with more information about students including their year, majors, background (first-generation college students or not) to make sure our analysis is reproducible.

6 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 29, 2021)

Sheather, S.J. (2009). *A Modern Approach to Regression with R*. New York: Springer Science + Business Media LLC.

36-617 Project2 Technical Appendix

Sifeng Li

12/6/2021

```
library(arm)
```

```
## Loading required package: MASS
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.11-2, built: 2020-7-27)
```

```
## Working directory is /Users/sifengli/Desktop/CMU/Fall 2021/Applied Linear Models
```

```
library(MASS)
library(kableExtra)
library(lme4)
library(ggplot2)
library(plyr)
library(stats)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::arrange()    masks plyr::arrange()
## x purrr::compact()    masks plyr::compact()
## x dplyr::count()      masks plyr::count()
## x tidyr::expand()     masks Matrix::expand()
## x dplyr::failwith()   masks plyr::failwith()
## x dplyr::filter()     masks stats::filter()
## x dplyr::group_rows() masks kableExtra::group_rows()
## x dplyr::id()         masks plyr::id()
## x dplyr::lag()        masks stats::lag()
## x dplyr::mutate()     masks plyr::mutate()
## x tidyr::pack()       masks Matrix::pack()
## x dplyr::rename()     masks plyr::rename()
## x dplyr::select()     masks MASS::select()
## x dplyr::summarise()  masks plyr::summarise()
## x dplyr::summarize()  masks plyr::summarize()
## x tidyr::unpack()     masks Matrix::unpack()
```

```
library(nlme)
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## The following object is masked from 'package:lme4':
##
##     lmList
```

```
library(dplyr)
library(quanteda)
```

```
## Package version: 3.1.0
## Unicode version: 13.0
## ICU version: 67.1
```

```
## Parallel computing: 8 of 8 threads used.
```

```
## See https://quanteda.io for tutorials and examples.
```

```
library(foreign)
library(quanteda.textstats)
library(alr3)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## The following object is masked from 'package:arm':
##
##     logit
```

```
##
## Attaching package: 'alr3'
```

```
## The following object is masked from 'package:MASS':
##
##     forbes
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following object is masked from 'package:quanteda':
##
##     index
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(ggfortify)
library(leaps)
library(glmnet)
```

```
## Loaded glmnet 4.1-2
```

```
library(boot)
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:alr3':
##
##     wool
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
## The following object is masked from 'package:arm':
##
##     logit
```

```
library(matrixStats)
```

```
##
## Attaching package: 'matrixStats'
```

```
## The following object is masked from 'package:dplyr':
##
##     count
```

```
## The following object is masked from 'package:plyr':
##
##     count
```

```
library(grid)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(plyr)
```

```
# load in the datafile - ratings
ratings<-read.csv('/Users/sifengli/Desktop/CMU/Fall 2021/Applied Linear Models/ratings.csv', header=TRU
head(ratings)
```

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

```
## 1 1     3      1       5     Fall   M      3       3       2       2         2
## 2 2     3      2       7     Fall   F      3       3       3       3         3
## 3 3     3      3       9   Spring   F      2       1       3       2         3
## 4 4     3      4       8   Spring   M      2       2       2       1         1
## 5 5     3      5      NA     Fall  --      3       3       3       3         3
## 6 6     3      6      NA     Fall   M      2       1       2       2         2
##   VisOrg TxtOrg Artifact Repeated
## 1      2      3       O5        1
## 2      3      3       O7        1
## 3      3      3       O9        1
## 4      1      1       O8        1
## 5      3      3        5        0
## 6      2      2        6        0
```

```
str(ratings)
```

```
## 'data.frame':    117 obs. of  15 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Rater    : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ Sample   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Overlap  : int  5 7 9 8 NA NA NA NA NA 10 ...
##  $ Semester : chr  "Fall" "Fall" "Spring" "Spring" ...
##  $ Sex      : chr  "M" "F" "F" "M" ...
##  $ RsrchQ   : int  3 3 2 2 3 2 2 2 3 2 ...
##  $ CritDes  : int  3 3 1 2 3 1 1 1 1 1 ...
##  $ InitEDA  : int  2 3 3 2 3 2 3 2 2 2 ...
##  $ SelMeth  : int  2 3 2 1 3 2 2 2 2 2 ...
##  $ InterpRes: int  2 3 3 1 3 2 2 2 2 3 ...
##  $ VisOrg   : int  2 3 3 1 3 2 2 2 2 2 ...
##  $ TxtOrg   : int  3 3 3 1 3 2 2 2 2 3 ...
##  $ Artifact : chr  "O5" "O7" "O9" "O8" ...
##  $ Repeated : int  1 1 1 1 0 0 0 0 0 1 ...
```

```
# load in the datafile - tall
tall<-read.csv('/Users/sifengli/Desktop/CMU/Fall 2021/Applied Linear Models/tall.csv',header=TRUE)
```

```
# make summary for both ratings and tall
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
```

```
sd(ratings$Sample)
```

```
## [1] 34.09186
```

```
sd(ratings$RsrchQ)
```

```
## [1] 0.5918446
```

```
sd(ratings$CritDes, na.rm=TRUE)
```

```
## [1] 0.8395669
```

```
sd(ratings$InitEDA)
```

```
## [1] 0.6995641
```

```
sd(ratings$SelMeth)
```

```
## [1] 0.486481
```

```
sd(ratings$InterpRes)
```

```
## [1] 0.6104744
```

```
sd(ratings$VisOrg, na.rm=TRUE)
```

```
## [1] 0.67333
```

```
sd(ratings$TxtOrg)
```

```
## [1] 0.6955503
```

```
# make a subset of the data for only the 13 artifacts seen by all three raters
allThreeRatings <- ratings %>%
  filter(ratings$Repeated == 1)

# summary of the subset
summary(allThreeRatings)
```

```
##        X              Rater       Sample          Overlap     Semester
##  Min.   :  1.00   Min.   :1   Min.   :  1.00   Min.   : 1   Length:39
##  1st Qu.: 23.50   1st Qu.:1   1st Qu.: 24.50   1st Qu.: 4   Class :character
##  Median : 51.00   Median :2   Median : 52.00   Median : 7   Mode  :character
##  Mean   : 53.46   Mean   :2   Mean   : 54.28   Mean   : 7
##  3rd Qu.: 81.50   3rd Qu.:3   3rd Qu.: 82.50   3rd Qu.:10
##  Max.   :109.00   Max.   :3   Max.   :110.00   Max.   :13
##      Sex                RsrchQ         CritDes         InitEDA
##  Length:39          Min.   :1.000   Min.   :1.000   Min.   :1.000
##  Class :character   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000
##  Mode  :character   Median :2.000   Median :2.000   Median :2.000
##                     Mean   :2.282   Mean   :1.718   Mean   :2.385
##                     3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000
##                     Max.   :3.000   Max.   :3.000   Max.   :3.000
##     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   :2.051   Mean   :2.513   Mean   :2.282   Mean   :2.667
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :3.000   Max.   :4.000   Max.   :3.000   Max.   :4.000
##    Artifact            Repeated
##  Length:39          Min.   :1
##  Class :character   1st Qu.:1
##  Mode  :character   Median :1
##                     Mean   :1
##                     3rd Qu.:1
##                     Max.   :1
```

```
# make all rubric-related variables to categorical variables
allThreeRatings$RsrchQ <- as.factor(allThreeRatings$RsrchQ)
allThreeRatings$CritDes <- as.factor(allThreeRatings$CritDes)
allThreeRatings$InitEDA <- as.factor(allThreeRatings$InitEDA)
allThreeRatings$SelMeth <- as.factor(allThreeRatings$SelMeth)
allThreeRatings$InterpRes <- as.factor(allThreeRatings$InterpRes)
allThreeRatings$VisOrg <- as.factor(allThreeRatings$VisOrg)
allThreeRatings$TxtOrg <- as.factor(allThreeRatings$TxtOrg)
```

```
sd(as.numeric(allThreeRatings$Sample))
```

```
## [1] 34.32963
```

```
sd(as.numeric(allThreeRatings$RsrchQ))
```

```
## [1] 0.5595448
```

```
sd(as.numeric(allThreeRatings$CritDes, na.rm=TRUE))
```

```
## [1] 0.7236137
```

```
sd(as.numeric(allThreeRatings$InitEDA))
```

```
## [1] 0.5436419
```

```
sd(as.numeric(allThreeRatings$SelMeth))
```

```
## [1] 0.5103517
```

```
sd(as.numeric(allThreeRatings$InterpRes))
```

```
## [1] 0.6013929
```

```
sd(as.numeric(allThreeRatings$VisOrg, na.rm=TRUE))
```

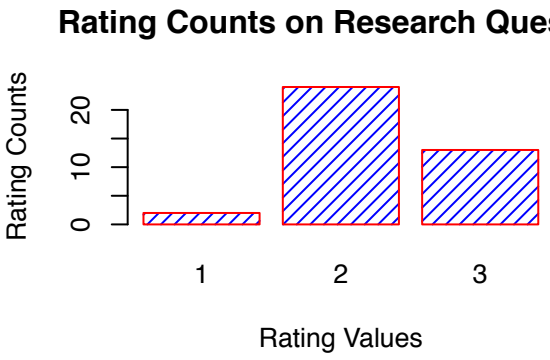
```
## [1] 0.6047495
```

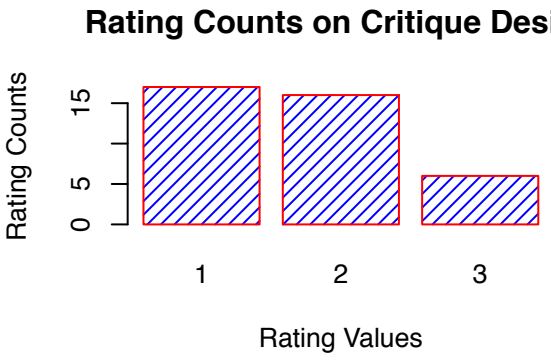
```
sd(as.numeric(allThreeRatings$TxtOrg))
```

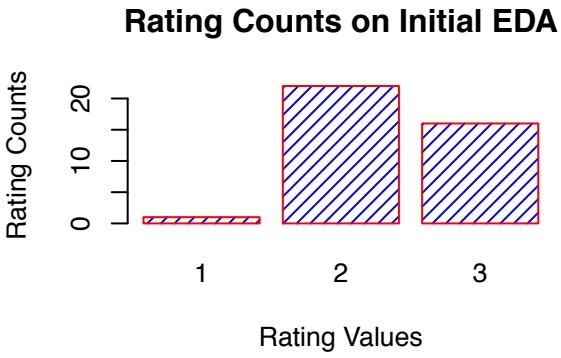
```
## [1] 0.6212607
```

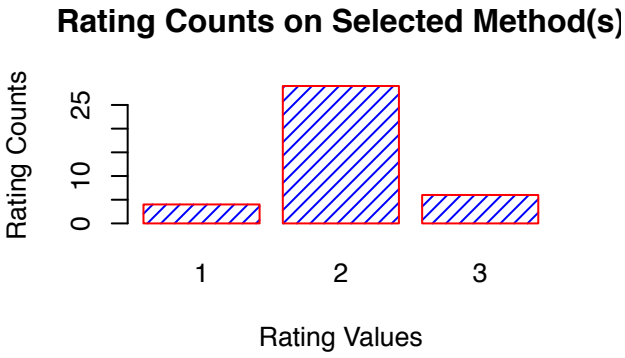
Researching on Distributions of Ratings

```
# distributions of ratings for each rubric
par(mfrow=c(2,2))
barplot(table(allThreeRatings$RsrchQ),main="Rating Counts on Research Question",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(allThreeRatings$CritDes),main="Rating Counts on Critique Design",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(allThreeRatings$InitEDA),main="Rating Counts on Initial EDA",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(allThreeRatings$SelMeth),main="Rating Counts on Selected Method(s)",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```

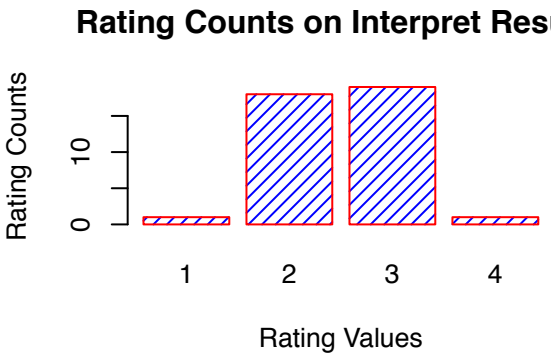


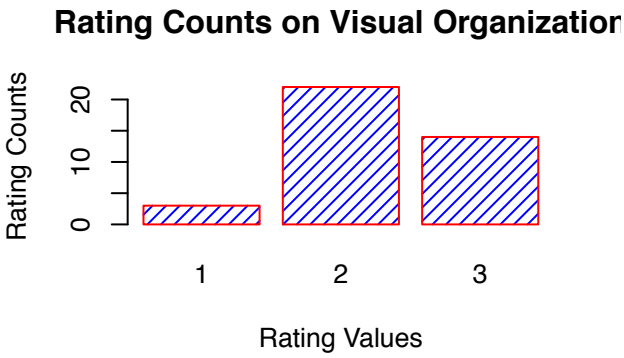


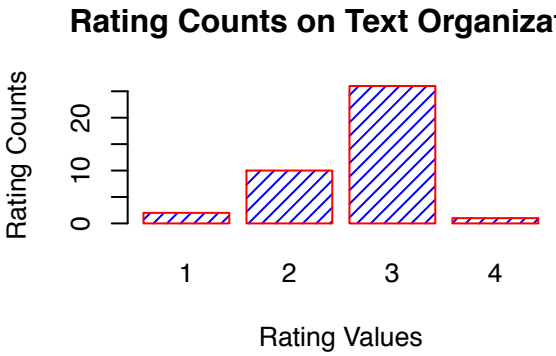





```
barplot(table(allThreeRatings$InterpRes),main="Rating Counts on Interpret Results",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(allThreeRatings$VisOrg),main="Rating Counts on Visual Organization",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(allThreeRatings$TxtOrg),main="Rating Counts on Text Organization",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```








```
# show the table of ratings given each rubric
RsrchQ<-table(allThreeRatings$RsrchQ)
addmargins(RsrchQ)
```

```
##
##   1   2   3 Sum
##   2  24  13  39
```

```
# percentage of RsrchQ
round(prop.table(RsrchQ)*100,digits=0)
```

```
##
##  1  2  3
##  5 62 33
```

```
CritDes<-table(allThreeRatings$CritDes)
addmargins(CritDes)
```

```
##
##   1   2   3 Sum
##  17  16   6  39
```

```
# percentage of CritDes
round(prop.table(CritDes)*100,digits=0)
```

```
##
##  1  2  3
## 44 41 15
```

```
InitEDA<-table(allThreeRatings$InitEDA)
addmargins(InitEDA)
```

```
##
##   1   2   3 Sum
##   1  22  16  39
```

```
# percentage of InitEDA
round(prop.table(InitEDA)*100,digits=0)
```

```
##
##  1  2  3
##  3 56 41
```

```
SelMeth<-table(allThreeRatings$SelMeth)
addmargins(SelMeth)
```

```
##
##   1   2   3 Sum
##   4  29   6  39
```

```
# percentage of SelMeth
round(prop.table(SelMeth)*100,digits=0)
```

```
##
##  1  2  3
## 10 74 15
```

```
InterpRes<-table(allThreeRatings$InterpRes)
addmargins(InterpRes)
```

```
##
##   1   2   3   4 Sum
##   1  18  19   1  39
```

```
# percentage of InterpRes
round(prop.table(InterpRes)*100,digits=0)
```

```
##
##  1  2  3  4
##  3 46 49  3
```

```
VisOrg<-table(allThreeRatings$VisOrg)
addmargins(VisOrg)
```

```
##
##   1   2   3 Sum
##   3  22  14  39
```

```
# percentage of VisOrg
round(prop.table(VisOrg)*100,digits=0)
```

```
##
##  1  2  3
##  8 56 36
```

```
TxtOrg<-table(allThreeRatings$TxtOrg)
addmargins(TxtOrg)
```

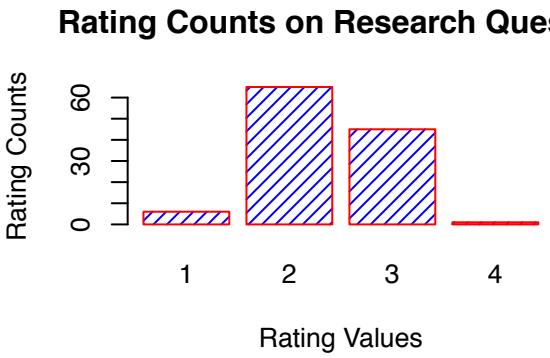
```
##
##   1   2   3   4 Sum
##   2  10  26   1  39
```

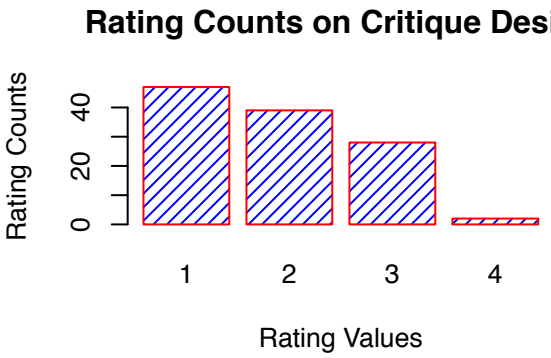
```r
# percentage of TxtOrg
round(prop.table(TxtOrg)*100,digits=0)
```

