Factors that can affect Course Ratings for CMU Freshman Statistics

Yuqing Xu

Abstract

This project mainly focuses on exploring the relationship between the ratings and other various factors (Rater, Semester, Sex, Repeated, Rubric) in this experiment on rating work in Freshman Statistics by raters from across the college. The data including these variables is from ?????? and includes ratings on 1 project-papers that were randomly sampled from a Fall and Spring section from three raters based on seven rubrics. The general method on model selection is to use ANOVA test to select among the intercept-only models, models with fixed effects, models with interactions, and models with random effects. For further analysis, we should try to fit more models to find the best one and validity of the model selected should be a concern in the future.

still working on results so the summary is incomplete for this part.

Introduction

Recently the Dietrich College at Carnegie Mellon University has been experimenting with rating work in Freshman Statistics by raters from across the college in order to prepare for rating student work performed in the new program. Here in this project, we want to explore the question about when difference raters are asked to rate project papers-refered to as "artifacts" based on seven different rubrics, if there is any relationship between the final ratings and the variables described below, or I will say if there exists any variable that can significantly affect the ratings. The questions will be addressed related to the topic in this project are: 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 experiement (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 file we are going to use is ratings.csv and tall.csv from ?????????. The data provides that for 91 project-papers that were randomly sampled from a Fall and Spring section of Freshman Statistics, three raters from three different departments were asked to rate these artifacts on seven rubrics as shown in Table 1, with rating scale for all rubrics shown in Table 2.

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. *NOTE: These are* **not** *the rubrics used by instructors or TA's in Freshman Statistics. They are* **only** *approved to be used in this experiment.*

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. *NOTE: This is* **not** *the rating scale used by instructors or TA's in Freshman Statistics. It is* **only** *approved to be used in this experiment.*

13 of the 91 artifacts were rated by all three raters, and the remaining were rated by only one. ratings.csv and tall.csv contain same information but were organized differently. We can see below in Table 3, variables for analysis are:

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

Table 3: Variables in the file `ratings.csv`. Variables that are **not** expected to be useful for analysis are shown in parentheses.

Method

(1)

To address the first question asking about if there is distinguishable distribution of ratings for each rubrics and each raters, we will need to extract the artifacts with all 3 raters to check the distributions seperately. Firstly, we want to create a bar plot for each rubric to see the difference in distribution of rubric with different ratings for the two seperated datasets(full and 13 artifacts). Also, a table of counts is a supplement for the plots to see if ratings are distributed similarly for all rubrics for both datasets. Secondly, to compare distributions across raters, same method is used. A bar plot is made for each rater from both full datasets and the datasets with 13 artifacts. And tables of counts are also used to help recheck the distribution of ratings based on different raters.

(2)

For the research question talking about the agreement among raters, we still seperate the dataset as the full dataset and the one with 13 artifacts. The basic idea is to use ICC value (intra-class correlation coefficient value), which measures the reliability of two different raters to measure subjects similarity. And we will calculate the ICC value for all ratings for each rubric and all rateings for each rubric from three raters pairwisely. Also, 2-way tables of counts for the ratings of each pair of raters, on each rubric recording the percent exact agreement between the two raters will be used to help to determine disagreement and agreement.

(3)(i)

We will start with addressing the 13 common artifacts with all 3 raters' ratings, adding fixed effect to the seven rubric-specific models by fitting linear mixed effect model (lmer). Variable "Repeated" will be removed because Repeated will be all the same for these 13 artifacts. Then, a backwards-elimination process is applied to the model so that only significant fixed effects are left by usiong fitLMER.fnc(). After that, we will use ANOVA test to compare the model with only intercept and the model with fixed effects, and see if more interactions are needed.

(3)(ii)

After finishing dealing with the 13 common artifacts, we will start working on the full dataset with same process. For the reason that we should use same dataset for every model fitting and comparison, missing data will be eliminated. We will add fixed effect to the seven rubric-specific models by fitting linear mixed effect model (lmer). Then, a backwards-elimination process is applied to the model so that only significant fixed effects are left by usiong fitLMER.fnc(). After that, we will use ANOVA test to compare the model with only intercept and the model with fixed effects, and see if more interactions are needed.

(3)(iii)

For those rubrics whose selected models from the previous step are not just the simple intercept-only models, we will examine each of these to see if the fixed effect make sense and if there are any interactions or additional random effects. For each of them, firstly, refit the model and check t-values for all variables to decide if they are significant; secondly, we will use ANOVA test to compare the model and model without "Rater" to see if we really need "Rater" as a factor for this rubric model; thirdly, we will add fixed-effect interactions between each pair of variables left in the best model selected from previous steps to the model, and use ANOVA test to compare the new model and original model to check if we need the interactions; finally, random effects on the models should be considered and compared to get the final model.

(3)(iv)

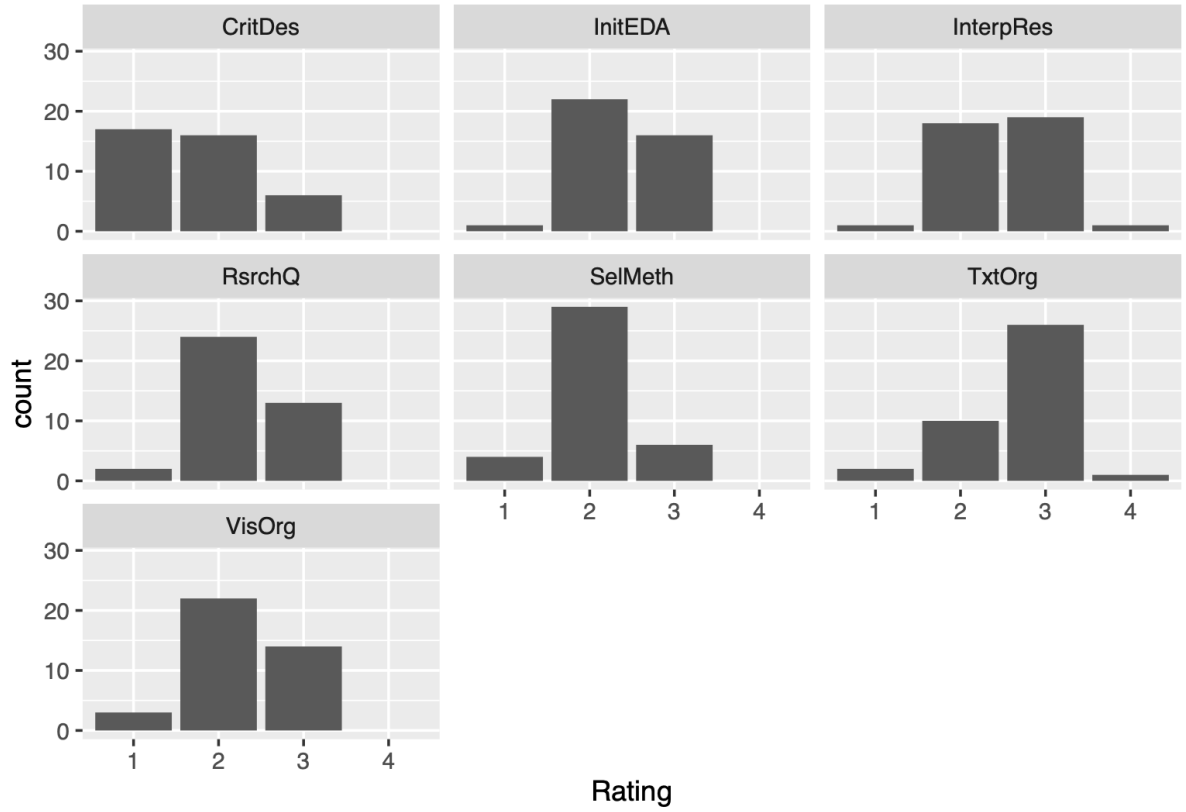
Finally, we combine all rubrics and consider it as a whole and try to add fixed effects, interactions, and new random effects to the "combined" intercept-only model using all the data. After adding fixed effect, interactions, and random effects to the model, we will use ANOVA to select the final model.

Result

(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?

Code Appendix part (1): page ?

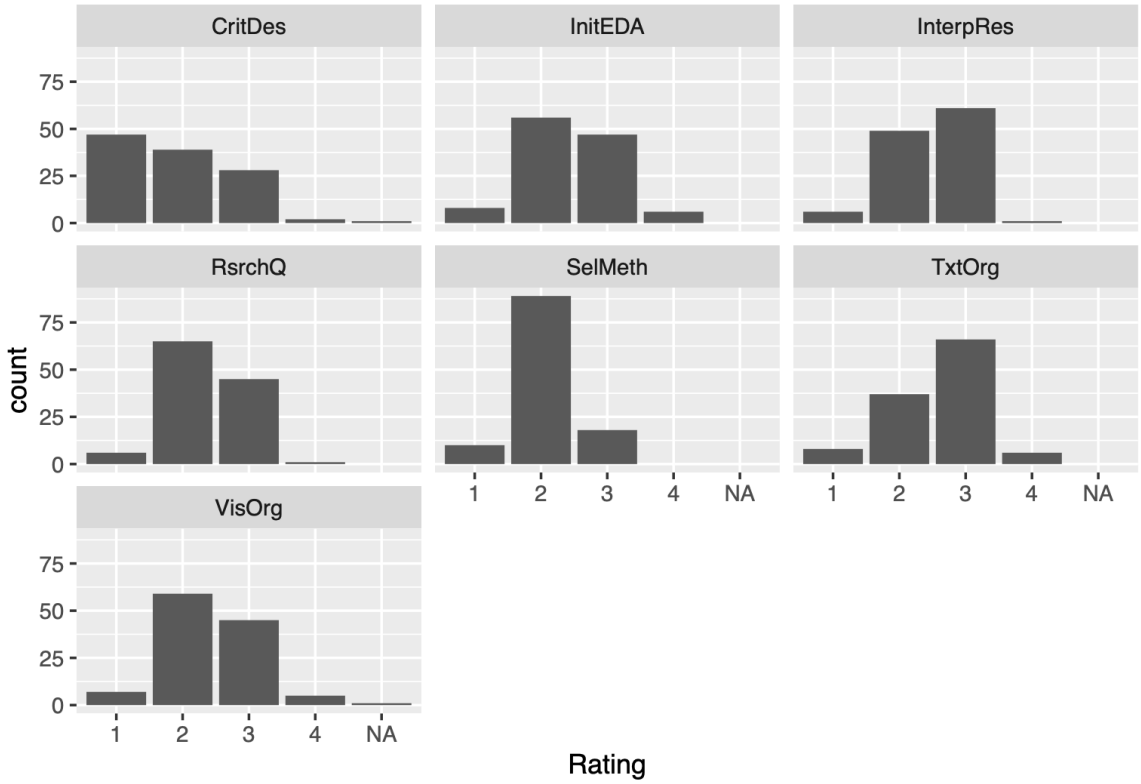
By plot shown below, which shown the distribution of 7 rubrics in the extracted dataset that contains 13 artifacts with all 3 raters' ratings, we can see that CritDes is the only one rubric which has most ratings at 1 than all the other ratings, as other rubrics have highest counts at score 2 or 3. Also, rubrics InitEDA, RsrchQ, SelMeth, and Visorg do not have any ratings at 4, but one reason might be the number of observations is small for this subset.



Also, we can find similar pattern based on the table of counts of the same dataset.

```
##          CritDes InitEDA InterpRes RsrchQ SelMeth TxtOrg VisOrg
## Rating 1      17       1         1      2       4      2      3
## Rating 2      16      22        18     24      29     10     22
## Rating 3       6      16        19     13       6     26     14
## Rating 4       0       0         1      0       0      1      0
```

By plot shown below, which shown the distribution of 7 rubrics in the full dataset , we can see that CritDes is the only one rubric which has most ratings at 1 than all the other ratings, as other rubrics have highest counts at score 2 or 3. Also, rubrics SelMeth does not have any ratings at 4.

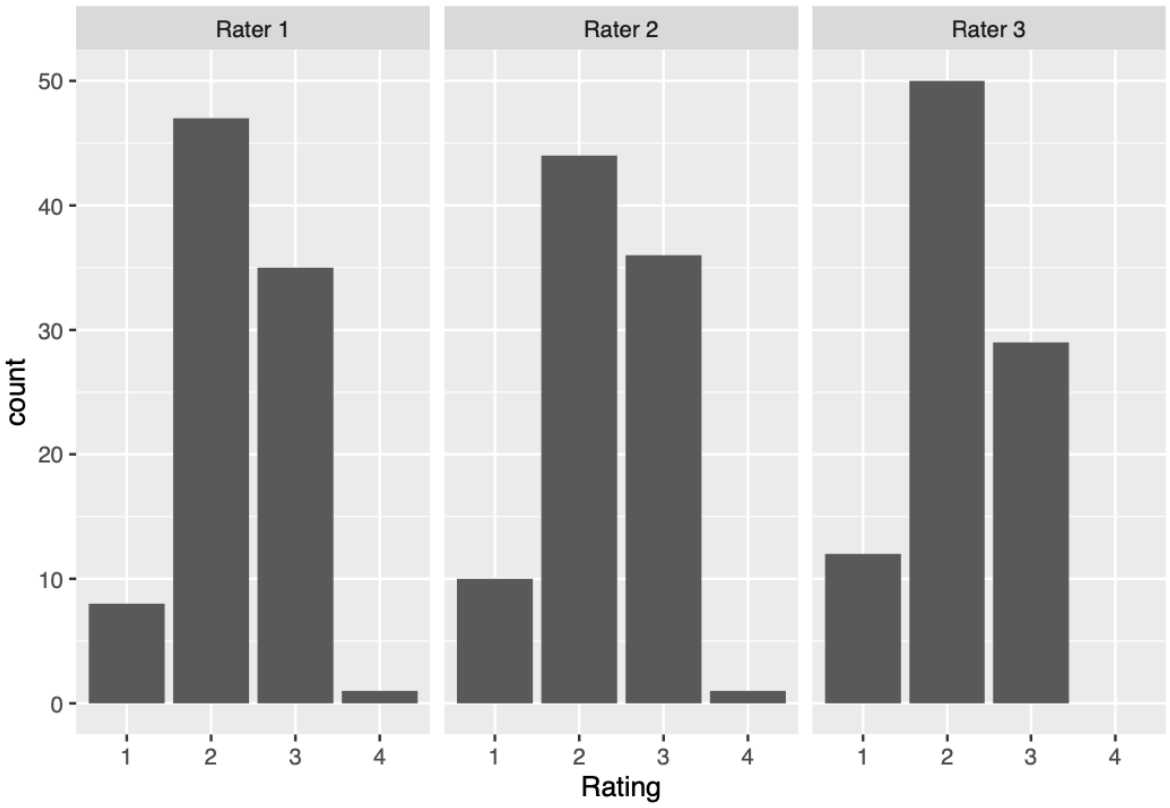


Also, we can find similar pattern based on the table of counts of the same dataset.

```
##          CritDes InitEDA InterpRes RsrchQ SelMeth TxtOrg VisOrg
## Rating 1      47       8         6      6      10      8      7
## Rating 2      39      56        49     65      89     37     59
## Rating 3      28      47        61     45      18     66     45
## Rating 4       2       6         1      1       0      6      5
## <NA>           1       0         0      0       0      0      1
```

From everything we have above, I will say that the distribution of CritDes is distinguishable from other rubrics and distributions of other rubrics are indistinguishable. For rubric CritDes, we can see that largest proportion of students get score 0; however, for all other rubrics, most students get score 2 or 3 and only few get 0 or 4, and this difference makes CritDes distinguishable from others. Raters tend to give lower score for rubric CritDes than other rubrics.

