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

Yanlin Li, yanlinli@andrew.cmu.edu

Abstract

(still working on this part)

Introduction

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

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

Data

The data is about the rating information of the 91 project papers (referred to as "artifacts"), which were randomly sampled from a Fall and Spring section of the Gen Ed program in Freshman Statistics. Three raters from different departments were assigned to do the ratings based on 7 rubrics. Only 13 of the 91 artifacts were graded by all the three raters. The rest 78 artifacts were rated by only one rater. Details about the rubrics and rating scale are in the tables below:

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.).

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.

The data set contains seven important variables:

1. `Rater`: Which of the three raters gave a rating. There are three raters, labeled by 1, 2, and 3.
2. `Artifact`: Unique identifier for each artifact.
3. `Repeated`: 1 = this is one of the 13 artifacts seen by all 3 raters
4. `Semester`: Which semester the artifact came from (Fall or Spring)
5. `Sex`: Sex or gender of student who created the artifact (M for male, and F for female)
6. `Rubric`: One of the seven rubrics described above
7. `Rating`: Rating on the specific rubric from the specific rater

Method

Question 1

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

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

To begin with, I made a summary table containing the counts of different ratings (1, 2, 3, and 4) for each rubric. The distribution of ratings for each rubric was also evaluated in the summary table with their mean

and standard deviation. Besides, seven bar charts of ratings regarding different rubrics were plotted for comparison. We use these visualizations and tables to compare the ratings for different rubrics.

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

Then, I switched to the subset of 13 artifacts seen by all three raters to do the same set of analysis. I compared the results made by the subset with the results for full data to see whether the 13 artifacts are representative of the whole set of 91 artifacts.

Question 2

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

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

First, intraclass correlation (ICC) was used to evaluate the agreement among the raters. ICC is the common correlation among a certain group of values. If the raters generally agree with each other, we expect the the ICC's between the ratings for each artifact is high, and the ICC's within all the ratings made by each rater is low. That means, ratings are similar between groups of artifacts, and are not related to raters. The ICC values will be evaluated using random intercept models for the ratings.

Second, pairwise rating agreement of the three raters were evaluated under the seven rubrics. A two-way table was made for each pair of raters (rater 1 & rater 2, or rater 2 & rater 3, or rater 1 & rater 3) for each rubric. The percentage of artifacts that the two raters gave the same ratings for that rubric is the agreement rate. We investigate the rates and see whether there is any rater whose agreement rate with both other raters are low. If that happens, we claim that the rater is disagree with the other raters with the ratings regarding that rubric.

Question 3

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

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

The random effect was selected automatically by `fitLMER.fnc` function. This function selects the random effects by forward-fitting. All the possible random effects were feeded into the function for selection.

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

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

4. If some p-values are greater than 0.05, then we delete the variable with the largest p-value and continue to run the loop.

(Still working on interaction. . .)

Question 4

Is there anything else interesting to say about this data?

For this part, I did some more analysis to evaluate how the ratings were linked to gender of students and semester. Some graphs were made for an exploratory data analysis.

(Still working on the possible models for this analysis. . .)