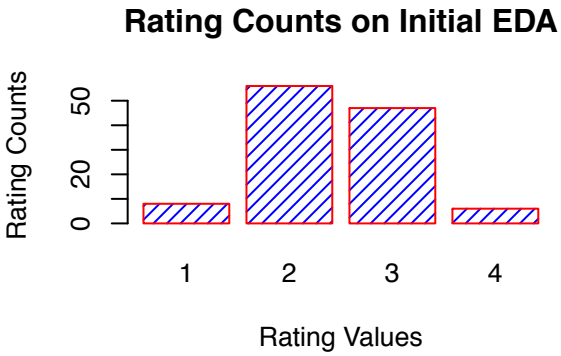
```
##
##  1  2  3  4
##  5 26 67  3
```

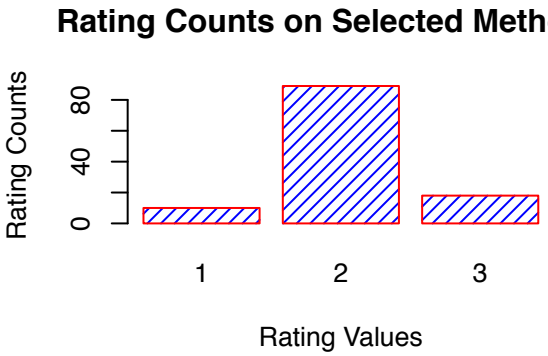
After observing the barplots and the frequency tables / percentage of ratings given each rubric, we can notice that the distribution of ratings for each rubrics is pretty much not indistinguishable from the other rubrics except for the rating on critique design and the rating on text organization. For rubrics other than rating on critique design and the rating on text organization, we can observe that raters give score 2 most frequently on artifacts. For the rating on critique design, we can observe that raters give score 1 most frequently on artifacts. For the rating on text organization, we can observe that raters give score 3 most frequently on artifacts.

```r
# work back to the full data
# barplot of ratings for each rubric
par(mfrow=c(2,2))
barplot(table(ratings$RsrchQ),main="Rating Counts on Research Question",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings$CritDes),main="Rating Counts on Critique Design",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings$InitEDA),main="Rating Counts on Initial EDA",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings$SelMeth),main="Rating Counts on Selected Method(s)",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```

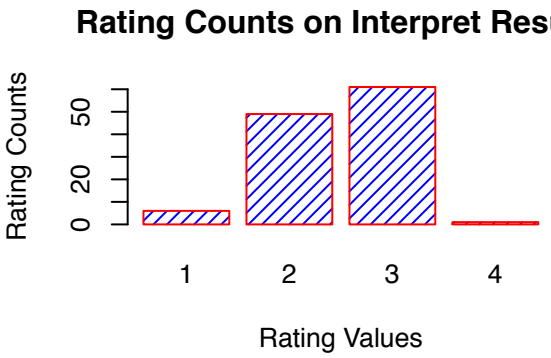


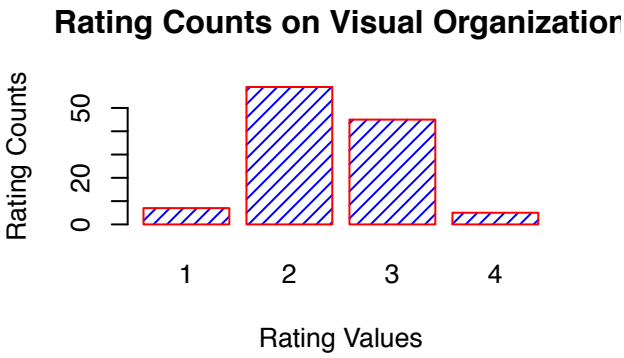


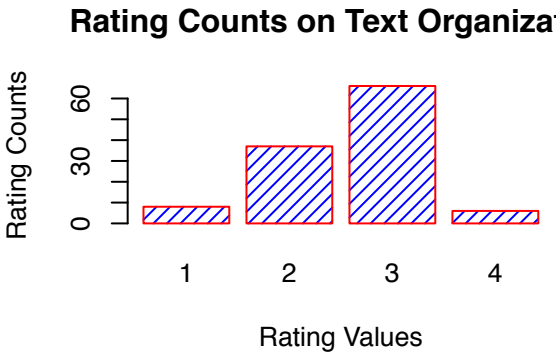





```
barplot(table(ratings$InterpRes),main="Rating Counts on Interpret Results",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings$VisOrg),main="Rating Counts on Visual Organization",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings$TxtOrg),main="Rating Counts on Text Organization",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```



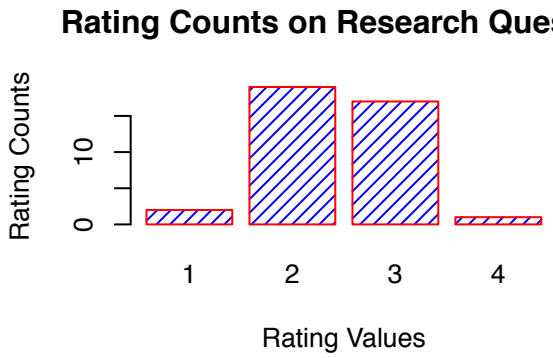


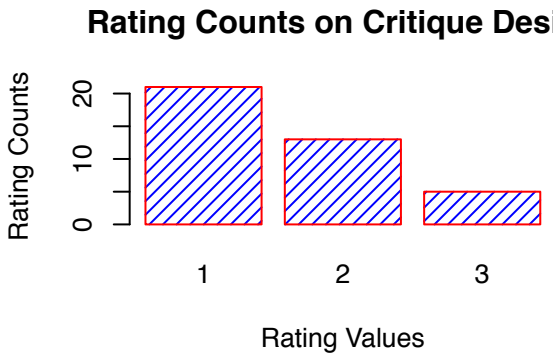


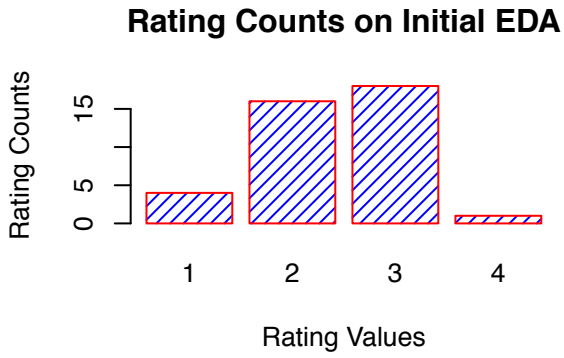
Comparing the full data with the subset, we can observe that the distribution of these ratings in the subset of the data are comparable to those in the full dataset. However, for the distribution of rating on Interpret Results and Text Organizations, they are more indistinguishable from each other. It is obvious that Rating on Critique Design tends to get especially low ratings. We believe that the thirtenn artifacts are representative of the whole set of 91 artifacts.

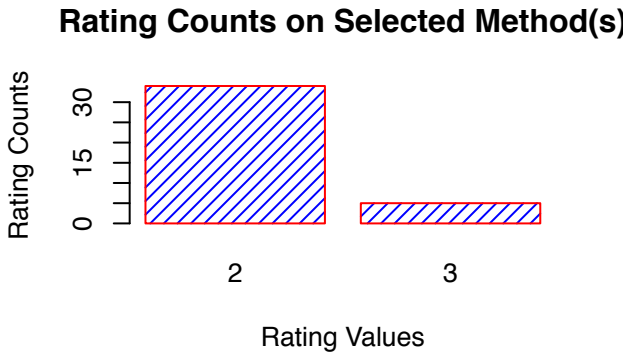
```
# rater 1
# the distribution of how rater 1 rates on different rubrics
ratings_score1 <- ratings %>%
  filter(ratings$Rater == 1)
```

```
par(mfrow=c(2,2))
barplot(table(ratings_score1$RsrchQ),main="Rating Counts on Research Question",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score1$CritDes),main="Rating Counts on Critique Design",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score1$InitEDA),main="Rating Counts on Initial EDA",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score1$SelMeth),main="Rating Counts on Selected Method(s)",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```

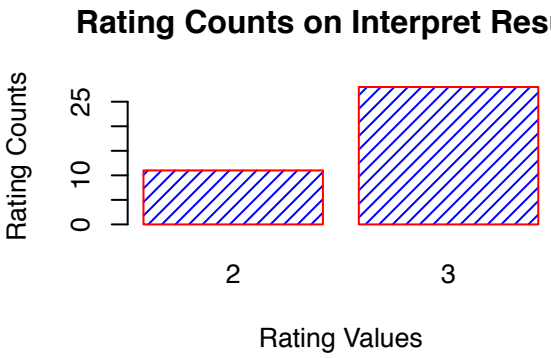


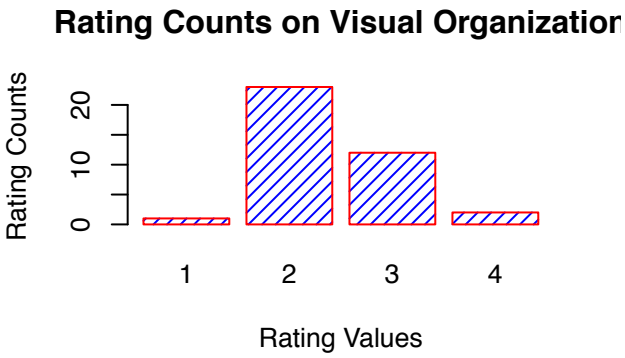


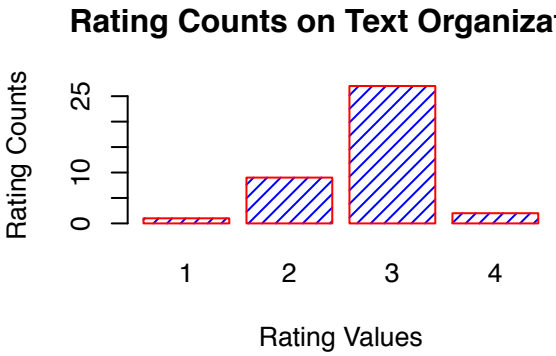





```
barplot(table(ratings_score1$InterpRes),main="Rating Counts on Interpret Results",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score1$VisOrg),main="Rating Counts on Visual Organization",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score1$TxtOrg),main="Rating Counts on Text Organization",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```

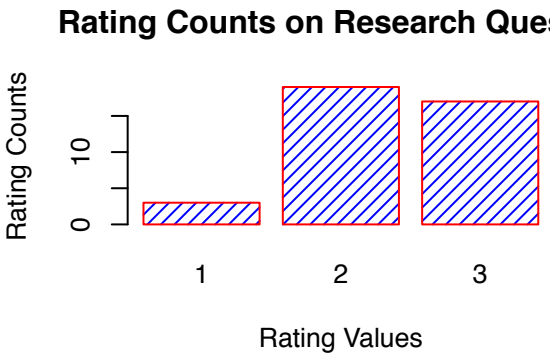


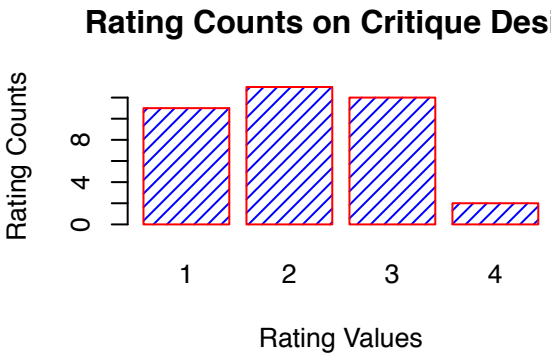


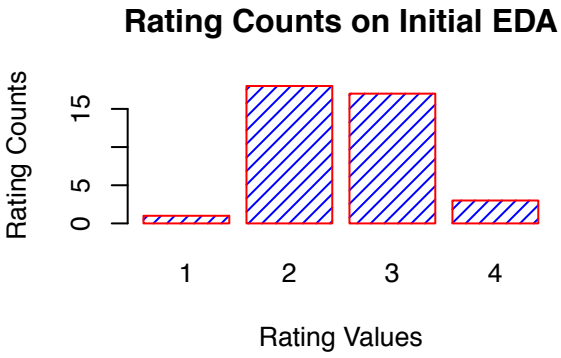


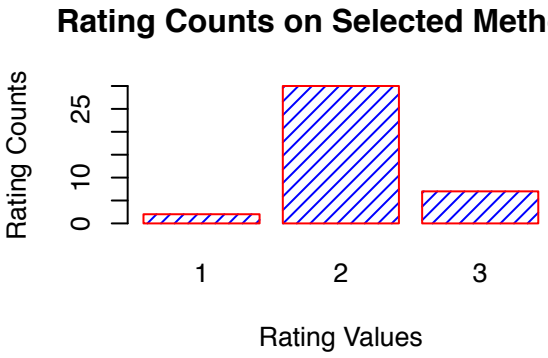

```
# rater 2
# the distribution of how rater 2 rates on different rubrics
ratings_score2 <- ratings %>%
  filter(ratings$Rater == 2)
```

```
par(mfrow=c(2,2))
barplot(table(ratings_score2$RsrchQ),main="Rating Counts on Research Question",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score2$CritDes),main="Rating Counts on Critique Design",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score2$InitEDA),main="Rating Counts on Initial EDA",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score2$SelMeth),main="Rating Counts on Selected Method(s)",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```

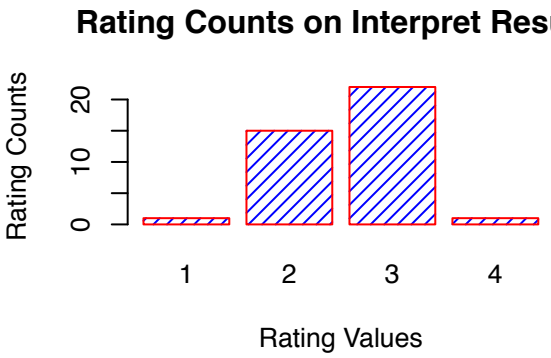


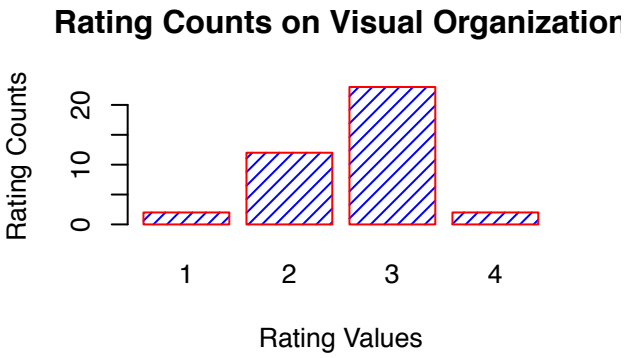


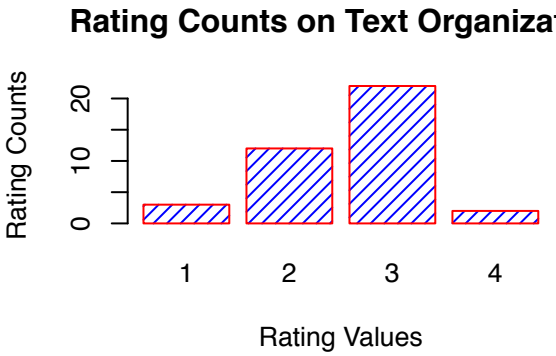





```
barplot(table(ratings_score2$InterpRes),main="Rating Counts on Interpret Results",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score2$VisOrg),main="Rating Counts on Visual Organization",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score2$TxtOrg),main="Rating Counts on Text Organization",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```