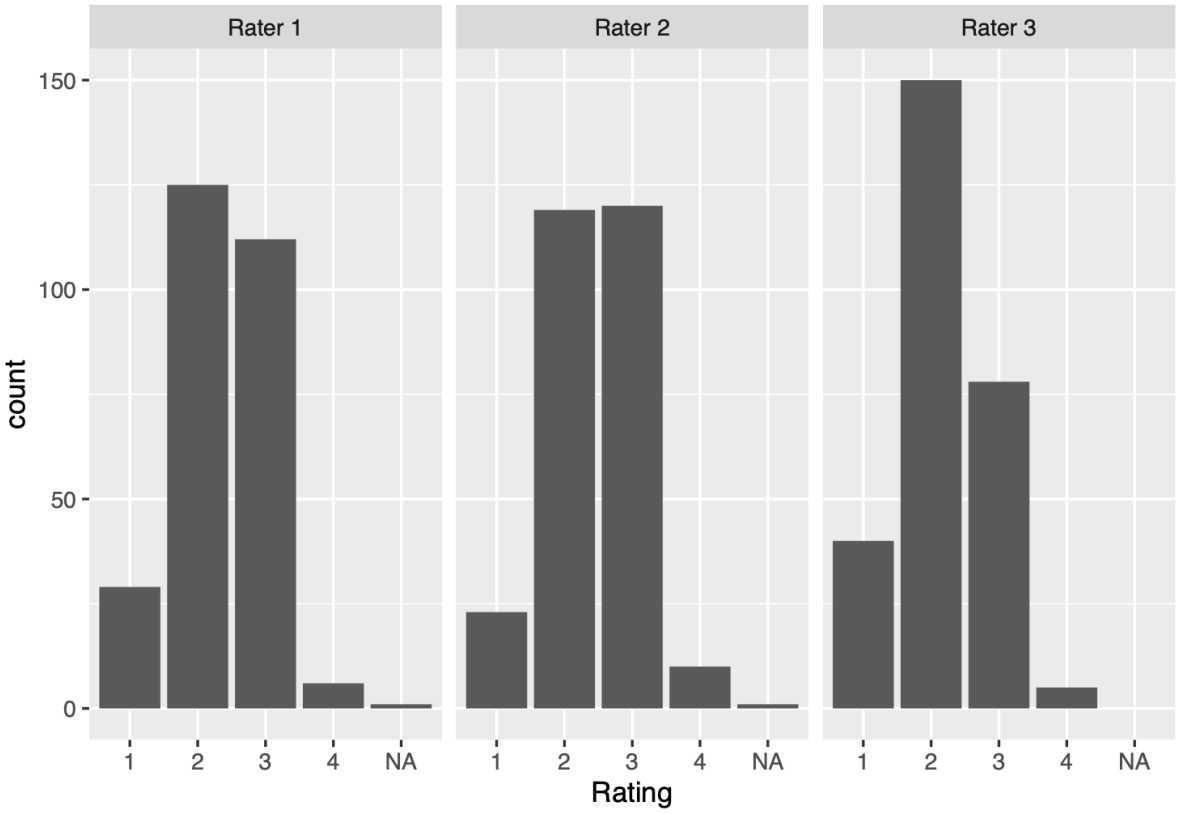
By plot shown below, which shown the distribution of 3 raters in the extracted dataset that contains 13 artifacts with all 3 raters' ratings, we can see that they all have similar patterns that they rate at score 2 more than other scores, and they rate at score 4 least, especially that rater 3 does not give any students score 4.



Also, we can find similar pattern based on the table of counts of the same dataset.

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

By plot shown below, which shown the distribution of 3 raters in the full dataset, we can see that they all have similar patterns that they rate at score 2 or 3 more than other scores, and they rate at score 4 least.



Also, we can find similar pattern based on the table of counts of the same dataset.

```
##          Rater 1 Rater 2 Rater 3
## Rating 1      29      23      40
## Rating 2     125     119     150
## Rating 3     112     120      78
## Rating 4       6      10       5
## <NA>           1       1       0
```

From everything we have above, I will say that all 3 raters are indistinguishable as they all have really similar patterns and none of them tends to give a lower/higher score.

As we do not have any missing value in the smaller 13-rubirc dataset, we do not need to deal with it. From the table below, we can see that there are missing values for Rating. We may need to remove or address it when we use these data in the model later.

```
##       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>
```

For the one missing sex value shown at the table below, for the reason that we do not want to lose this data

and also we cannot decide sex for this data easily based on what we have now, we will just leave it as a third sex category "–".

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

(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?

Code Appendix part (2): page ?

From the table below showing the ICC value for ratings on each rubric and the value for ratings from raters pairwisely from the subset containing only the 13 artifacts, we can see that InterpResm RsrchQ, and TxtOrg have somehow low ICC values indicating their especially low reliability on each other raters. Also, I will say that for the overall ICC value for each rubric, they are all not high.

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

From the table below showing the ICC value for ratings on each rubric and the value for ratings from raters pairwisely from the subset containing only the 13 artifacts, we can see that InterpResm RsrchQ, and TxtOrg have somehow low ICC values indicating their especially low reliability on each other raters. Also, I will say that for the overall ICC value for each rubric, they are all not high. And ICC for all data and ICC for subset are quite similar.

```
##            ICC.alldata ICC.common        a12  a23  a13
## CritDes           0.67       0.57       0.54 0.69 0.62
## InitEDA           0.69       0.49       0.69 0.85 0.54
## InterpRes         0.22       0.23       0.62 0.62 0.54
## RsrchQ            0.21       0.19       0.38 0.54 0.77
## SelMeth           0.47       0.52       0.92 0.69 0.62
## TxtOrg            0.19       0.14       0.69 0.54 0.62
## VisOrg            0.66       0.59       0.54 0.77 0.77
```

From the tables above, even though that the reliability among raters for each rubric is not high, we cannot say that they disagree with each other. Based on the tables below, I will say that most times the raters agree

with each other as the percent exact agreement between each pair of raters for each rubric is not low and most time their ratings match the ratings from the other.

will add some more 2-way tables including percent exact agreement between the two raters later

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

Code Appendix part (3)(i): page ?

For the reduced dataset containing 13 common artifacts with all 3 raters' ratings, to check if fixed effects are needed for each model of each rubric, the intercept-only model for each rubric is compared with the model adding fixed effect Rater, Semester, and Sex, with backwards-elimination applied. The result of ANOVA test for each pair of models is larger than 0.05, which tells that for each rubric, the intercept-only model is adequate. Thus, the final model chosen for each rubric is:

```
## $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)
```

Code Appendix part (3)(ii): page ?

For the full dataset, to check if fixed effects are needed for each model of each rubric, the intercept-only model for each rubric is compared with the model adding fixed effect Rater, Semester, and Sex, with

backwards-elimination applied. The result of ANOVA test for some pair of models is larger than 0.05 and for some pair is less than 0,05, which tells that for rubric InitEDA, RsrchQ, and TxtOrg, the intercept-only model is adequate, but for all other four rubrics CritDes, InterpRes, SelMeth, and VisOrg, the model with some fixed effects is better than intercept-only model. Thus, the final model chosen for each rubric is:

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


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

Code Appendix part (3)(iii): page ?

For those rubrics CritDes, InterpRes, SelMeth, and VisOrg whose selected models from the previous step are not just the simple intercept-only models, we will examine each of these to see if the fixed effect make sense and if there are any interactions or additional random effects.

SelMeth: After refitting the model, from the table below giving the t-values for all variables, we can see that absolute values of t-values are large enough, indicating that the variables in this model are all significant and none of them needs to be removed.

```
##                    Estimate Std. Error t value
## as.factor(Rater)1      2.25       0.08   29.99
## as.factor(Rater)2      2.23       0.07   29.99
## as.factor(Rater)3      2.03       0.08   27.03
## SemesterS19           -0.36       0.10   -3.66
```

With this model including variables Rater and Semester, ANOVA test is applied on this model and model without Rater. The result p-value is less than 0.05, indicating that model without Rater is not adequate for

the data. Thus, old model from previous step (the one with Rater) is selected in this step.

```
## refitting model(s) with ML (instead of REML)
## Data: tall.nonmissing[tall.nonmissing$Rubric == "SelMeth", ]
## Models:
## tmp.single_intercept: as.numeric(Rating) ~ Semester + (1 | Artifact)
## tmp: as.numeric(Rating) ~ as.factor(Rater) + Semester + (1 | Artifact) - 1
##                      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    4 145.07 156.08 -68.534   137.07
## tmp                     6 142.05 158.58 -65.027   130.05 7.0146  2    0.02998 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After, interactions between fixed effects are added to the model. In this case, Rater*Semester is added. We use ANOVA test to select between the new model with interaction and the old model. The result p-value is larger than 0.05, telling that the model without interactions is adequate. And the old model is selected.

```
## Data: tall.nonmissing[tall.nonmissing$Rubric == "SelMeth", ]
## Models:
## tmp: as.numeric(Rating) ~ as.factor(Rater) + Semester + (1 | Artifact) - 1
## tmp.fixed_interactions: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) + as.factor(Rater):S
##                        npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## tmp                       6 142.05 158.58 -65.027   130.05
## tmp.fixed_interactions    8 143.46 165.49 -63.731   127.46 2.592  2     0.2736
```

Finally, we will consider if random effects (Semester|Artifact), (as.factor(Rater)|Artifact) are needed. For the reason that lmer() cannot fitthese two new models, we cannot add any random effects on this model. Thus, final model selected is the one below.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + Semester + (1 | Artifact) -
##     1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "SelMeth", ]
##
## REML criterion at convergence: 143.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0480 -0.3923 -0.0551  0.2674  2.5827
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.08973  0.2996
##  Residual             0.10842  0.3293
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1  2.25037    0.07503  29.992
## as.factor(Rater)2  2.22653    0.07424  29.991
## as.factor(Rater)3  2.03316    0.07521  27.033
## SemesterS19       -0.35860    0.09796  -3.661
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2 a.(R)3
## as.fctr(R)2  0.285
## as.fctr(R)3  0.287  0.280
## SemesterS19 -0.413 -0.391 -0.394
```

add interpretation of final model

add model for other 3 rubrics

Code Appendix part (3)(iv): page ?

still need to work on

Finally, we combine all rubrics and consider it as a whole and try to add fixed effects, interactions, and new random effects to the "combined" intercept-only model using all the data. After adding fixed effect, interactions, and random effects to the model, we will use ANOVA to select the final model.

Try adding fixed effects with no interactions. Backwards-elimination on the model with fixed effects. Add interactions to the model above. Backwards-elimination. ANOVA test on models above.

Add random effects (3 for raters, 2 for semesters, 7 for rubrics, and 21 for the interaction) ANOVA to select. Interpret model.

Discussion

(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?

The distribution of CritDes is distinguishable from other rubrics and distributions of other rubrics are indistinguishable. For rubric CritDes, we can see that largest proportion of students get score 0; however, for all other rubrics, most students get score 2 or 3 and only few get 0 or 4, and this difference makes CritDes distinguishable from others. Raters tend to give lower score for rubric CritDes than other rubrics. From everything we have above, I will say that all 3 raters are indistinguishable as they all have really similar patterns and none of them tends to give a lower/higher score.

(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?

Even though that the reliability among raters for each rubric is not high, we cannot say that they disagree with each other. I will say that most times the raters agree with each other as the percent exact agreement between each pair of raters for each rubric is not low and most time their ratings match the ratings from the other.

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

still working on the last part of this question

(4) Is there anything else interesting to say about this data?

Maybe we can also try to see if there is any tendency on the gender that if female or male tend to get distinguishable higher/lower ratings for these different rubrics and maybe later for different artifacts and other variables. From the distribution plots, roughly same amount of female and male can get 2/3 for all rubrics, but also from some rubrics, only male/female or mostly male/female get 4.0.

strengths: Thorough exploration is made on checking the effect of raters and rubrics on the ratings. Model is selected from a lot of and different tests to check the significance of variables and the necessity of fixed effects, interactions, and random effects.

limitations: Have not find a way to figure out the difference in the models fitted to the data from the 13 common items and the models fitting to all the data. Have not find ways to deal with missing values. Have not really check the validity of the models selected.

To further improve the model and the understanding of this project, maybe later in the future we can try other difference models after the missing values and other problems are addressed, and models should be selected also based on the validity.