Result

Question 1

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

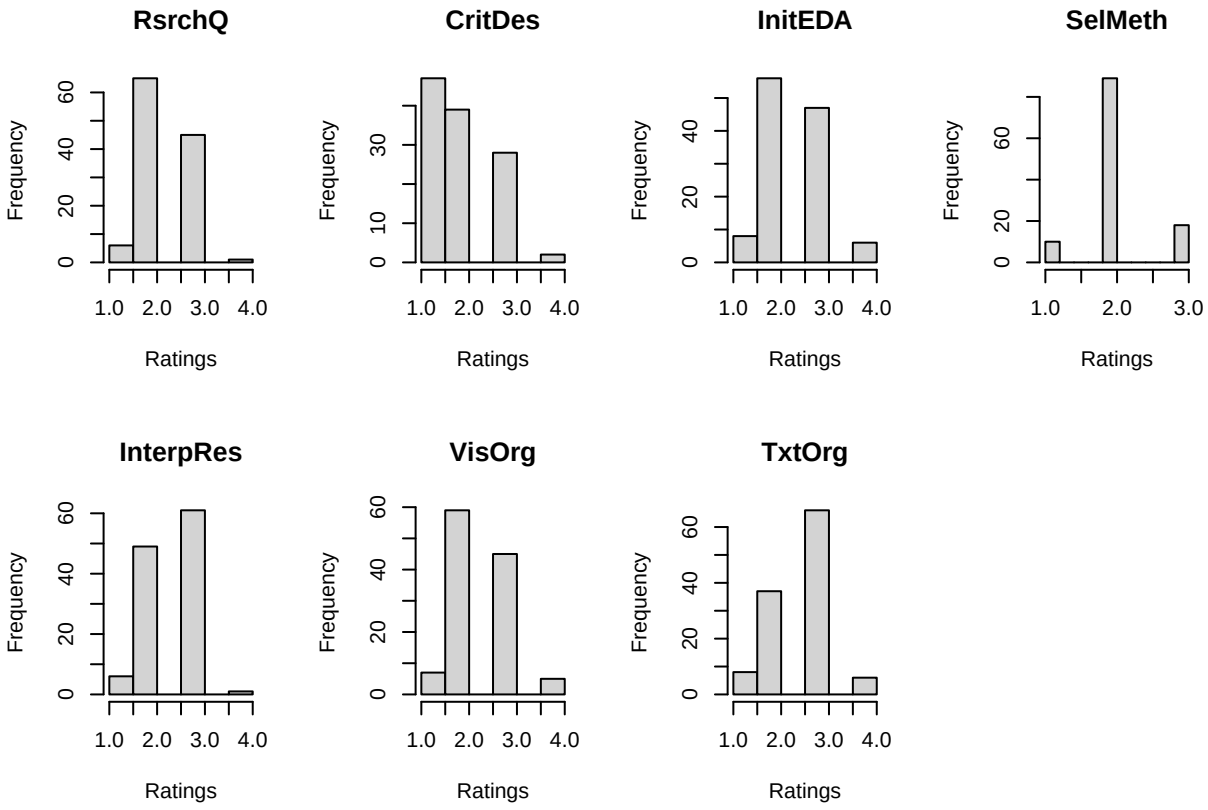
Rubrics

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

Table 1: Summary table: Rubics

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

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



Here are the histograms for the seven rubrics. We can see that only Critique Design skews to the right, and all the other are highest with grade 2 or 3.

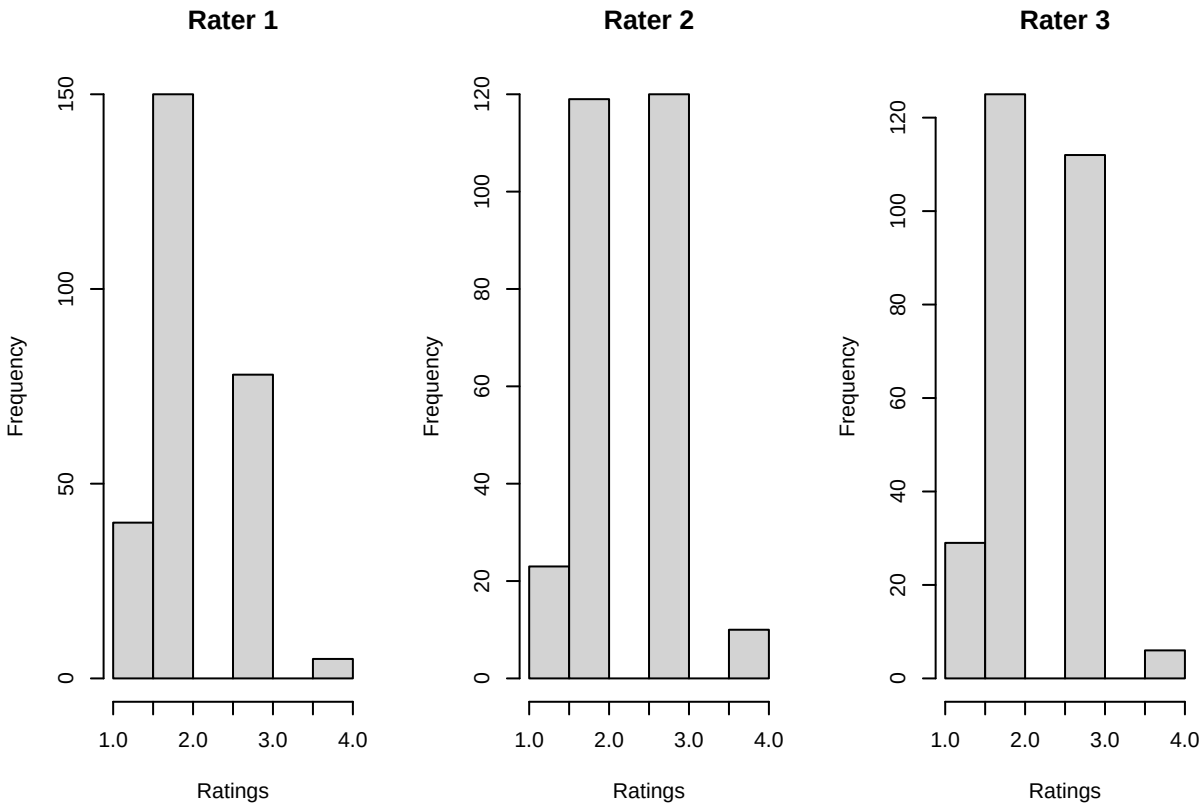
Raters

Here is a summary table for ratings produced by the three raters.

Table 2: Summary table: Raters

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

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



The same pattern can be observed from the three histograms above.

Then we did the same analysis for the subset of 13 artifacts seen by all three raters. The results are similar. (See Appendix 1 Distribution of subset of artifacts seen by all three raters Page ? for details) Thus, we conclude that these thirteen artifacts are representative of the whole set of 91 artifacts.

Question 2

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

Overall agreement

Table 3: ICC analysis for different rubrics regarding raters & artifacts

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

We can see that the ICC values for raters are all very small (< 0.05). The ICC value means correlation between ratings on any two different artifacts by the same rater. The low value indicates that for anyone of the three raters, the link between the ratings by him is very low. Thus, scores are not expected to be affacted

by raters. The ICC value is the highest for rating on Interpret Results, which is understandable because the raters can have their preference in case of interpretations.

