Analysis on Ratings for Freshman Statistics Projects

Zhuoheng Han
Department of Statistics and Data Science, Carnegie Mellon University
zhuohenh@andrew.cmu.edu

Abstract

In this paper, we explore the ratings for Freshman Statistics projects from Carnegie Mellon University Dietrich College. We used two data sets, ratings and tall, to help us analyze the ratings. Methods such as exploratory data analysis, intraclass correlation, 2-way table, and multilevel model are used. We find that distributions of ratings for each rubric or each rater are not indistinguishable from the distributions of other rubrics or other raters; discover that ratings given by raters do not agree for all rubrics; build a multi-level model consists of fixed effects, random effects and interaction terms between the variables Rater, Semester, and Rubric; notice that distribution of ratings for each semester or each sex are indistinguishable. In order to improve our analysis, we can investigate the missing values by checking the student records and add data set in 2020.

Introduction

Dietrich College at Carnegie Mellon University is in the process of implementing a new "General Education" program for undergraduates. In order to determine whether the new program is successful, the college hopes to rate student work. In this paper, we are discussing how ratings are related to various factors in this experiment such as rater, semester, sex and 7 rating rubrics from ratings data set and tall data set. We address four research questions:

- Is the distribution of ratings for each rubrics pretty much indistinguishable from the other rubrics? Is the distribution of ratings given by each rater pretty much indistinguishable from the other raters?
- For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees with the others? Or do they all disagree?
- 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?
- Is the distribution of ratings between two semesters indistinguishable? What about male and female?

Data

The data set comes from a new experiment performed by Dietrich College at Carnegie Mellon University. To evaluate 91 artifacts, 3 raters were asked to rate the artifacts using 7 separate rubrics, as shown in **Table 1**. The rating scale 1-4 for all rubrics is 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

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

The rating dataset contains total 15 columns and 117 rows. Each line of the dataset provides information for one rating piece. The definition of each variable is given below:
1. (X): Row number in the data set
2. Rater: Which of the three raters gave a rating
3. (Sample): Sample number
4. (Overlap): Unique identifier for artifact seen by all 3 raters Which semester the
5. Semester: Which semester the artifact came from
6. Sex: Sex of student who created the artifact Rating on Research
7. RsrchQ: Rating on Research Question
8. CritDes: Rating on Critique Design
9. InitEDA: Rating on Initial EDA
10. SelMeth: Rating on Select Method(s)
11. InterpRes: Rating on Interpret Results
12. VisOr: Rating on Visual Organization
13. TxtOrg : Rating on Text Organization
14. Artifact: Unique identifier for each artifact
15. Repeated: 1 = this is one of the 13 artifacts seen by all 3 raters

The tall data set contains the same data, but organized so that each row contains just one rating, in the column labelled Rating, and the rubric for that rating is listed in the column labelled Rubric. In this way, the dimension of the data set is 819×8.
Table 3 is the summary statistics for numeric variables. Specifically, it shows the median and mean of ratings of each rubric. The mean of ratings of critical design is relatively lower than the means of ratings of other rubrics. The median of ratings of interpret results and text organization are 3 while other rubric ratings' median are 2.

```
    RsrchQ        CritDes         InitEDA         SelMeth        InterpRes
Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:2.00   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
Median :2.00   Median :2.000   Median :2.000   Median :2.000   Median :3.000
Mean   :2.35   Mean   :1.871   Mean   :2.436   Mean   :2.068   Mean   :2.487
3rd Qu.:3.00   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000
Max.   :4.00   Max.   :4.000   Max.   :4.000   Max.   :3.000   Max.   :4.000
               NA's   :1
    VisOrg          TxtOrg
Min.   :1.000   Min.   :1.000
1st Qu.:2.000   1st Qu.:2.000
Median :2.000   Median :3.000
Mean   :2.414   Mean   :2.598
3rd Qu.:3.000   3rd Qu.:3.000
Max.   :4.000   Max.   :4.000
NA's   :1
```

Table 3. Summary statistics for ratings of each rubric

There are some NA's in the data sets, two missing rating values and one missing sex information. Since we are focusing on ratings and cannot find another group for sex, so we decide to drop those NA's.
After removing missing values, we make bar plots of counts of each rating for each rubric. We can find that for critique design, the most ratings are 1, which are totally different with other rubrics. For other rubrics except critique design, the top 2 ratings are 2 and 3 and the count of ratings 1 and 4 are relatively low.

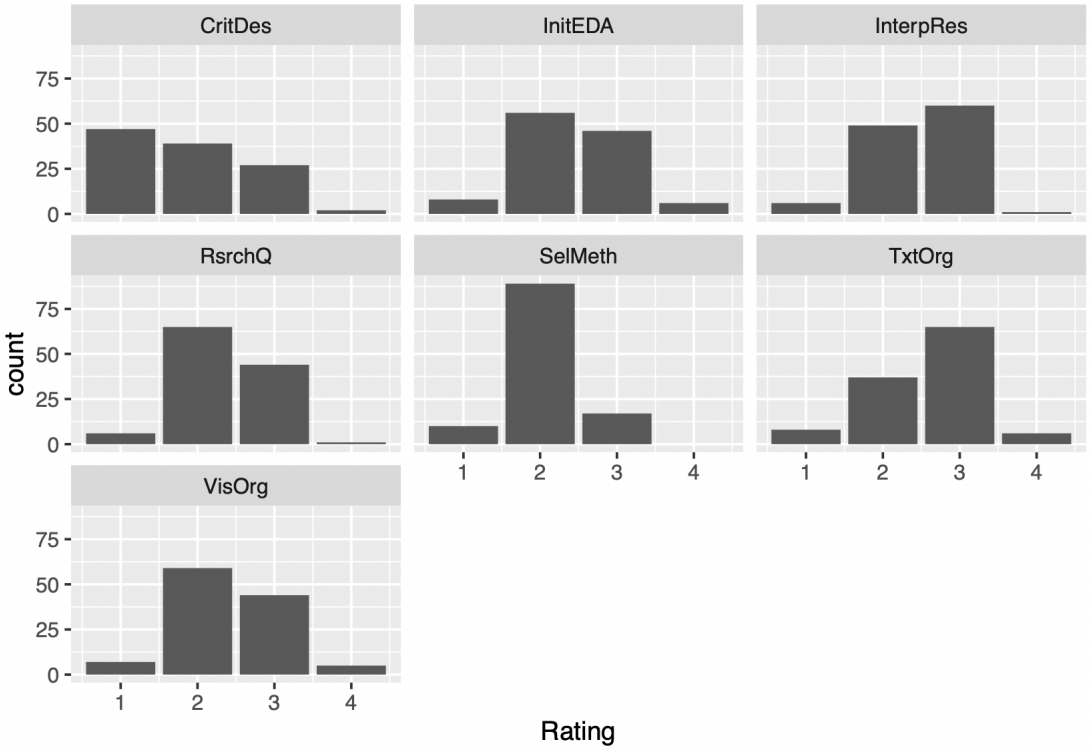


Figure 1. Counts of rating scales of each rubric

Methods

Distribution of ratings by rubrics and raters

We use exploratory data analysis, mainly plotting bar plots. We get a subset of the data for just the 13 artifacts seen by all raters, and make bar plots of counts of ratings for each rubric based on the subset of the data. We compare with the bar plots for the whole data set shown in the data section and determine whether these thirteen artifacts are representative of the whole set of 91 artifacts.

Agreement on ratings

We treat each artifact as a cluster of three ratings, and fit seven random-intercept models, one for each rubric, on the repeated data set and the full data set. Then, we calculate intraclass

correlations for ech model. To find which raters might be contributing to disagreement, we make 2-way tables of counts for the ratings of each pair of raters on each rubric. Then, we can calculate the percent exact agreement, which equals to diagonal sum of the table / total sum. That helps us to determine who is agreeing with whom on each rubric.

Factors related to the ratings

We add fixed effects for rater, semester, sex to the random intercept models for the full data set and do the backward elimination variable selection. In order to explore interactions with rubric, we begin with the model Rating $\sim$ (0 + Rubric|Artifact), and then add fixed effects (and possibly interactions) for all of the variables rater, semester, sex, repeated and/or rubric. We apply ANOVA to compare models with and without interactions after variable selection and finally get the best one.

More analysis on distribution of ratings

We again use exploratory data analysis on semester and sex in order to figure out whether distribution of ratings are distinguishable and why the final model contains semester but not sex.

Results

Distribution of ratings by rubrics and raters

First of all, we filter out 13 artifacts rated by all three raters. Here are the summary statistics (**Table 4**) and bar plots (**Figure 2**) of ratings of each rubric based on repeated artifacts. Compared with the summary statistics and bar plots in the Data section, we can find distributions are quite similar for each rubric between the full data set and the subset of data set.

RsrchQ	CritDes	InitEDA	SelMeth	InterpRes
Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000
1st Qu.:2.000	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 :2.000	Median :3.000
Mean :2.282	Mean :1.718	Mean :2.385	Mean :2.051	Mean :2.513
3rd Qu.:3.000	3rd Qu.:2.000	3rd Qu.:3.000	3rd Qu.:2.000	3rd Qu.:3.000
Max. :3.000	Max. :3.000	Max. :3.000	Max. :3.000	Max. :4.000

VisOrg	TxtOrg
Min. :1.000	Min. :1.000
1st Qu.:2.000	1st Qu.:2.000
Median :2.000	Median :3.000
Mean :2.282	Mean :2.667
3rd Qu.:3.000	3rd Qu.:3.000
Max. :3.000	Max. :4.000