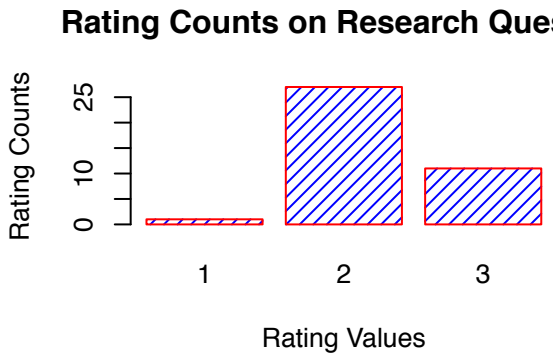


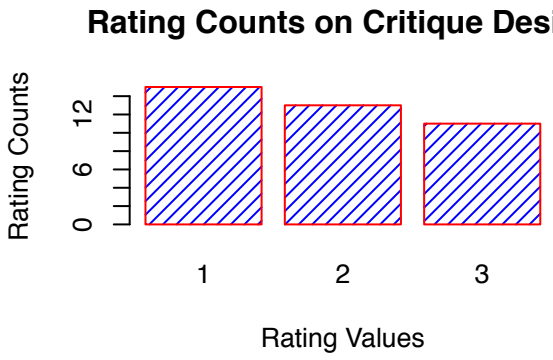


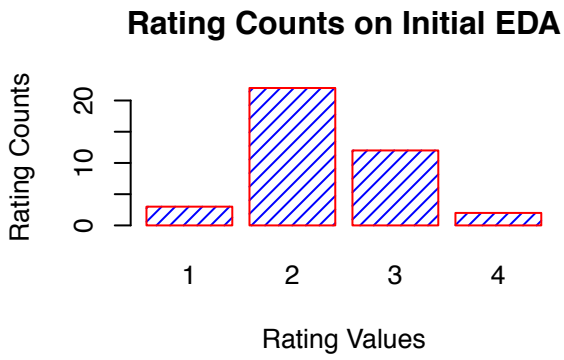


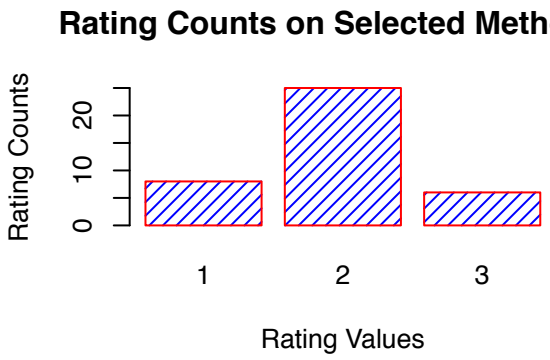

```
# rater 3
# the distribution of how rater 3 rates on different rubrics
ratings_score3 <- ratings %>%
  filter(ratings$Rater == 3)
```

```
par(mfrow=c(2,2))
barplot(table(ratings_score3$RsrchQ),main="Rating Counts on Research Question",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score3$CritDes),main="Rating Counts on Critique Design",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score3$InitEDA),main="Rating Counts on Initial EDA",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score3$SelMeth),main="Rating Counts on Selected Method(s)",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```

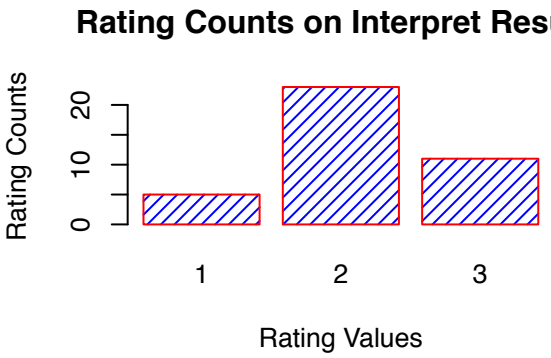


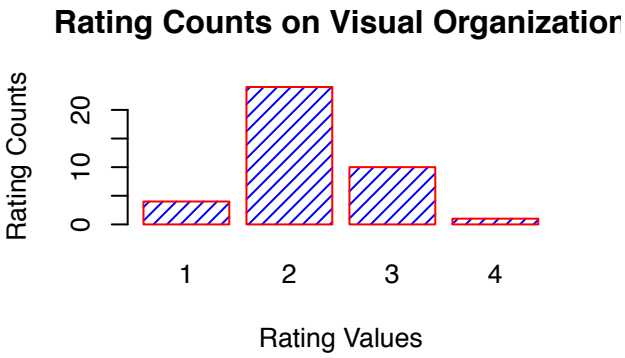


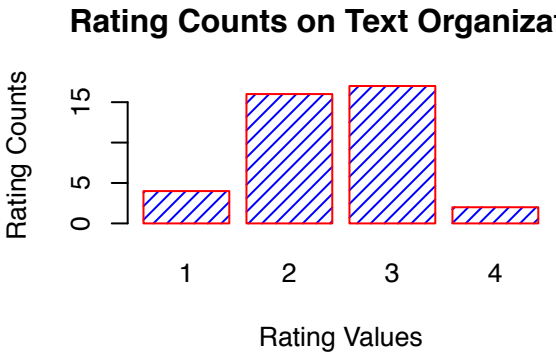





```
barplot(table(ratings_score3$InterpRes),main="Rating Counts on Interpret Results",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score3$VisOrg),main="Rating Counts on Visual Organization",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_score3$TxtOrg),main="Rating Counts on Text Organization",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```







Since we believe that the 13 artifacts are representative of the whole set of 91 artifacts, we continue using the subset in the following analysis.

```
# consider the subset
# count the number of ratings at each level given rater 1
ratings_sub_score1 <- allThreeRatings %>%
  filter(allThreeRatings$Rater == 1)

ratings_sub_score2 <- allThreeRatings %>%
  filter(allThreeRatings$Rater == 2)

ratings_sub_score3 <- allThreeRatings %>%
  filter(allThreeRatings$Rater == 3)
```

```
# extract 7 rubrics
# for rater 1, grouped by score 1-4
ratings_sub_score1_rub <- ratings_sub_score1[7:13]
apply(X=ratings_sub_score1_rub, 2, FUN=function(x) length(which(x==1))) #8
```

```
##    RsrchQ   CritDes   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg
##         0         6         1         0         0         1         0
```

```
apply(X=ratings_sub_score1_rub, 2, FUN=function(x) length(which(x==2))) #47
```

```
##    RsrchQ   CritDes   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg
##         8         6         4        11         5         9         4
```

```
apply(X=ratings_sub_score1_rub, 2, FUN=function(x) length(which(x==3))) #35
```

```
##    RsrchQ   CritDes   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg
##         5         1         8         2         8         3         8
```

```
apply(X=ratings_sub_score1_rub, 2, FUN=function(x) length(which(x==4))) #1
```

```
##    RsrchQ   CritDes   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg
##         0         0         0         0         0         0         1
```

```
rater_1 <- data.frame(
  rating = factor(c("1","2","3","4")),
  count = c(8,47,35,1)
)
rater_1
```

```
##   rating count
## 1      1     8
## 2      2    47
## 3      3    35
## 4      4     1
```

```
# extract 7 rubrics
# for rater 2, grouped by score 1-4
ratings_sub_score2_rub <- ratings_sub_score2[7:13]
apply(X=ratings_sub_score2_rub, 2, FUN=function(x) length(which(x==1))) #10
```

```
##    RsrchQ   CritDes   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg
##         2         5         0         1         0         1         1
```

```
apply(X=ratings_sub_score2_rub, 2, FUN=function(x) length(which(x==2))) #44
```

```
##    RsrchQ   CritDes   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg
##         7         5         8        10         6         5         3
```

```
apply(X=ratings_sub_score2_rub, 2, FUN=function(x) length(which(x==3))) #36
```

```
##    RsrchQ   CritDes   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg
##         4         3         5         2         6         7         9
```

```
apply(X=ratings_sub_score2_rub, 2, FUN=function(x) length(which(x==4))) #1
```

```
##    RsrchQ   CritDes   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg
##         0         0         0         0         1         0         0
```

```
rater_2 <- data.frame(
  rating = factor(c("1","2","3","4")),
  count = c(10,44,36,1)
)
rater_2
```

```
##   rating count
## 1      1    10
## 2      2    44
## 3      3    36
## 4      4     1
```

```
# extract 7 rubrics
# for rater 3, grouped by score 1-4
ratings_sub_score3_rub <- ratings_sub_score3[7:13]
apply(X=ratings_sub_score3_rub, 2, FUN=function(x) length(which(x==1))) #12
```

```
##    RsrchQ   CritDes   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg
##         0         6         0         3         1         1         1
```

```r
apply(X=ratings_sub_score3_rub, 2, FUN=function(x) length(which(x==2))) #50
```

```
##    RsrchQ   CritDes   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg
##         9         5        10         8         7         8         3
```

```r
apply(X=ratings_sub_score3_rub, 2, FUN=function(x) length(which(x==3))) #29
```

```
##    RsrchQ   CritDes   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg
##         4         2         3         2         5         4         9
```

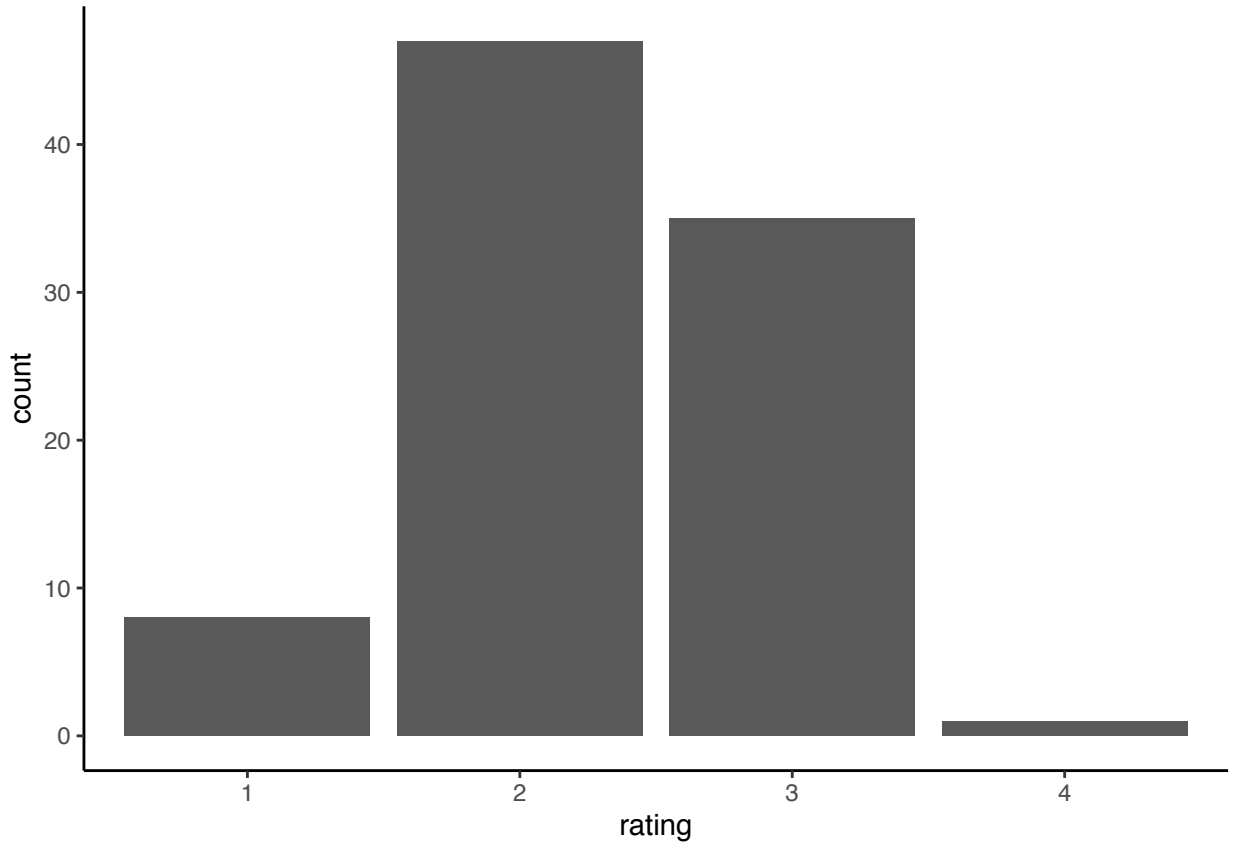
```r
apply(X=ratings_sub_score3_rub, 2, FUN=function(x) length(which(x==4))) #0
```

```
##    RsrchQ   CritDes   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg
##         0         0         0         0         0         0         0
```

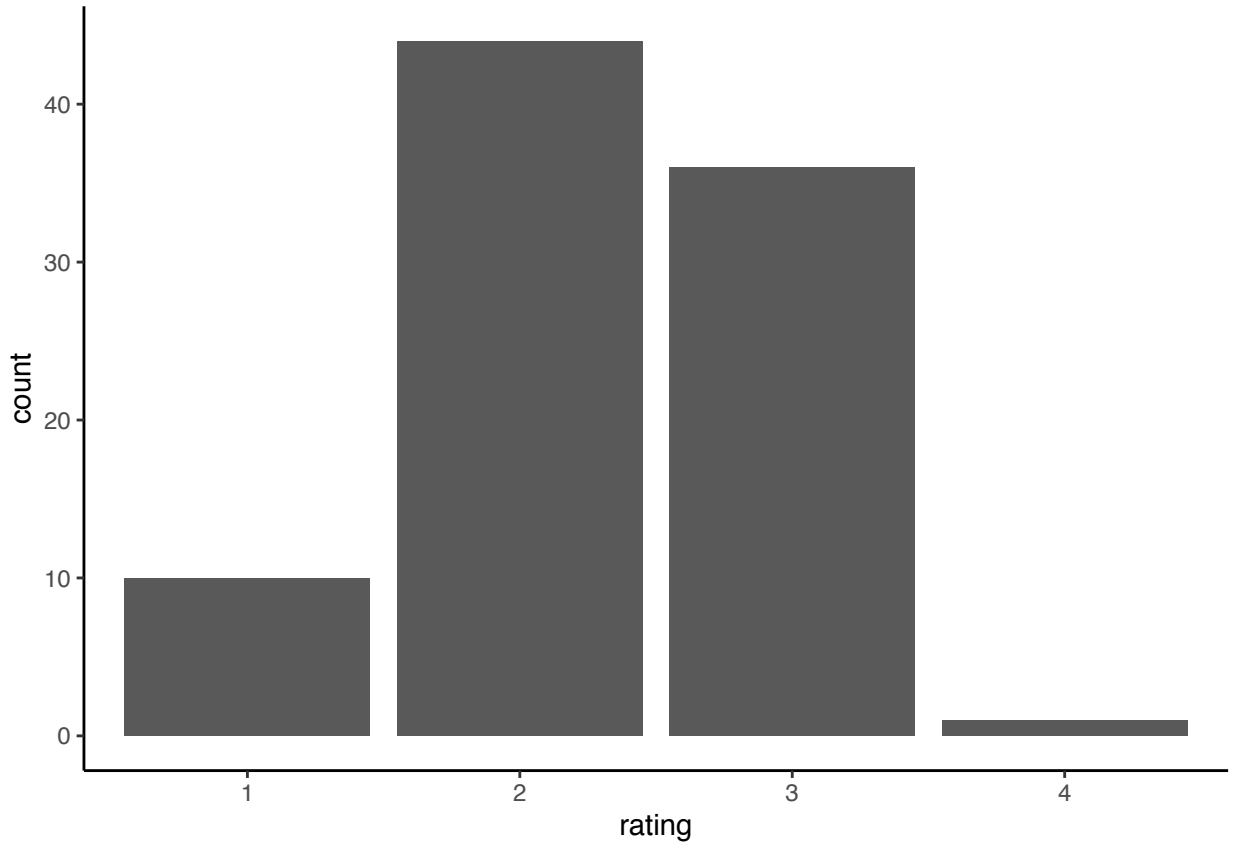
```r
rater_3 <- data.frame(
  rating = factor(c("1","2","3","4")),
  count = c(12,50,29,0)
)
rater_3
```

```
##   rating count
## 1      1    12
## 2      2    50
## 3      3    29
## 4      4     0
```

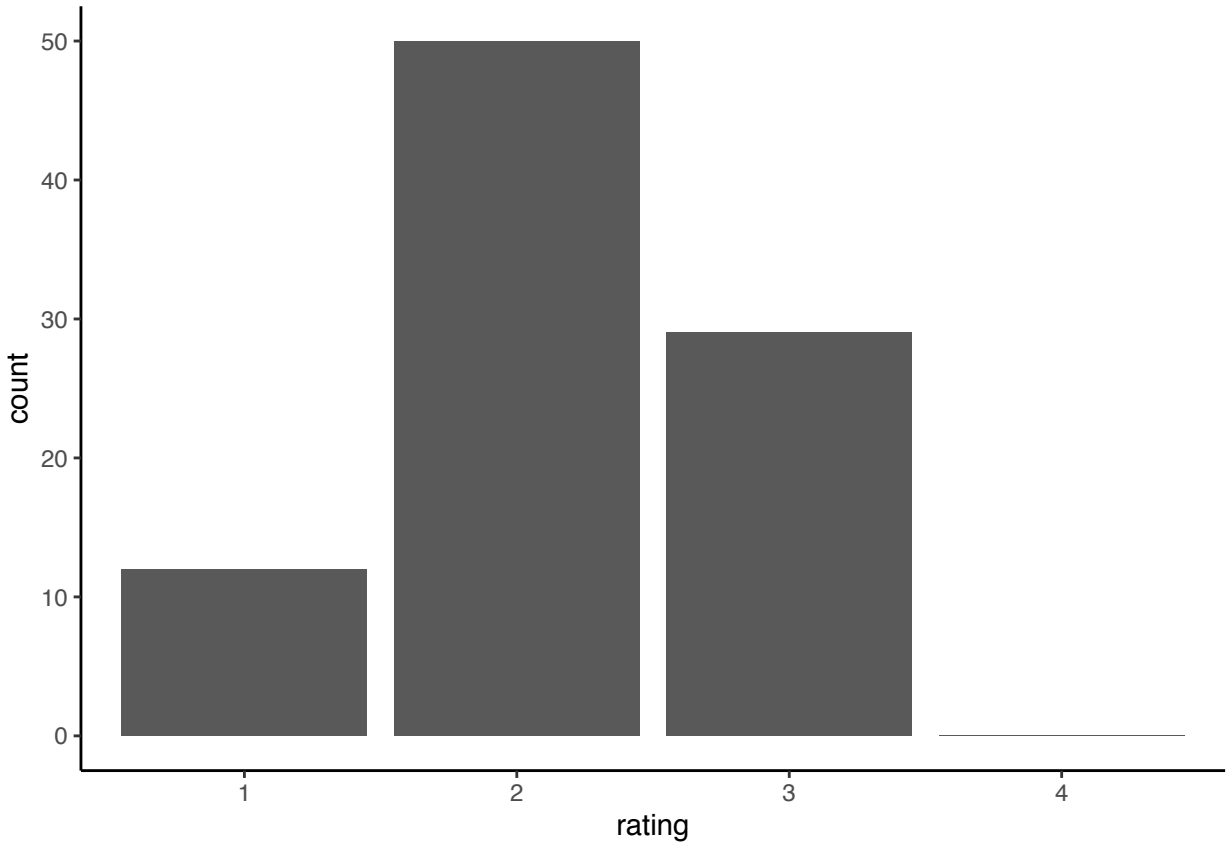
```r
# the distribution of ratings by each rater
par(mfrow=c(2,2))

ggplot(data=rater_1, aes(x=rating, y=count)) +
  geom_bar(stat="identity") + theme_classic()
```




```
ggplot(data=rater_2, aes(x=rating, y=count)) +
  geom_bar(stat="identity") + theme_classic()
```



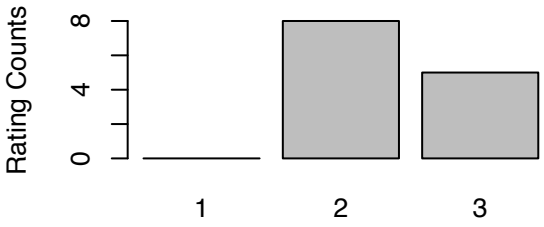

```
ggplot(data=rater_3, aes(x=rating, y=count)) +
  geom_bar(stat="identity") + theme_classic()
```



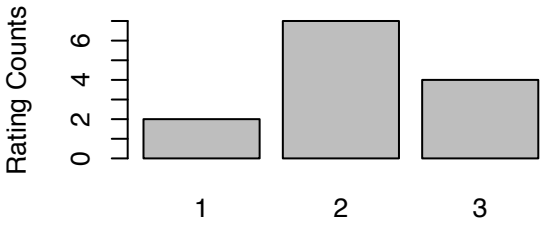
Comparing the distribution of three raters rating on different rubrics, we can observe that the distribution of these ratings given by each rater is pretty much indistinguishable from the other users. No rater tends to give especially high or low ratings.