Technical Appendix

Yuqing Xu

11/17/2021

```
library(arm)
```

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

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

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

```
##
## arm (Version 1.12-2, built: 2021-10-15)
```

```
## Working directory is /Users/abcdefg/Documents/applied linear models
```

```
library(plyr)
library(ggplot2)
```

Project 2

(1)

```
ratings = read.csv("/Users/abcdefg/Documents/applied linear models/ratings.csv")
tall = read.csv("/Users/abcdefg/Documents/applied linear models/tall.csv")
```

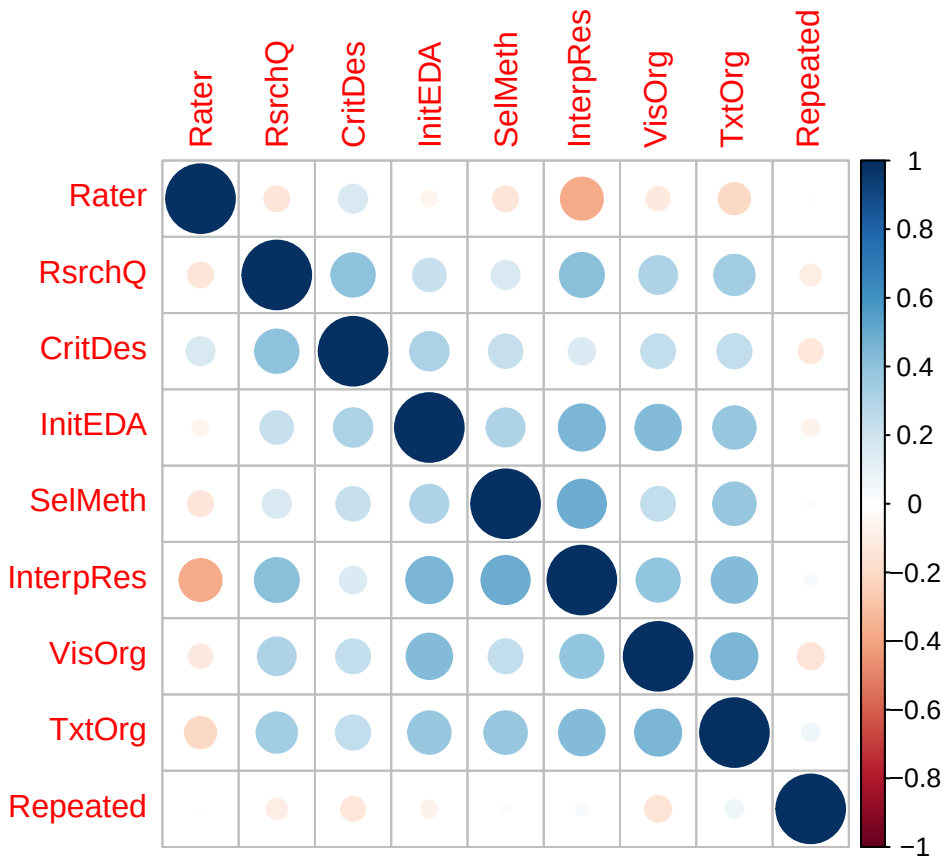
```
#View(ratings)
```

```
ratings = ratings[-c(44,99),-c(1,3,4)]
cor(ratings[,-c(2,3,11)])
```

```
##                  Rater      RsrchQ    CritDes     InitEDA     SelMeth
## Rater      1.000000000 -0.13193332  0.1676545 -0.05226415 -0.13399431
## RsrchQ    -0.131933322  1.00000000  0.4000197  0.22670065  0.16802772
## CritDes    0.167654547  0.40001972  1.0000000  0.31455017  0.23834350
## InitEDA   -0.052264154  0.22670065  0.3145502  1.00000000  0.30731794
## SelMeth   -0.133994306  0.16802772  0.2383435  0.30731794  1.00000000
## InterpRes -0.374615213  0.41008204  0.1505755  0.45157590  0.49236154
```

```
## VisOrg    -0.117266504  0.30314717  0.2435068  0.43808496  0.24429722
## TxtOrg    -0.207864667  0.34610611  0.2492529  0.38221154  0.38503177
## Repeated  -0.007612999 -0.09002461 -0.1229248 -0.06022036 -0.01428738
##              InterpRes     VisOrg      TxtOrg     Repeated
## Rater     -0.37461521 -0.1172665 -0.20786467 -0.007612999
## RsrchQ     0.41008204  0.3031472  0.34610611 -0.090024613
## CritDes    0.15057546  0.2435068  0.24925291 -0.122924815
## InitEDA    0.45157590  0.4380850  0.38221154 -0.060220357
## SelMeth    0.49236154  0.2442972  0.38503177 -0.014287383
## InterpRes  1.00000000  0.3952725  0.43904367  0.030394942
## VisOrg     0.39527250  1.0000000  0.45021670 -0.144226245
## TxtOrg     0.43904367  0.4502167  1.00000000  0.068682449
## Repeated   0.03039494 -0.1442262  0.06868245  1.000000000
```

```
corrplot::corrplot(cor(ratings[,-c(2,3,11)]))
```

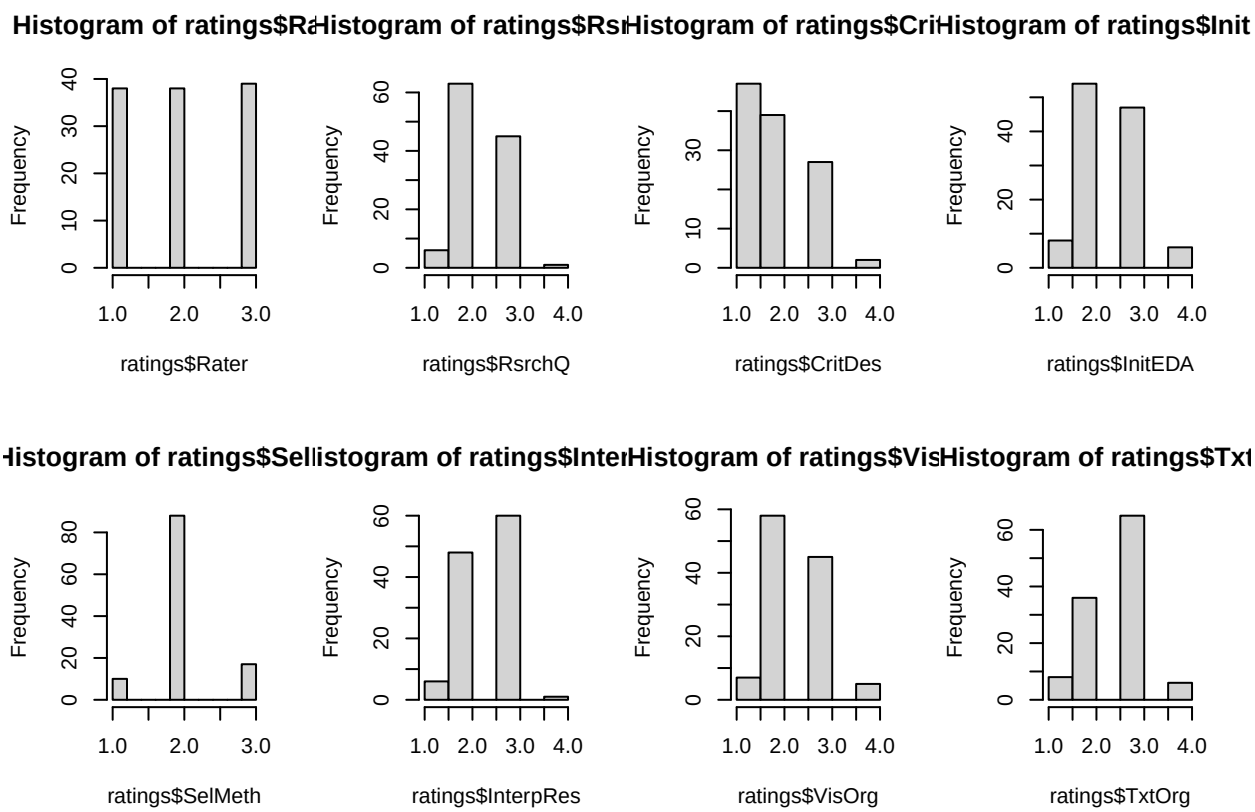



```
summary(ratings)
```

```
##      Rater         Semester            Sex              RsrchQ
##  Min.   :1.000   Length:115         Length:115         Min.   :1.000
##  1st Qu.:1.000   Class :character   Class :character   1st Qu.:2.000
##  Median :2.000   Mode  :character   Mode  :character   Median :2.000
##  Mean   :2.009                                         Mean   :2.357
##  3rd Qu.:3.000                                         3rd Qu.:3.000
##  Max.   :3.000                                         Max.   :4.000
```

```
##     CritDes         InitEDA         SelMeth        InterpRes
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :2.000   Median :3.000
##  Mean   :1.861   Mean   :2.443   Mean   :2.061   Mean   :2.487
##  3rd Qu.:2.500   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :4.000   Max.   :4.000   Max.   :3.000   Max.   :4.000
##      VisOrg          TxtOrg      Artifact            Repeated
##  Min.   :1.000   Min.   :1.0   Length:115         Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:2.0   Class :character   1st Qu.:0.0000
##  Median :2.000   Median :3.0   Mode  :character   Median :0.0000
##  Mean   :2.417   Mean   :2.6                      Mean   :0.3391
##  3rd Qu.:3.000   3rd Qu.:3.0                      3rd Qu.:1.0000
##  Max.   :4.000   Max.   :4.0                      Max.   :1.0000
```

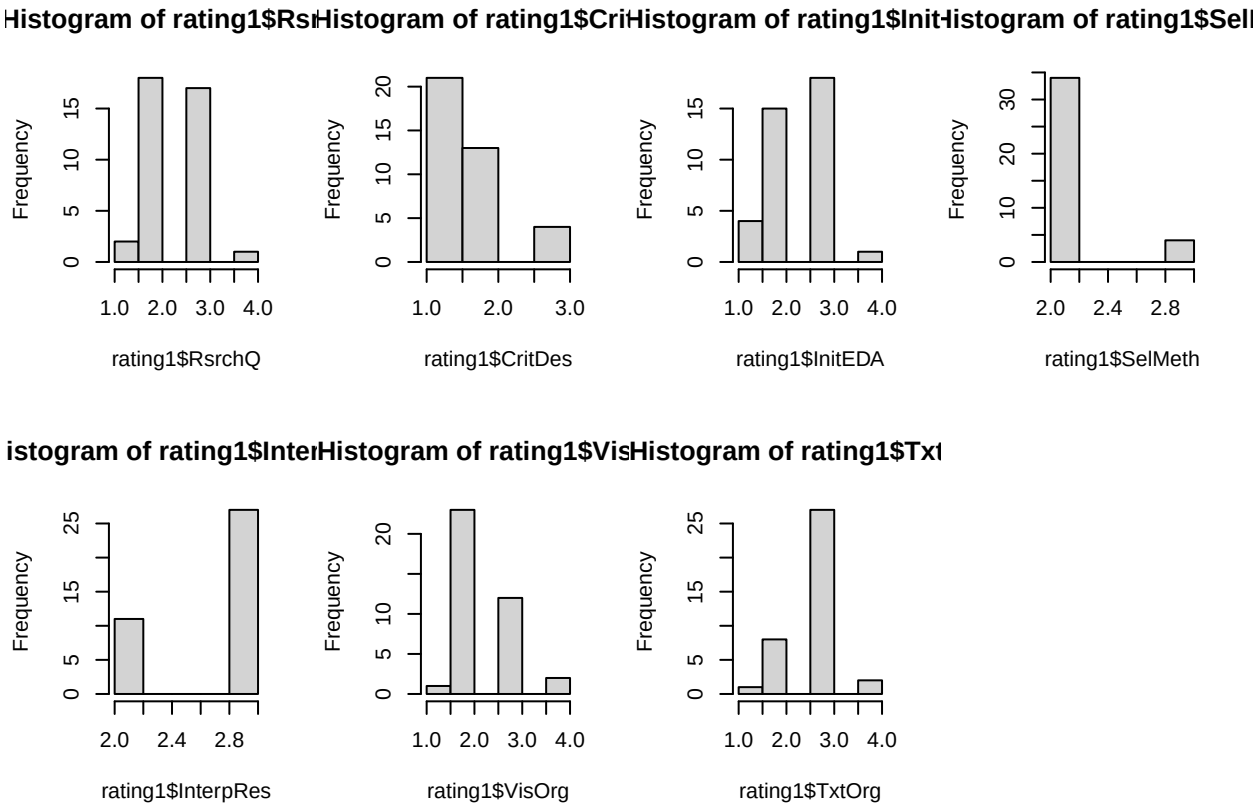
```r
par(mfrow = c(2,4))
hist(ratings$Rater)
hist(ratings$RsrchQ)
hist(ratings$CritDes)
hist(ratings$InitEDA)
hist(ratings$SelMeth)
hist(ratings$InterpRes)
hist(ratings$VisOrg)
hist(ratings$TxtOrg)
```



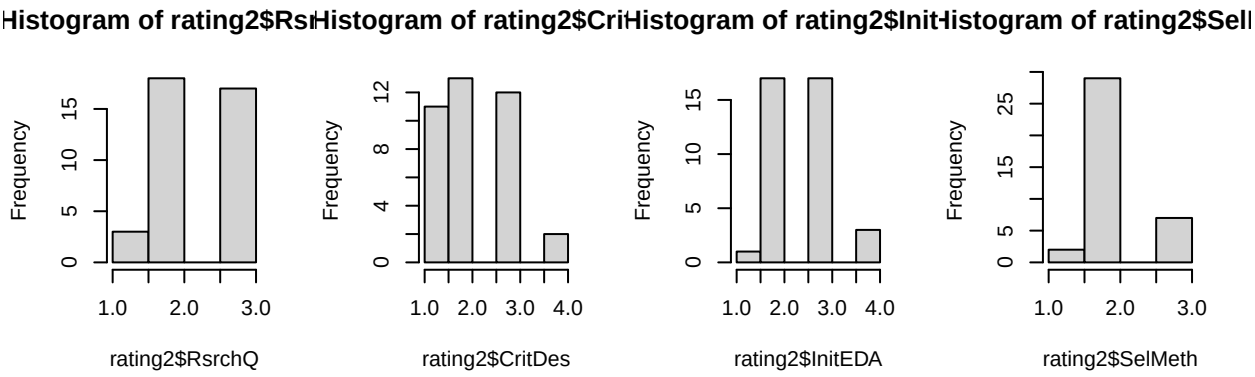
For all these rubrics, RsrchQ, InitEDA, SelMeth, InterpRes, VisOrg, TxtOrg look similar with each other