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

Agreement for different raters

Table 4: Pairwise agreement rate for different rubrics

Rubrics	Agreement.12	Agreement.23	Agreement.13	Disagree.rater
RsrchQ	0	0	1	2
CritDes	1	1	1	None
InitEDA	1	1	1	1
SelMeth	1	0	0	3
InterpRes	1	0	0	None
VisOrg	1	1	1	None
TxtOrg	0	1	0	None

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

Question 3

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

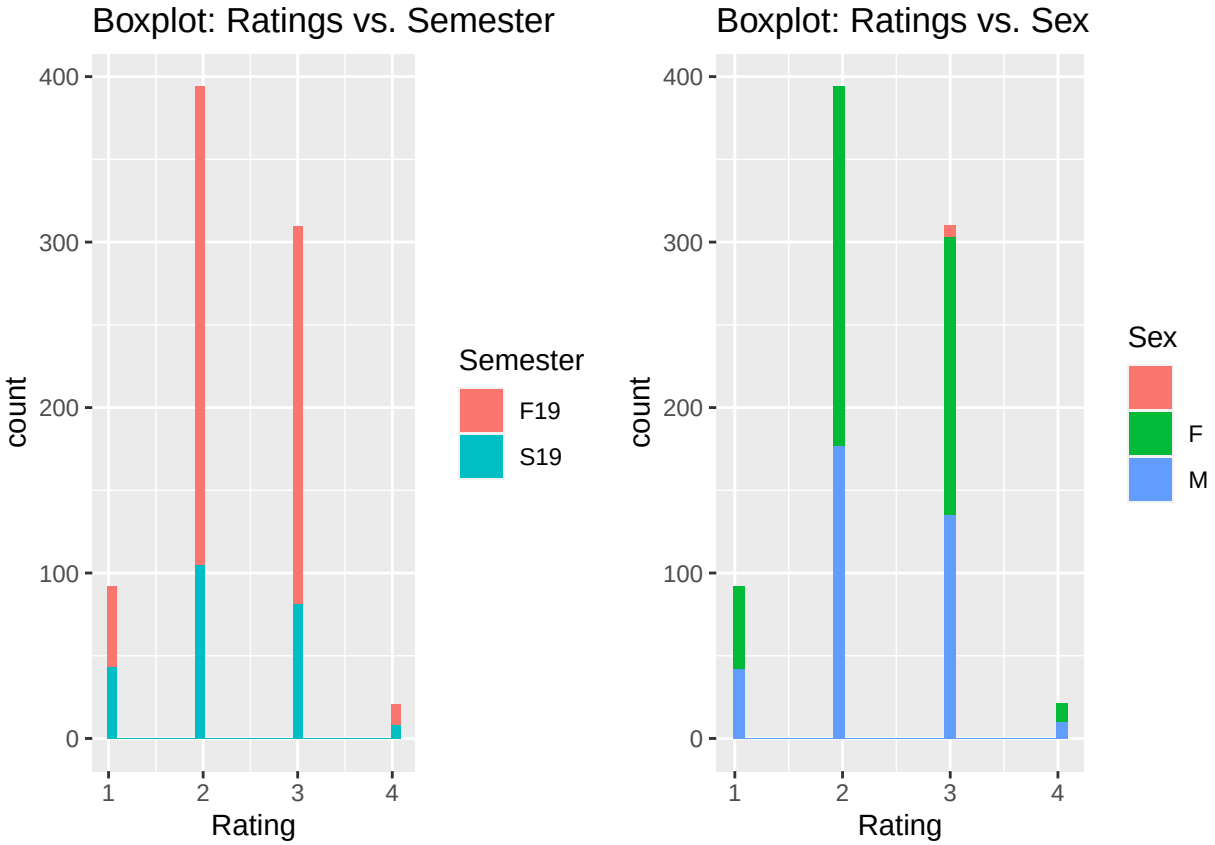
For the first step, the random effect of artifact on rubric and rater was selected. Fixed effects of rubric, rater, and repeated were then added to the model.

(still working on it. . .)