Table 4. Summary statistics for ratings of each rubric

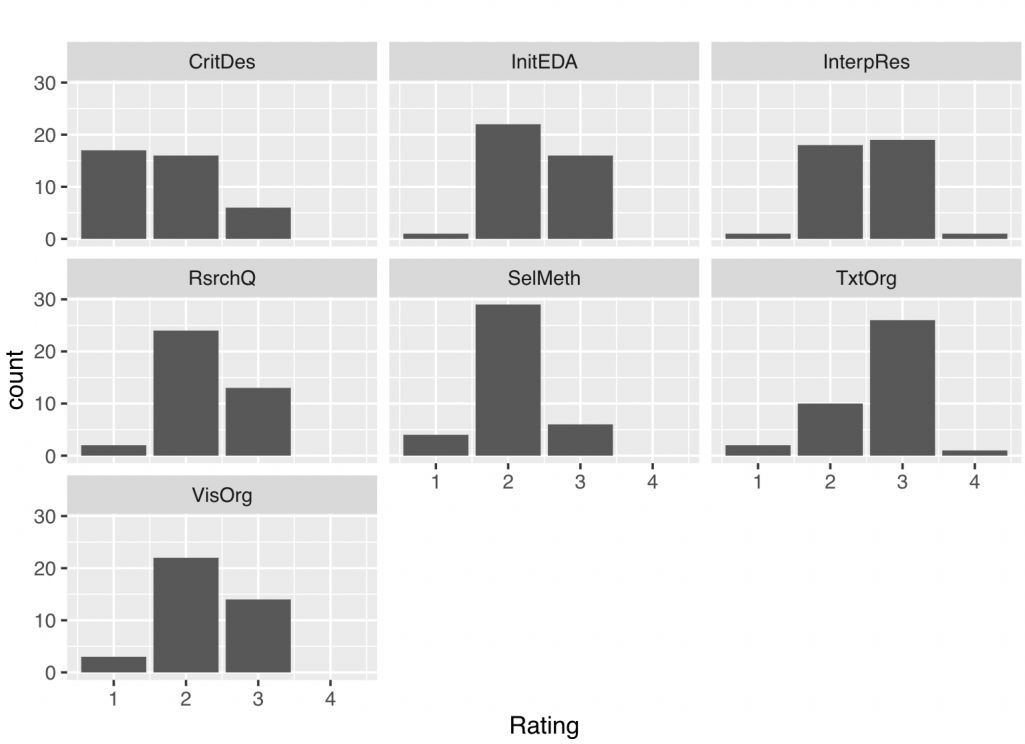


Figure 2. Counts of rating scales of each rubric in repeated data set

For critique design, the most ratings are 1, which are totally different with other rubrics. For rubrics initial EDA, research question, and visual organization, the ratings distribution are quite similar, where counts of rating 2 are the most and counts of rating 3 are quite close to counts rating 2. For interpretation results and text organization, the ratings distribution are quite similar, where counts of rating 3 are the most and then counts of rating 2. As for select methods, the ratings are mostly 2. Counts of rating 1,3, and 4 are extremely fewer than rating 2. Thus, distributions of ratings for each rubrics are not indistinguishable from the distributions of other rubrics.

In order to find the distribution of ratings given by different raters, we also make bar plots based on full data set and repeated data set. Compared with two sets of bar plots, we can find distributions of ratings are quite similar for each rater.

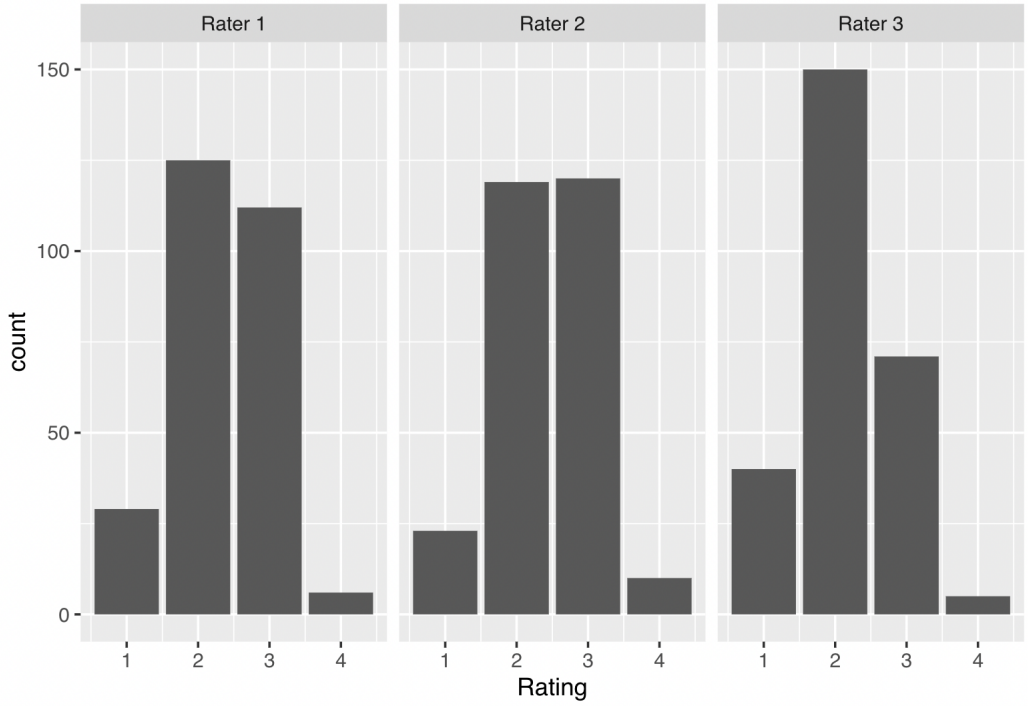


Figure 3. Counts of rating scales of each rater using full data set

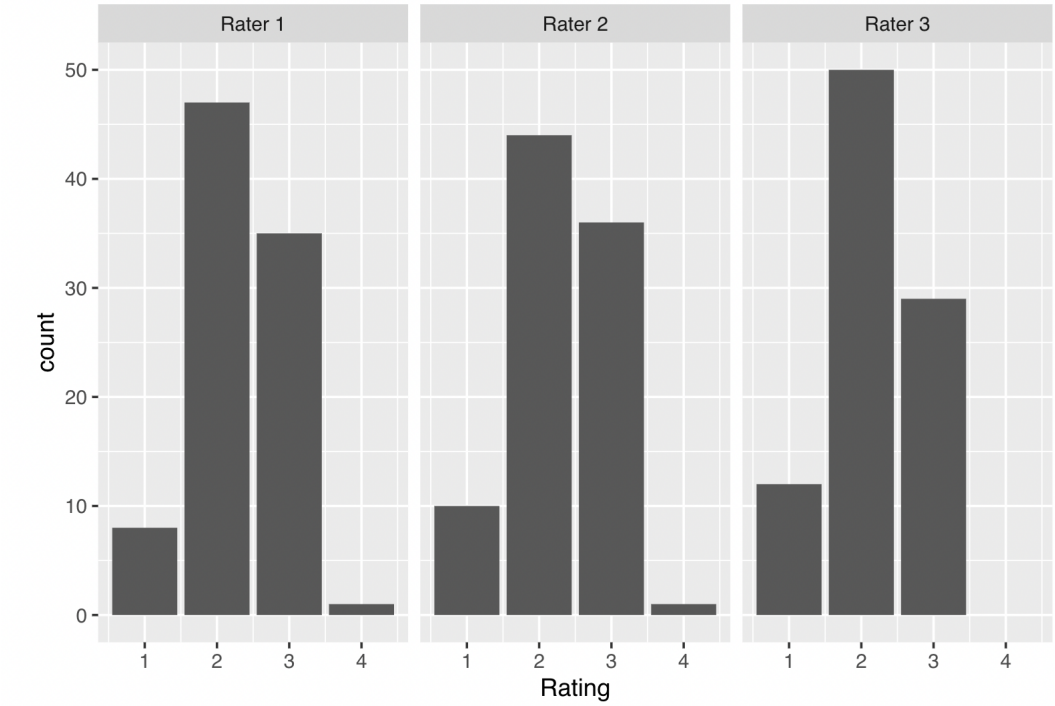


Figure 4. Counts of rating scales of each rater using repeated data set

For the rater 1 and 2, the most ratings they give are 2 and 3 and counts are close. For the rater 3, he/she gives rating 2 most, which is about twice as much as giving rating 3. In this

way, distributions of ratings for each rater are not indistinguishable from the distributions of other raters.

Agreement on ratings

Since we want to find out whether raters generally agree on their scores for each rubrics, we apply intra-cluster correlation (ICC) on the repeated data set and the full data set. **Table 5**, as shown below, includes ICC_repeated (intraclass correlation for all rubrics using 13 repeated artifacts data set), ICC_full (intraclass correlation for all rubrics using full data set), agreement12 (percentage of exact agreement on rater 1 and 2), agreement13 (percentage of exact agreement on rater 1 and 3), and agreement23 (percentage of exact agreement on rater 2 and 3).

```
##           ICC_repeated  ICC_full agreement12 agreement13 agreement23
## RsrchQ       0.1891892 0.2072956   0.3846154   0.7692308   0.5384615
## CritDes      0.5725594 0.6705262   0.5384615   0.6153846   0.6923077
## InitEDA      0.4929577 0.6880645   0.6923077   0.5384615   0.8461538
## SelMeth      0.5212766 0.4636330   0.9230769   0.6153846   0.6923077
## InterpRes    0.2295720 0.2212442   0.6153846   0.5384615   0.6153846
## VisOrg       0.5924529 0.6614900   0.5384615   0.7692308   0.7692308
## TxtOrg       0.1428571 0.1914696   0.6923077   0.6153846   0.5384615
```

Table 5. Agreement on each rubric

We can find that ICCs of rubric research question, interpret result, and text organization are low, which indicates raters do not agree with their rates for those rubrics. ICCs of rubric critical design, initial EDA, select results, and visual organization are relatively higher, which means raters are more in agreement with their rates for those rubrics, but there are still disagreements.
Then, we investigate the percentages of the exact agreement between each pair of raters to check whether it agrees the conclusion based on ICCs. The percentage of exact agreement between rater 1 and rater 2 on rubric research question is equal to 0.38, which is the only one percentage of agreement below 0.5. That means rater 1 and rater 2 have disagreement on this rubric. Most percentages are in the range from 0.5 to 0.8, which indicates some disagreement between two raters, but mostly they agree with each other.
Thus, the results from ICCs and percentages of the exact agreement between two raters show that 3 raters do not agree with their ratings for all rubrics.