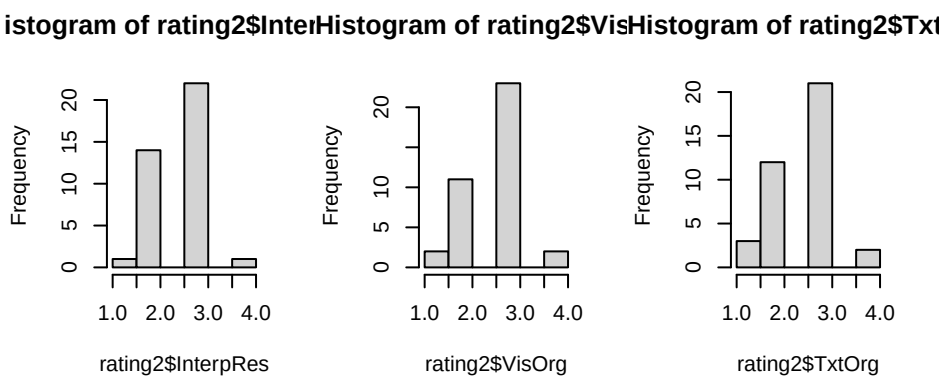
when comparing to Rater and CritDes. They have more values at the middle like 2.0 and 3.0, and far less values at 1.0 and 4.0. CritDes is tight skewed that it has higher frequency at value 1.0 and 2.0 other than 3.0 and 4.0. I will consider those similar rubrics as not distinguishable and others as distinguishable.

```
par(mfrow = c(2,4))
rating1 <- ratings[ratings$Rater == 1, ]
#View(rating1)
hist(rating1$RsrchQ)
hist(rating1$CritDes)
hist(rating1$InitEDA)
hist(rating1$SelMeth)
hist(rating1$InterpRes)
hist(rating1$VisOrg)
hist(rating1$TxtOrg)
```

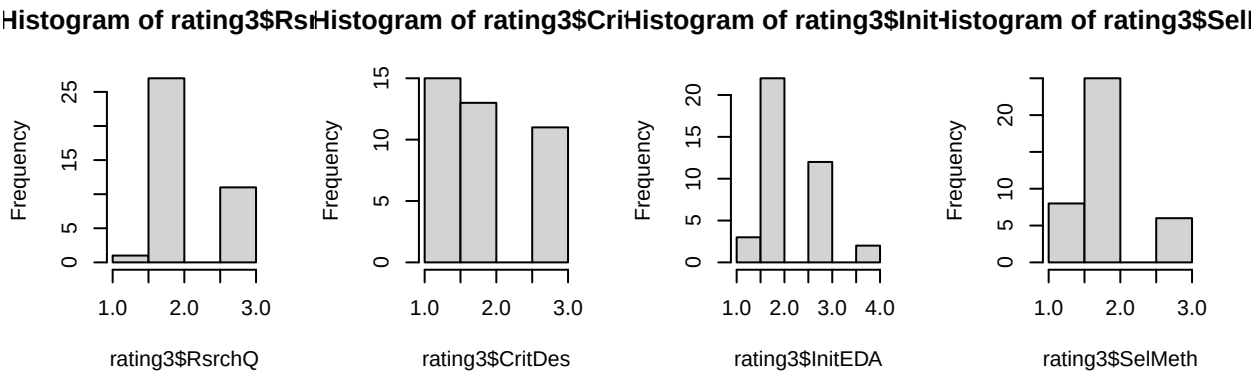


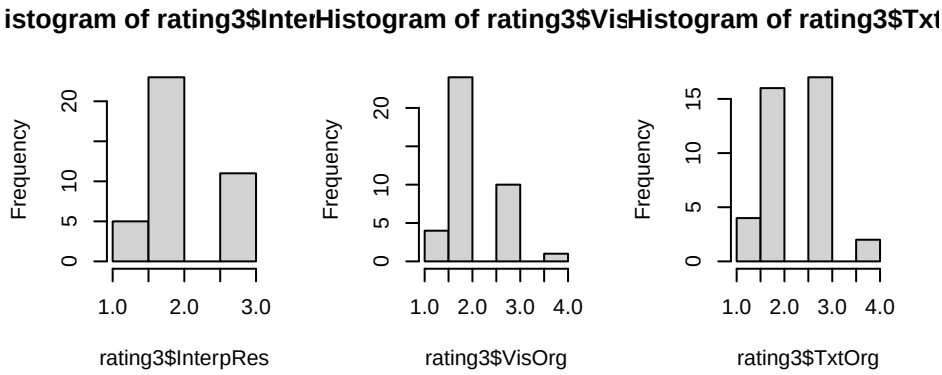

```
par(mfrow = c(2,4))
rating2 <- ratings[ratings$Rater == 2, ]
#View(rating1)
hist(rating2$RsrchQ)
hist(rating2$CritDes)
hist(rating2$InitEDA)
hist(rating2$SelMeth)
hist(rating2$InterpRes)
hist(rating2$VisOrg)
hist(rating2$TxtOrg)
```






```
par(mfrow = c(2,4))
rating3 <- ratings[ratings$Rater == 3, ]
#View(rating1)
hist(rating3$RsrchQ)
hist(rating3$CritDes)
hist(rating3$InitEDA)
hist(rating3$SelMeth)
hist(rating3$InterpRes)
hist(rating3$VisOrg)
hist(rating3$TxtOrg)
```





By the distribution of each rubric from each rater, I do not think there is rater that gives distinguishable ratings, as each set of rubrics for all raters tends to have similar distributions, which means that they give similar ratings to rubrics. Also, from the correlation plot, we cannot see any string correlation between rater and other rubrics. Thus, I do not think there is rater that tends to give especially high or low ratings/distinguishable ratings.

(2)

Measure the intraclass correlation, ICC value, to find out if the raters agree with each other.

```
#View(tall)
c <- tall[grep("O",tall$Artifact),]
```

```
RsrchQ_ratings <- c[c$Rubric == "RsrchQ",]
CritDes_ratings <- c[c$Rubric == "CritDes",]
InitEDA_ratings <- c[c$Rubric == "InitEDA",]
SelMeth_ratings <- c[c$Rubric == "SelMeth",]
InterpRes_ratings <- c[c$Rubric == "InterpRes",]
VisOrg_ratings <- c[c$Rubric == "VisOrg",]
TxtOrg_ratings <- c[c$Rubric == "TxtOrg",]
```

```
RsrchQ_mod = lmer(Rating ~ 1 + (1|Rater), data = RsrchQ_ratings)
```

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

```
CritDes_mod = lmer(Rating ~ 1 + (1|Rater), data = CritDes_ratings)
```

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

```
InitEDA_mod = lmer(Rating ~ 1 + (1|Rater), data = InitEDA_ratings)
SelMeth_mod = lmer(Rating ~ 1 + (1|Rater), data = SelMeth_ratings)
```

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

```
InterpRes_mod = lmer(Rating ~ 1 + (1|Rater), data = InterpRes_ratings)
VisOrg_mod = lmer(Rating ~ 1 + (1|Rater), data = VisOrg_ratings)
```

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

```
TxtOrg_mod = lmer(Rating ~ 1 + (1|Rater), data = TxtOrg_ratings)
```

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

RsrchQ

```
Repeated <- ratings[ratings$Repeated==1,]

RsrchQ_r1_r2 <- data.frame(r1 = Repeated$RsrchQ[Repeated$Rater == 1], r2 = Repeated$RsrchQ[Repeated$Rat
                           a1 = Repeated$Artifact[Repeated$Rater == 1], a2 = Repeated$Artifact[Repeated
r1 <- factor(RsrchQ_r1_r2$r1, levels = 1:4)
r2 <- factor(RsrchQ_r1_r2$r2, levels = 1:4)
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
```

From the table above, we can see that for rater 1 and rater 2, at the rubric RsrchQ, they have the same rate of 5/13, 4 same on 2 and 1 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for about half of them. And only few are rated really differently. Thus, I will day that they usually agree on each other's scores.

```
RsrchQ_r2_r3 <- data.frame(r2 = Repeated$RsrchQ[Repeated$Rater == 2], r3 = Repeated$RsrchQ[Repeated$Rat
                           a2 = Repeated$Artifact[Repeated$Rater == 2], a3 = Repeated$Artifact[Repeated
r2 <- factor(RsrchQ_r2_r3$r2, levels = 1:4)
r3 <- factor(RsrchQ_r2_r3$r3, levels = 1:4)
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
```

From the table above, we can see that for rater 2 and rater 3, at the rubric RsrchQ, they have the same rate of 7/13, 5 same on 2 and 2 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for about almost rest of them. Thus, I will say that they usually agree on each other's scores. Thus, overall I will say for all 13 artifacts and for rubric RsrchQ, the raters usually agree with each other's ratings.

```
library(performance)
```

```
##
## Attaching package: 'performance'
```

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

```
RsrchQ_ratings <- c[c$Rubric == "RsrchQ",]
RsrchQ_mod = lmer(Rating ~ 1 + (1|Rater), data=RsrchQ_ratings)
```

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

```
summary(RsrchQ_mod)
```

```
## 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
```

```
performance::icc(RsrchQ_mod)
```

```
## Warning: Can't compute random effect variances. Some variance components equal zero. Your
##   model may suffer from singularity (see '?lme4::isSingular' and
##   '?performance::check_singularity').
##   Solution: Respecify random structure! You may also decrease the 'tolerance'
##   level to enforce the calculation of random effect variances.

## [1] NA
```

By the results given above, we can see that the differnce among the ratings by raters are small for rubric RsrchQ. Then I will say they have roughly uniform ratings on rubric RsrchQ.

CritDes

```
CritDes_r1_r2 <- data.frame(r1 = Repeated$CritDes[Repeated$Rater == 1], r2 = Repeated$CritDes[Repeated$
                            a1 = Repeated$Artifact[Repeated$Rater == 1], a2 = Repeated$Artifact[Repeated
r1 <- factor(CritDes_r1_r2$r1, levels = 1:4)
r2 <- factor(CritDes_r1_r2$r2, levels = 1:4)
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
```

From the table above, we can see that for rater 1 and rater 2, at the rubric CritDes, they have the same rate of 7/13, 3 same on 1, 3 same on 2 and 1 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for most part of it, and there is only 1 has grade 1 from one and 3 from the other. Only few are rated really differently. Thus, I will say that they usually agree on each other's scores.

```
CritDes_r2_r3 <- data.frame(r2 = Repeated$CritDes[Repeated$Rater == 2], r3 = Repeated$CritDes[Repeated$
                            a2 = Repeated$Artifact[Repeated$Rater == 2], a3 = Repeated$Artifact[Repeated
r2 <- factor(CritDes_r2_r3$r2, levels = 1:4)
r3 <- factor(CritDes_r2_r3$r3, levels = 1:4)
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
```

From the table above, we can see that for rater 2 and rater 3, at the rubric CritDes, they have the same rate of 9/13, 5 same on 1 and 3 same on 2, and 1 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for about almost rest of them. Thus, I will say that they usually agree on each other's scores. Thus, overall I will say for all 13 artifacts and for rubric CritDes, the raters usually agree with each other's ratings.

```
CritDes_ratings <- c[c$Rubric == "CritDes",]
CritDes_mod = lmer(Rating ~ 1 + (1|Rater), data=CritDes_ratings)
```

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

```
summary(CritDes_mod)
```

```
## 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
```

```
performance::icc(CritDes_mod)
```

```
## Warning: Can't compute random effect variances. Some variance components equal zero. Your
##   model may suffer from singularity (see '?lme4::isSingular' and
##   '?performance::check_singularity').
##   Solution: Respecify random structure! You may also decrease the 'tolerance'
##   level to enforce the calculation of random effect variances.
```

```
## [1] NA
```

By the results given above, we can see that the differnce among the ratings by raters are small for rubric CritDes Then I will say they have roughly uniform ratings on rubric CritDes.

InitEDA

```
InitEDA_r1_r2 <- data.frame(r1 = Repeated$InitEDA[Repeated$Rater == 1], r2 = Repeated$InitEDA[Repeated$
                            a1 = Repeated$Artifact[Repeated$Rater == 1], a2 = Repeated$Artifact[Repeate
r1 <- factor(InitEDA_r1_r2$r1, levels = 1:4)
r2 <- factor(InitEDA_r1_r2$r2, levels = 1:4)
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
```

From the table above, we can see that for rater 1 and rater 2, at the rubric InitEDA, they have the same rate of 9/13, 4 same on 2, 5 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for almost all the rest. Thus, I will say that they usually agree on each other's scores.

```
InitEDA_r2_r3 <- data.frame(r2 = Repeated$InitEDA[Repeated$Rater == 2], r3 = Repeated$InitEDA[Repeated$
                            a2 = Repeated$Artifact[Repeated$Rater == 2], a3 = Repeated$Artifact[Repeate
r2 <- factor(InitEDA_r2_r3$r2, levels = 1:4)
r3 <- factor(InitEDA_r2_r3$r3, levels = 1:4)
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
```

From the table above, we can see that for rater 2 and rater 3, at the rubric InitEDA, they have the same rate of 11/13, 8 same on 2 and 3 same on 3, and 1 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for about almost rest of them. Thus, I will say that they usually agree on each other's scores. Thus, overall I will say for all 13 artifacts and for rubric InitEDA, the raters usually agree with each other's ratings.

```
InitEDA_ratings <- c[c$Rubric == "InitEDA",]
InitEDA_mod = lmer(Rating ~ 1 + (1|Rater), data=InitEDA_ratings)
summary(InitEDA_mod)
```

```
## 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
```

```
performance::icc(InitEDA_mod)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.003
##   Conditional ICC: 0.003
```

By the results given above, we can see that the differnce among the ratings by raters are small for rubric InitEDA Then I will say they have roughly uniform ratings on rubric InitEDA.