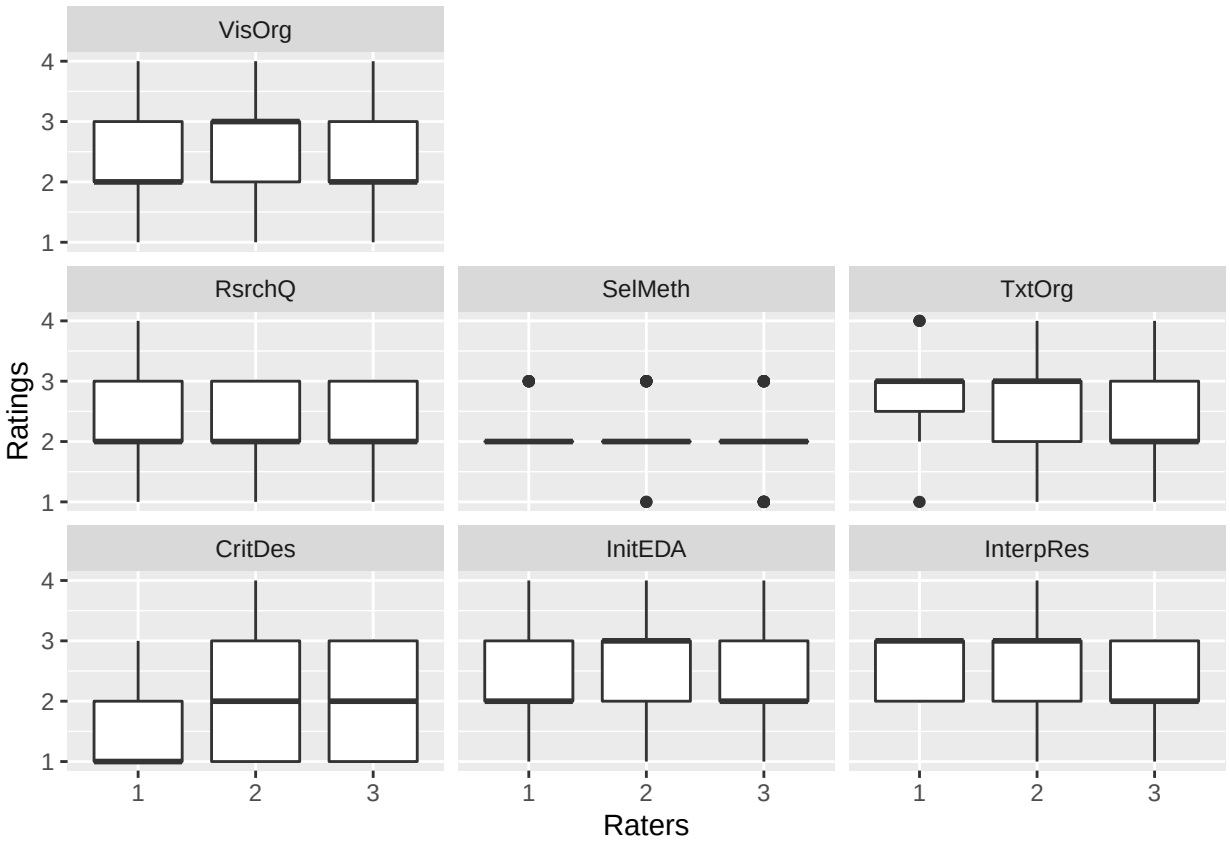
Question 4

Is there anything else interesting to say about this data?

Exploratory data analysis



From the histograms above, we can see that the distribution of ratings are different across semesters. The ratings are the same for different genders of students.



From the box plots above, we can see that the median ratings for all rubrics are all around 2 and 3. There exists some difference in ratings for different raters for some rubrics.

Models

(still working on it. . .)

Discussion

(still working on this part. . .)

Question 1

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

Question 2

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

Question 3

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

Question 4

Is there anything else interesting to say about this data?

Reference

Technical Appendix

1.

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

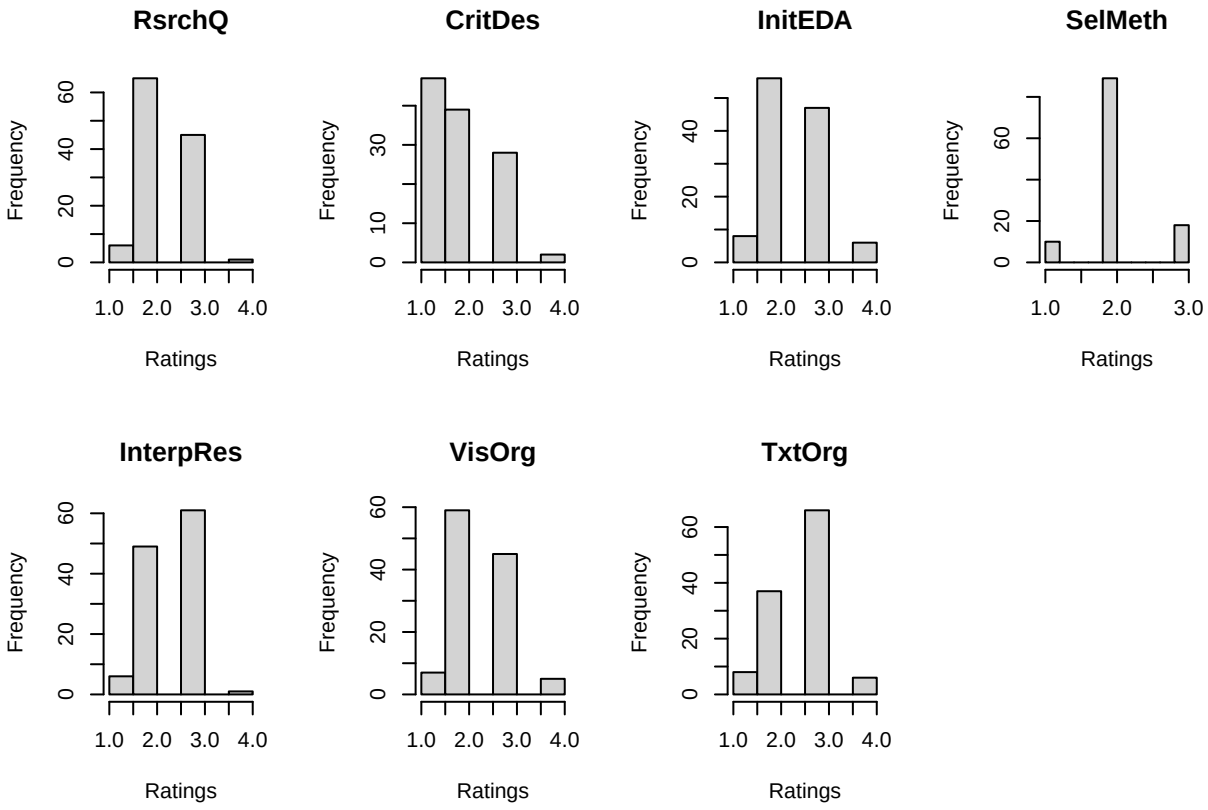
Rubrics

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

Table 5: Summary table: Rubics

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

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



Here are the histograms for the seven rubrics. We can see that only Critique Design skews to the right, and all the other are highest with grade 2 or 3.