Factors related to the ratings

To start with, we add fixed effects to the seven rubric-specific models using repeated data. Our seven models start with

$$\text{Rating} \sim -1 + \text{Rater} + \text{Semester} + \text{Sex} + (1|\text{Artifact}).$$

We apply backward elimination to each model and **Table 6** shows the final models for each rubric. From the result below, we don't need to add any fixed effects or interactions to the models for each rubric when using 13 artifacts data set.

```
$RsrchQ
Rating ~ (1 | Artifact)

$CritDes
Rating ~ (1 | Artifact)

$InitEDA
Rating ~ (1 | Artifact)

$SelMeth
Rating ~ (1 | Artifact)

$InterpRes
Rating ~ (1 | Artifact)

$VisOrg
Rating ~ (1 | Artifact)

$TxtOrg
Rating ~ (1 | Artifact)
```

Table 6. Models for each rubric based on repeated data set

Then, we add fixed effects to the seven rubric-specific models using full data. Applying same procedure as above, we get final models for each rubric, as shown in **Table 7**. From the result below, we don't need to add any fixed effects or interactions to the models for rubric research questions, initial EDA, and text organization. As for other rubrics, we need to examine each of these 4 models to see if the fixed effects make sense to us and if there are any interactions or additional random effects to consider.

```
$RsrchQ
Rating ~ (1 | Artifact)

$CritDes
Rating ~ Rater + (1 | Artifact) - 1

$InitEDA
Rating ~ (1 | Artifact)

$SelMeth
Rating ~ Rater + Semester + (1 | Artifact) - 1

$InterpRes
Rating ~ Rater + (1 | Artifact) - 1

$VisOrg
Rating ~ Rater + (1 | Artifact) - 1

$TxtOrg
Rating ~ (1 | Artifact)
```

Table 7. Models for each rubric based on full data set

Since rubric selected method has a model with two fixed effects, we check the fixed effect interactions for that model. According to the ANOVA table (**Table 8**) showed below, we can find that

$$\text{Rating} \sim -1 + \text{Rater} + \text{Semester} + (1|\text{Artifact}).$$

is the best model since it has the lowest p-value. In addition, AIC, BIC and logLik of model without interactions are all smaller than the model including the interaction. Thus, the fixed-effect interactions are not needed.

```
Data: tall[tall$Rubric == "SelMeth", ]
Models:
tmp.single_intercept: Rating ~ Semester + (1 | Artifact)
tmp: Rating ~ Rater + Semester + (1 | Artifact) - 1
tmp.fixed_interactions: Rating ~ Rater + (1 | Artifact) + Rater:Semester - 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 *
tmp.fixed_interactions    8 143.46 165.49 -63.731   127.46 2.5920  2    0.27362
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 8. ANOVA table

After investigating interaction, we check for random effects for those 4 models. For each model, we should add random effects that are also present as fixed effects. For rubric critical design model, we should try adding (Rater|Artifact); For rubric select method model, we should try adding (Rater|Artifact) and (Semester|Artifact); For rubric interpret results model, we should try adding (Rater|Artifact); For rubric visual organization model, we should try adding (Rater|Artifact). From Appendix D.iii), we can find that we do not need to add any random effects since models added with random effects even do not exist, which means no testings are needed.
Finally, we are going to add fixed effects, interactions, and new random effects to "combined" model Rating $\sim 1 + (0 + \text{Rubric}|\text{Artifact})$ using full data set.
To start with, we are going to select fixed effects. We add all fixed effects without interaction to the intercept-only "combined" model and do a backward elimination. According to Appendix D.4.1, we can get the model

$$\text{Rating} \sim (0 + \text{Rubric}|\text{Artifact}) + \text{Rater} + \text{Semester} + \text{Rubric}$$

as our final model. Then we try adding interaction to that model to see whether there are any interactions we need to consider. We add interactions to the model we get above and also do a backward elimination. According to Appendix D.4.2 We then get the model

$$\text{Rating} \sim (0 + \text{Rubric}|\text{Artifact}) + \text{Rater} + \text{Semester} + \text{Rubric} + \text{Rater} : \text{Rubric}$$

We again apply ANOVA (**Table 9**) to get the best model among those models. We find that the model with interaction between fixed effect rater and rubric is the best model.

```
Data: tall
Models:
comb.back_elim: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric
comb.inter_elim: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric + Rater:Rubric
comb.inter.u: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric + Rater:Semester +
Rater:Rubric + Semester:Rubric + Rater:Semester:Rubric
                npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
comb.back_elim    39 1464.0 1647.2 -693.02   1386.0
comb.inter_elim   51 1454.5 1694.1 -676.26   1352.5 33.526 12   0.000801 ***
comb.inter.u      71 1471.4 1804.8 -664.68   1329.4 23.161 20   0.280962
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 9. ANOVA table

At last, we consider adding random effects to the best model we get so far. The fixed-effects terms we have to work with are: Rater, Semester and Rater:Rubric. We want to add each of these without a random intercept, to preserve the structure of the model (separate random intercepts for each rubric). Based on Appendix D.4.4), we can find that we need to include $(0 + \text{Rater}|\text{Artifact})$ in the model. In this way, our final model should be

Rating $\sim$ (0+Rubric|Artifact)+(0+Rater|Artifact)+Rater+Semester+Rubric+Rater : Rubric

To help interpret each component in the final model, an overview of the variable's meaning are provided:
Semester - This is a fixed effect, which means rating is affected by semester.
$(0 + \text{Rater}|\text{Artifact}) + \text{Rater}$ - There is a kind of rater and artifact interaction. Each Rater's rating on each Artifact differs from what we would expect from the fixed effects alone by a small random effect that depends on the Artifact.
$(0 + \text{Rubric}|\text{Artifact}) + \text{Rubric}$ – There is an interaction between the rubrics and the artifacts is what we might expect. There are different average scores on each rubric, but the rubric averages also vary a bit from one Artifact to the next by a small random effect that depends on Artifact.
$\text{Rater} + \text{Rubric} + \text{Rater} : \text{Rubric}$ – There is a Rater and Rubric interaction: each Rater uses each rubric in a way that is not like, or even parallel to, other rater's rubric usage. The interaction suggests that the raters are not all interpreting the rubrics in the same way.

More analysis on distribution of ratings

From **Figure 5**, we can find the distribution of ratings is quite similar between two semesters. Rating 2 is the most frequent rating in both semester and then rating 3 is the second most. However, There are huge differences in counts between two semesters. For rating 2, the count in Fall 2019 is around three times as large as the count in Spring 2019; For rating 3, the count in Fall 2019 is about twice as large as the count in Spring 2019. This can be a reason that our final model includes semester.

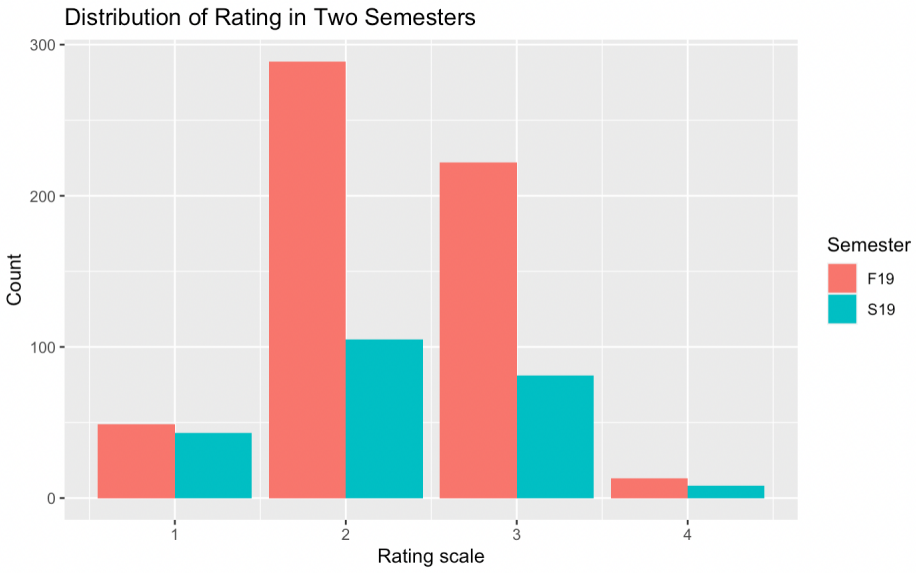


Figure 5. Counts of rating scales in two semesters

From **Figure 6**, we can find the distribution of ratings is quite similar in male and female. Rating 2 is the most frequent rating and then rating 3 is the second most. Since the dis-

tribution of ratings are indistinguishable between male and female, that can be the reason that our final model does not include sex as fixed effect.

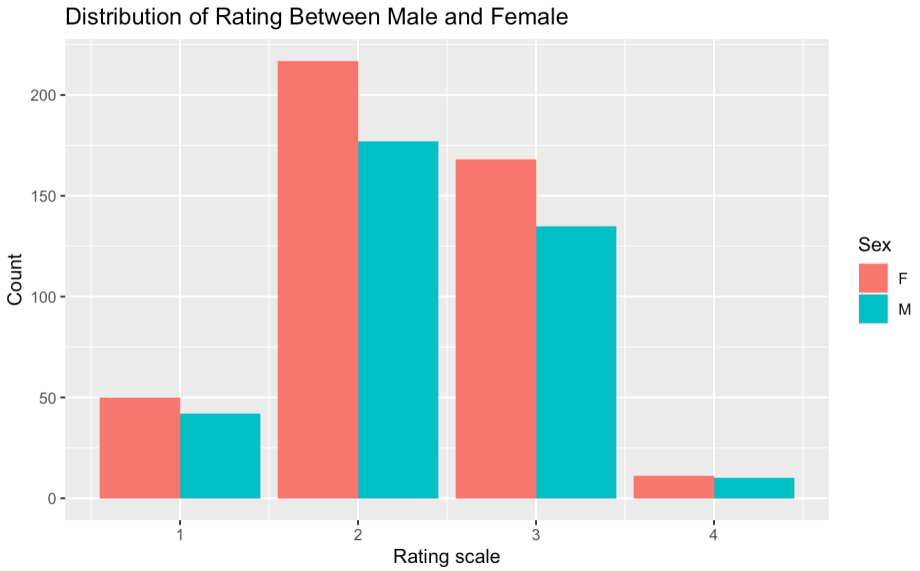


Figure 6. Counts of rating scales in two semesters

Discussion

There are 2 NA's in ratings and 1 missing sex in our data sets. In this paper, we dropped those 2 ratings, which might cause the imbalance of data set since two students will only have 6 rubrics rated. In addition, we also dropped all data related to the freshman with the missing sex. The best way to address the missing values is to investigate the records on those students. In this way, we can fill out those NA's and get the full data set. What's more, we find that the data contains more freshmen in Fall 2019 than those in Spring 2019. Even though we do not know the exact number of freshmen in each semester, it might be more helpful to balance the data in each semester and add new data set in 2020. We can reevaluate this experiment in order to make more comprehensive analysis.
Findings from Result section, the rater and rubric interaction in our final model and exploratory data analysis both suggest that the raters are not all interpreting the rubrics in the same way. Thus, we recommend that perhaps the raters should be trained more in order to make the raters' ratings more similar to each other. In that way, this general education program can be more successful than it is now.