SelMeth

```
SelMeth_r1_r2 <- data.frame(r1 = Repeated$SelMeth[Repeated$Rater == 1], r2 = Repeated$SelMeth[Repeated$
                            a1 = Repeated$Artifact[Repeated$Rater == 1], a2 = Repeated$Artifact[Repeate
r1 <- factor(SelMeth_r1_r2$r1, levels = 1:4)
r2 <- factor(SelMeth_r1_r2$r2, levels = 1:4)
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
```

From the table above, we can see that for rater 1 and rater 2, at the rubric SelMeth, they have the same rate of 12/13, 10 same on 2, 2 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for almost all the rest. Thus, I will say that they usually agree on each other's scores.

```
SelMeth_r2_r3 <- data.frame(r2 = Repeated$SelMeth[Repeated$Rater == 2], r3 = Repeated$SelMeth[Repeated$
                            a2 = Repeated$Artifact[Repeated$Rater == 2], a3 = Repeated$Artifact[Repeate
r2 <- factor(SelMeth_r2_r3$r2, levels = 1:4)
r3 <- factor(SelMeth_r2_r3$r3, levels = 1:4)
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
```

From the table above, we can see that for rater 2 and rater 3, at the rubric SelMeth, they have the same rate of 8/13, 7 same on 2 and 1 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for about almost rest of them. Thus, I will say that they usually agree on each other's scores. Thus, overall I will say for all 13 artifacts and for rubric SelMeth, the raters usually agree with each other's ratings.

```
SelMeth <- c[c$Rubric == "SelMeth",]
SelMeth_mod = lmer(Rating ~ 1 + (1|Rater), data=SelMeth_ratings)
```

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

```
summary(SelMeth_mod)
```

```
## 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
```

```
performance::icc(SelMeth_mod)
```

```
## Warning: Can't compute random effect variances. Some variance components equal zero. Your
##   model may suffer from singularity (see '?lme4::isSingular' and
##   '?performance::check_singularity').
##   Solution: Respecify random structure! You may also decrease the 'tolerance'
##   level to enforce the calculation of random effect variances.
```

```
## [1] NA
```

By the results given above, we can see that the differnce among the ratings by raters are small for rubric SelMeth Then I will say they have roughly uniform ratings on rubric SelMeth.

InterpRes

```
InterpRes_r1_r2 <- data.frame(r1 = Repeated$InterpRes[Repeated$Rater == 1], r2 = Repeated$InterpRes[Repe
                              a1 = Repeated$Artifact[Repeated$Rater == 1], a2 = Repeated$Artifact[Repeate
r1 <- factor(InterpRes_r1_r2$r1, levels = 1:4)
r2 <- factor(InterpRes_r1_r2$r2, levels = 1:4)
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
```

From the table above, we can see that for rater 1 and rater 2, at the rubric InterpRes, they have the same rate of 8/13, 3 same on 2, 5 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for almost all the rest except for one that it is rated at 2 by one rater and 4 by the other. Thus, I will say that they usually agree on each other's scores.

```
InterpRes_r2_r3 <- data.frame(r2 = Repeated$InterpRes[Repeated$Rater == 2], r3 = Repeated$InterpRes[Rep
                              a2 = Repeated$Artifact[Repeated$Rater == 2], a3 = Repeated$Artifact[Repeate
r2 <- factor(InterpRes_r2_r3$r2, levels = 1:4)
r3 <- factor(InterpRes_r2_r3$r3, levels = 1:4)
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
```

From the table above, we can see that for rater 2 and rater 3, at the rubric InterpRes, they have the same rate of 8/13, 4 same on 2 and 4 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for about almost rest of them except for one that it is grade at 2 by one rater and 4 by another rater. Thus, I will say that they usually agree on each other's scores. Thus, overall I will say for all 13 artifacts and for rubric InterpRes, the raters usually agree with each other's ratings.

```
InterpRes <- c[c$Rubric == "InterpRes",]
InterpRes_mod = lmer(Rating ~ 1 + (1|Rater), data=InterpRes_ratings)
summary(InterpRes_mod)
```

```
## 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
```

```
performance::icc(InterpRes_mod)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.011
##   Conditional ICC: 0.011
```

By the results given above, we can see that the differnce among the ratings by raters are small for rubric InterpRes Then I will say they have roughly uniform ratings on rubric InterpRes.

VisOrg

```
VisOrg_r1_r2 <- data.frame(r1 = Repeated$VisOrg[Repeated$Rater == 1], r2 = Repeated$VisOrg[Repeated$Rat
                           a1 = Repeated$Artifact[Repeated$Rater == 1], a2 = Repeated$Artifact[Repeate
r1 <- factor(VisOrg_r1_r2$r1, levels = 1:4)
r2 <- factor(VisOrg_r1_r2$r2, levels = 1:4)
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
```

From the table above, we can see that for rater 1 and rater 2, at the rubric VisOrg, they have the same rate of 6/13, 4 same on 2, 2 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for almost all the rest. Thus, I will say that they usually agree on each other's scores.

```
VisOrg_r2_r3 <- data.frame(r2 = Repeated$VisOrg[Repeated$Rater == 2], r3 = Repeated$VisOrg[Repeated$Rat
                           a2 = Repeated$Artifact[Repeated$Rater == 2], a3 = Repeated$Artifact[Repeate
r2 <- factor(VisOrg_r2_r3$r2, levels = 1:4)
r3 <- factor(VisOrg_r2_r3$r3, levels = 1:4)
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
```

From the table above, we can see that for rater 2 and rater 3, at the rubric VisOrg, they have the same rate of 9/13, 5 same on 2 and 4 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for about almost rest of them. Thus, I will say that they usually agree on each other's scores. Thus, overall I will say for all 13 artifacts and for rubric VisOrg, the raters usually agree with each other's ratings.

```
VisOrg <- c[c$Rubric == "VisOrg",]
VisOrg_mod = lmer(Rating ~ 1 + (1|Rater), data=VisOrg_ratings)
```

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

```
summary(VisOrg_mod)
```

```
## 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
```

```
performance::icc(VisOrg_mod)
```

```
## Warning: Can't compute random effect variances. Some variance components equal zero. Your
##   model may suffer from singularity (see '?lme4::isSingular' and
##   '?performance::check_singularity').
##   Solution: Respecify random structure! You may also decrease the 'tolerance'
##   level to enforce the calculation of random effect variances.
```

```
## [1] NA
```

By the results given above, we can see that the differnce among the ratings by raters are small for rubric VisOrg Then I will say they have roughly uniform ratings on rubric VisOrg.

TxtOrg

```
TxtOrg_r1_r2 <- data.frame(r1 = Repeated$TxtOrg[Repeated$Rater == 1], r2 = Repeated$TxtOrg[Repeated$Rat
                           a1 = Repeated$Artifact[Repeated$Rater == 1], a2 = Repeated$Artifact[Repeate
r1 <- factor(TxtOrg_r1_r2$r1, levels = 1:4)
r2 <- factor(TxtOrg_r1_r2$r2, levels = 1:4)
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
```

From the table above, we can see that for rater 1 and rater 2, at the rubric TxtOrg, they have the same rate of 9/13, 2 same on 2, 7 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for almost all the rest except for one. Thus, I will say that they usually agree on each other's scores.

```
TxtOrg_r2_r3 <- data.frame(r2 = Repeated$TxtOrg[Repeated$Rater == 2], r3 = Repeated$TxtOrg[Repeated$Rat
                           a2 = Repeated$Artifact[Repeated$Rater == 2], a3 = Repeated$Artifact[Repeate
r2 <- factor(TxtOrg_r2_r3$r2, levels = 1:4)
r3 <- factor(TxtOrg_r2_r3$r3, levels = 1:4)
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
```

From the table above, we can see that for rater 2 and rater 3, at the rubric TxtOrg, they have the same rate of 7/13, 74 same on 3. And for other artifacts that they are rated differently, they are still rated similarly for about almost rest of them. Thus, I will say that they usually agree on each other's scores. Thus, overall I will say for all 13 artifacts and for rubric TxtOrg, the raters usually agree with each other's ratings.

```
TxtOrg <- c[c$Rubric == "TxtOrg",]
TxtOrg_mod = lmer(Rating ~ 1 + (1|Rater), data=TxtOrg_ratings)
```

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

```
summary(TxtOrg_mod)
```

```
## 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
```

```
performance::icc(TxtOrg_mod)
```

```
## Warning: Can't compute random effect variances. Some variance components equal zero. Your
##   model may suffer from singularity (see '?lme4::isSingular' and
##   '?performance::check_singularity').
##   Solution: Respecify random structure! You may also decrease the 'tolerance'
##   level to enforce the calculation of random effect variances.

## [1] NA
```

By the results given above, we can see that the differnce among the ratings by raters are small for rubric TxtOrg Then I will say they have roughly uniform ratings on rubric TxtOrg.

random effect in each artifact

```
performance::icc(lmer(Rating ~ 1 + (1|Artifact), data=RsrchQ_ratings))
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.189
##   Conditional ICC: 0.189
```

```
performance::icc(lmer(Rating ~ 1 + (1|Artifact), data=CritDes_ratings))
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.573
##   Conditional ICC: 0.573
```

```
performance::icc(lmer(Rating ~ 1 + (1|Artifact), data=InitEDA_ratings))
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.493
##   Conditional ICC: 0.493
```

```
performance::icc(lmer(Rating ~ 1 + (1|Artifact), data=SelMeth_ratings))
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.521
##   Conditional ICC: 0.521
```

```
performance::icc(lmer(Rating ~ 1 + (1|Artifact), data=InterpRes_ratings))
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.230
##   Conditional ICC: 0.230
```

```
performance::icc(lmer(Rating ~ 1 + (1|Artifact), data=VisOrg_ratings))
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.592
##   Conditional ICC: 0.592
```

```
performance::icc(lmer(Rating ~ 1 + (1|Artifact), data=TxtOrg_ratings))
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.143
##   Conditional ICC: 0.143
```

Here we have the ICC value for all rubrics with raondom effect artifact. We can see that the value for RsrchQ, InterpRes and TxtOrg are about 0.2, which gives that rating for different artifacts do not change a lot, and the ratings change comparatively larger for other rubrics with higher ICC values.

(3)

```
lmer_3 <- lmer(Rating ~ 1 + (0 + Rubric|Artifact), data = tall)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00236116 (tol = 0.002, component 1)
```

```
lmer_3_1 <- update(lmer_3, .~. + Semester)
```

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

```
anova(lmer_3, lmer_3_1)
```

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

```
## Data: tall
## Models:
## lmer_3: Rating ~ 1 + (0 + Rubric | Artifact)
## lmer_3_1: Rating ~ (0 + Rubric | Artifact) + Semester
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer_3     30 1537.2 1678.3 -738.58   1477.2
## lmer_3_1   31 1535.1 1681.0 -736.57   1473.1 4.0182  1    0.04501 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
lmer_3_2 <- update(lmer_3, .~. + Rater)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## unable to evaluate scaled gradient
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge: degenerate Hessian with 2 negative eigenvalues
```

```r
anova(lmer_3, lmer_3_2)
```

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

```
## Data: tall
## Models:
## lmer_3: Rating ~ 1 + (0 + Rubric | Artifact)
## lmer_3_2: Rating ~ (0 + Rubric | Artifact) + Rater
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer_3     30 1537.2 1678.3 -738.58   1477.2
## lmer_3_2   31 1530.9 1676.8 -734.45   1468.9 8.2508  1   0.004073 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
lmer_3_3 <- update(lmer_3, .~. + Sex)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00739175 (tol = 0.002, component 1)
```

```r
anova(lmer_3, lmer_3_3)
```

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

```
## Data: tall
## Models:
## lmer_3: Rating ~ 1 + (0 + Rubric | Artifact)
## lmer_3_3: Rating ~ (0 + Rubric | Artifact) + Sex
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer_3     30 1537.2 1678.3 -738.58   1477.2
## lmer_3_3   32 1536.9 1687.5 -736.43   1472.9 4.2923  2     0.1169
```

```r
lmer_3_4 <- update(lmer_3, .~. + Repeated)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00629081 (tol = 0.002, component 1)
```

```r
anova(lmer_3, lmer_3_4)
```

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

```
## Data: tall
## Models:
## lmer_3: Rating ~ 1 + (0 + Rubric | Artifact)
## lmer_3_4: Rating ~ (0 + Rubric | Artifact) + Repeated
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer_3     30 1537.2 1678.3 -738.58   1477.2
## lmer_3_4   31 1538.1 1684.0 -738.05   1476.1 1.0476  1     0.3061
```

```
RsrchQ_tall <- tall[tall$Rubric == "RsrchQ",]
CritDes_tall <- tall[tall$Rubric == "CritDes",]
InitEDA_tall <- tall[tall$Rubric == "InitEDA",]
SelMeth_tall <- tall[tall$Rubric == "SelMeth",]
InterpRes_tall <- tall[tall$Rubric == "InterpRes",]
VisOrg_tall <- tall[tall$Rubric == "VisOrg",]
TxtOrg_tall <- tall[tall$Rubric == "TxtOrg",]
```

RsrchQ

```
RsrchQ_3 <- lmer(Rating ~ 1 + (1 + 1|Artifact), data = RsrchQ_tall)
```

```
RsrchQ_3_1 <- update(RsrchQ_3, .~. + Semester)
anova(RsrchQ_3, RsrchQ_3_1)
```

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

```
## Data: RsrchQ_tall
## Models:
## RsrchQ_3: Rating ~ 1 + (1 + 1 | Artifact)
## RsrchQ_3_1: Rating ~ (1 + 1 | Artifact) + Semester
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## RsrchQ_3      3 213.19 221.48 -103.60   207.19
## RsrchQ_3_1    4 214.57 225.62 -103.28   206.57 0.6253  1     0.4291
```

```
RsrchQ_3_2 <- update(RsrchQ_3, .~. + Rater)
anova(RsrchQ_3, RsrchQ_3_2)
```