Researching on Whether Raters Agree on the Score

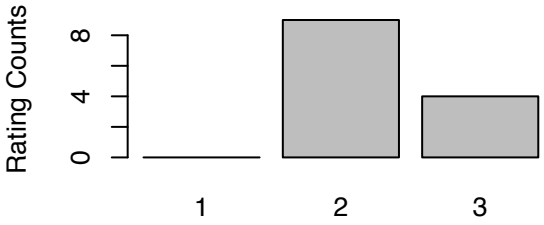
```
# rating on Research Question by each rater
par(mfrow=c(2,2))
plot(ratings_sub_score1$RsrchQ,
     xlab="Rating on Research Question by Rater1",ylab="Rating Counts")
plot(ratings_sub_score2$RsrchQ,
     xlab="Rating on Research Question by Rater2",ylab="Rating Counts")
plot(ratings_sub_score3$RsrchQ,
     xlab="Rating on Research Question by Rater3",ylab="Rating Counts")
```



Rating on Research Question by Rater1

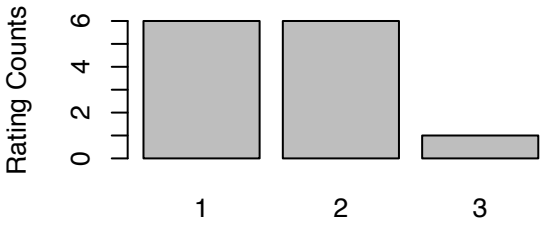


Rating on Research Question by Rater2

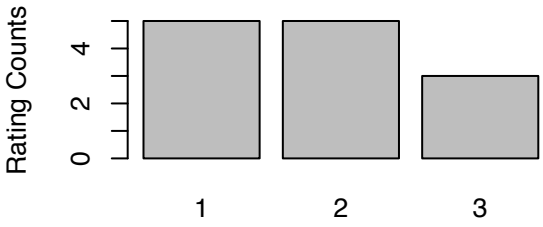


Rating on Research Question by Rater3

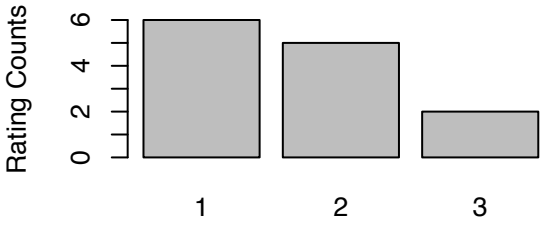
```
# rating on Critique Design by each rater
par(mfrow=c(2,2))
plot(ratings_sub_score1$CritDes,
     xlab="Rating on Critique Design by Rater1",ylab="Rating Counts")
plot(ratings_sub_score2$CritDes,
     xlab="Rating on Critique Design by Rater2",ylab="Rating Counts")
plot(ratings_sub_score3$CritDes,
     xlab="Rating on Critique Design by Rater3",ylab="Rating Counts")
```



Rating on Critique Design by Rater1

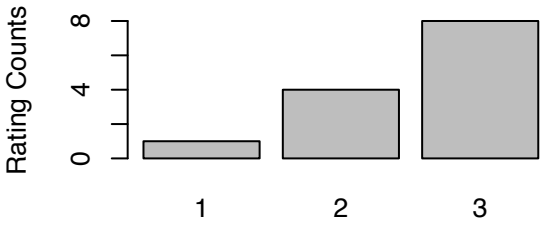


Rating on Critique Design by Rater2

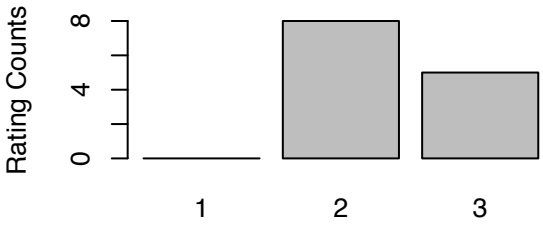


Rating on Critique Design by Rater3

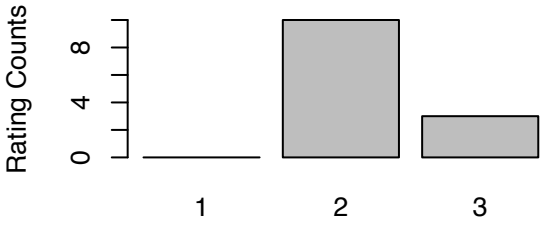
```
# rating on Initial EDA by each rater
par(mfrow=c(2,2))
plot(ratings_sub_score1$InitEDA,
     xlab="Rating on Initial EDA by Rater1",ylab="Rating Counts")
plot(ratings_sub_score2$InitEDA,
     xlab="Rating on Initial EDA by Rater2",ylab="Rating Counts")
plot(ratings_sub_score3$InitEDA,
     xlab="Rating on Initial EDA by Rater3",ylab="Rating Counts")
```



Rating on Initial EDA by Rater1

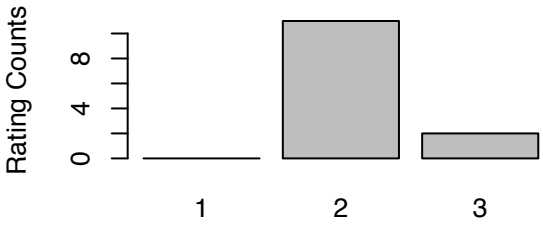


Rating on Initial EDA by Rater2

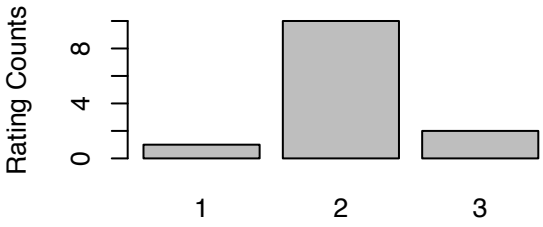


Rating on Initial EDA by Rater3

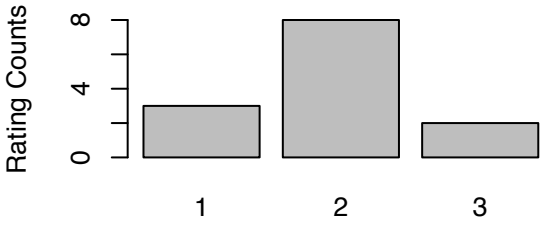
```
# rating on Select Method(s) by each rater
par(mfrow=c(2,2))
plot(ratings_sub_score1$SelMeth,
     xlab="Rating on Select Method(s) by Rater1",ylab="Rating Counts")
plot(ratings_sub_score2$SelMeth,
     xlab="Rating on Select Method(s) by Rater2",ylab="Rating Counts")
plot(ratings_sub_score3$SelMeth,
     xlab="Rating on Select Method(s) by Rater3",ylab="Rating Counts")
```



Rating on Select Method(s) by Rater1

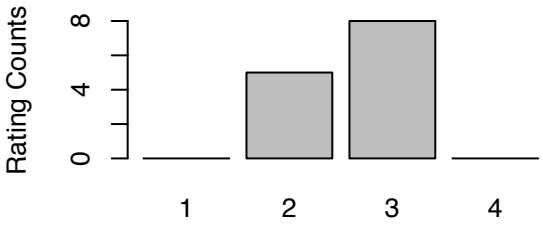


Rating on Select Method(s) by Rater2

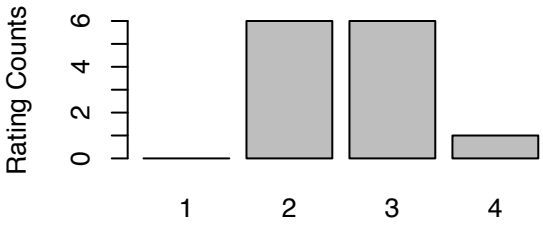


Rating on Select Method(s) by Rater3

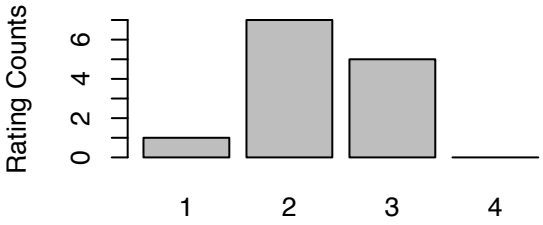
```
# rating on Interpret Results by each rater
par(mfrow=c(2,2))
plot(ratings_sub_score1$InterpRes,
     xlab="Rating on Interpret Results by Rater1",ylab="Rating Counts")
plot(ratings_sub_score2$InterpRes,
     xlab="Rating on Interpret Results by Rater2",ylab="Rating Counts")
plot(ratings_sub_score3$InterpRes,
     xlab="Rating on Interpret Results by Rater3",ylab="Rating Counts")
```



Rating on Interpret Results by Rater1

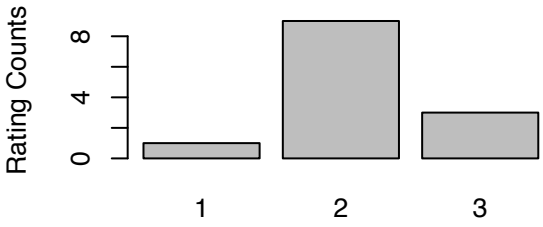


Rating on Interpret Results by Rater2

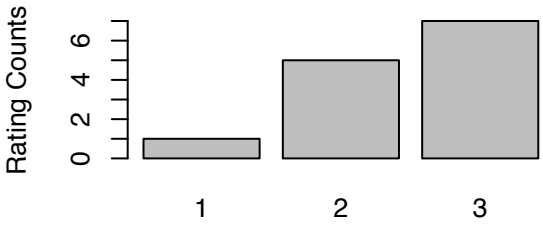


Rating on Interpret Results by Rater3

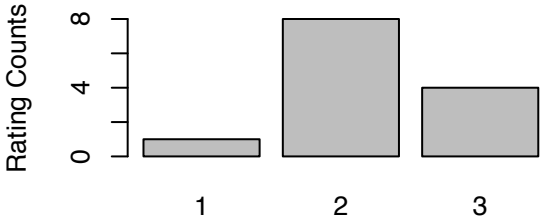
```
# rating on Visual Organization by each rater
par(mfrow=c(2,2))
plot(ratings_sub_score1$VisOrg,
     xlab="Rating on Visual Organization by Rater1",ylab="Rating Counts")
plot(ratings_sub_score2$VisOrg,
     xlab="Rating on Visual Organization by Rater2",ylab="Rating Counts")
plot(ratings_sub_score3$VisOrg,
     xlab="Rating on Visual Organization by Rater3",ylab="Rating Counts")
```



Rating on Visual Organization by Rater1

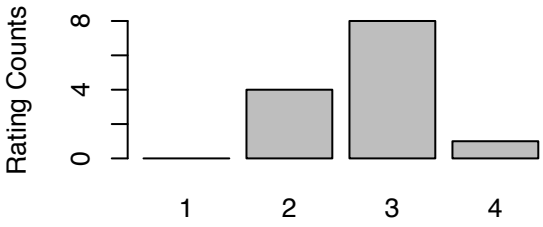


Rating on Visual Organization by Rater2

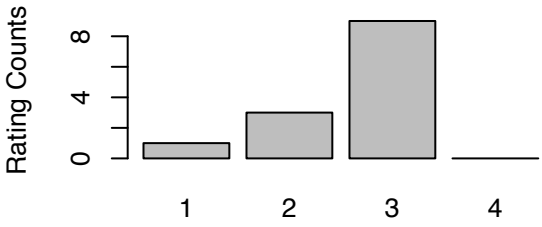


Rating on Visual Organization by Rater3

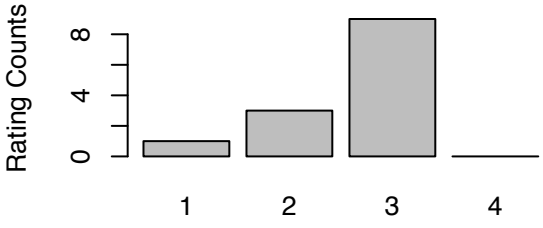
```
# rating on Text Organization by each rater
par(mfrow=c(2,2))
plot(ratings_sub_score1$TxtOrg,
     xlab="Rating on Text Organization by Rater1",ylab="Rating Counts")
plot(ratings_sub_score2$TxtOrg,
     xlab="Rating on Text Organization by Rater2",ylab="Rating Counts")
plot(ratings_sub_score3$TxtOrg,
     xlab="Rating on Text Organization by Rater3",ylab="Rating Counts")
```



Rating on Text Organization by Rater1



Rating on Text Organization by Rater2



Rating on Text Organization by Rater3

```
# calculate ICC's as a measure of rater agreement
names(tall)
```

```
## [1] "X"        "Rater"    "Artifact" "Repeated" "Semester" "Sex"      "Rubric"
## [8] "Rating"
```

```
# group the ratings
common <- tall[grep("O",tall$Artifact),]
head(common)
```

```
##     X Rater Artifact Repeated Semester Sex Rubric Rating
## 1   1     3       O5        1      F19   M RsrchQ      3
## 2   2     3       O7        1      F19   F RsrchQ      3
## 3   3     3       O9        1      S19   F RsrchQ      2
## 4   4     3       O8        1      S19   M RsrchQ      2
## 10 10     3      O10        1      F19   F RsrchQ      2
## 11 11     3      O13        1      F19   M RsrchQ      2
```

```
dim(common)
```

```
## [1] 273   8
```

Calculating ICC on each rubric (Research Question 2)

```
common$Rater <- as.factor(common$Rater)
common$Artifact <- as.factor(common$Artifact)
common$Semester <- as.factor(common$Semester)
common$Sex <- as.factor(common$Sex)
```

```
# ICC on Research Question
RsrchQ.ratings <- common[common$Rubric=="RsrchQ",]
RsrchQ_1 <- lmer(Rating ~ 1 + (1|Rater), data=RsrchQ.ratings)
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(RsrchQ_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Rater)
##    Data: RsrchQ.ratings
##
## REML criterion at convergence: 67.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2912 -0.5041 -0.5041  1.2831  1.2831
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Rater    (Intercept) 0.0000   0.0000
##  Residual             0.3131   0.5595
## Number of obs: 39, groups:  Rater, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.2820     0.0896   25.47
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
RsrchQ_ICC_1 <- (0.0000)/(0.0000+0.3131)
RsrchQ_ICC_1
```

```
## [1] 0
```

Here, the correlation is very low, since knowing the rating on one student's artifact should not be a good predictor of the rating on another student's artifact.

```
# ICC on Research Question
RsrchQ_2 <- lmer(Rating ~ 1 + (1|Artifact), data=RsrchQ.ratings)
summary(RsrchQ_2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: RsrchQ.ratings
##
## REML criterion at convergence: 66.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3025 -0.5987 -0.3276  0.9696  1.6472
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.05983  0.2446
##  Residual             0.25641  0.5064
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.2821     0.1057   21.59
```

```
RsrchQ_ICC_2 <- (0.05983)/(0.05983+0.25641)
RsrchQ_ICC_2
```

```
## [1] 0.1891918
```