Reference

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

Appendix

```
library(lme4)
library(arm)
library(ggplot2)
library(tidyverse)
library(LMERConvenienceFunctions)
library(RLRsim)
```

A. Data Cleaning

Import the dataset and delete variables that are not expected to be useful for analysis. Then, we drop NA's.

```
ratings = read.csv("~/Desktop/ratings.csv")
tall = read.csv("~/Desktop/tall.csv")
```

Dimension of two data sets.

```
dim(ratings)
```

```
## [1] 117  15
```

```
dim(tall)
```

```
## [1] 819   8
```

```
ratings = ratings[-c(1,3,4)]
```

See the summary statistics of the numeric variable.

```
summary(ratings[- which(names(ratings) %in% c("Rater", "Semester", "Sex", "Artifact", "Repeated"))])
```

```
##      RsrchQ         CritDes         InitEDA         SelMeth         InterpRes
##  Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.00   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.00   Median :2.000   Median :2.000   Median :2.000   Median :3.000
##  Mean   :2.35   Mean   :1.871   Mean   :2.436   Mean   :2.068   Mean   :2.487
##  3rd Qu.:3.00   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :4.00   Max.   :4.000   Max.   :4.000   Max.   :3.000   Max.   :4.000
##                 NA's   :1
##      VisOrg          TxtOrg
##  Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :3.000
##  Mean   :2.414   Mean   :2.598
##  3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :4.000   Max.   :4.000
##  NA's   :1
```

See the summary statistics of the categrical variable.

```
apply(ratings[which(names(ratings) %in% c("Rater", "Semester", "Sex", "Repeated"))], 2, table)
```

```
## $Rater
##
##  1  2  3
## 39 39 39
##
## $Semester
##
##   Fall Spring
##     83     34
##
## $Sex
##
## --  F  M
##  1 64 52
##
## $Repeated
##
##  0  1
## 78 39
```

Check NA's

```
tall[is.na(tall["Rating"]),]
```

```
##       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
```

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

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

Drop NA's in both data sets

```
ratings <- ratings %>% drop_na()
ratings <- ratings[-c(5),]
```

```
tall <- tall %>% drop_na()
tall <- tall[-c(5,122,238,355,472,589,705),]
```

Change the type of variable

```
tall$Rating = as.numeric(tall$Rating)
tall$Rater = as.factor(tall$Rater)
```

Get the subset of full data set, which contains the information of 13 common artifacts

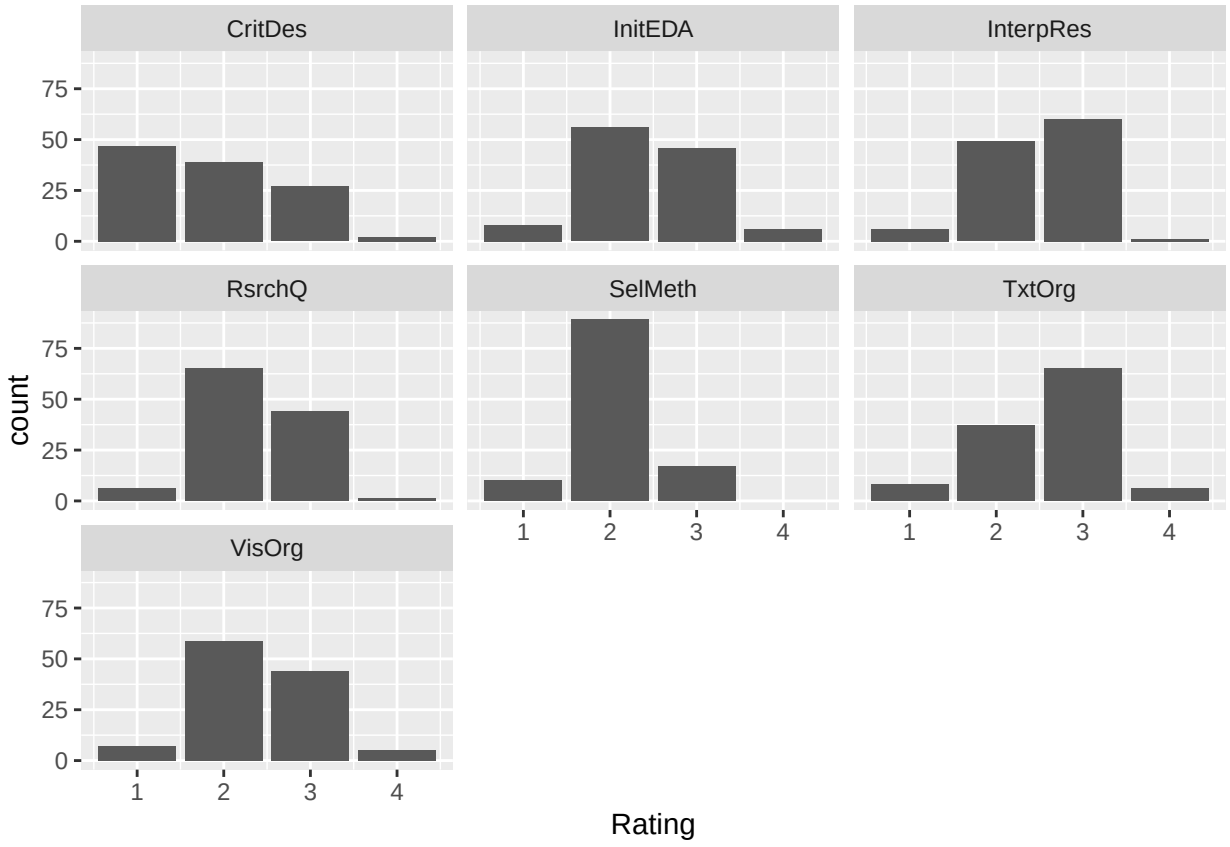
```
ratings.repeat = ratings[ratings$Repeated == 1,]
```

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

B. Research Question 1

Counts of ratings of each rubric from the full data set

```
ggplot(tall,aes(x = Rating)) +
  facet_wrap( ~ Rubric) +
  geom_bar()
```



Summary statistics of ratings of each rubric from the full data set

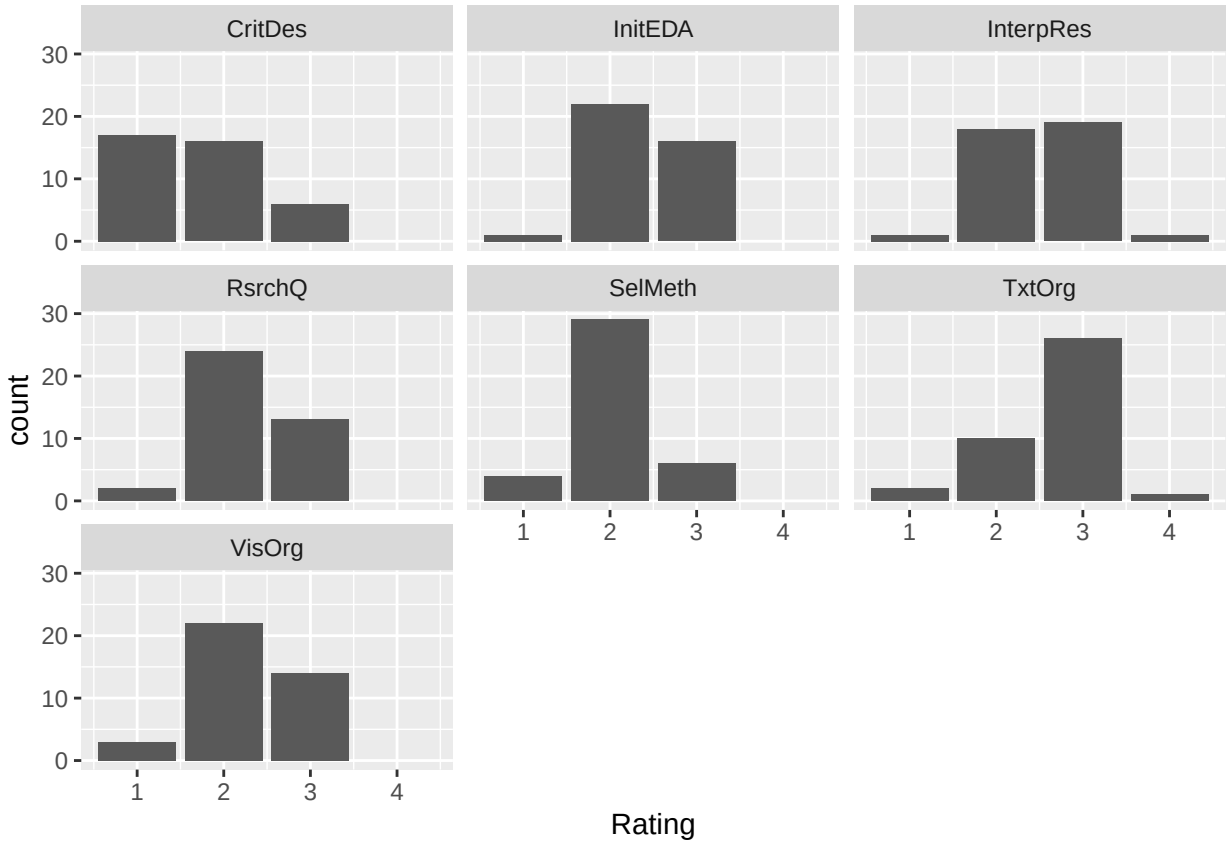
```
summary(ratings.repeat[- which(names(ratings.repeat) %in% c("Rater", "Semester", "Sex", "Artifact", "Rep
```

```
##      RsrchQ          CritDes         InitEDA         SelMeth
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000
```

```
##  Median :2.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :2.282   Mean   :1.718   Mean   :2.385   Mean   :2.051
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :3.000   Max.   :3.000   Max.   :3.000   Max.   :3.000
##    InterpRes         VisOrg          TxtOrg
##  Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000   Median :2.000   Median :3.000
##  Mean   :2.513   Mean   :2.282   Mean   :2.667
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :4.000   Max.   :3.000   Max.   :4.000
```

Counts of ratings of each rubric from the subset of data set

```
ggplot(tall.repeat,aes(x = Rating)) +
  facet_wrap( ~ Rubric) +
  geom_bar()
```



Summary statistics of ratings of each rubric by different raters from the subset of full data set

```
summary(ratings.repeat[ratings.repeat$Rater == 1,][- which(names(ratings.repeat) %in% c("Semester", "Se
```

```
##      Rater       RsrchQ         CritDes         InitEDA         SelMeth
##  Min.   :1   Min.   :2.000   Min.   :1.000   Min.   :1.000   Min.   :2.000
##  1st Qu.:1   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000
```

```
##  Median :1   Median :2.000   Median :2.000   Median :3.000   Median :2.000
##  Mean   :1   Mean   :2.385   Mean   :1.615   Mean   :2.538   Mean   :2.154
##  3rd Qu.:1   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :1   Max.   :3.000   Max.   :3.000   Max.   :3.000   Max.   :3.000
##    InterpRes         VisOrg          TxtOrg
##  Min.   :2.000   Min.   :1.000   Min.   :2.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000   Median :2.000   Median :3.000
##  Mean   :2.615   Mean   :2.154   Mean   :2.769
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :3.000   Max.   :3.000   Max.   :4.000
```

```
summary(ratings.repeat[ratings.repeat$Rater == 2,][- which(names(ratings.repeat) %in% c("Semester", "Se
```