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

```
## Data: RsrchQ_tall
## Models:
## RsrchQ_3: Rating ~ 1 + (1 + 1 | Artifact)
## RsrchQ_3_2: Rating ~ (1 + 1 | Artifact) + Rater
##            npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
## RsrchQ_3      3 213.19 221.48 -103.6   207.19
## RsrchQ_3_2    4 213.39 224.44 -102.7   205.39 1.8008  1     0.1796
```

```
RsrchQ_3_3<- update(RsrchQ_3, .~. + Sex)
anova(RsrchQ_3, RsrchQ_3_3)
```

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

## Data: RsrchQ_tall
## Models:
## RsrchQ_3: Rating ~ 1 + (1 + 1 | Artifact)
## RsrchQ_3_3: Rating ~ (1 + 1 | Artifact) + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## RsrchQ_3      3 213.19 221.48 -103.60   207.19
## RsrchQ_3_3    5 215.37 229.18 -102.68   205.37 1.8253  2     0.4015
```

```
RsrchQ_3_4 <- update(RsrchQ_3, .~. + Repeated)
anova(RsrchQ_3, RsrchQ_3_4)
```

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

## Data: RsrchQ_tall
## Models:
## RsrchQ_3: Rating ~ 1 + (1 + 1 | Artifact)
## RsrchQ_3_4: Rating ~ (1 + 1 | Artifact) + Repeated
##            npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## RsrchQ_3      3 213.19 221.48 -103.60   207.19
## RsrchQ_3_4    4 214.57 225.62 -103.28   206.57 0.627  1     0.4285
```

CritDes

```
CritDes_3 <- lmer(Rating ~ 1 + (1 + 1|Artifact), data = CritDes_tall)
```

```
CritDes_3_1 <- update(CritDes_3, .~. + Semester)
anova(CritDes_3, CritDes_3_1)
```

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

## Data: CritDes_tall
## Models:
## CritDes_3: Rating ~ 1 + (1 + 1 | Artifact)
## CritDes_3_1: Rating ~ (1 + 1 | Artifact) + Semester
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## CritDes_3      3 280.86 289.12 -137.43   274.86
## CritDes_3_1    4 282.58 293.60 -137.29   274.58 0.2751  1     0.5999
```

```
CritDes_3_2 <- update(CritDes_3, .~. + Rater)
anova(CritDes_3, CritDes_3_2)
```

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

## Data: CritDes_tall
## Models:
## CritDes_3: Rating ~ 1 + (1 + 1 | Artifact)
## CritDes_3_2: Rating ~ (1 + 1 | Artifact) + Rater
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## CritDes_3      3 280.86 289.12 -137.43   274.86
## CritDes_3_2    4 280.76 291.77 -136.38   272.76 2.0985  1     0.1474
```

```
CritDes_3_3<- update(CritDes_3, .~. + Sex)
anova(CritDes_3, CritDes_3_3)
```

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

```
## Data: CritDes_tall
## Models:
## CritDes_3: Rating ~ 1 + (1 + 1 | Artifact)
## CritDes_3_3: Rating ~ (1 + 1 | Artifact) + Sex
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## CritDes_3      3 280.86 289.12 -137.43   274.86
## CritDes_3_3    5 282.65 296.42 -136.33   272.65 2.2017  2     0.3326
```

```
CritDes_3_4 <- update(CritDes_3, .~. + Repeated)
anova(CritDes_3, CritDes_3_4)
```

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

```
## Data: CritDes_tall
## Models:
## CritDes_3: Rating ~ 1 + (1 + 1 | Artifact)
## CritDes_3_4: Rating ~ (1 + 1 | Artifact) + Repeated
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## CritDes_3      3 280.86 289.12 -137.43   274.86
## CritDes_3_4    4 281.85 292.87 -136.93   273.85 1.0045  1     0.3162
```

InitEDA

```
InitEDA_3 <- lmer(Rating ~ 1 + (1 + 1|Artifact), data = InitEDA_tall)
```

```
InitEDA_3_1 <- update(InitEDA_3, .~. + Semester)
anova(InitEDA_3, InitEDA_3_1)
```

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

```
## Data: InitEDA_tall
## Models:
## InitEDA_3: Rating ~ 1 + (1 + 1 | Artifact)
## InitEDA_3_1: Rating ~ (1 + 1 | Artifact) + Semester
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## InitEDA_3      3 243.42 251.71 -118.71   237.42
## InitEDA_3_1    4 245.38 256.43 -118.69   237.38 0.0391  1     0.8432
```

```
InitEDA_3_2 <- update(InitEDA_3, .~. + Rater)
anova(InitEDA_3, InitEDA_3_2)
```

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

```
## Data: InitEDA_tall
## Models:
## InitEDA_3: Rating ~ 1 + (1 + 1 | Artifact)
## InitEDA_3_2: Rating ~ (1 + 1 | Artifact) + Rater
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## InitEDA_3      3 243.42 251.71 -118.71   237.42
## InitEDA_3_2    4 243.26 254.31 -117.63   235.26 2.1635  1     0.1413
```

```
InitEDA_3_3<- update(InitEDA_3, .~. + Sex)
anova(InitEDA_3, InitEDA_3_3)
```

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

## Data: InitEDA_tall
## Models:
## InitEDA_3: Rating ~ 1 + (1 + 1 | Artifact)
## InitEDA_3_3: Rating ~ (1 + 1 | Artifact) + Sex
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## InitEDA_3      3 243.42 251.71 -118.71   237.42
## InitEDA_3_3    5 246.75 260.56 -118.38   236.75 0.6718  2     0.7147
```

```
InitEDA_3_4 <- update(InitEDA_3, .~. + Repeated)
anova(InitEDA_3, InitEDA_3_4)
```

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

## Data: InitEDA_tall
## Models:
## InitEDA_3: Rating ~ 1 + (1 + 1 | Artifact)
## InitEDA_3_4: Rating ~ (1 + 1 | Artifact) + Repeated
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## InitEDA_3      3 243.42 251.71 -118.71   237.42
## InitEDA_3_4    4 245.27 256.32 -118.63   237.27 0.1544  1     0.6944
```

SelMeth

```
SelMeth_3 <- lmer(Rating ~ 1 + (1 + 1|Artifact), data = SelMeth_tall)
```

```
SelMeth_3_1 <- update(SelMeth_3, .~. + Semester)
anova(SelMeth_3, SelMeth_3_1)
```

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

## Data: SelMeth_tall
## Models:
## SelMeth_3: Rating ~ 1 + (1 + 1 | Artifact)
## SelMeth_3_1: Rating ~ (1 + 1 | Artifact) + Semester
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## SelMeth_3      3 159.53 167.82 -76.768   153.53
## SelMeth_3_1    4 148.64 159.69 -70.322   140.64 12.891  1  0.0003301 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SelMeth_3_2 <- update(SelMeth_3, .~. + Rater)
anova(SelMeth_3, SelMeth_3_2)
```

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

```
## Data: SelMeth_tall
## Models:
## SelMeth_3: Rating ~ 1 + (1 + 1 | Artifact)
## SelMeth_3_2: Rating ~ (1 + 1 | Artifact) + Rater
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## SelMeth_3      3 159.53 167.82 -76.768   153.53
## SelMeth_3_2    4 157.43 168.48 -74.714   149.43 4.1064  1    0.04272 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SelMeth_3_3<- update(SelMeth_3, .~. + Sex)
anova(SelMeth_3, SelMeth_3_3)
```

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

```
## Data: SelMeth_tall
## Models:
## SelMeth_3: Rating ~ 1 + (1 + 1 | Artifact)
## SelMeth_3_3: Rating ~ (1 + 1 | Artifact) + Sex
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## SelMeth_3      3 159.53 167.82 -76.768   153.53
## SelMeth_3_3    5 155.32 169.13 -72.660   145.32 8.2155  2    0.01644 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SelMeth_3_4 <- update(SelMeth_3, .~. + Repeated)
anova(SelMeth_3, SelMeth_3_4)
```

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

```
## Data: SelMeth_tall
## Models:
## SelMeth_3: Rating ~ 1 + (1 + 1 | Artifact)
## SelMeth_3_4: Rating ~ (1 + 1 | Artifact) + Repeated
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## SelMeth_3      3 159.53 167.82 -76.768   153.53
## SelMeth_3_4    4 161.49 172.54 -76.745   153.49 0.0453  1     0.8314
```

InterpRes_tall

```
InterpRes_3 <- lmer(Rating ~ 1 + (1 + 1|Artifact), data = InterpRes_tall)
```

```r
InterpRes_3_1 <- update(InterpRes_3, .~. + Semester)
anova(InterpRes_3, InterpRes_3_1)
```

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

## Data: InterpRes_tall
## Models:
## InterpRes_3: Rating ~ 1 + (1 + 1 | Artifact)
## InterpRes_3_1: Rating ~ (1 + 1 | Artifact) + Semester
##               npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## InterpRes_3      3 220.09 228.38 -107.05   214.09
## InterpRes_3_1    4 221.76 232.81 -106.88   213.76 0.3386  1     0.5606
```

```r
InterpRes_3_2 <- update(InterpRes_3, .~. + Rater)
anova(InterpRes_3, InterpRes_3_2)
```

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

## Data: InterpRes_tall
## Models:
## InterpRes_3: Rating ~ 1 + (1 + 1 | Artifact)
## InterpRes_3_2: Rating ~ (1 + 1 | Artifact) + Rater
##               npar    AIC    BIC   logLik deviance  Chisq Df Pr(>Chisq)
## InterpRes_3      3 220.09 228.38 -107.048   214.09
## InterpRes_3_2    4 203.79 214.84  -97.897   195.79 18.302  1  1.885e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
InterpRes_3_3<- update(InterpRes_3, .~. + Sex)
anova(InterpRes_3, InterpRes_3_3)
```

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

## Data: InterpRes_tall
## Models:
## InterpRes_3: Rating ~ 1 + (1 + 1 | Artifact)
## InterpRes_3_3: Rating ~ (1 + 1 | Artifact) + Sex
##               npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## InterpRes_3      3 220.09 228.38 -107.05   214.09
## InterpRes_3_3    5 223.14 236.95 -106.57   213.14 0.9519  2     0.6213
```

```r
InterpRes_3_4 <- update(InterpRes_3, .~. + Repeated)
anova(InterpRes_3, InterpRes_3_4)
```

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

## Data: InterpRes_tall
## Models:
## InterpRes_3: Rating ~ 1 + (1 + 1 | Artifact)
## InterpRes_3_4: Rating ~ (1 + 1 | Artifact) + Repeated
##               npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## InterpRes_3      3 220.09 228.38 -107.05   214.09
## InterpRes_3_4    4 222.01 233.06 -107.01   214.01 0.0812  1     0.7757
```

VisOrg

```
VisOrg_3 <- lmer(Rating ~ 1 + (1 + 1|Artifact), data = VisOrg_tall)
```

```
VisOrg_3_1 <- update(VisOrg_3, .~. + Semester)
anova(VisOrg_3, VisOrg_3_1)
```

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

## Data: VisOrg_tall
## Models:
## VisOrg_3: Rating ~ 1 + (1 + 1 | Artifact)
## VisOrg_3_1: Rating ~ (1 + 1 | Artifact) + Semester
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## VisOrg_3      3 228.95 237.21 -111.47   222.95
## VisOrg_3_1    4 229.33 240.34 -110.67   221.33 1.6196  1     0.2031
```

```
VisOrg_3_2 <- update(VisOrg_3, .~. + Rater)
anova(VisOrg_3, VisOrg_3_2)
```

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

## Data: VisOrg_tall
## Models:
## VisOrg_3: Rating ~ 1 + (1 + 1 | Artifact)
## VisOrg_3_2: Rating ~ (1 + 1 | Artifact) + Rater
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## VisOrg_3      3 228.95 237.21 -111.47   222.95
## VisOrg_3_2    4 230.40 241.42 -111.20   222.40 0.5461  1     0.4599
```

```
VisOrg_3_3<- update(VisOrg_3, .~. + Sex)
anova(VisOrg_3, VisOrg_3_3)
```

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

## Data: VisOrg_tall
## Models:
## VisOrg_3: Rating ~ 1 + (1 + 1 | Artifact)
## VisOrg_3_3: Rating ~ (1 + 1 | Artifact) + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## VisOrg_3      3 228.95 237.21 -111.47   222.95
## VisOrg_3_3    5 231.47 245.23 -110.73   221.47 1.4831  2     0.4764
```

```
VisOrg_3_4 <- update(VisOrg_3, .~. + Repeated)
anova(VisOrg_3, VisOrg_3_4)
```

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

```
## Data: VisOrg_tall
## Models:
## VisOrg_3: Rating ~ 1 + (1 + 1 | Artifact)
## VisOrg_3_4: Rating ~ (1 + 1 | Artifact) + Repeated
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## VisOrg_3      3 228.95 237.21 -111.47   222.95
## VisOrg_3_4    4 229.76 240.77 -110.88   221.76 1.1894  1     0.2754
```

TxtOrg

```
TxtOrg_3 <- lmer(Rating ~ 1 + (1 + 1|Artifact), data = TxtOrg_tall)
```

```
TxtOrg_3_1 <- update(TxtOrg_3, .~. + Semester)
anova(TxtOrg_3, TxtOrg_3_1)
```

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

## Data: TxtOrg_tall
## Models:
## TxtOrg_3: Rating ~ 1 + (1 + 1 | Artifact)
## TxtOrg_3_1: Rating ~ (1 + 1 | Artifact) + Semester
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## TxtOrg_3      3 251.45 259.74 -122.73   245.45
## TxtOrg_3_1    4 251.92 262.97 -121.96   243.92 1.5339  1     0.2155
```