Now the ICC is the correlation between any two rater's ratings on the same artifact. If the raters are consistent with one another in how they rate, we would expect this correlation to be higher, Moreover, the between-raters correlation does tell us something useful about rater agreement: raters agree more when their correlations are higher. The ICC value of 0.189 here indicates that these raters do not agree much on rating the Research Question since the correlation is not relatively high.

```
# ICC on Critique Design
CritDes.ratings <- common[common$Rubric=="CritDes",]
CritDes_1 <- lmer(Rating ~ 1 + (1|Rater), data=CritDes.ratings)
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(CritDes_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Rater)
##    Data: CritDes.ratings
##
## REML criterion at convergence: 86.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.9922 -0.9922  0.3898  0.3898  1.7717
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Rater    (Intercept) 0.0000   0.0000
##  Residual             0.5236   0.7236
## Number of obs: 39, groups:  Rater, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   1.7179     0.1159   14.83
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
CritDes_ICC_1 <- (0.0000)/(0.0000+0.5236)
CritDes_ICC_1
```

```
## [1] 0
```

Here, the correlation is very low, since knowing the rating on one student's artifact should not be a good predictor of the rating on another student's artifact.

```
# ICC on Critique Design
CritDes_2 <- lmer(Rating ~ 1 + (1|Artifact), data=CritDes.ratings)
summary(CritDes_2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: CritDes.ratings
##
## REML criterion at convergence: 75.1
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9647 -0.4386 -0.2978  0.5318  2.1987
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3091   0.5560
##  Residual             0.2308   0.4804
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   1.7179     0.1723   9.969
```

```
CritDes_ICC_2 <- (0.3091)/(0.3091+0.2308)
CritDes_ICC_2
```

```
## [1] 0.5725134
```

The ICC value of 0.573 here indicates that these raters do not agree much on rating the Critique Design since the correlation is relatively low.

```
# ICC on Initial EDA
InitEDA.ratings <- common[common$Rubric=="InitEDA",]
InitEDA_1 <- lmer(Rating ~ 1 + (1|Rater), data=InitEDA.ratings)
summary(InitEDA_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Rater)
##    Data: InitEDA.ratings
##
## REML criterion at convergence: 65.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5616 -0.7083 -0.6965  1.1215  1.1451
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  Rater    (Intercept) 0.0009862 0.0314
##  Residual             0.2948718 0.5430
## Number of obs: 39, groups:  Rater, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.38462    0.08882   26.85
```

```
InitEDA_ICC_1 <- (0.0009862)/(0.0009862+0.2948718)
InitEDA_ICC_1
```

```
## [1] 0.003333356
```

Here, the correlation is very low, since knowing the rating on one student's artifact should not be a good predictor of the rating on another student's artifact.

```
# ICC on Initial EDA
InitEDA_2 <- lmer(Rating ~ 1 + (1|Artifact), data=InitEDA.ratings)
```

```
summary(InitEDA_2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InitEDA.ratings
##
## REML criterion at convergence: 56.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1670 -0.2504 -0.2504  0.4006  1.6663
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.1496   0.3867
##  Residual             0.1538   0.3922
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.3846     0.1243   19.18
```

```
InitEDA_ICC_2 <- (0.1496)/(0.1496+0.1538)
InitEDA_ICC_2
```

```
## [1] 0.4930784
```

The ICC value of 0.493 here indicates that these raters do not agree much on rating the Initial EDA since the correlation is relatively low.

```
# ICC on Select Method(s)
SelMeth.ratings <- common[common$Rubric=="SelMeth",]
SelMeth_1 <- lmer(Rating ~ 1 + (1|Rater), data=SelMeth.ratings)
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(SelMeth_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Rater)
##    Data: SelMeth.ratings
##
## REML criterion at convergence: 60.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0599 -0.1005 -0.1005 -0.1005  1.8590
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Rater    (Intercept) 0.0000   0.0000
##  Residual             0.2605   0.5104
## Number of obs: 39, groups:  Rater, 3
##
## Fixed effects:
##             Estimate Std. Error t value
```

```
## (Intercept)  2.05128    0.08172    25.1
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
SelMeth_ICC_1 <- (0.0000)/(0.0000+0.2605)
SelMeth_ICC_1
```

```
## [1] 0
```

Here, the correlation is very low, since knowing the rating on one student's artifact should not be a good predictor of the rating on another student's artifact.

```
# ICC on Select Method(s)
SelMeth_2 <- lmer(Rating ~ 1 + (1|Artifact), data=SelMeth.ratings)
summary(SelMeth_2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: SelMeth.ratings
##
## REML criterion at convergence: 50.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.11366 -0.03357 -0.03357  0.62101  2.04652
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.1396   0.3736
##  Residual             0.1282   0.3581
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.0513     0.1184   17.32
```

```
SelMeth_ICC_2 <- (0.1396)/(0.1396+0.1282)
SelMeth_ICC_2
```

```
## [1] 0.5212845
```

The ICC value of 0.521 here indicates that these raters do not agree much on rating the Select Method(s) since the correlation is relatively low.

```
# ICC on Interpret Results
InterpRes.ratings <- common[common$Rubric=="InterpRes",]
InterpRes_1 <- lmer(Rating ~ 1 + (1|Rater), data=InterpRes.ratings)
summary(InterpRes_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Rater)
##    Data: InterpRes.ratings
##
## REML criterion at convergence: 72.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4822 -0.8773  0.7917  0.7917  2.4608
```

```
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Rater    (Intercept) 0.003945 0.06281
##  Residual             0.358974 0.59914
## Number of obs: 39, groups:  Rater, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.5128     0.1026    24.5
```

```
InterpRes_ICC_1 <- (0.003945)/(0.003945+0.358974)
InterpRes_ICC_1
```

```
## [1] 0.01087019
```

Here, the correlation is very low, since knowing the rating on one student's artifact should not be a good predictor of the rating on another student's artifact.

```
# ICC on Interpret Results
InterpRes_2 <- lmer(Rating ~ 1 + (1|Artifact), data=InterpRes.ratings)
summary(InterpRes_2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InterpRes.ratings
##
## REML criterion at convergence: 71.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0965 -0.8061  0.4844  0.7806  2.6635
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.08405  0.2899
##  Residual             0.28205  0.5311
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    2.513      0.117   21.47
```

```
InterpRes_ICC_2 <- (0.08405)/(0.08405+0.28205)
InterpRes_ICC_2
```

```
## [1] 0.2295821
```

The ICC value of 0.230 here indicates that these raters do not agree much on rating the Interpret Results since the correlation is relatively not high.

```
# ICC on Visual Organization
VisOrg.ratings <- common[common$Rubric=="VisOrg",]
VisOrg_1 <- lmer(Rating ~ 1 + (1|Rater), data=VisOrg.ratings)
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(VisOrg_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Rater)
##    Data: VisOrg.ratings
##
## REML criterion at convergence: 73.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1200 -0.4664 -0.4664  1.1872  1.1872
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Rater    (Intercept) 0.0000   0.0000
##  Residual             0.3657   0.6047
## Number of obs: 39, groups:  Rater, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.28205    0.09684   23.57
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
VisOrg_ICC_1 <- (0.0000)/(0.0000+0.3657)
VisOrg_ICC_1
```

```
## [1] 0
```

Here, the correlation is very low, since knowing the rating on one student's artifact should not be a good predictor of the rating on another student's artifact.

```
# ICC on Visual Organization
VisOrg_2 <- lmer(Rating ~ 1 + (1|Artifact), data=VisOrg.ratings)
summary(VisOrg_2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: VisOrg.ratings
##
## REML criterion at convergence: 60.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5168 -0.7176 -0.1341  0.3414  1.7241
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.2236   0.4729
##  Residual             0.1538   0.3922
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.2821     0.1454   15.69
```

```
VisOrg_ICC_2 <- (0.2236)/(0.2236+0.1538)
VisOrg_ICC_2
```

```
## [1] 0.5924748
```

The ICC value of 0.592 here indicates that these raters do agree much on rating the Visual Organization since the correlation is relatively high.

```
# ICC on Text Organization
TxtOrg.ratings <- common[common$Rubric=="TxtOrg",]
TxtOrg_1 <- lmer(Rating ~ 1 + (1|Rater), data=TxtOrg.ratings)
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(TxtOrg_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Rater)
##    Data: TxtOrg.ratings
##
## REML criterion at convergence: 75.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6827 -1.0731  0.5365  0.5365  2.1462
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Rater    (Intercept) 0.000    0.0000
##  Residual             0.386    0.6213
## Number of obs: 39, groups:  Rater, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.66667    0.09948   26.81
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
TxtOrg_ICC_1 <- (0.0000)/(0.0000+0.386)
TxtOrg_ICC_1
```

```
## [1] 0
```

Here, the correlation is very low, since knowing the rating on one student's artifact should not be a good predictor of the rating on another student's artifact.

```
# ICC on Text Organization
TxtOrg_2 <- lmer(Rating ~ 1 + (1|Artifact), data=TxtOrg.ratings)
summary(TxtOrg_2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: TxtOrg.ratings
##
## REML criterion at convergence: 74.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.6943 -0.7698  0.3849  0.3849  2.5019
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.05556  0.2357
##  Residual             0.33333  0.5774
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.6667     0.1132   23.55
```

```
TxtOrg_ICC_2 <- (0.05556)/(0.05556+0.33333)
TxtOrg_ICC_2
```

```
## [1] 0.1428682
```

The ICC value of 0.143 here indicates that these raters do not agree much on rating the Text Organization since the correlation is not relatively high.

The ICC's can help us determine whether the raters are generally in agreement on each rubric, but they cannot tell us which raters might be contributing to disagreement.Then, we want to make a 2-way table of counts for the ratings of each pair of raters, on each rubric to identify which rater is agree with which rater on each rubric.

Calculating PEA on each rubric (Research Question 2)

```
# compute exact agreement between any two raters and on each rubric
# cross-classifying the ratings that each pair of raters gives
# on the subset of 13 artifacts seen by each rater
repeated <- ratings[ratings$Repeated==1,]
```

```
# rating on research questions
raters_1_and_2_on_RsrchQ <- data.frame(r1=repeated$RsrchQ[repeated$Rater==1],
                                       r2=repeated$RsrchQ[repeated$Rater==2],
                                       a1=repeated$Artifact[repeated$Rater==1],
                                       a2=repeated$Artifact[repeated$Rater==2]
                                       )
```

```
r1 <- factor(raters_1_and_2_on_RsrchQ$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_RsrchQ$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
##   1 0 0 0 0
##   2 1 4 3 0
##   3 1 3 1 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Research Question between rater 1 and rater 2 is (4+1)/(1+4+3+1+3+1)=38.5

```
# rating on research questions
raters_1_and_3_on_RsrchQ <- data.frame(r1=repeated$RsrchQ[repeated$Rater==1],
                                       r3=repeated$RsrchQ[repeated$Rater==3],
                                       a1=repeated$Artifact[repeated$Rater==1],
```

```
                                        a3=repeated$Artifact[repeated$Rater==3]
                                        )
```

```
r1 <- factor(raters_1_and_3_on_RsrchQ$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_RsrchQ$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 0 0 0 0
##   2 0 7 1 0
##   3 0 2 3 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Research Question between rater 1 and rater 3 is (7+3)/(7+1+2+3)=76.9

```
# rating on research questions
raters_2_and_3_on_RsrchQ <- data.frame(r2=repeated$RsrchQ[repeated$Rater==2],
                                        r3=repeated$RsrchQ[repeated$Rater==3],
                                        a2=repeated$Artifact[repeated$Rater==2],
                                        a3=repeated$Artifact[repeated$Rater==3]
                                        )
```

```
r2 <- factor(raters_2_and_3_on_RsrchQ$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_RsrchQ$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 0 2 0 0
##   2 0 5 2 0
##   3 0 2 2 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Research Question between rater 2 and rater 3 is (5+2)/(2+5+2+2+2)=53.8

For rating on research question, rater 1 does not quite agree with rater 2.

```
# rating on critique design
raters_1_and_2_on_CritDes <- data.frame(r1=repeated$CritDes[repeated$Rater==1],
                                        r2=repeated$CritDes[repeated$Rater==2],
                                        a1=repeated$Artifact[repeated$Rater==1],
                                        a2=repeated$Artifact[repeated$Rater==2]
                                        )
```

```
r1 <- factor(raters_1_and_2_on_CritDes$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_CritDes$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
##   1 3 2 1 0
##   2 2 3 1 0
##   3 0 0 1 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Critique Design between rater 1 and rater 2 is (3+3+1)/(3+2+1+2+3+1+1)=53.8

```
# rating on critique design
raters_1_and_3_on_CritDes <- data.frame(r1=repeated$CritDes[repeated$Rater==1],
                                        r3=repeated$CritDes[repeated$Rater==3],
                                        a1=repeated$Artifact[repeated$Rater==1],
                                        a3=repeated$Artifact[repeated$Rater==3]
                                        )
```

```
r1 <- factor(raters_1_and_3_on_CritDes$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_CritDes$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 4 2 0 0
##   2 2 3 1 0
##   3 0 0 1 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Critique Design between rater 1 and rater 3 is (4+3+1)/(4+2+1+2+3+1)=61.5

```
# rating on critique design
raters_2_and_3_on_CritDes <-data.frame(r2=repeated$CritDes[repeated$Rater==2],
                                        r3=repeated$CritDes[repeated$Rater==3],
                                        a3=repeated$Artifact[repeated$Rater==3]
                                        )
```

```
r2 <- factor(raters_2_and_3_on_CritDes$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_CritDes$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 5 0 0 0
##   2 1 3 1 0
##   3 0 2 1 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Critique Design between rater 2 and rater 3 is (5+3+1)/(5+2+1+1+3+1)=69.2

For rating on critique design, there is no obvious disagreement between raters.

```
# rating on initial EDA
raters_1_and_2_on_InitEDA <-data.frame(r1=repeated$InitEDA[repeated$Rater==1],
                                        r2=repeated$InitEDA[repeated$Rater==2],
                                       a1=repeated$Artifact[repeated$Rater==1],
                                        a2=repeated$Artifact[repeated$Rater==2]
                                        )
```

```
r1 <- factor(raters_1_and_2_on_InitEDA$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_InitEDA$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
```

```
##   1 0 1 0 0
##   2 0 4 0 0
##   3 0 3 5 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Initial EDA between rater 1 and rater 2 is (5+4)/(5+4+1+3)=69.2

```
# rating on initial EDA
raters_1_and_3_on_InitEDA <-data.frame(r1=repeated$InitEDA[repeated$Rater==1],
                                       r3=repeated$InitEDA[repeated$Rater==3],
                                      a1=repeated$Artifact[repeated$Rater==1],
                                       a3=repeated$Artifact[repeated$Rater==3]
                                       )
```

```
r1 <- factor(raters_1_and_3_on_InitEDA$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_InitEDA$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 0 1 0 0
##   2 0 4 0 0
##   3 0 5 3 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Initial EDA between rater 1 and rater 3 is (3+4)/(5+4+1+3)=53.8

```
# rating on initial EDA
raters_2_and_3_on_InitEDA <-data.frame(r2=repeated$InitEDA[repeated$Rater==2],
                                       r3=repeated$InitEDA[repeated$Rater==3],
                                       a3=repeated$Artifact[repeated$Rater==3]
                                       )
```

```
r2 <- factor(raters_2_and_3_on_InitEDA$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_InitEDA$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 0 0 0 0
##   2 0 8 0 0
##   3 0 2 3 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Initial EDA between rater 2 and rater 3 is (3+8)/(8+2+3)=84.6

For rating on initial EDA, there is no obvious disagreement between raters.

```
# rating on select method(s)
raters_1_and_2_on_SelMeth <-data.frame(r1=repeated$SelMeth[repeated$Rater==1],
                                       r2=repeated$SelMeth[repeated$Rater==2],
                                      a1=repeated$Artifact[repeated$Rater==1],
                                       a2=repeated$Artifact[repeated$Rater==2]
                                       )
```

```
r1 <- factor(raters_1_and_2_on_SelMeth$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_SelMeth$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
```

```
## r1   1  2  3  4
##   1  0  0  0  0
##   2  1 10  0  0
##   3  0  0  2  0
##   4  0  0  0  0
```

The percent of exact agreement rate on rating Select Method(s) between rater 1 and rater 2 is (10+2)/(10+2+1)=92.3

```
# rating on select method(s)
raters_1_and_3_on_SelMeth <-data.frame(r1=repeated$SelMeth[repeated$Rater==1],
                                       r3=repeated$SelMeth[repeated$Rater==3],
                                      a1=repeated$Artifact[repeated$Rater==1],
                                       a3=repeated$Artifact[repeated$Rater==3]
                                       )
```

```
r1 <- factor(raters_1_and_3_on_SelMeth$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_SelMeth$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 0 0 0 0
##   2 3 7 1 0
##   3 0 1 1 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Select Method(s) between rater 1 and rater 3 is (7+1)/(3+7+1+1+1)=61.5

```
# rating on select method(s)
raters_2_and_3_on_SelMeth <-data.frame(r2=repeated$SelMeth[repeated$Rater==2],
                                       r3=repeated$SelMeth[repeated$Rater==3],
                                      a2=repeated$Artifact[repeated$Rater==2],
                                       a3=repeated$Artifact[repeated$Rater==3]
                                       )
```

```
r2 <- factor(raters_2_and_3_on_SelMeth$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_SelMeth$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 1 0 0 0
##   2 2 7 1 0
##   3 0 1 1 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Select Method(s) between rater 1 and rater 3 is (1+7+1)/(1+2+7+1+1+1)=69.2

For rating on select method(s), there is no obvious disagreement between raters.