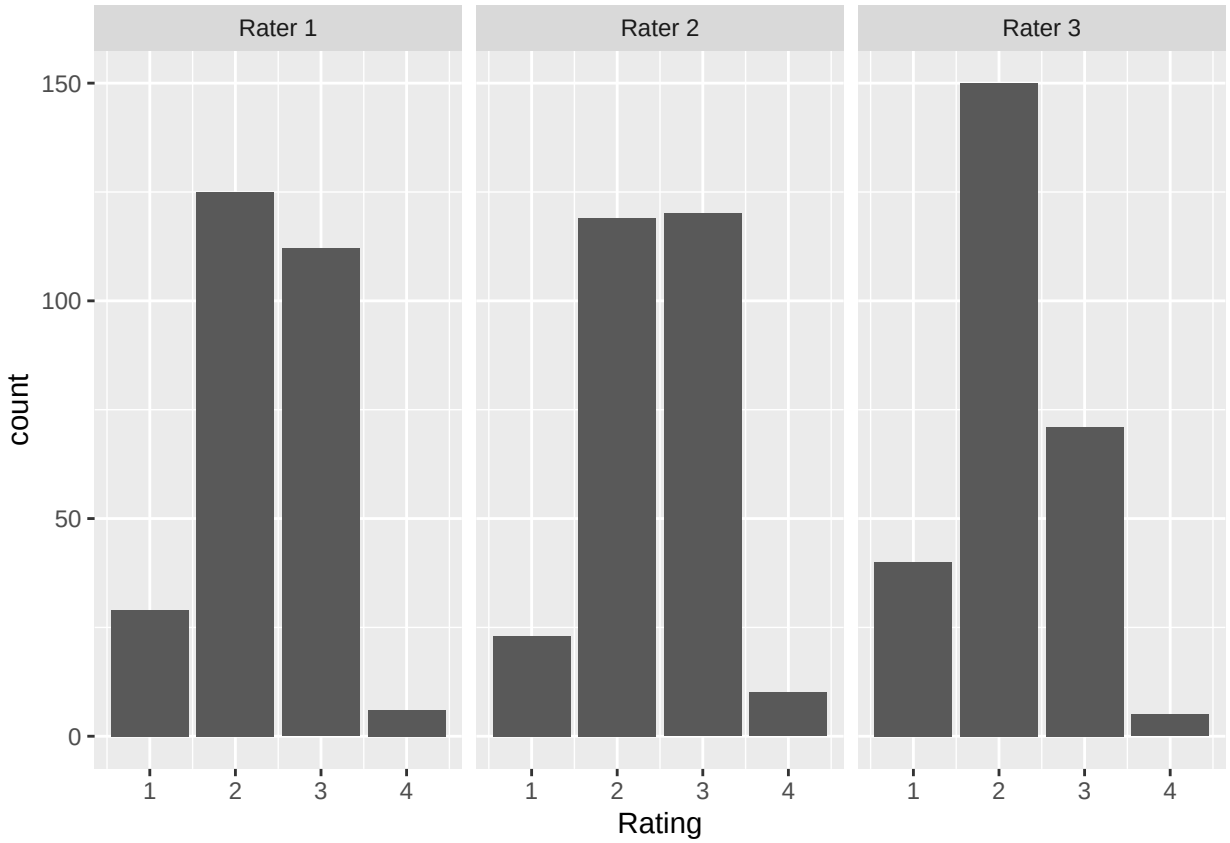
```
##      Rater       RsrchQ         CritDes         InitEDA         SelMeth
##  Min.   :2   Min.   :1.000   Min.   :1.000   Min.   :2.000   Min.   :1.000
##  1st Qu.:2   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2   Median :2.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :2   Mean   :2.154   Mean   :1.846   Mean   :2.385   Mean   :2.077
##  3rd Qu.:2   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :2   Max.   :3.000   Max.   :3.000   Max.   :3.000   Max.   :3.000
##    InterpRes         VisOrg          TxtOrg
##  Min.   :2.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000   Median :3.000   Median :3.000
##  Mean   :2.615   Mean   :2.462   Mean   :2.615
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :4.000   Max.   :3.000   Max.   :3.000
```

```
summary(ratings.repeat[ratings.repeat$Rater == 3,][- which(names(ratings.repeat) %in% c("Semester", "Se
```

```
##      Rater       RsrchQ         CritDes         InitEDA         SelMeth
##  Min.   :3   Min.   :2.000   Min.   :1.000   Min.   :2.000   Min.   :1.000
##  1st Qu.:3   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :3   Median :2.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :3   Mean   :2.308   Mean   :1.692   Mean   :2.231   Mean   :1.923
##  3rd Qu.:3   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :3   Max.   :3.000   Max.   :3.000   Max.   :3.000   Max.   :3.000
##    InterpRes         VisOrg          TxtOrg
##  Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :3.000
##  Mean   :2.308   Mean   :2.231   Mean   :2.615
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :3.000   Max.   :3.000   Max.   :3.000
```

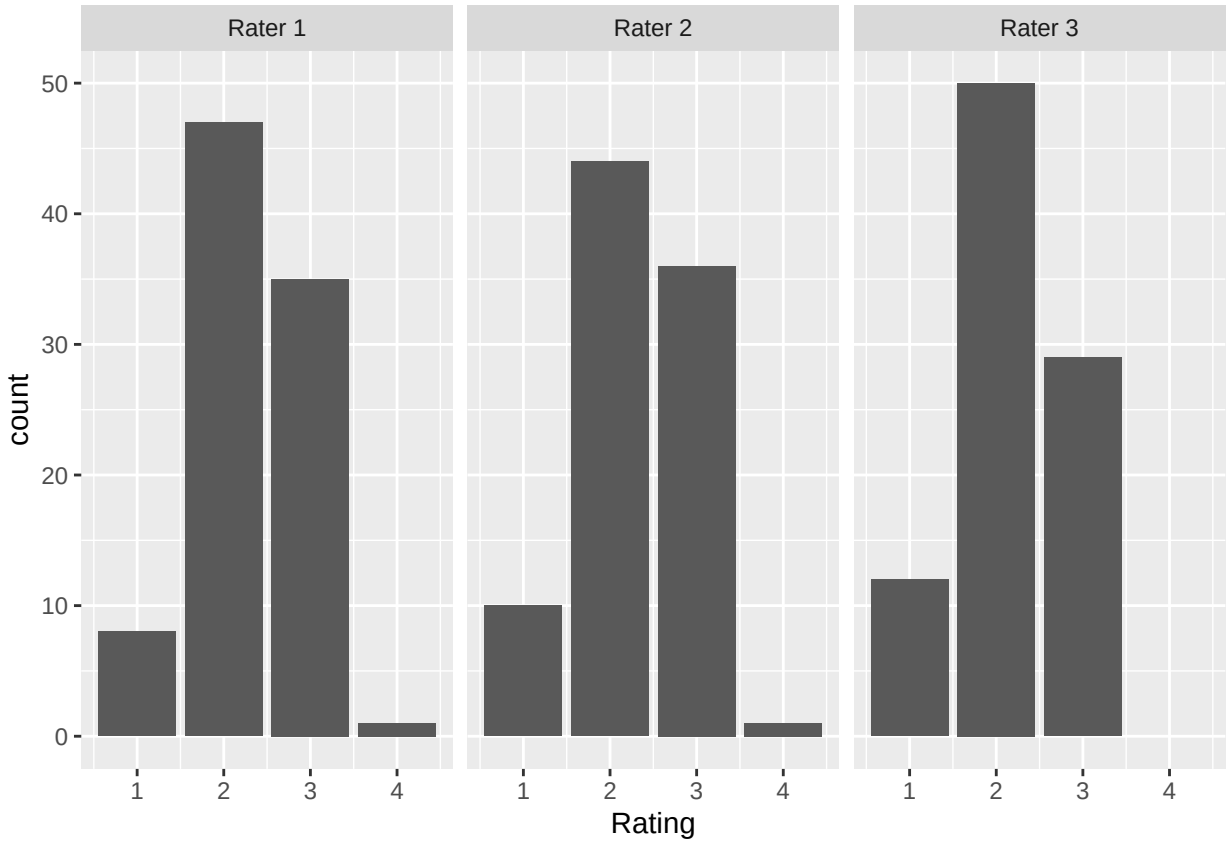
Counts of ratings of each rater from the full data set

```
rater.name <- function(x) { paste("Rater",x) }
ggplot(tall,aes(x = Rating)) +
  facet_wrap( ~ Rater, labeller=labeller(Rater=rater.name)) +
  geom_bar()
```



Counts of ratings of each rater from the subset of full data set

```
ggplot(tall.repeat,aes(x = Rating)) +
  facet_wrap( ~ Rater, labeller=labeller(Rater=rater.name)) +
  geom_bar()
```



C. Research Question 2

Calculate ICC for each rubric using subset data set and full data set.

```
Rubric.names <- c("RsrchQ","CritDes","InitEDA","SelMeth","InterpRes","VisOrg","TxtOrg")
ICC.vec <- NULL
for (i in Rubric.names) {
  tmp <- lmer(Rating ~ 1+(1|Artifact), data=tall.repeat[tall.repeat$Rubric==i,])
  sig2 <- summary(tmp)$sigma^2
  tau2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
  ICC <- tau2/(tau2+sig2)
  ICC.vec <- append(ICC.vec,ICC)
}
names(ICC.vec) <- Rubric.names
```

```
ICC.vec.full <- NULL
for (i in Rubric.names) {
  tmp <- lmer(Rating ~ 1+(1|Artifact), data=tall[tall$Rubric==i,])
  sig2 <- summary(tmp)$sigma^2
  tau2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
  ICC <- tau2/(tau2+sig2)
  ICC.vec.full <- append(ICC.vec.full,ICC)
}
names(ICC.vec.full) <- Rubric.names
```