```
TxtOrg_3_2 <- update(TxtOrg_3, .~. + Rater)
anova(TxtOrg_3, TxtOrg_3_2)
```

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

## Data: TxtOrg_tall
## Models:
## TxtOrg_3: Rating ~ 1 + (1 + 1 | Artifact)
## TxtOrg_3_2: Rating ~ (1 + 1 | Artifact) + Rater
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## TxtOrg_3      3 251.45 259.74 -122.73   245.45
## TxtOrg_3_2    4 248.88 259.93 -120.44   240.88 4.5725  1    0.03249 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TxtOrg_3_3<- update(TxtOrg_3, .~. + Sex)
anova(TxtOrg_3, TxtOrg_3_3)
```

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

## Data: TxtOrg_tall
## Models:
## TxtOrg_3: Rating ~ 1 + (1 + 1 | Artifact)
## TxtOrg_3_3: Rating ~ (1 + 1 | Artifact) + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## TxtOrg_3      3 251.45 259.74 -122.73   245.45
## TxtOrg_3_3    5 254.99 268.80 -122.50   244.99 0.4621  2     0.7937
```

```
TxtOrg_3_4 <- update(TxtOrg_3, .~. + Repeated)
anova(TxtOrg_3, TxtOrg_3_4)
```

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

```
## Data: TxtOrg_tall
## Models:
## TxtOrg_3: Rating ~ 1 + (1 + 1 | Artifact)
## TxtOrg_3_4: Rating ~ (1 + 1 | Artifact) + Repeated
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## TxtOrg_3      3 251.45 259.74 -122.73   245.45
## TxtOrg_3_4    4 252.99 264.04 -122.49   244.99 0.4656  1      0.495
```

random effect

```
library(LMERConvenienceFunctions)
test_model <- lmer(Rating ~ 1 + Rater + Semester + Rubric + Sex + Repeated + (0+Rubric|Artifact), data =
```

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

```
test_model1 <- fitLMER.fnc(test_model, ran.effects=c("(Rater|Artifact)", "(Semester|Artifact)",  "(Sex|A
```

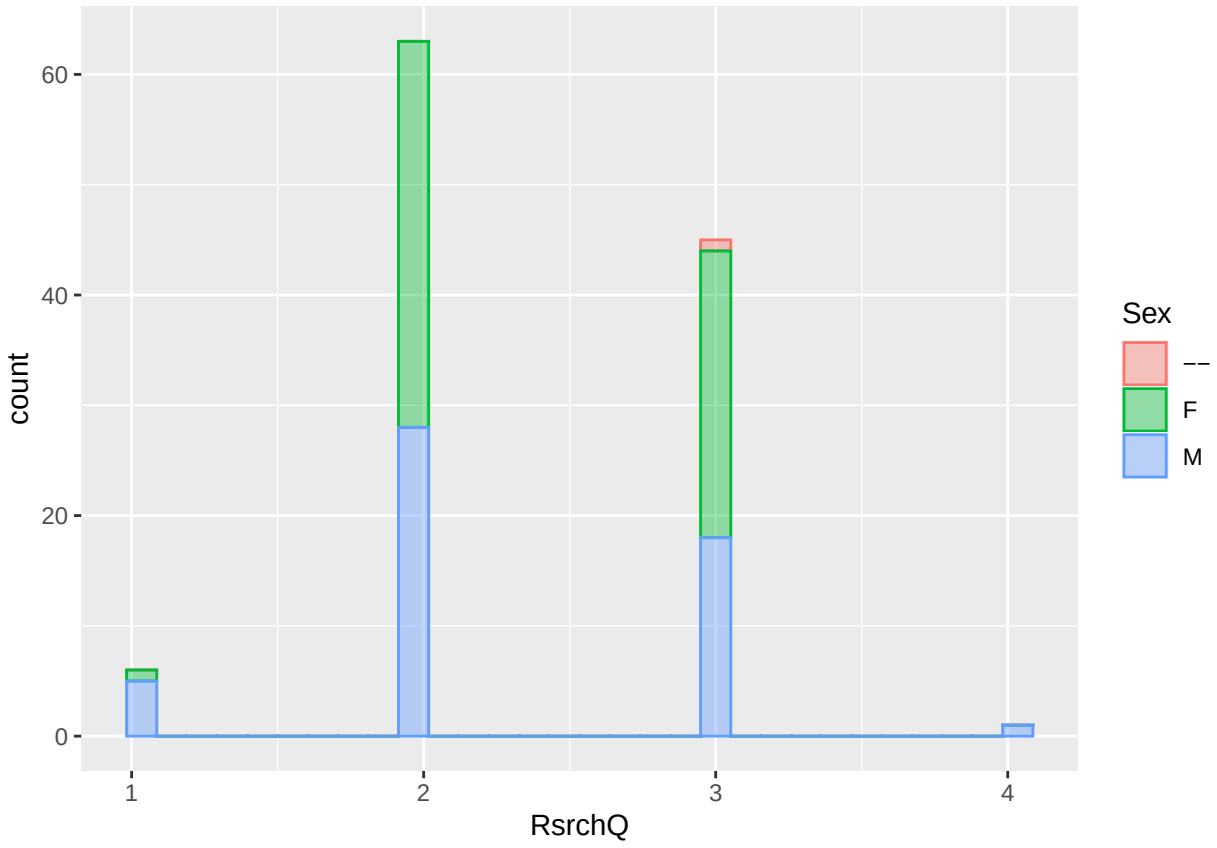
```
## ============================================================
## ===              backfitting fixed effects             ===
## ============================================================
## setting REML to FALSE
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Repeated" = 0.4811 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     BIC simple = 1664; BIC complex = 1670; decrease = -6 < 5
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.118 >= 0.05
##     not part of higher-order interaction
##     BIC simple = 1655; BIC complex = 1664; decrease = -9 < 5
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ============================================================
## ===            forwardfitting random effects           ===
## ============================================================
## evaluating addition of (Rater|Artifact) to model

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

```
##  log-likelihood ratio test p-value = 0.0005778204
##  adding (Rater|Artifact) to model
## evaluating addition of (Semester|Artifact) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 0.9899901
##  not adding (Semester|Artifact) to model
## evaluating addition of (Sex|Artifact) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 0.4960985
##  not adding (Sex|Artifact) to model
## evaluating addition of (Repeated|Artifact) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 0.9017367
##  not adding (Repeated|Artifact) to model
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## setting REML to FALSE

## boundary (singular) fit: see ?isSingular

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.0696 >= 0.05
##     not part of higher-order interaction

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

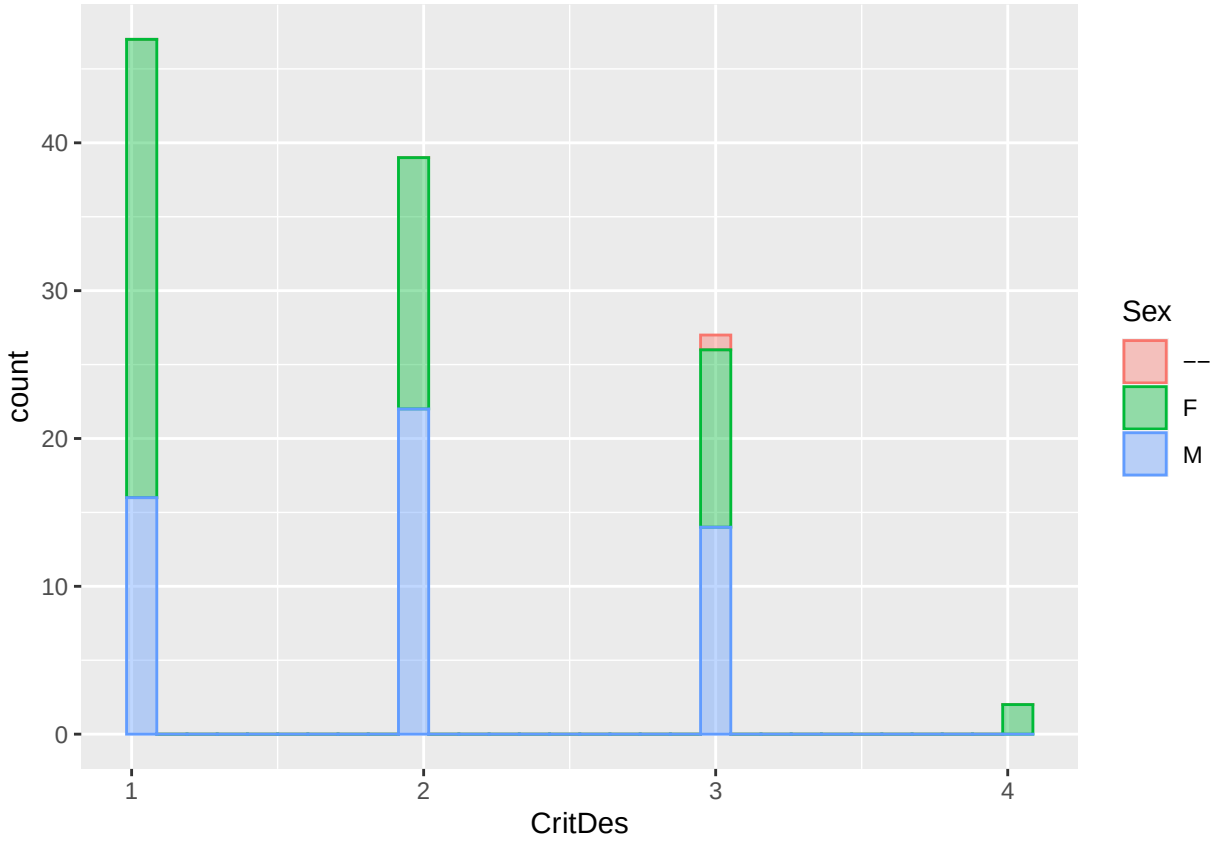
```
##     BIC simple = 1659; BIC complex = 1658; decrease = 1 < 5
##     removing term

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## resetting REML to TRUE

## boundary (singular) fit: see ?isSingular

## pruning random effects structure ...
##   nothing to prune
## log file is mylogfile.txt
```

(4)

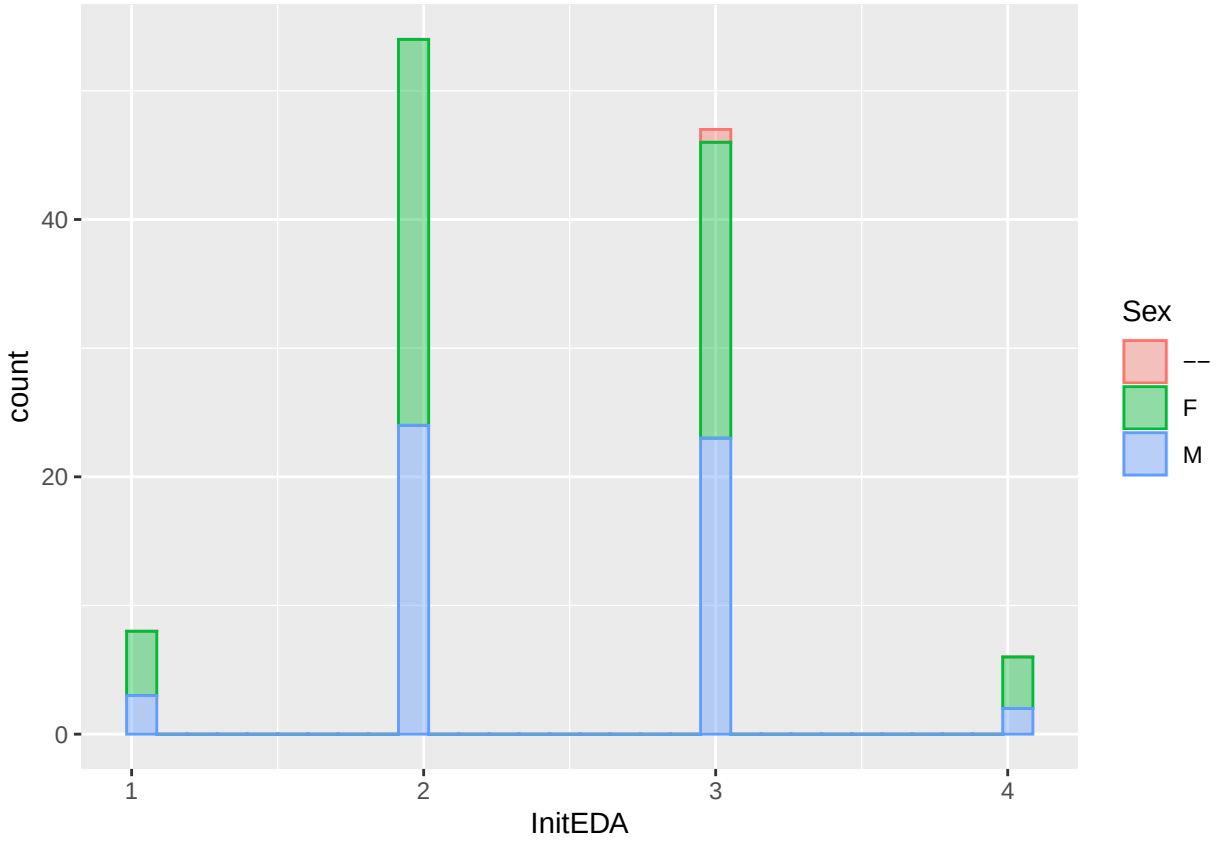
```
ggplot(ratings, aes(x = RsrchQ)) + geom_histogram(aes(color = Sex, fill = Sex), bins = 30, alpha = 0.4)
```



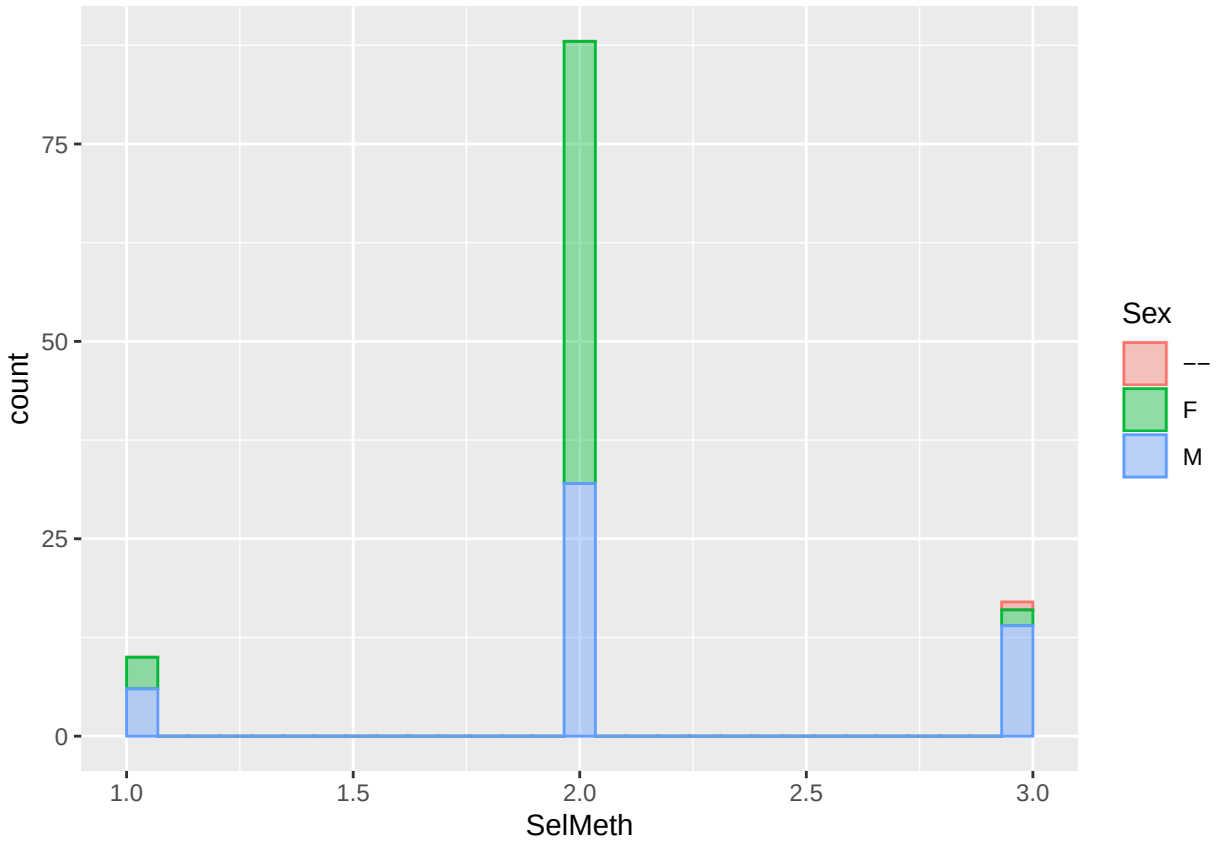