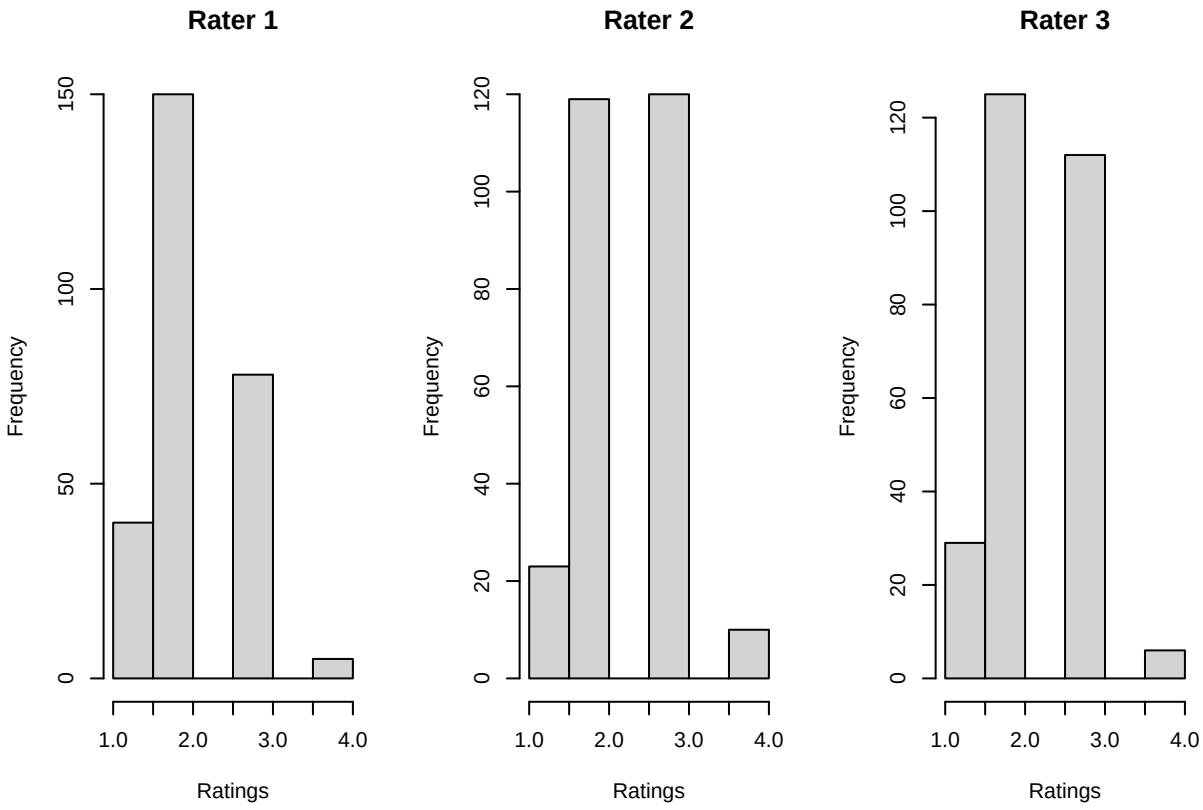
Raters

Here is a summary table for ratings produced by the three raters.

Table 6: Summary table: Raters

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

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



The same pattern can be observed from the three histograms above.

Distribution of subset of artifacts seen by all three raters

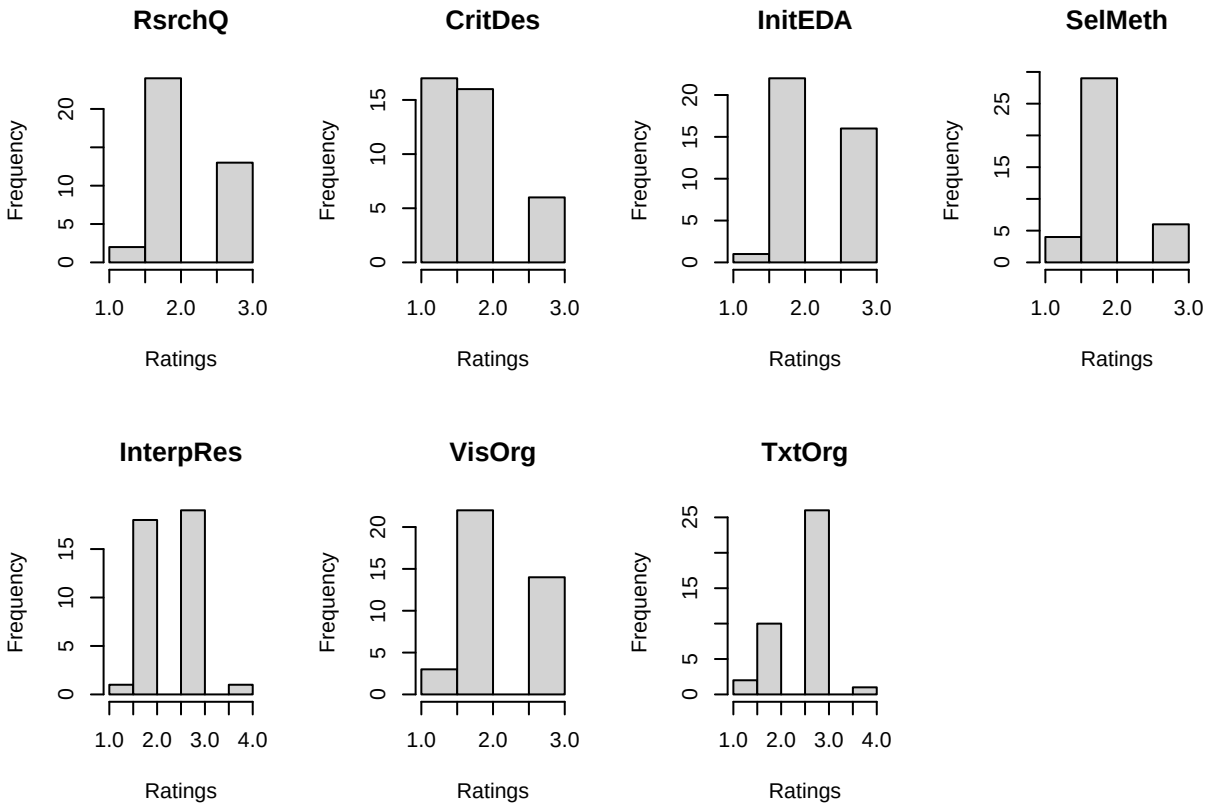
In this part, we will do all the analysis appeared above to evaluate whether a subset of 13 artifacts seen by all three raters are representative of the whole set of 91 artifacts.