```
# rating on interpret results
raters_1_and_2_on_InterpRes<-
  data.frame(r1=repeated$InterpRes[repeated$Rater==1],
             r2=repeated$InterpRes[repeated$Rater==2],
             a1=repeated$Artifact[repeated$Rater==1],
             a2=repeated$Artifact[repeated$Rater==2]
```

```
            )
```

```
r1 <- factor(raters_1_and_2_on_InterpRes$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_InterpRes$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
##   1 0 0 0 0
##   2 0 3 1 1
##   3 0 3 5 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Interpret Results between rater 1 and rater 2 is (3+5)/(3+5+3+1+1)=61.5

```
# rating on interpret results
raters_1_and_3_on_InterpRes<-
  data.frame(r1=repeated$InterpRes[repeated$Rater==1],
             r3=repeated$InterpRes[repeated$Rater==3],
             a1=repeated$Artifact[repeated$Rater==1],
             a3=repeated$Artifact[repeated$Rater==3]
             )
```

```
r1 <- factor(raters_1_and_3_on_InterpRes$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_InterpRes$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 0 0 0 0
##   2 1 3 1 0
##   3 0 4 4 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Interpret Results between rater 1 and rater 2 is (3+4)/(3+4+4+1+1)=53.8

```
# rating on interpret results
raters_2_and_3_on_InterpRes<-
  data.frame(r2=repeated$InterpRes[repeated$Rater==2],
             r3=repeated$InterpRes[repeated$Rater==3],
             a2=repeated$Artifact[repeated$Rater==2],
             a3=repeated$Artifact[repeated$Rater==3]
             )
```

```
r2 <- factor(raters_2_and_3_on_InterpRes$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_InterpRes$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 0 0 0 0
##   2 1 4 1 0
##   3 0 2 4 0
##   4 0 1 0 0
```

The percent of exact agreement rate on rating Interpret Results between rater 2 and rater 3 is

(4+4)/(1+4+1+2+4+1)=61.5

For rating on interpret results, there is no obvious disagreement between raters.

```
# rating on visual organization
raters_1_and_2_on_VisOrg<-
  data.frame(r1=repeated$VisOrg[repeated$Rater==1],
             r2=repeated$VisOrg[repeated$Rater==2],
             a1=repeated$Artifact[repeated$Rater==1],
             a2=repeated$Artifact[repeated$Rater==2]
             )
```

```
r1 <- factor(raters_1_and_2_on_VisOrg$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_VisOrg$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
##   1 1 0 0 0
##   2 0 4 5 0
##   3 0 1 2 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Visual Organization between rater 1 and rater 2 is (1+4+2)/(1+4+5+2+1)=53.8

```
# rating on visual organization
raters_1_and_3_on_VisOrg<-
  data.frame(r1=repeated$VisOrg[repeated$Rater==1],
             r3=repeated$VisOrg[repeated$Rater==3],
             a1=repeated$Artifact[repeated$Rater==1],
             a3=repeated$Artifact[repeated$Rater==3]
             )
```

```
r1 <- factor(raters_1_and_3_on_VisOrg$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_VisOrg$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 1 0 0 0
##   2 0 7 2 0
##   3 0 1 2 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Visual Organization between rater 1 and rater 3 is (1+7+2)/(1+7+2+2+1)=76.9

```
# rating on visual organization
raters_2_and_3_on_VisOrg<-
  data.frame(r2=repeated$VisOrg[repeated$Rater==2],
             r3=repeated$VisOrg[repeated$Rater==3],
             a2=repeated$Artifact[repeated$Rater==2],
             a3=repeated$Artifact[repeated$Rater==3]
             )
```

```
r2 <- factor(raters_2_and_3_on_VisOrg$r2,levels=1:4)
r3 <- factor(raters_1_and_3_on_VisOrg$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 1 0 0 0
##   2 0 5 0 0
##   3 0 3 4 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Visual Organization between rater 2 and rater 3 is (1+5+4)/(1+5+3+4)=76.9

For rating on visual organizations, there is no obvious disagreement between raters.

```
# rating on text organization
raters_1_and_2_on_TxtOrg<-
  data.frame(r1=repeated$TxtOrg[repeated$Rater==1],
             r2=repeated$TxtOrg[repeated$Rater==2],
             a1=repeated$Artifact[repeated$Rater==1],
             a2=repeated$Artifact[repeated$Rater==2]
             )
```

```
r1 <- factor(raters_1_and_2_on_TxtOrg$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_TxtOrg$r2,levels=1:4)
(t12 <- table(r1,r2))
```

```
##    r2
## r1  1 2 3 4
##   1 0 0 0 0
##   2 0 2 2 0
##   3 0 1 7 0
##   4 1 0 0 0
```

The percent of exact agreement rate on rating Text Organization between rater 1 and rater 2 is (2+7)/(2+2+1+7+1)=69.2

```
# rating on text organization
raters_1_and_3_on_TxtOrg<-
  data.frame(r1=repeated$TxtOrg[repeated$Rater==1],
             r3=repeated$TxtOrg[repeated$Rater==3],
             a1=repeated$Artifact[repeated$Rater==1],
             a3=repeated$Artifact[repeated$Rater==3]
             )
```

```
r1 <- factor(raters_1_and_3_on_TxtOrg$r1,levels=1:4)
r3 <- factor(raters_1_and_3_on_TxtOrg$r3,levels=1:4)
(t13 <- table(r1,r3))
```

```
##    r3
## r1  1 2 3 4
##   1 0 0 0 0
##   2 1 1 2 0
##   3 0 1 7 0
##   4 0 1 0 0
```

The percent of exact agreement rate on rating Text Organization between rater 1 and rater 3 is (1+7)/(1+2+1+1+7+1)=61.5

```
# rating on text organization
raters_2_and_3_on_TxtOrg<-
  data.frame(r2=repeated$TxtOrg[repeated$Rater==2],
```

```
            r3=repeated$TxtOrg[repeated$Rater==3],
            a2=repeated$Artifact[repeated$Rater==2],
            a3=repeated$Artifact[repeated$Rater==3]
            )
```

```
r2 <- factor(raters_2_and_3_on_TxtOrg$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_TxtOrg$r3,levels=1:4)
(t23 <- table(r2,r3))
```

```
##    r3
## r2  1 2 3 4
##   1 0 1 0 0
##   2 1 0 2 0
##   3 0 2 7 0
##   4 0 0 0 0
```

The percent of exact agreement rate on rating Text Organization between rater 1 and rater 3 is (7)/(1+2+1+2+7)=53.8

For rating on text organizations, there is no obvious disagreement between raters.

Researching on how Various Factors Related to the Ratings

Part 1: Adding fixed effects to the seven rubric-specific models using just the data from the 13 common artifacts that all three raters saw

```
#install.packages("LMERConvenienceFunctions")
#install.packages("RLRsim")
library(LMERConvenienceFunctions)
library(RLRsim)
```

```
tall.13 <- tall[grep("O",tall$Artifact),]

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

tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)
```

```
## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE
```

```
## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.7355 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.279 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
```

```
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
```

```
# backwards elimination with fitLMER.fnc() yields a model with raters only
formula(tmp.back_elim)
```

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

```
# 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)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall.13[tall.13$Rubric == "RsrchQ", ]
## Models:
## tmp.int_only: as.numeric(Rating) ~ (1 | Artifact)
## tmp.back_elim: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##               npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp.int_only     3 69.457 74.447 -31.728   63.457
## tmp.back_elim    5 72.018 80.335 -31.009   62.018 1.4391  2      0.487
```

```
# p-value
anova(tmp.int_only,tmp.back_elim)$"Pr(>Chisq)"[2]
```

```
## refitting model(s) with ML (instead of REML)
```

```
## [1] 0.4869707
```

We can observe that the intercept-only model is adequate here (the p-value is much greater than 0.05). 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 loop for every rubric case
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]
```

```r
  # choose the best model by comparing p-value
  if (pval<=0.05) {
    tmp_final <- tmp.back_elim
    } else {
      tmp_final <- tmp.single_intercept
    }
  # add the best model to list
  model.formula.13[[i]] <- formula(tmp_final)
  }
```

```
## ============================================================
## ===              backfitting fixed effects             ===
## ============================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.2229 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.1826 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ============================================================
## ===            forwardfitting random effects           ===
## ============================================================
##  ===       random slopes       ===
## ============================================================
## ===            re-backfitting fixed effects            ===
## ============================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
## refitting model(s) with ML (instead of REML)
## ============================================================
## ===              backfitting fixed effects             ===
## ============================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.8137 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.6429 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ============================================================
```

```
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
## refitting model(s) with ML (instead of REML)
## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.8294 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.2947 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
## refitting model(s) with ML (instead of REML)
## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.7355 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.279 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
```

```
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
## refitting model(s) with ML (instead of REML)
## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.9383 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.4287 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
## refitting model(s) with ML (instead of REML)
## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.5358 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.1319 >= 0.05
##     not part of higher-order interaction
```

```
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
```

```
## refitting model(s) with ML (instead of REML)
```

```
## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.1922 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.1078 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
```

```
## refitting model(s) with ML (instead of REML)
```

```
# print out the terms included in each rubric-specific model
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)
```

For the seven rubric-specific models using just the data from the 13 common artifacts that all three raters saw, we find that adding fixed effects does not improve the fit of any of the seven models. With the fact that we do not find any fixed effects are significant, we decide not to add any interaction terms or new random effects further, using only the data reduced to the 13 common rubrics.

Part 2: adding fixed effects to the seven rubric-specific models using tall data (all the data)

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

# eliminate two observations with missing data
# only do fitting and comparison on non missing data

# check these two rows contain missing data
tall[c(161,684),]
```

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

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

# for loop for every rubric case
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 by comparing p-value
  if (pval<=0.05) {
    tmp_final <- tmp.back_elim
    } else {
      tmp_final <- tmp.single_intercept
```

```
    }
  # add the best model to list
  model.formula.alldata[[i]] <- formula(tmp_final)
}
```

```
## ============================================================
## ===              backfitting fixed effects             ===
## ============================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.6474 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.3309 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ============================================================
## ===            forwardfitting random effects           ===
## ============================================================
##  ===       random slopes       ===
## ============================================================
## ===            re-backfitting fixed effects            ===
## ============================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
## refitting model(s) with ML (instead of REML)
## ============================================================
## ===              backfitting fixed effects             ===
## ============================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.8292 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.6014 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ============================================================
## ===            forwardfitting random effects           ===
## ============================================================
##  ===       random slopes       ===
## ============================================================
## ===            re-backfitting fixed effects            ===
## ============================================================
```

```
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
## refitting model(s) with ML (instead of REML)
## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.4701 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.2935 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
## refitting model(s) with ML (instead of REML)
## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.4446 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.3417 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
```

```
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
## refitting model(s) with ML (instead of REML)
## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
## refitting model(s) with ML (instead of REML)
## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.5925 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.1874 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
```

```
## pruning random effects structure ...
##   nothing to prune
```

```
## refitting model(s) with ML (instead of REML)
```

```
## ========================================================
## ===              backfitting fixed effects          ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.2186 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.1977 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
```

```
## refitting model(s) with ML (instead of REML)
```

```
# print out the terms included in each seven-rubric model
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 + Sex + (1 |
##     Artifact) - 1
##
## $TxtOrg
## as.numeric(Rating) ~ (1 | Artifact)
##
```

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

For the seven rubric specific models using the entire dataset, we find that adding fixed effects perform some different results. For InitEDA, RsrchQ, and TxtOrg, adding fixed effects does not improve the fit of those models, i.e. those models are just simple random-intercept models. However, for CritDes, InterpRes, and VisOrg, adding Rater and removing the intercept improves the fit of those models; for SelMeth, adding Semester and removing the intercept improves the fit of the model. Therefore, we think that for rubrics CritDes, InterpRes, and VisOrg, Rater is related to Ratings; but for only one rubric SelMeth, Semester is related to Ratings.

Part 3: trying interactions and new random effects for the seven rubric specific models using tall data (all the data)

For InitEDA, RsrchQ and SelMeth, 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 (1) if the fixed effects make sense to us; and (2) if there are any interactions or additional random effects to consider.

```
# Examine Selected Method(s)
selmeth_fla <- formula(model.formula.alldata[["SelMeth"]])
selmeth_tmp <- lmer(selmeth_fla,data=tall.nonmissing[tall.nonmissing$Rubric=="SelMeth",])
round(summary(selmeth_tmp)$coef,2) ## fixed effects and their t-values
```

```
##                   Estimate Std. Error t value
## as.factor(Rater)1     3.22       0.45    7.11
## as.factor(Rater)2     3.19       0.45    7.05
## as.factor(Rater)3     3.00       0.44    6.75
## SemesterS19          -0.32       0.10   -3.12
## SexF                 -1.04       0.45   -2.28
## SexM                 -0.91       0.45   -2.02
```

```
# now check to make sure we really need "Rater" as a factor...
selmeth_tmp.single_intercept <- update(selmeth_tmp, . ~ . + 1 - as.factor(Rater))
anova(selmeth_tmp.single_intercept,selmeth_tmp)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall.nonmissing[tall.nonmissing$Rubric == "SelMeth", ]
## Models:
## selmeth_tmp.single_intercept: as.numeric(Rating) ~ Semester + Sex + (1 | Artifact)
## selmeth_tmp: as.numeric(Rating) ~ as.factor(Rater) + Semester + Sex + (1 | Artifact) - 1
##                              npar    AIC    BIC  logLik deviance  Chisq Df
## selmeth_tmp.single_intercept    6 147.94 164.51 -67.968   135.94
## selmeth_tmp                     8 144.52 166.62 -64.260   128.52 7.4154  2
##                              Pr(>Chisq)
## selmeth_tmp.single_intercept
## selmeth_tmp                     0.02453 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## now check for fixed-effect interactions
## Since only Rater and Semester are involved, we only need to examine Rater*Semester
selmeth_tmp.fixed_interactions <- update(selmeth_tmp, . ~ . + as.factor(Rater)*Semester - Semester)

anova(selmeth_tmp,selmeth_tmp.fixed_interactions)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall.nonmissing[tall.nonmissing$Rubric == "SelMeth", ]
```

```
## Models:
## selmeth_tmp: as.numeric(Rating) ~ as.factor(Rater) + Semester + Sex + (1 | Artifact) - 1
## selmeth_tmp.fixed_interactions: as.numeric(Rating) ~ as.factor(Rater) + Sex + (1 | Artifact) + as.fa
##                                npar    AIC    BIC  logLik deviance  Chisq Df
## selmeth_tmp                       8 144.52 166.62 -64.260   128.52
## selmeth_tmp.fixed_interactions   10 145.77 173.40 -62.887   125.77 2.7467  2
##                                Pr(>Chisq)
## selmeth_tmp
## selmeth_tmp.fixed_interactions     0.2533
```

Here, it shows that 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, we should try (Rater|Artifact) and (Semester|Artifact).

Note what the first one, for model mA is: there are more random effects than there are observations in the data set, meaning that lmer() cannot fit a model. Thus, the model as.numeric(Rating) -1 + as.factor(Rater) + Semester + (1 | Artifact) + (Semseter | Artifact) isn't even possible, so no testing is needed.

Again, the model as.numeric(Rating) -1 + as.factor(Rater) + Semester + (1 | Artifact) + (as.factor(Rater) | Artifact) isn't even possible, so no testing is needed.

```
# final model for SelMeth
summary(selmeth_tmp)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + Semester + Sex + (1 |
##     Artifact) - 1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "SelMeth", ]
##
## REML criterion at convergence: 144.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.09631 -0.34555 -0.06849  0.33489  2.66067
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.09013  0.3002
##  Residual             0.10714  0.3273
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1   3.2227     0.4531   7.113
## as.factor(Rater)2   3.1946     0.4530   7.051
## as.factor(Rater)3   3.0000     0.4441   6.755
## SemesterS19        -0.3195     0.1025  -3.119
## SexF               -1.0352     0.4536  -2.282
## SexM               -0.9136     0.4523  -2.020
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2 a.(R)3 SmsS19 SexF
## as.fctr(R)2  0.981
## as.fctr(R)3  0.980  0.980
## SemesterS19  0.000  0.002  0.000
## SexF        -0.980 -0.980 -0.979 -0.097
```

```
## SexM        -0.981 -0.982 -0.982 -0.035  0.978
```

```
# Examine Critique Design
critdes_fla <- formula(model.formula.alldata[["CritDes"]])
critdes_tmp <- lmer(critdes_fla,data=tall.nonmissing[tall.nonmissing$Rubric=="CritDes",])
round(summary(critdes_tmp)$coef,2) ## fixed effects and their t-values
```

```
##                   Estimate Std. Error t value
## as.factor(Rater)1     1.69       0.12   13.99
## as.factor(Rater)2     2.12       0.12   17.34
## as.factor(Rater)3     1.91       0.12   15.83
```

```
# now check to make sure we really need "Rater" as a factor...
critdes_tmp.single_intercept <- update(critdes_tmp, . ~ . + 1 - as.factor(Rater))
anova(critdes_tmp.single_intercept,critdes_tmp)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall.nonmissing[tall.nonmissing$Rubric == "CritDes", ]
## Models:
## critdes_tmp.single_intercept: as.numeric(Rating) ~ (1 | Artifact)
## critdes_tmp: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##                              npar    AIC    BIC  logLik deviance  Chisq Df
## critdes_tmp.single_intercept    3 280.86 289.12 -137.43   274.86
## critdes_tmp                     5 276.86 290.62 -133.43   266.86 7.9996  2
##                              Pr(>Chisq)
## critdes_tmp.single_intercept
## critdes_tmp                     0.01832 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now check for fixed-effect interactions. Since only Rater is involved, then no checking needed.