```
ggplot(ratings, aes(x = CritDes)) + geom_histogram(aes(color = Sex, fill = Sex), bins = 30, alpha = 0.4)
```



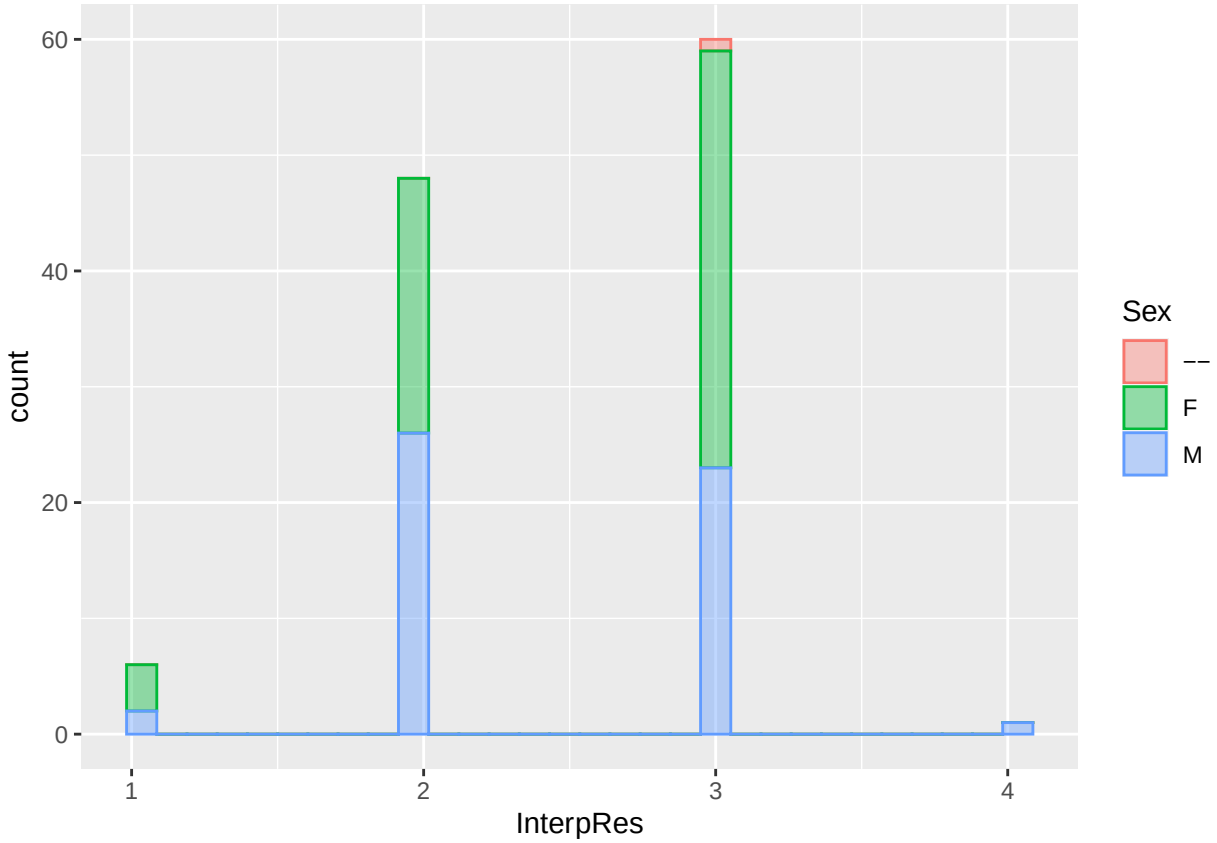

```
ggplot(ratings, aes(x = InitEDA)) + geom_histogram(aes(color = Sex, fill = Sex), bins = 30, alpha = 0.4)
```



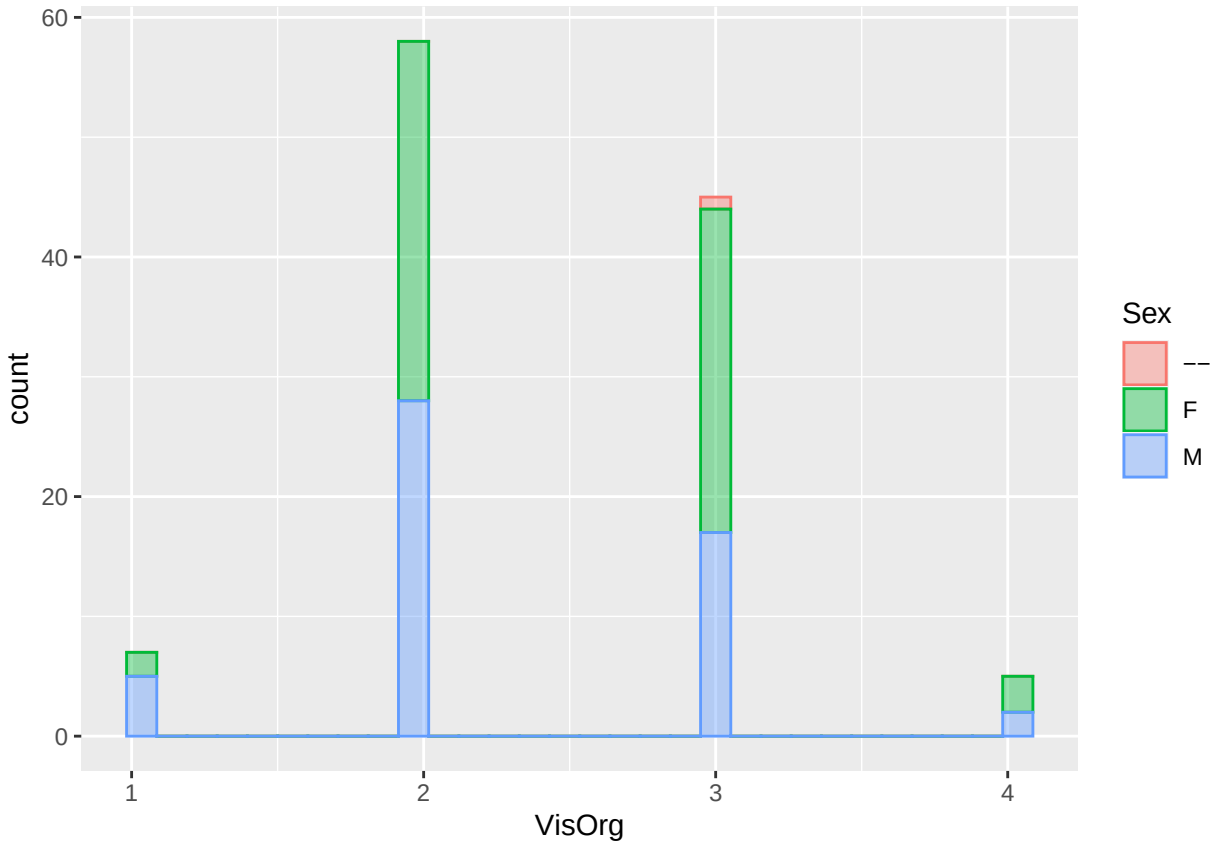

```
ggplot(ratings, aes(x = SelMeth)) + geom_histogram(aes(color = Sex, fill = Sex), bins = 30, alpha = 0.4)
```



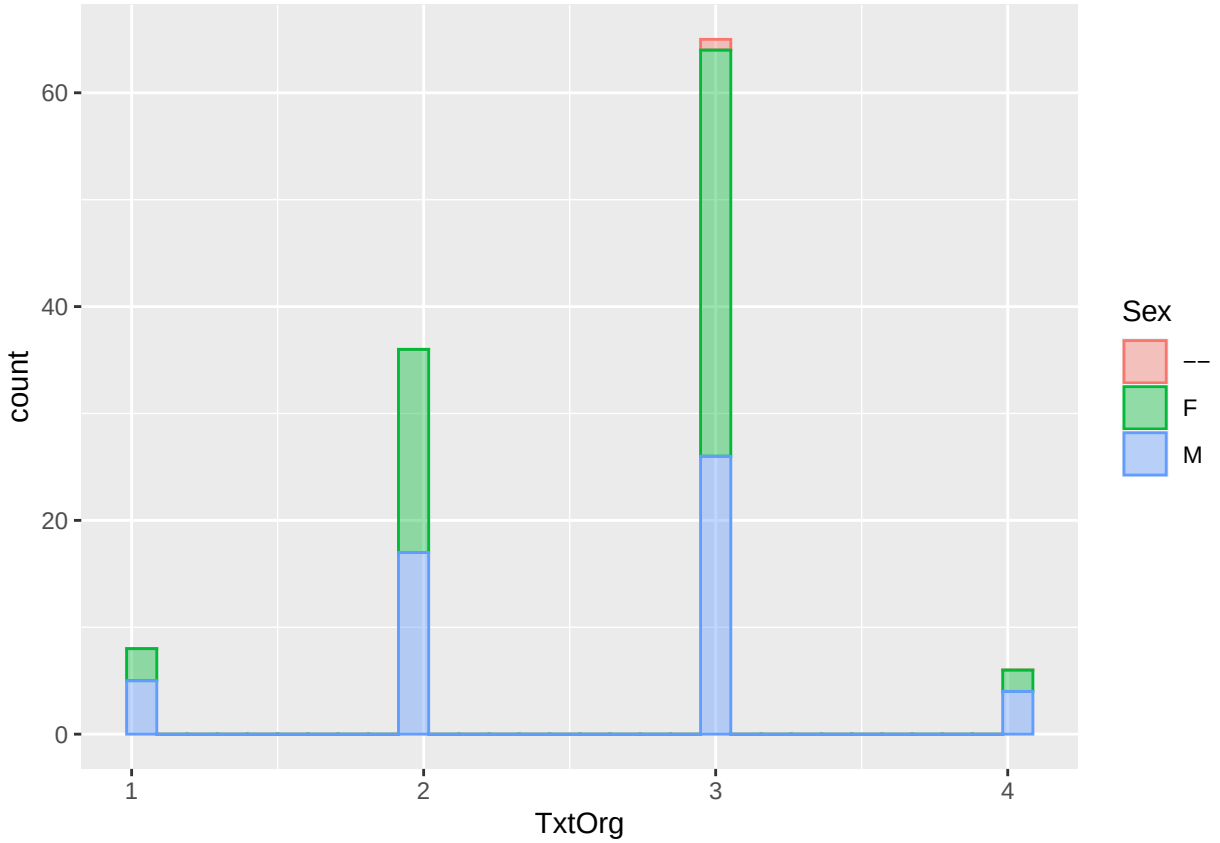

```
ggplot(ratings, aes(x = InterpRes)) + geom_histogram(aes(color = Sex, fill = Sex), bins = 30, alpha = 0
```




```
ggplot(ratings, aes(x = VisOrg)) + geom_histogram(aes(color = Sex, fill = Sex), bins = 30, alpha = 0.4)
```




```
ggplot(ratings, aes(x = TxtOrg)) + geom_histogram(aes(color = Sex, fill = Sex), bins = 30, alpha = 0.4)
```



Maybe we can also try to see if there is any tendency on the gender that if female or male tend to get distinguishable higher/lower ratings for these different rubrics and maybe later for different artifacts and other variables. From the plots above that roughly same amount of female and male can get 2/3 for all rubrics, but also from some rubrics, only male/female or mostly male/female get 4.0.