Rubrics

Table 7: Summary table: Rubics - Subset

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

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



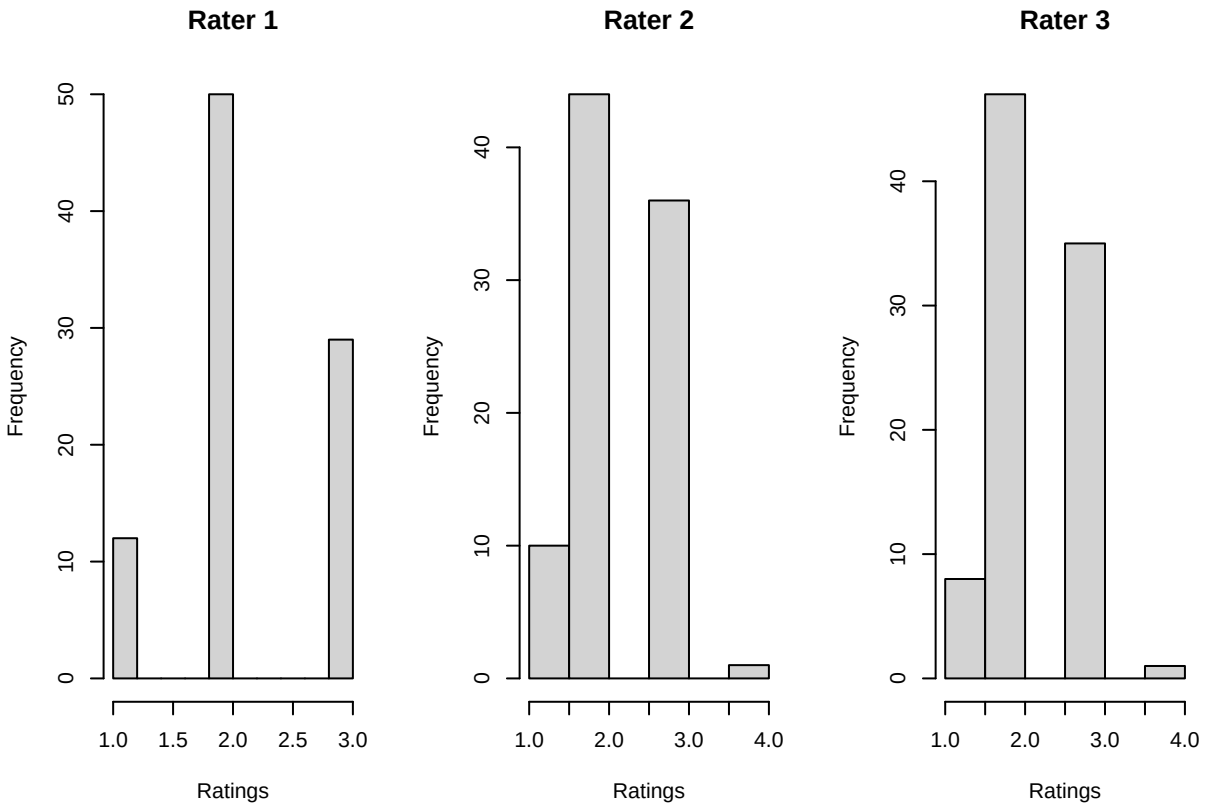
We can see the similar patterns from these histograms and the histograms from the full data.

Raters

Table 8: Summary table: Raters

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

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



The patterns in these plots are approximately the same with the plots produced by full data

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

2

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

Overall agreement

Table 9: ICC analysis for different rubrics regarding raters & artifacts

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

We can see that the ICC values for raters are all very small (< 0.05). The ICC value means correlation between ratings on any two different artifacts by the same rater. The low value indicates that for anyone of the three raters, the link between the ratings by him is very low. Thus, scores are not expected to be affacted by raters. The ICC value is the highest for rating on Interpret Results, which is understandable because the

raters can have their preference in case of interpretations.