Finally we check for random effects. Note what the first one, for model mA is: there are more random effects than there are observations in the data set, meaning that lmer() cannot fit a model. Thus, the model isn't even possible, so no testing is needed.

```
# final model for CritDes
summary(critdes_tmp)
```

```
## 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: 274.2
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.54697 -0.50107 -0.08068  0.63782  1.61697
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.4401   0.6634
##  Residual             0.2475   0.4975
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##                   Estimate Std. Error t value
```

```
## as.factor(Rater)1   1.6926     0.1210   13.99
## as.factor(Rater)2   2.1184     0.1222   17.34
## as.factor(Rater)3   1.9144     0.1210   15.83
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2
## as.fctr(R)2 0.245
## as.fctr(R)3 0.243  0.245
```

```
# Examine Interpret Result
interpres_fla <- formula(model.formula.alldata[["InterpRes"]])
interpres_tmp <- lmer(interpres_fla,data=tall.nonmissing[tall.nonmissing$Rubric=="InterpRes",])
round(summary(interpres_tmp)$coef,2) ## fixed effects and their t-values
```

```
##                   Estimate Std. Error t value
## as.factor(Rater)1     2.71       0.09   30.19
## as.factor(Rater)2     2.59       0.09   28.87
## as.factor(Rater)3     2.16       0.09   24.12
```

```
# now check to make sure we really need "Rater" as a factor...
interpres_tmp.single_intercept <- update(interpres_tmp, . ~ . + 1 - as.factor(Rater))
anova(interpres_tmp.single_intercept,interpres_tmp)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall.nonmissing[tall.nonmissing$Rubric == "InterpRes", ]
## Models:
## interpres_tmp.single_intercept: as.numeric(Rating) ~ (1 | Artifact)
## interpres_tmp: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##                                 npar    AIC    BIC   logLik deviance  Chisq Df
## interpres_tmp.single_intercept    3 220.09 228.38 -107.048   214.09
## interpres_tmp                     5 203.66 217.47  -96.831   193.66 20.433  2
##                                 Pr(>Chisq)
## interpres_tmp.single_intercept
## interpres_tmp                    3.657e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now check for fixed-effect interactions. Since only Rater is involved, no need for checking.

Finally, we check for new random effects. Note that the first one, for model mA is: there are more random effects than there are observations in the data set, meaning that lmer() cannot fit a model. Thus, the model isn't even possible, so no testing is needed.

```
# final model for InterpRes
summary(interpres_tmp)
```

```
## 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: 202.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5101 -0.7484  0.3763  0.6532  2.6479
##
## Random effects:
```

```
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.06471  0.2544
##  Residual             0.25381  0.5038
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1  2.70517    0.08961   30.19
## as.factor(Rater)2  2.58701    0.08961   28.87
## as.factor(Rater)3  2.16116    0.08961   24.12
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2
## as.fctr(R)2 0.063
## as.fctr(R)3 0.063  0.063
```

```
# Examine Visual Organization
visorg_fla <- formula(model.formula.alldata[["VisOrg"]])
visorg_tmp <- lmer(visorg_fla,data=tall.nonmissing[tall.nonmissing$Rubric=="VisOrg",])
round(summary(visorg_tmp)$coef,2) ## fixed effects and their t-values
```

```
##                   Estimate Std. Error t value
## as.factor(Rater)1     2.38        0.1   24.67
## as.factor(Rater)2     2.65        0.1   27.75
## as.factor(Rater)3     2.30        0.1   24.06
```

```
# now check to make sure we really need "Rater" as a factor...
visorg_tmp.single_intercept <- update(visorg_tmp, . ~ . + 1 - as.factor(Rater))
anova(visorg_tmp.single_intercept,visorg_tmp)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall.nonmissing[tall.nonmissing$Rubric == "VisOrg", ]
## Models:
## visorg_tmp.single_intercept: as.numeric(Rating) ~ (1 | Artifact)
## visorg_tmp: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##                             npar    AIC    BIC  logLik deviance  Chisq Df
## visorg_tmp.single_intercept    3 228.95 237.21 -111.47   222.95
## visorg_tmp                     5 222.97 236.74 -106.48   212.97 9.9784  2
##                             Pr(>Chisq)
## visorg_tmp.single_intercept
## visorg_tmp                    0.006811 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now let's check for fixed-effect interactions. Since only Rater is involved, no need for checking.

Finally, we try adding new random effects. Note what the first one, for model mA is: there are more random effects than there are observations in the data set, meaning that lmer() cannot fit a model. Thus, the model isn't even possible, so no testing is needed.

```
# final model for VisOrg
summary(visorg_tmp)
```

```
## 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: 221.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5008 -0.3334 -0.2599  0.4108  1.8726
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.2937   0.5420
##  Residual             0.1454   0.3813
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1  2.38148    0.09652   24.67
## as.factor(Rater)2  2.65269    0.09558   27.75
## as.factor(Rater)3  2.29935    0.09558   24.06
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2
## as.fctr(R)2 0.265
## as.fctr(R)3 0.265  0.264
```

Based on the results of likelihood ratio test and t-values, we find that for rubric CritDes, InterpRes, and VisOrg, including Rater in the model is important. Since Rater is the only fixed effect included in those models, there is no need to try fixed effects interactions. Moreover, since there are more random effects than number of observations in the dataset, the model with the random intercept of Rater grouped by Artifact cannot be fit, so we decide not to include any new random intercepts into the model. Therefore, for rubric CritDes, InterpRes, and VisOrg, the final model includes Rater as a fixed effect, but no additional fixed interactions or random effects included.

Part 4: Trying to add fixed effects, interactions, and new random effects to the "combined" model Rating 1 + (0 + Rubric|Artifact), using tall data (all the data).

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (0 + Rubric | Artifact)
##    Data: tall.nonmissing
##
## REML criterion at convergence: 1481.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0247 -0.4970 -0.0754  0.5166  3.7824
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.6484   0.8053
##           RubricInitEDA   0.3779   0.6147   0.27
##           RubricInterpRes 0.2525   0.5025   0.02 0.79
##           RubricRsrchQ    0.1733   0.4163   0.40 0.51 0.74
##           RubricSelMeth   0.1034   0.3216   0.58 0.39 0.42 0.29
##           RubricTxtOrg    0.3946   0.6282   0.04 0.69 0.80 0.64 0.25
```

```
##           RubricVisOrg    0.3153   0.5615   0.19 0.78 0.77 0.60 0.31 0.79
##  Residual                 0.1942   0.4407
## Number of obs: 817, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.24698    0.04048   55.51
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00260717 (tol = 0.002, component 1)
```

```
# Try adding fixed effects with no interactions
comb.full <- update(comb.0, . ~ . + as.factor(Rater) + Semester + Sex + Repeated + Rubric)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0368127 (tol = 0.002, component 1)
```

```
summary(comb.full)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) +
##     Semester + Sex + Repeated + Rubric
##    Data: tall.nonmissing
##
## REML criterion at convergence: 1436.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1142 -0.5053 -0.0216  0.5145  3.8024
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.54865  0.7407
##           RubricInitEDA   0.34962  0.5913   0.47
##           RubricInterpRes 0.17506  0.4184   0.23 0.75
##           RubricRsrchQ    0.16854  0.4105   0.59 0.44 0.71
##           RubricSelMeth   0.06827  0.2613   0.40 0.61 0.74 0.41
##           RubricTxtOrg    0.26198  0.5118   0.34 0.62 0.71 0.57 0.67
##           RubricVisOrg    0.25592  0.5059   0.35 0.74 0.68 0.52 0.42 0.76
##  Residual                 0.18839  0.4340
## Number of obs: 817, groups:  Artifact, 91
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)         2.820361   0.388467   7.260
## as.factor(Rater)2   0.002027   0.054805   0.037
## as.factor(Rater)3  -0.174718   0.054961  -3.179
## SemesterS19        -0.174745   0.087851  -1.989
## SexF               -0.802780   0.383735  -2.092
## SexM               -0.792390   0.382742  -2.070
## Repeated           -0.074479   0.098554  -0.756
## RubricInitEDA       0.541301   0.094934   5.702
## RubricInterpRes     0.580815   0.100065   5.804
## RubricRsrchQ        0.456028   0.086782   5.255
## RubricSelMeth       0.162899   0.093287   1.746
## RubricTxtOrg        0.685792   0.098768   6.943
```

```
## RubricVisOrg       0.524270   0.098304   5.333

##
## Correlation matrix not shown by default, as p = 13 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it

## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.0368127 (tol = 0.002, component 1)
```

```
comb.back_elim <- fitLMER.fnc(comb.full, log.file.name = FALSE)
```

```
## Warning in fitLMER.fnc(comb.full, log.file.name = FALSE): Argument "ran.effects" is empty, which mea
## TRUE

## ============================================================
## ===              backfitting fixed effects             ===
## ============================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.091 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     removing term
##   iteration 2
##     p-value for term "Repeated" = 0.0861 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     removing term
## pruning random effects structure ...
##   nothing to prune
## ============================================================
## ===            forwardfitting random effects           ===
## ============================================================
##  ===       random slopes       ===
## ============================================================
## ===            re-backfitting fixed effects            ===
## ============================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE

## boundary (singular) fit: see ?isSingular

## pruning random effects structure ...
##   nothing to prune
```

```
# Continue to try interactions
comb.inter <- update(comb.back_elim, . ~ . + as.factor(Rater)*Semester*Rubric)
```

```
## boundary (singular) fit: see ?isSingular
```

```
ss <- getME(comb.inter,c("theta","fixef"))
comb.inter.u<- update(comb.inter,start=ss,
                      control=lmerControl(optimizer="bobyqa",
                                          optCtrl=list(maxfun=2e5)))
```

```
## boundary (singular) fit: see ?isSingular
```

```
# Backward elimination
comb.inter_elim <- fitLMER.fnc(comb.inter.u, log.file.name = FALSE)
```

```
## Warning in fitLMER.fnc(comb.inter.u, log.file.name = FALSE): Argument "ran.effects" is empty, which r
## TRUE
```

```
## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 3
##   iteration 1
##     p-value for term "as.factor(Rater):Semester:Rubric" = 0.5402 >= 0.05
##     not part of higher-order interaction
```

```
## boundary (singular) fit: see ?isSingular
```

```
##     removing term
## processing model terms of interaction level 2
##   iteration 2
##     p-value for term "as.factor(Rater):Semester" = 0.5569 >= 0.05
##     not part of higher-order interaction
```

```
## boundary (singular) fit: see ?isSingular
```

```
##     removing term
##   iteration 3
##     p-value for term "Semester:Rubric" = 0.0696 >= 0.05
##     not part of higher-order interaction
```

```
## boundary (singular) fit: see ?isSingular
```

```
##     removing term
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 2
##   all terms of interaction level 2 significant
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
```

```
## boundary (singular) fit: see ?isSingular
```

```
## pruning random effects structure ...
##   nothing to prune
```

```
anova(comb.back_elim,comb.inter_elim,comb.inter.u)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall.nonmissing
## Models:
## comb.back_elim: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric
## comb.inter_elim: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric
## comb.inter.u: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric + a
##                 npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## comb.back_elim    39 1475.2 1658.7 -698.58   1397.2
## comb.inter_elim   51 1465.5 1705.5 -681.76   1363.5 33.653 12   0.000765 ***
## comb.inter.u      71 1481.8 1815.9 -669.91   1339.8 23.694 20   0.256027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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

```
# consider random effects
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)
```

```
## boundary (singular) fit: see ?isSingular
```

```
# compare models we've selected before to the one with random effects added
anova(m0,mA)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Warning in commonArgs(par, fn, control, environment()): maxfun < 10 *
## length(par)^2 is not recommended.
```

```
## Data: tall.nonmissing
## Models:
## m0: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric + as.factor(
## mA: as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) | Artifact) + as.factor(Rate
##    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m0   51 1465.5 1705.5 -681.76   1363.5
## mA   57 1425.9 1694.1 -655.94   1311.9 51.624  6  2.219e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
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)
```

```
## boundary (singular) fit: see ?isSingular
```

```
anova(m0,mA)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: tall.nonmissing
## Models:
## m0: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric + as.factor(
## mA: as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + Semester | Artifact) + as.factor(Rater) + Se
```

```
##    npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## m0   51 1465.5 1705.5 -681.76   1363.5
## mA   54 1472.9 1727.0 -682.47   1364.9     0  3          1
```

```
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)
```

```
## boundary (singular) fit: see ?isSingular
```

```
# formula of the final model
formula(comb.final)
```

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

```
# summary of the final model
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: 1380.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.07857 -0.46641 -0.03094  0.45414  2.74724
##
## Random effects:
##  Groups     Name               Variance Std.Dev. Corr
##  Artifact   RubricCritDes      0.49340  0.7024
##             RubricInitEDA      0.31065  0.5574    0.32
##             RubricInterpRes    0.09975  0.3158    0.15  0.67
##             RubricRsrchQ       0.17689  0.4206    0.50  0.19  0.54
##             RubricSelMeth      0.03792  0.1947    0.16  0.22  0.38 -0.23
##             RubricTxtOrg       0.24190  0.4918    0.27  0.43  0.35  0.30  0.19
##             RubricVisOrg       0.22674  0.4762    0.18  0.50  0.44  0.27 -0.16
##  Artifact.1 as.factor(Rater)1 0.01407  0.1186
##             as.factor(Rater)2 0.11491  0.3390   -0.40
##             as.factor(Rater)3 0.10664  0.3266    0.40  0.68
##  Residual                     0.13438  0.3666
##
##
##
##
##
##
##
##   0.53
##
##
##
```

```
##
## Number of obs: 817, groups:  Artifact, 91
##
## Fixed effects:
##                                  Estimate Std. Error t value
## (Intercept)                       1.76432    0.11379  15.505
## as.factor(Rater)2                 0.36868    0.13912   2.650
## as.factor(Rater)3                 0.21242    0.12964   1.639
## SemesterS19                      -0.16354    0.07712  -2.121
## RubricInitEDA                     0.73728    0.12941   5.697
## RubricInterpRes                   0.98939    0.12713   7.783
## RubricRsrchQ                      0.72389    0.11747   6.162
## RubricSelMeth                     0.40801    0.12409   3.288
## RubricTxtOrg                      1.01338    0.12950   7.826
## RubricVisOrg                      0.65225    0.13287   4.909
## as.factor(Rater)2:RubricInitEDA  -0.29989    0.15574  -1.926
## as.factor(Rater)3:RubricInitEDA  -0.30213    0.15540  -1.944
## as.factor(Rater)2:RubricInterpRes -0.51407    0.15310  -3.358
## as.factor(Rater)3:RubricInterpRes -0.71655    0.15266  -4.694
## as.factor(Rater)2:RubricRsrchQ   -0.48810    0.14687  -3.323
## as.factor(Rater)3:RubricRsrchQ   -0.32775    0.14627  -2.241
## as.factor(Rater)2:RubricSelMeth  -0.38747    0.14989  -2.585
## as.factor(Rater)3:RubricSelMeth  -0.37982    0.14868  -2.555
## as.factor(Rater)2:RubricTxtOrg   -0.55191    0.15612  -3.535
## as.factor(Rater)3:RubricTxtOrg   -0.45490    0.15578  -2.920
## as.factor(Rater)2:RubricVisOrg   -0.10629    0.15817  -0.672
## as.factor(Rater)3:RubricVisOrg   -0.28019    0.15782  -1.775

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

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

Based on the results of likelihood ratio test and t-values, we find that for rubric SelMeth, including Semester in the model is important. Since there are more random effects than number of observations in the dataset, the model with the random intercept of Rater grouped by Artifact and the model with the random intercept of Semester grouped by Artifact cannot be fit, so we decide not to include any new random intercepts into the model. Therefore, for rubric SelMeth, the final model includes Rater and Semester as the fixed effects, but no additional fixed interactions or random effects included.

Based on the result of likelihood ratio tests and AIC values, we find that the final combined result includes Rater, Semester, Rubric, and the interaction of Rater and Rubric as fixed effects and Rubric and Rater as random effects, grouped by artifacts.

From the above table, we can interpret the fixed effect results as: Compared to Rater 1, we would expect that the ratings given by Rater 2 are 0.37 units higher on average and the ratings given by Rater 3 are 0.21 units higher on average, with keeping other predictors constant. Compared to fall semester, we would expect that the ratings given by spring semester are 0.26 units lower on average, with keeping other predictors constant. Compared to rubric CritDes, we would expect that the ratings on rubric InitEDA are 0.74 units higher on average, the ratings on rubric InterpRes are 0.99 units higher on average, the ratings on rubric RsrchQ are 0.73 units higher on average, the ratings on rubric SelMeth are 0.41 units higher on average, the ratings on rubric TxtOrg are 1.02 units higher on average, and the ratings on rubric VisOrg are 0.65 units higher on average, with keeping other predictors constant.

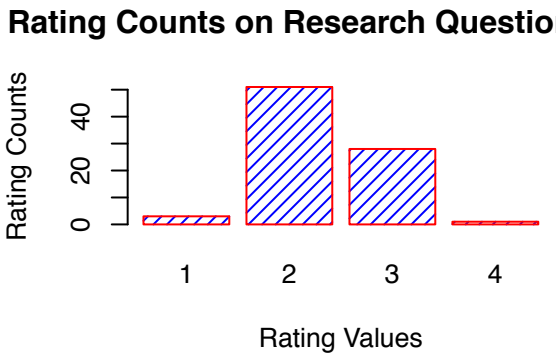
From the above table, we can interpret the interaction term results as: Compared to the rating on rubric CritDes given by Rater 1, we would expect that the ratings on rubric InitEDA given by Rater 2 are 0.30 units lower on average and the ratings on rubric InitEDA given by Rater 3 are 0.29 units lower on average. Compared to the rating on rubric CritDes given by Rater 1, we would expect that the ratings on rubric InterpRes given by Rater 2 are 0.51 units lower on average and the ratings on rubric InterpRes given by Rater 3 are 0.71 units lower on average. Compared to the rating on rubric CritDes given by Rater 1, we would expect that the ratings on rubric RsrchQ given by Rater 2 are 0.49 units lower on average and the ratings on rubric RsrchQ given by Rater 3 are 0.32 units lower on average. Compared to the rating on rubric CritDes given by Rater 1, we would expect that the ratings on rubric SelMeth given by Rater 2 are 0.39 units lower on average and the ratings on rubric SelMeth given by Rater 3 are 0.39 units lower on average. Compared to the rating on rubric CritDes given by Rater 1, we would expect that the ratings on rubric TxtOrg given by Rater 2 are 0.55 units lower on average and the ratings on rubric TxtOrg given by Rater 3 are 0.45 units lower on average. Compared to the rating on rubric CritDes given by Rater 1, we would expect that the ratings on rubric VisOrg given by Rater 2 are 0.10 units lower on average and the ratings on rubric VisOrg given by Rater 3 are 0.28 units lower on average.

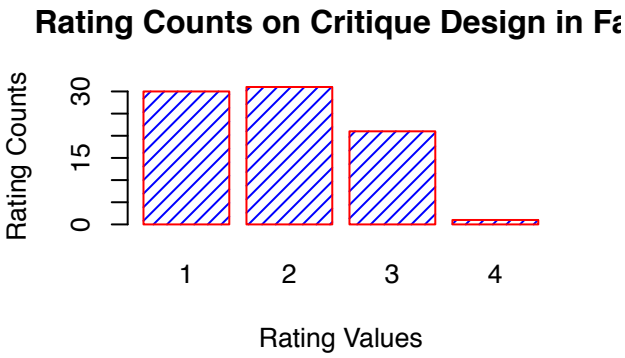
Interesting Things on the Dataset

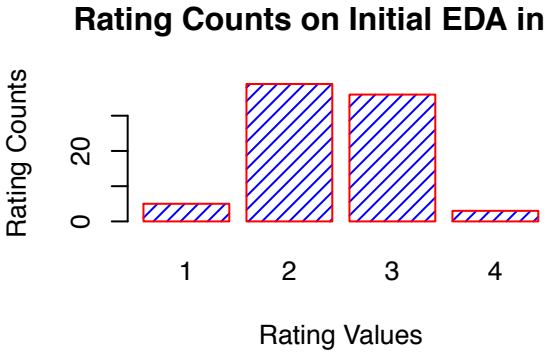
I would like to research on interesting facts based on semester since we do not cover the differentiation on this variable in the previous analysis.

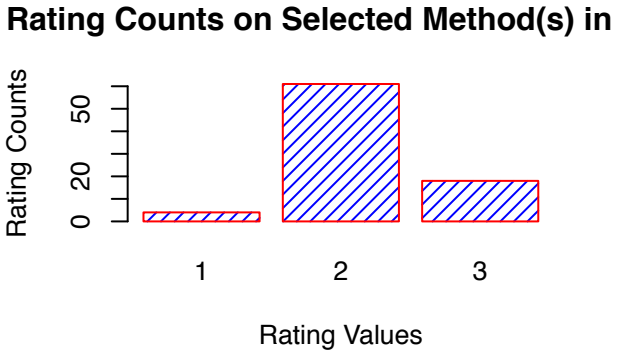
```
# filter two subsets with Fall and Spring, respectively
ratings_sem1 <- ratings %>%
  filter(ratings$Semester == "Fall")
ratings_sem2 <- ratings %>%
  filter(ratings$Semester == "Spring")
```