Calculate percentage of exact agreement for each rubric as follow.

```
agreement.results <- cbind(ICC_repeated=ICC.vec, ICC_full=ICC.vec.full, agreement12=0, agreement13=0, ag
agreement.tables <- as.list(rep(NA,7))
for (i in Rubric.names) {
  tmp.data = data.frame(r1=ratings.repeat[ratings.repeat$Rater==1, i],
                        r2=ratings.repeat[ratings.repeat$Rater==2, i],
                        r3=ratings.repeat[ratings.repeat$Rater==3, i],
                        a1=ratings.repeat[ratings.repeat$Rater==1, "Artifact"],
                        a2=ratings.repeat[ratings.repeat$Rater==2, "Artifact"],
                        a3=ratings.repeat[ratings.repeat$Rater==3, "Artifact"])

  t1 <- factor(tmp.data$r1,levels=1:4)
  t2 <- factor(tmp.data$r2,levels=1:4)
  t3 <- factor(tmp.data$r3,levels=1:4)
  t12 <- table(t1,t2)
  agreement12 <- (t12[1,1]+t12[2,2]+t12[3,3]+t12[4,4])/sum(t12)
  t13 <- table(t1,t3)
  agreement13 <- (t13[1,1]+t13[2,2]+t13[3,3]+t13[4,4])/sum(t13)
  t23 <- table(t2,t3)
  agreement23 <- (t23[1,1]+t23[2,2]+t23[3,3]+t23[4,4])/sum(t23)
  agreement.results[i,3:5] <- c(agreement12, agreement13, agreement23)
}
agreement.results
```

```
##           ICC_repeated  ICC_full agreement12 agreement13 agreement23
## RsrchQ       0.1891892 0.2072956   0.3846154   0.7692308   0.5384615
## CritDes      0.5725594 0.6705262   0.5384615   0.6153846   0.6923077
## InitEDA      0.4929577 0.6880645   0.6923077   0.5384615   0.8461538
## SelMeth      0.5212766 0.4636330   0.9230769   0.6153846   0.6923077
## InterpRes    0.2295720 0.2212442   0.6153846   0.5384615   0.6153846
## VisOrg       0.5924529 0.6614900   0.5384615   0.7692308   0.7692308
## TxtOrg       0.1428571 0.1914696   0.6923077   0.6153846   0.5384615
```

D. Research Question 3

1. Adding fixed effects to the seven rubric-specific models using repeated data

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

for (i in Rubric.names) {
  rubric.data <- tall.repeat[tall.repeat$Rubric==i,]
  tmp <- lmer(Rating ~ -1+Rater+Semester+Sex+(1|Artifact),
              data=tall.repeat[tall.repeat$Rubric==i,], REML=FALSE)
  tmp.back_elim <- fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE)
  tmp.single_intercept <-  lmer(Rating ~ (1|Artifact),
                                data=tall.repeat[tall.repeat$Rubric==i,],REML=FALSE)
  pval <- anova(tmp.single_intercept,tmp.back_elim)$"Pr(>Chisq)"[2]
  if (pval<=0.05) {
    tmp_final <- tmp.back_elim
  }
  else {
    tmp_final <- tmp.single_intercept
```

```
  }

model.formula.repeat[[i]] <- formula(tmp_final)
}
```

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

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

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

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

```
## ============================================================
## ===              backfitting fixed effects             ===
## ============================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.2229 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.1826 >= 0.05
##     not part of higher-order interaction
##     removing term
```

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

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

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

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

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

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

## ========================================================
```

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

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

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

## ====================================================
## ===              backfitting fixed effects         ===
## ====================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.8294 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.2947 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ====================================================
## ===            forwardfitting random effects       ===
## ====================================================
##  ===       random slopes       ===
## ====================================================
## ===            re-backfitting fixed effects        ===
## ====================================================
## processing model terms of interaction level 1
```

```
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune

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

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

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

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

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

## ============================================================
## ===              backfitting fixed effects             ===
## ============================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.5358 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.1319 >= 0.05
```

```
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune

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

```
model.formula.repeat
```

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

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

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

for (i in Rubric.names) {
  tmp <- lmer(Rating ~ -1+Rater+Semester+Sex+(1|Artifact),
              data=tall[tall$Rubric==i,], REML=FALSE)
  tmp.back_elim <- fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE)
  tmp.single_intercept <-  lmer(Rating ~ (1|Artifact),
```

```
                                data=tall[tall$Rubric==i,],REML=FALSE)
  pval <- anova(tmp.single_intercept,tmp.back_elim)$"Pr(>Chisq)"[2]
  if (pval<=0.05) {
    tmp_final <- tmp.back_elim
  }
  else {
    tmp_final <- tmp.single_intercept
  }

model.formula[[i]] <- formula(tmp_final)
}
```

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

```
## ============================================================
## ===              backfitting fixed effects             ===
## ============================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.6166 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.3987 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ============================================================
## ===            forwardfitting random effects           ===
## ============================================================
##  ===       random slopes       ===
## ============================================================
## ===            re-backfitting fixed effects            ===
## ============================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
```

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

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

```
## ============================================================
## ===              backfitting fixed effects             ===
## ============================================================
## processing model terms of interaction level 1
##   iteration 1
```

```
##     p-value for term "Semester" = 0.7154 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.5297 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune

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

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

## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.8802 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.7402 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
```

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

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

## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.1935 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune

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

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

## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.608 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.5312 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
```

```
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune

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

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

## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.2158 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.3523 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune

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

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

## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.5041 >= 0.05
##     not part of higher-order interaction
##     removing term
```

```
##   iteration 2
##     p-value for term "Semester" = 0.205 >= 0.05
##     not part of higher-order interaction
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
##  ===       random slopes       ===
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune

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

```
model.formula
```

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

3. Trying interactions and new random effects for the 4 rubric specific models using all data set.

CritDes

```
tmp <- lmer(model.formula[["CritDes"]],data=tall[tall$Rubric=="CritDes",])
tmp.single_intercept <- update(tmp, . ~ . + 1 - Rater)
anova(tmp,tmp.single_intercept)
```

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

```
## Data: tall[tall$Rubric == "CritDes", ]
## Models:
## tmp.single_intercept: Rating ~ (1 | Artifact)
## tmp: Rating ~ Rater + (1 | Artifact) - 1
##                       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    3 277.68 285.91 -135.84   271.68
## tmp                     5 273.62 287.35 -131.81   263.62 8.0535  2    0.01783 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m0 <- tmp
mA <- update(m0, . ~ . + (Rater|Artifact))
```

```
## Error: number of observations (=115) <= number of random effects (=267) for term (Rater | Artifact);
```

```
m <- update(mA, . ~ . - (1|Artifact))
```

```
## Error in h(simpleError(msg, call)): error in evaluating the argument 'object' in selecting a method
```

```
exactRLRT(m0=m0,mA=mA,m=m)
```

```
## Error in exactRLRT(m0 = m0, mA = mA, m = m): object 'm' not found
```

```
summary(tmp)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ Rater + (1 | Artifact) - 1
##    Data: tall[tall$Rubric == "CritDes", ]
##
## REML criterion at convergence: 271
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.55495 -0.50027 -0.08228  0.64663  1.60935
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.4349   0.6595
##  Residual             0.2473   0.4972
## Number of obs: 115, groups:  Artifact, 89
##
## Fixed effects:
##        Estimate Std. Error t value
## Rater1   1.6863     0.1207   13.98
## Rater2   2.1129     0.1219   17.34
## Rater3   1.8908     0.1219   15.51
##
## Correlation of Fixed Effects:
##        Rater1 Rater2
## Rater2 0.244
## Rater3 0.244  0.246
```

SelMeth

```
tmp <- lmer(model.formula[["SelMeth"]],data=tall[tall$Rubric=="SelMeth",])
tmp.single_intercept <- update(tmp, . ~ . + 1 - Rater)
tmp.fixed_interactions <- update(tmp, . ~ . + Rater*Semester - Semester)
anova(tmp,tmp.single_intercept,tmp.fixed_interactions)
```

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

## Data: tall[tall$Rubric == "SelMeth", ]
## Models:
## tmp.single_intercept: Rating ~ Semester + (1 | Artifact)
## tmp: Rating ~ Rater + Semester + (1 | Artifact) - 1
## tmp.fixed_interactions: Rating ~ Rater + (1 | Artifact) + Rater:Semester - 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
## tmp.fixed_interactions     8 143.46 165.49 -63.731   127.46 2.5920  2    0.27362
##
## tmp.single_intercept
## tmp                     *
## tmp.fixed_interactions
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing (Semester|Artifact)

```
m0 <- tmp
mA <- update(m0, . ~ . + (Semester|Artifact))
```

```
## Error: number of observations (=116) <= number of random effects (=180) for term (Semester | Artifac
```

```
m <- update(mA, . ~ . - (1|Artifact))
```

```
## Error in h(simpleError(msg, call)): error in evaluating the argument 'object' in selecting a method
```

```
exactRLRT(m0=m0,mA=mA,m=m)
```

```
## Error in exactRLRT(m0 = m0, mA = mA, m = m): object 'm' not found
```

Testing (Rater|Artifact)

```
m0 <- tmp
mA <- update(m0, . ~ . + (Rater|Artifact))
```

```
## Error: number of observations (=116) <= number of random effects (=270) for term (Rater | Artifact);
```

```
m <- update(mA, . ~ . - (1|Artifact))
```

```
## Error in h(simpleError(msg, call)): error in evaluating the argument 'object' in selecting a method
```

```
exactRLRT(m0=m0,mA=mA,m=m)
```

```
## Error in exactRLRT(m0 = m0, mA = mA, m = m): object 'm' not found
```

```
summary(tmp)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ Rater + Semester + (1 | Artifact) - 1
##    Data: tall[tall$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
## Rater1       2.25037    0.07503  29.992
## Rater2       2.22653    0.07424  29.991
## Rater3       2.03316    0.07521  27.033
## SemesterS19 -0.35860    0.09796  -3.661
##
## Correlation of Fixed Effects:
##             Rater1 Rater2 Rater3
## Rater2       0.285
## Rater3       0.287  0.280
## SemesterS19 -0.413 -0.391 -0.394
```

InterpRes

```
tmp <- lmer(model.formula[["InterpRes"]],data=tall[tall$Rubric=="InterpRes",])
tmp.single_intercept <- update(tmp, . ~ . + 1 - Rater)
anova(tmp,tmp.single_intercept)
```

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

```
## Data: tall[tall$Rubric == "InterpRes", ]
## Models:
## tmp.single_intercept: Rating ~ (1 | Artifact)
## tmp: Rating ~ Rater + (1 | Artifact) - 1
##                      npar    AIC    BIC   logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    3 218.53 226.79 -106.263   212.53
## tmp                     5 200.66 214.43  -95.331   190.66 21.864  2  1.787e-05
##
```

```
## tmp.single_intercept
## tmp                  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m0 <- tmp
mA <- update(m0, . ~ . + (Rater|Artifact))
```

```
## Error: number of observations (=116) <= number of random effects (=270) for term (Rater | Artifact);
```

```
m <- update(mA, . ~ . - (1|Artifact))
```

```
## Error in h(simpleError(msg, call)): error in evaluating the argument 'object' in selecting a method
```