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

Agreement for different raters

Table 10: Pairwise agreement rate for different rubrics

Rubrics	Agreement.12	Agreement.23	Agreement.13	Disagree.rater
RsrchQ	0	0	1	2
CritDes	1	1	1	None
InitEDA	1	1	1	1
SelMeth	1	0	0	3
InterpRes	1	0	0	None
VisOrg	1	1	1	None
TxtOrg	0	1	0	None

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

Full data

Here is the ICC analysis done with the full data.

Table 11: ICC analysis for different rubrics regarding raters & artifacts - Full data

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

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

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

3

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

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

random slope on it.

We first make a selections from the possible random effect terms. Basically, we consider Artifacts and Raters as the groups.

Model selection

We see that we should only keep the random slope for Rubric and Rater with respect to Artifacts. We start by selecting the random effect and then fixed effect. The random effect was selected automatically by `fitLMER.fnc` function.

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

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

```
## Rating ~ 1 + (0 + Rubric | Artifact) + (0 + Rater | Artifact) +
##     Rubric + Rater + Repeated
```

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

Model interpretation

```
## lmer(formula = Rating ~ 1 + (0 + Rubric | Artifact) + (0 + Rater |
##     Artifact) + Rubric + Rater + Repeated, data = na.omit(tall))
##                 coef.est coef.se
## (Intercept)      2.11     0.11
## RubricInitEDA    0.54     0.09
## RubricInterpRes  0.58     0.10
## RubricRsrchQ     0.46     0.09
## RubricSelMeth    0.16     0.09
## RubricTxtOrg     0.68     0.10
## RubricVisOrg     0.53     0.10
## Rater           -0.09     0.04
## Repeated        -0.08     0.09
##
## Error terms:
##  Groups     Name            Std.Dev. Corr
##  Artifact   RubricCritDes   0.73
##             RubricInitEDA   0.52      0.39
##             RubricInterpRes 0.27      0.01  0.65
##             RubricRsrchQ    0.35      0.53  0.21  0.53
##             RubricSelMeth   0.15      0.30  0.36  0.24 -0.16
##             RubricTxtOrg    0.42      0.21  0.43  0.45  0.28  0.21
##             RubricVisOrg    0.44      0.24  0.62  0.46  0.32 -0.08  0.60
##  Artifact.1 Rater           0.13
```

```
##  Residual                   0.42
## ---
## number of obs: 817, groups: Artifact, 91
## AIC = 1498.8, DIC = 1351
## deviance = 1385.9
```

From the summary, we can see that all the rubrics will result in a higher rating except Critique Design, which means that this is the part that have lower ratings.

There are some insignificant effects. First, raters with higher code will give a slightly smaller rating. Second, ratings from the artifacts seen by all three raters can score less. These two results do not make much sense in reality either.

ICC

The ICC values for the full model is in the table below:

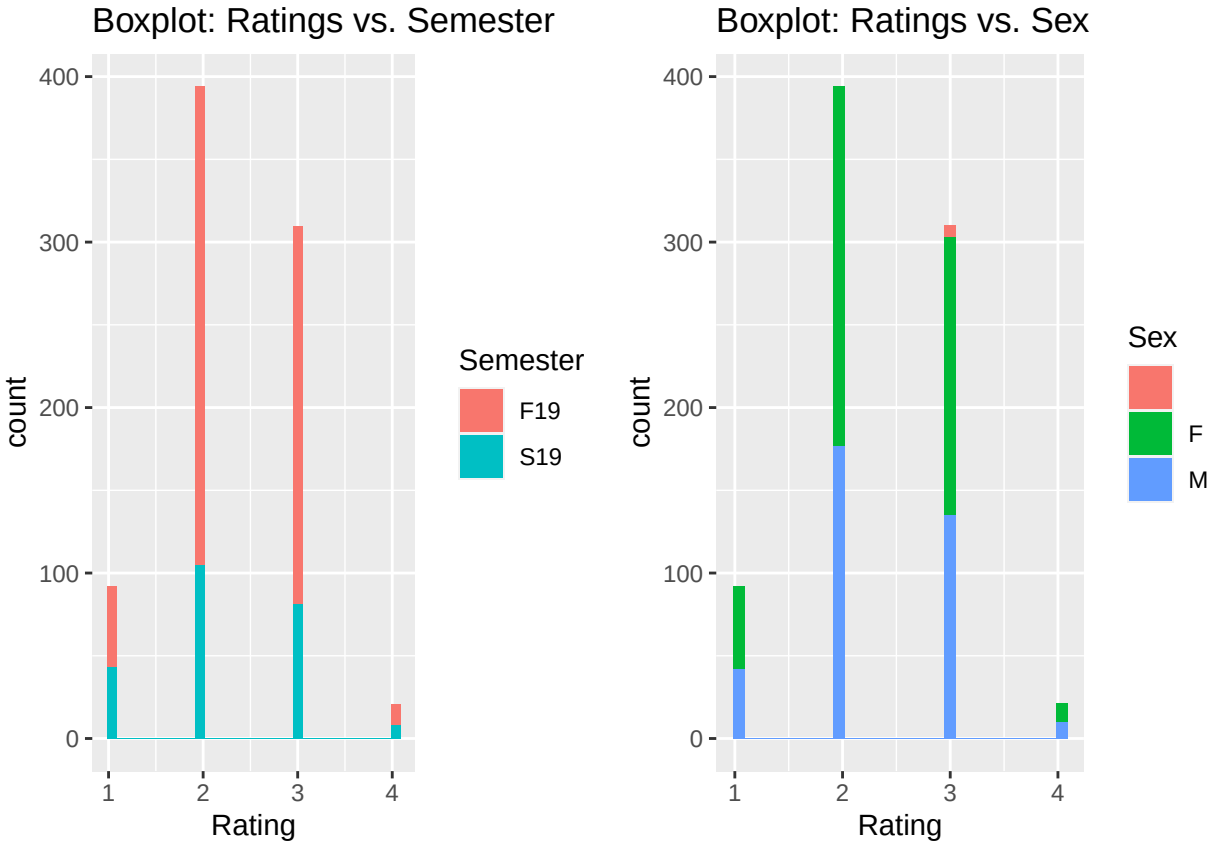
Table 12: ICC analysis for Artifacts under full model