```
# for fall semester
# distributions of ratings for each rubric
par(mfrow=c(2,2))
barplot(table(ratings_sem1$RsrchQ),main="Rating Counts on Research Question in Fall",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_sem1$CritDes),main="Rating Counts on Critique Design in Fall",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_sem1$InitEDA),main="Rating Counts on Initial EDA in Fall",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_sem1$SelMeth),main="Rating Counts on Selected Method(s) in Fall",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```

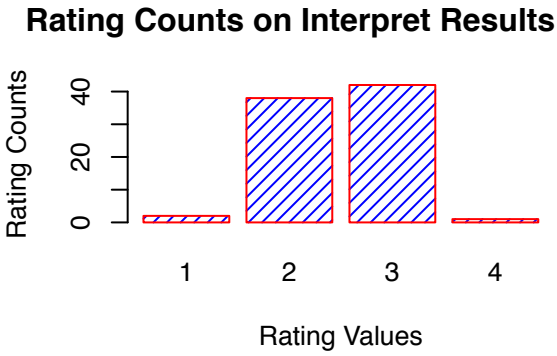


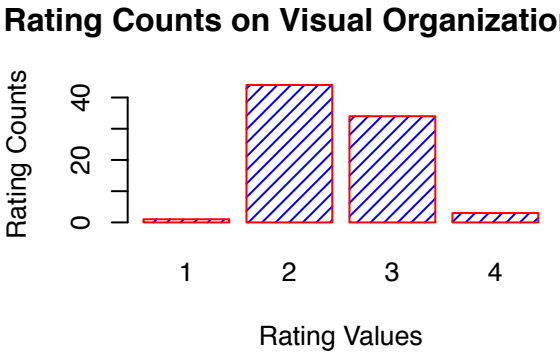


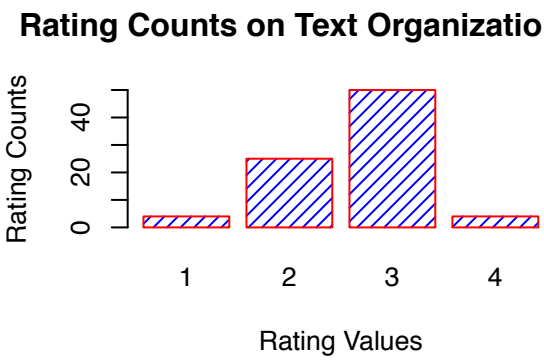




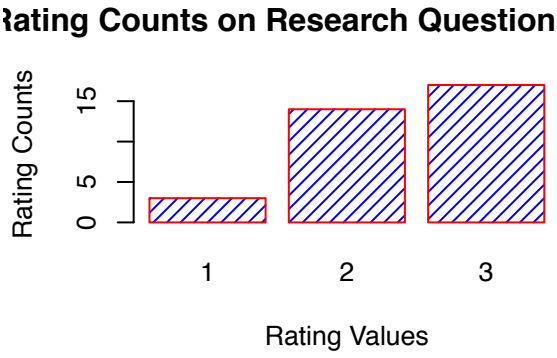

```
barplot(table(ratings_sem1$InterpRes),main="Rating Counts on Interpret Results in Fall",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_sem1$VisOrg),main="Rating Counts on Visual Organization in Fall",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_sem1$TxtOrg),main="Rating Counts on Text Organization in Fall",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```

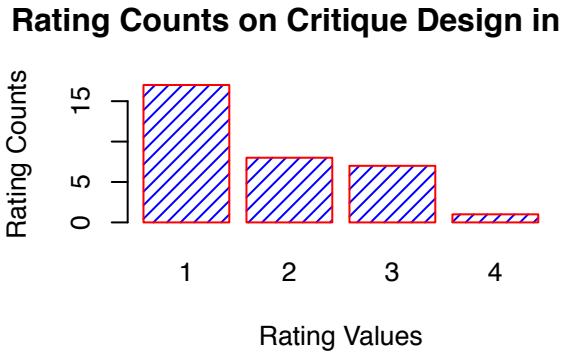


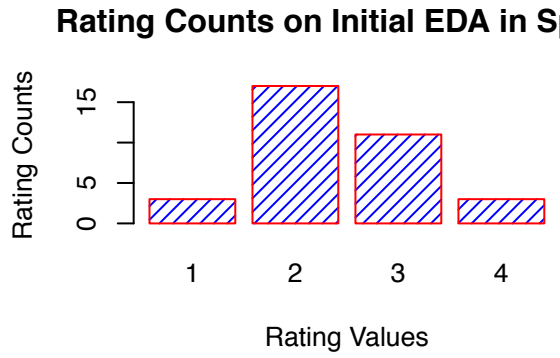


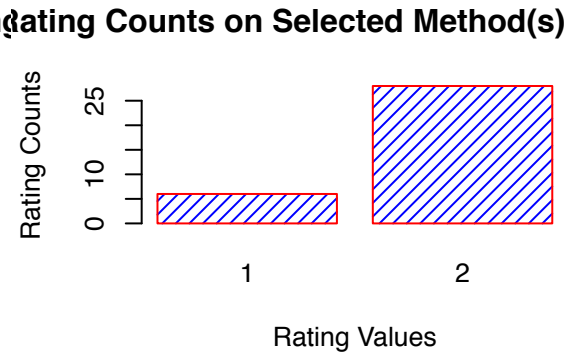



```
# for spring semester
# distributions of ratings for each rubric
par(mfrow=c(2,2))
barplot(table(ratings_sem2$RsrchQ),main="Rating Counts on Research Question in Spring",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_sem2$CritDes),main="Rating Counts on Critique Design in Spring",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_sem2$InitEDA),main="Rating Counts on Initial EDA in Spring",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_sem2$SelMeth),main="Rating Counts on Selected Method(s) in Spring",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```

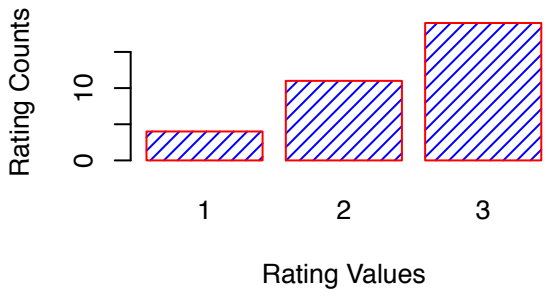


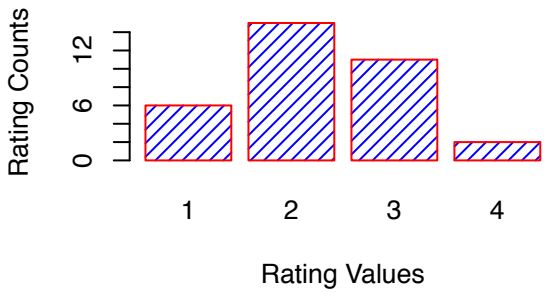


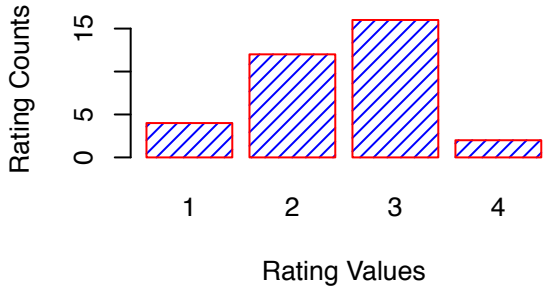





```
barplot(table(ratings_sem2$InterpRes),main="Rating Counts on Interpret Results in Spring",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_sem2$VisOrg),main="Rating Counts on Visual Organization in Spring",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
barplot(table(ratings_sem2$TxtOrg),main="Rating Counts on Text Organization in Spring",
        xlab="Rating Values", ylab="Rating Counts",border="red",
        col="blue",density=20)
```








```
# for fall semester
# show the table of ratings given each rubric
RsrchQ<-table(ratings_sem1$RsrchQ)
addmargins(RsrchQ)
```

```
##
##   1   2   3   4 Sum
##   3  51  28   1  83
```

```
# percentage of RsrchQ
round(prop.table(RsrchQ)*100,digits=0)
```

```
##
##  1  2  3  4
##  4 61 34  1
```

```
CritDes<-table(ratings_sem1$CritDes)
addmargins(CritDes)
```

```
##
##   1   2   3   4 Sum
##  30  31  21   1  83
```

```
# percentage of CritDes
round(prop.table(CritDes)*100,digits=0)
```

```
##
##  1  2  3  4
## 36 37 25  1
```

```
InitEDA<-table(ratings_sem1$InitEDA)
addmargins(InitEDA)
```

```
##
##   1   2   3   4 Sum
##   5  39  36   3  83
```

```
# percentage of InitEDA
round(prop.table(InitEDA)*100,digits=0)
```

```
##
##  1  2  3  4
##  6 47 43  4
```

```
SelMeth<-table(ratings_sem1$SelMeth)
addmargins(SelMeth)
```

```
##
##   1   2   3 Sum
##   4  61  18  83
```

```
# percentage of SelMeth
round(prop.table(SelMeth)*100,digits=0)
```

```
##
##  1  2  3
##  5 73 22
```

```
InterpRes<-table(ratings_sem1$InterpRes)
addmargins(InterpRes)
```

```
##
##   1   2   3   4 Sum
##   2  38  42   1  83
```

```
# percentage of InterpRes
round(prop.table(InterpRes)*100,digits=0)
```

```
##
##  1  2  3  4
##  2 46 51  1
```

```
VisOrg<-table(ratings_sem1$VisOrg)
addmargins(VisOrg)
```

```
##
##   1   2   3   4 Sum
##   1  44  34   3  82
```

```
# percentage of VisOrg
round(prop.table(VisOrg)*100,digits=0)
```

```
##
##  1  2  3  4
##  1 54 41  4
```

```
TxtOrg<-table(ratings_sem1$TxtOrg)
addmargins(TxtOrg)
```

```
##
##   1   2   3   4 Sum
##   4  25  50   4  83
```

```
# percentage of TxtOrg
round(prop.table(TxtOrg)*100,digits=0)
```

```
##
##  1  2  3  4
##  5 30 60  5
```

```
# for spring semester
# show the table of ratings given each rubric
RsrchQ<-table(ratings_sem2$RsrchQ)
addmargins(RsrchQ)
```

```
##
##   1   2   3 Sum
##   3  14  17  34
```

```
# percentage of RsrchQ
round(prop.table(RsrchQ)*100,digits=0)
```

```
##
##  1  2  3
##  9 41 50
```

```
CritDes<-table(ratings_sem2$CritDes)
addmargins(CritDes)
```

```
##
##   1   2   3   4 Sum
##  17   8   7   1  33
```

```
# percentage of CritDes
round(prop.table(CritDes)*100,digits=0)
```

```
##
##  1  2  3  4
## 52 24 21  3
```

```
InitEDA<-table(ratings_sem2$InitEDA)
addmargins(InitEDA)
```

```
##
##   1   2   3   4 Sum
##   3  17  11   3  34
```

```
# percentage of InitEDA
round(prop.table(InitEDA)*100,digits=0)
```

```
##
##  1  2  3  4
##  9 50 32  9
```

```
SelMeth<-table(ratings_sem2$SelMeth)
addmargins(SelMeth)
```

```
##
##   1   2 Sum
##   6  28  34
```

```
# percentage of SelMeth
round(prop.table(SelMeth)*100,digits=0)
```

```
##
##  1  2
## 18 82
```

```
InterpRes<-table(ratings_sem2$InterpRes)
addmargins(InterpRes)
```

```
##
##   1   2   3 Sum
##   4  11  19  34
```

```
# percentage of InterpRes
round(prop.table(InterpRes)*100,digits=0)
```

```
##
##  1  2  3
## 12 32 56
```

```
VisOrg<-table(ratings_sem2$VisOrg)
addmargins(VisOrg)
```

```
##
##   1   2   3   4 Sum
##   6  15  11   2  34
```

```
# percentage of VisOrg
round(prop.table(VisOrg)*100,digits=0)
```

```
##
##  1  2  3  4
## 18 44 32  6
```

```
TxtOrg<-table(ratings_sem2$TxtOrg)
addmargins(TxtOrg)
```

```
##
##   1   2   3   4 Sum
##   4  12  16   2  34
```

```
# percentage of TxtOrg
round(prop.table(TxtOrg)*100,digits=0)
```

```
##
##  1  2  3  4
## 12 35 47  6
```

By drawing the barplot and calculating the percentage of each score for each rubric (since fall and spring semester do not have same amount of artifacts in the dataset) based on Fall semester and Spring semester, I figured out that for rubric Research Question, raters give more score 2 in Fall semester but give more score 3 in Spring semester. For rubric Critique Design, raters give approximately same large amount of score 1 and score 2 in Fall semester but give obviously more score 1 in Spring semester. For rubric Initial EDA, raters give approximately same amount of score 2 and score 3 in Fall semester but give obviously more score 2 in Spring semester. For rubric Select Method(s), raters give obviously more score 2 in both Fall and Spring semester. For rubric Interpret Results, raters give obviously more score 3 in both Fall and Spring semester. For rubric Visual Organization, raters give obviously more score 2 in both Fall and Spring semester. For rubric Text Organization, raters give obviously more score 3 in both Fall and Spring semester.