```
exactRLRT(m0=m0,mA=mA,m=m)
```

```
## Error in exactRLRT(m0 = m0, mA = mA, m = m): object 'm' not found
```

```
summary(tmp)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ Rater + (1 | Artifact) - 1
##    Data: tall[tall$Rubric == "InterpRes", ]
##
## REML criterion at convergence: 199.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5317 -0.7627  0.2635  0.6614  2.6535
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.06224  0.2495
##  Residual             0.25250  0.5025
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##        Estimate Std. Error t value
## Rater1  2.70421    0.08912   30.34
## Rater2  2.58574    0.08912   29.01
## Rater3  2.13918    0.09027   23.70
##
## Correlation of Fixed Effects:
##        Rater1 Rater2
## Rater2 0.061
## Rater3 0.062  0.062
```

VisOrg

```
tmp <- lmer(model.formula[["VisOrg"]],data=tall[tall$Rubric=="VisOrg",])
tmp.single_intercept <- update(tmp, . ~ . + 1 - Rater)
anova(tmp,tmp.single_intercept)
```

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

## Data: tall[tall$Rubric == "VisOrg", ]
## Models:
## tmp.single_intercept: Rating ~ (1 | Artifact)
## tmp: Rating ~ Rater + (1 | Artifact) - 1
##                      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    3 227.21 235.44 -110.60   221.21
## tmp                     5 220.82 234.54 -105.41   210.82 10.392  2   0.005539
##
## tmp.single_intercept
## tmp                  **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m0 <- tmp
mA <- update(m0, . ~ . + (Rater|Artifact))
```

```
## Error: number of observations (=115) <= number of random effects (=267) for term (Rater | Artifact);
```

```
m <- update(mA, . ~ . - (1|Artifact))
```

```
## Error in h(simpleError(msg, call)): error in evaluating the argument 'object' in selecting a method
```

```
exactRLRT(m0=m0,mA=mA,m=m)
```

```
## Error in exactRLRT(m0 = m0, mA = mA, m = m): object 'm' not found
```

```
summary(tmp)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ Rater + (1 | Artifact) - 1
##    Data: tall[tall$Rubric == "VisOrg", ]
##
## REML criterion at convergence: 219.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5004 -0.3365 -0.2483  0.3841  1.8552
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.2907   0.5392
##  Residual             0.1467   0.3830
## Number of obs: 115, groups:  Artifact, 89
##
```

```
## Fixed effects:
##        Estimate Std. Error t value
## Rater1  2.37794    0.09658   24.62
## Rater2  2.64891    0.09564   27.70
## Rater3  2.28355    0.09658   23.64
##
## Correlation of Fixed Effects:
##        Rater1 Rater2
## Rater2 0.263
## Rater3 0.265  0.263
```

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

5. Start with adding all fixed effect.

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

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

```
summary(comb.0)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (0 + Rubric | Artifact)
##    Data: tall
##
## REML criterion at convergence: 1471.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0218 -0.4940 -0.0753  0.5271  3.7759
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.64070  0.8004
##           RubricInitEDA   0.38288  0.6188   0.26
##           RubricInterpRes 0.25658  0.5065   0.00 0.79
##           RubricRsrchQ    0.17398  0.4171   0.38 0.50 0.74
##           RubricSelMeth   0.09619  0.3102   0.56 0.37 0.41 0.26
##           RubricTxtOrg    0.40425  0.6358   0.03 0.69 0.80 0.64 0.24
##           RubricVisOrg    0.31878  0.5646   0.17 0.78 0.76 0.60 0.29 0.79
##  Residual                 0.19477  0.4413
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.23210    0.04013   55.63
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
comb.full <- update(comb.0, . ~ . + Rater + Semester + Sex + Repeated + Rubric)
summary(comb.full)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Sex + Repeated +
##     Rubric
##    Data: tall
##
## REML criterion at convergence: 1429.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1091 -0.5065 -0.0178  0.5242  3.7932
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.55311  0.7437
##           RubricInitEDA   0.35239  0.5936   0.47
##           RubricInterpRes 0.17512  0.4185   0.23 0.75
##           RubricRsrchQ    0.16997  0.4123   0.58 0.44 0.71
##           RubricSelMeth   0.06816  0.2611   0.39 0.60 0.74 0.41
##           RubricTxtOrg    0.26339  0.5132   0.34 0.62 0.70 0.56 0.67
##           RubricVisOrg    0.25809  0.5080   0.35 0.73 0.68 0.52 0.41 0.76
##  Residual                 0.18916  0.4349
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)      2.013748   0.109103  18.457
## Rater2           0.001977   0.054887   0.036
## Rater3          -0.174867   0.055045  -3.177
## SemesterS19     -0.175017   0.087850  -1.992
## SexM             0.010506   0.081271   0.129
## Repeated        -0.073586   0.098522  -0.747
## RubricInitEDA    0.547054   0.095710   5.716
## RubricInterpRes  0.587091   0.100893   5.819
## RubricRsrchQ     0.460875   0.087516   5.266
## RubricSelMeth    0.164863   0.094265   1.749
## RubricTxtOrg     0.692880   0.099523   6.962
## RubricVisOrg     0.530182   0.099136   5.348
##
## Correlation of Fixed Effects:
##             (Intr) Rater2 Rater3 SmsS19 SexM   Repetd RbIEDA RbrcIR RbrcRQ
## Rater2      -0.245
## Rater3      -0.237  0.499
## SemesterS19 -0.361  0.008  0.000
## SexM        -0.398 -0.026 -0.035  0.302
## Repeated    -0.154  0.001 -0.003  0.079  0.009
## RubrcIntEDA -0.552 -0.001  0.000 -0.001  0.000  0.007
## RbrcIntrpRs -0.660 -0.001  0.000 -0.001  0.000 -0.009  0.734
## RubrcRsrchQ -0.626 -0.001  0.000 -0.001  0.000 -0.039  0.585  0.756
## RubricSlMth -0.689 -0.001  0.000 -0.001  0.000 -0.088  0.659  0.777  0.689
## RubrcTxtOrg -0.611 -0.001  0.000 -0.001  0.000  0.005  0.674  0.751  0.682
## RubricVsOrg -0.607 -0.001 -0.001 -0.002 -0.001 -0.021  0.715  0.745  0.668
##             RbrcSM RbrcTO
## Rater2
```

```
## Rater3
## SemesterS19
## SexM
## Repeated
## RubrcIntEDA
## RbrcIntrpRs
## RubrcRsrchQ
## RubricSlMth
## RubrcTxtOrg  0.725
## RubricVsOrg  0.680  0.750
```

Apply backward elimination.

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

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

## ============================================================
## ===               backfitting fixed effects            ===
## ============================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.887 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     removing term
##   iteration 2
##     p-value for term "Repeated" = 0.0919 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

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

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

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

Summary of the final model without interaction.

```
summary(comb.back_elim)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric
##    Data: tall
##
## REML criterion at convergence: 1424.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1200 -0.5125 -0.0173  0.5302  3.7752
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.55495  0.7449
##           RubricInitEDA   0.35064  0.5921   0.47
##           RubricInterpRes 0.16892  0.4110   0.23 0.75
##           RubricRsrchQ    0.16777  0.4096   0.59 0.44 0.70
##           RubricSelMeth   0.06499  0.2549   0.40 0.60 0.74 0.40
##           RubricTxtOrg    0.25615  0.5061   0.33 0.61 0.69 0.55 0.66
##           RubricVisOrg    0.25894  0.5089   0.35 0.73 0.68 0.52 0.41 0.75
##  Residual                 0.18934  0.4351
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)      2.0084130  0.0987610  20.336
## Rater2           0.0003231  0.0547446   0.006
## Rater3          -0.1771062  0.0548892  -3.227
## SemesterS19     -0.1730357  0.0826927  -2.093
## RubricInitEDA    0.5474747  0.0957148   5.720
## RubricInterpRes  0.5864544  0.1008618   5.814
## RubricRsrchQ     0.4584082  0.0874179   5.244
## RubricSelMeth    0.1590770  0.0937771   1.696
## RubricTxtOrg     0.6930033  0.0995479   6.962
## RubricVisOrg     0.5289027  0.0990973   5.337
##
## Correlation of Fixed Effects:
##             (Intr) Rater2 Rater3 SmsS19 RbIEDA RbrcIR RbrcRQ RbrcSM RbrcTO
## Rater2      -0.281
## Rater3      -0.277  0.499
## SemesterS19 -0.264  0.017  0.011
## RubrcIntEDA -0.610 -0.001  0.000 -0.002
## RbrcIntrpRs -0.735 -0.001  0.000  0.000  0.734
## RubrcRsrchQ -0.701 -0.001  0.000  0.002  0.586  0.756
## RubricSlMth -0.782  0.000  0.000  0.006  0.662  0.779  0.688
## RubrcTxtOrg -0.679 -0.001  0.000 -0.001  0.674  0.751  0.682  0.728
## RubricVsOrg -0.675 -0.001 -0.001  0.000  0.715  0.745  0.667  0.681  0.750
```

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

2. Add interaction in our best model we get before.

```
comb.inter <- update(comb.back_elim, . ~ . + Rater*Semester*Rubric)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00431172 (tol = 0.002, component 1)
```