Rubrics	ICC.Artifact
CritDes	0.7482393
InitEDA	0.6061530
InterpRes	0.2985412
RsrchQ	0.4153027
SelMeth	0.1072792
TxtOrg	0.4974592
VisOrg	0.5226807

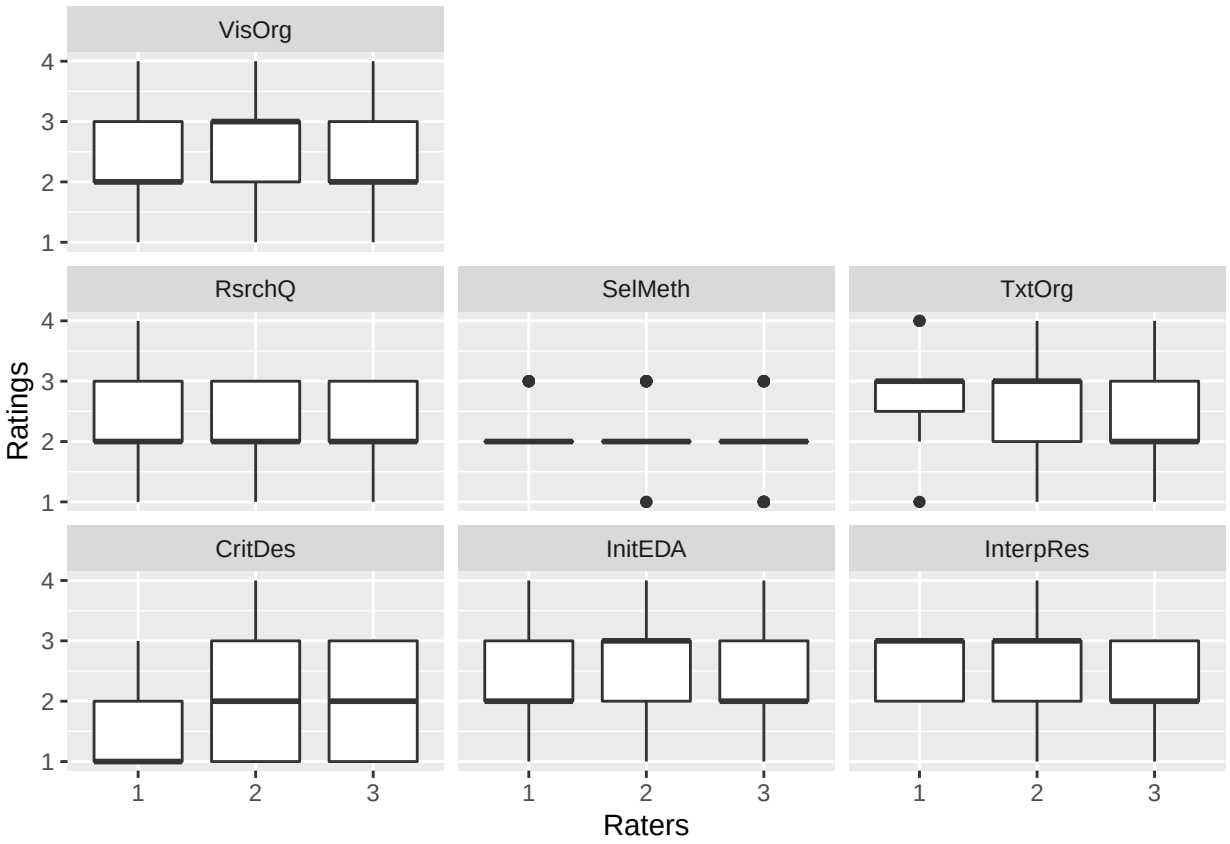
This time, the values of the ICC values are higher for most of the rubrics.

4

Additional analysis



From the histograms above, we can see that the distribution of ratings are different across semesters. The ratings are the same for different genders of students.



From the box plots above, we can see that the median ratings for all rubrics are all around 2 and 3. There exists some difference in ratings for different raters for some rubrics.

Weakness

1. The data set is small for analysis, especially when there are only 13 artifacts that have been graded by all raters. This is thus hard to draw a conclusion the fairness of rating process.
2. This analysis can only focus on the fairness of the rating process, because no data except sex was provided about the authors of those artifacts. In reality, the link between ratings and raters should be much lower than the one between ratings and authors. More information about authors should be provided if we want to predict ratings.

Codes