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

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

```
summary(comb.inter.u)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric +
##     Rater:Semester + Rater:Rubric + Semester:Rubric + Rater:Semester:Rubric
##    Data: tall
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))
##
## REML criterion at convergence: 1424.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9141 -0.5141 -0.0653  0.5023  3.6609
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.48550  0.6968
##           RubricInitEDA   0.35257  0.5938   0.42
##           RubricInterpRes 0.14619  0.3824   0.32 0.80
##           RubricRsrchQ    0.16444  0.4055   0.66 0.43 0.72
##           RubricSelMeth   0.06297  0.2509   0.45 0.64 0.78 0.49
##           RubricTxtOrg    0.25441  0.5044   0.44 0.65 0.67 0.60 0.62
##           RubricVisOrg    0.25527  0.5052   0.35 0.73 0.68 0.57 0.35 0.76
##  Residual                 0.18839  0.4340
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                     1.739538   0.136568  12.738
## Rater2                          0.302995   0.155107   1.953
## Rater3                          0.237851   0.155863   1.526
## SemesterS19                    -0.129077   0.250318  -0.516
## RubricInitEDA                   0.765215   0.165241   4.631
## RubricInterpRes                 0.979228   0.162160   6.039
```

```
## RubricRsrchQ                          0.710427   0.147386   4.820
## RubricSelMeth                         0.462750   0.155274   2.980
## RubricTxtOrg                          1.011251   0.160899   6.285
## RubricVisOrg                          0.647869   0.166603   3.889
## Rater2:SemesterS19                    0.268014   0.303883   0.882
## Rater3:SemesterS19                   -0.072789   0.301026  -0.242
## Rater2:RubricInitEDA                 -0.325018   0.204108  -1.592
## Rater3:RubricInitEDA                 -0.374190   0.205354  -1.822
## Rater2:RubricInterpRes               -0.469281   0.201051  -2.334
## Rater3:RubricInterpRes               -0.711515   0.202316  -3.517
## Rater2:RubricRsrchQ                  -0.447050   0.189326  -2.361
## Rater3:RubricRsrchQ                  -0.474411   0.190681  -2.488
## Rater2:RubricSelMeth                 -0.301450   0.193678  -1.556
## Rater3:RubricSelMeth                 -0.365656   0.194970  -1.875
## Rater2:RubricTxtOrg                  -0.449164   0.200927  -2.235
## Rater3:RubricTxtOrg                  -0.407754   0.202209  -2.016
## Rater2:RubricVisOrg                   0.009042   0.205059   0.044
## Rater3:RubricVisOrg                  -0.287443   0.206299  -1.393
## SemesterS19:RubricInitEDA            -0.050212   0.301475  -0.167
## SemesterS19:RubricInterpRes           0.127813   0.295706   0.432
## SemesterS19:RubricRsrchQ              0.133874   0.267750   0.500
## SemesterS19:RubricSelMeth            -0.089616   0.282837  -0.317
## SemesterS19:RubricTxtOrg              0.166097   0.293176   0.567
## SemesterS19:RubricVisOrg              0.146845   0.302496   0.485
## Rater2:SemesterS19:RubricInitEDA      0.020326   0.392376   0.052
## Rater3:SemesterS19:RubricInitEDA      0.252422   0.389961   0.647
## Rater2:SemesterS19:RubricInterpRes -0.266618   0.385390  -0.692
## Rater3:SemesterS19:RubricInterpRes -0.152392   0.383354  -0.398
## Rater2:SemesterS19:RubricRsrchQ    -0.217348   0.360414  -0.603
## Rater3:SemesterS19:RubricRsrchQ     0.354319   0.357388   0.991
## Rater2:SemesterS19:RubricSelMeth   -0.401035   0.370200  -1.083
## Rater3:SemesterS19:RubricSelMeth   -0.192670   0.367887  -0.524
## Rater2:SemesterS19:RubricTxtOrg    -0.542267   0.385011  -1.408
## Rater3:SemesterS19:RubricTxtOrg    -0.316395   0.382614  -0.827
## Rater2:SemesterS19:RubricVisOrg    -0.603626   0.392909  -1.536
## Rater3:SemesterS19:RubricVisOrg    -0.186749   0.390759  -0.478

##
## Correlation matrix not shown by default, as p = 42 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it

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

Apply backward elimination on the interaction model.

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

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

## ======================================================
```

```
## ===              backfitting fixed effects         ===
## ============================================================
## processing model terms of interaction level 3
##   iteration 1
##     p-value for term "Rater:Semester:Rubric" = 0.5526 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     removing term
## processing model terms of interaction level 2
##   iteration 2
##     p-value for term "Rater:Semester" = 0.598 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     removing term
##   iteration 3
##     p-value for term "Semester:Rubric" = 0.0761 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

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

## boundary (singular) fit: see ?isSingular

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

Summary of final model with interaction

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric +
##     Rater:Rubric
##    Data: tall
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))
##
## REML criterion at convergence: 1419.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9280 -0.5122 -0.0447  0.4827  3.5854
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.50348  0.7096
##           RubricInitEDA   0.35480  0.5956   0.44
##           RubricInterpRes 0.15192  0.3898   0.35 0.82
##           RubricRsrchQ    0.17953  0.4237   0.63 0.44 0.72
##           RubricSelMeth   0.06727  0.2594   0.42 0.60 0.74 0.36
##           RubricTxtOrg    0.26069  0.5106   0.42 0.64 0.67 0.55 0.64
##           RubricVisOrg    0.25491  0.5049   0.34 0.71 0.68 0.51 0.38 0.77
##  Residual                 0.18519  0.4303
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##                         Estimate Std. Error t value
## (Intercept)              1.75945    0.11785  14.929
## Rater2                   0.36537    0.13296   2.748
## Rater3                   0.21421    0.13297   1.611
## SemesterS19             -0.17780    0.08228  -2.161
## RubricInitEDA            0.74625    0.13676   5.457
## RubricInterpRes          1.01453    0.13479   7.527
## RubricRsrchQ             0.74926    0.12419   6.033
## RubricSelMeth            0.42672    0.13040   3.272
## RubricTxtOrg             1.04967    0.13551   7.746
## RubricVisOrg             0.68354    0.13947   4.901
## Rater2:RubricInitEDA    -0.30843    0.17249  -1.788
## Rater3:RubricInitEDA    -0.29522    0.17282  -1.708
## Rater2:RubricInterpRes  -0.53674    0.17008  -3.156
## Rater3:RubricInterpRes  -0.75247    0.17049  -4.414
## Rater2:RubricRsrchQ     -0.50157    0.16151  -3.106
## Rater3:RubricRsrchQ     -0.37068    0.16179  -2.291
## Rater2:RubricSelMeth    -0.39602    0.16467  -2.405
## Rater3:RubricSelMeth    -0.41324    0.16504  -2.504
## Rater2:RubricTxtOrg     -0.58380    0.17141  -3.406
## Rater3:RubricTxtOrg     -0.48649    0.17177  -2.832
## Rater2:RubricVisOrg     -0.14444    0.17442  -0.828
## Rater3:RubricVisOrg     -0.33380    0.17481  -1.910

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

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

3. Apply ANOVA to find out the best model.

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

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

```
## Data: tall
## Models:
## comb.back_elim: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric
## comb.inter_elim: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric + Rater:Rubric
## comb.inter.u: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric + Rater:Semester + Rater:
##                 npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## comb.back_elim    39 1464.0 1647.2 -693.02   1386.0
## comb.inter_elim   51 1454.5 1694.1 -676.26   1352.5 33.526 12   0.000801 ***
## comb.inter.u      71 1471.4 1804.8 -664.68   1329.4 23.161 20   0.280962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Try to add random effect.

(0 + Rater|Artifact)

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

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

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

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

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

```
## Data: tall
## Models:
## m0: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric + Rater:Rubric
## mA: Rating ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact) + Rater + Semester + Rubric + Rater:Ru
##    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m0   51 1454.5 1694.1 -676.26   1352.5
## mA   57 1415.9 1683.6 -650.94   1301.9 50.647  6  3.487e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(0 + Semester|Artifact)

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

```
## 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 1 negative eigenvalues
```

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

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

```
## Data: tall
## Models:
## m0: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric + Rater:Rubric
## mA: Rating ~ (0 + Rubric | Artifact) + (0 + Semester | Artifact) + Rater + Semester + Rubric + Rater
##    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m0   51 1454.5 1694.1 -676.26   1352.5
## mA   54 1458.4 1712.0 -675.18   1350.4 2.1534  3     0.5412
```

(0 + Rater:Rubric|Artifact)

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

```
## Error: number of observations (=810) <= number of random effects (=1890) for term (0 + Rater:Rubric
```

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

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

```
## Data: tall
## Models:
## m0: Rating ~ (0 + Rubric | Artifact) + Rater + Semester + Rubric + Rater:Rubric
## mA: Rating ~ (0 + Rubric | Artifact) + (0 + Semester | Artifact) + Rater + Semester + Rubric + Rater
##    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m0   51 1454.5 1694.1 -676.26   1352.5
## mA   54 1458.4 1712.0 -675.18   1350.4 2.1534  3     0.5412
```

5. Final model

```
comb.final <- lmer(Rating ~ (0 + Rubric|Artifact) + (0 + Rater|Artifact) + Rater +
             Semester + Rubric + Rater:Rubric, data=tall)
```

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

```
summary(comb.final)$varcor
```

```
##  Groups     Name            Std.Dev. Corr
##  Artifact   RubricCritDes   0.70466
##             RubricInitEDA   0.56380   0.318
##             RubricInterpRes 0.31948   0.142  0.674
##             RubricRsrchQ    0.42313   0.500  0.194  0.538
##             RubricSelMeth   0.19554   0.145  0.227  0.376 -0.240
##             RubricTxtOrg    0.50030   0.269  0.437  0.364  0.305  0.213
##             RubricVisOrg    0.48204   0.175  0.504  0.445  0.276 -0.160  0.537
##  Artifact.1 Rater1          0.11317
##             Rater2          0.33429  -0.486
##             Rater3          0.30682   0.332  0.663
##  Residual                   0.36699
```

```
summary(comb.final)$coef
```

```
##                           Estimate Std. Error    t value
## (Intercept)              1.7575637 0.11404553 15.4110695
## Rater2                   0.3660357 0.13918435  2.6298628
## Rater3                   0.1959108 0.12966732  1.5108727
## SemesterS19             -0.1591847 0.07647713 -2.0814681
## RubricInitEDA            0.7394956 0.12995561  5.6903703
## RubricInterpRes          0.9915251 0.12770499  7.7641838
## RubricRsrchQ             0.7261970 0.11792700  6.1580215
## RubricSelMeth            0.4106840 0.12470528  3.2932364
## RubricTxtOrg             1.0157859 0.12999662  7.8139408
## RubricVisOrg             0.6542619 0.13352531  4.8999093
## Rater2:RubricInitEDA    -0.2998031 0.15608617 -1.9207535
## Rater3:RubricInitEDA    -0.2947458 0.15634742 -1.8851978
## Rater2:RubricInterpRes  -0.5132313 0.15348164 -3.3439265
## Rater3:RubricInterpRes  -0.7148636 0.15363623 -4.6529623
## Rater2:RubricRsrchQ     -0.4874212 0.14721814 -3.3108771
## Rater3:RubricRsrchQ     -0.3224080 0.14726165 -2.1893547
## Rater2:RubricSelMeth    -0.3863728 0.15030727 -2.5705530
## Rater3:RubricSelMeth    -0.3871739 0.14961201 -2.5878532
## Rater2:RubricTxtOrg     -0.5510430 0.15646077 -3.5219242
## Rater3:RubricTxtOrg     -0.4449139 0.15673150 -2.8387015
## Rater2:RubricVisOrg     -0.1049002 0.15860588 -0.6613892
## Rater3:RubricVisOrg     -0.2752337 0.15884361 -1.7327340
```

E. Research Question 4

1. Semester

```
Fall = tall[tall$Semester == 'F19',]
Spring = tall[tall$Semester == 'S19',]
```

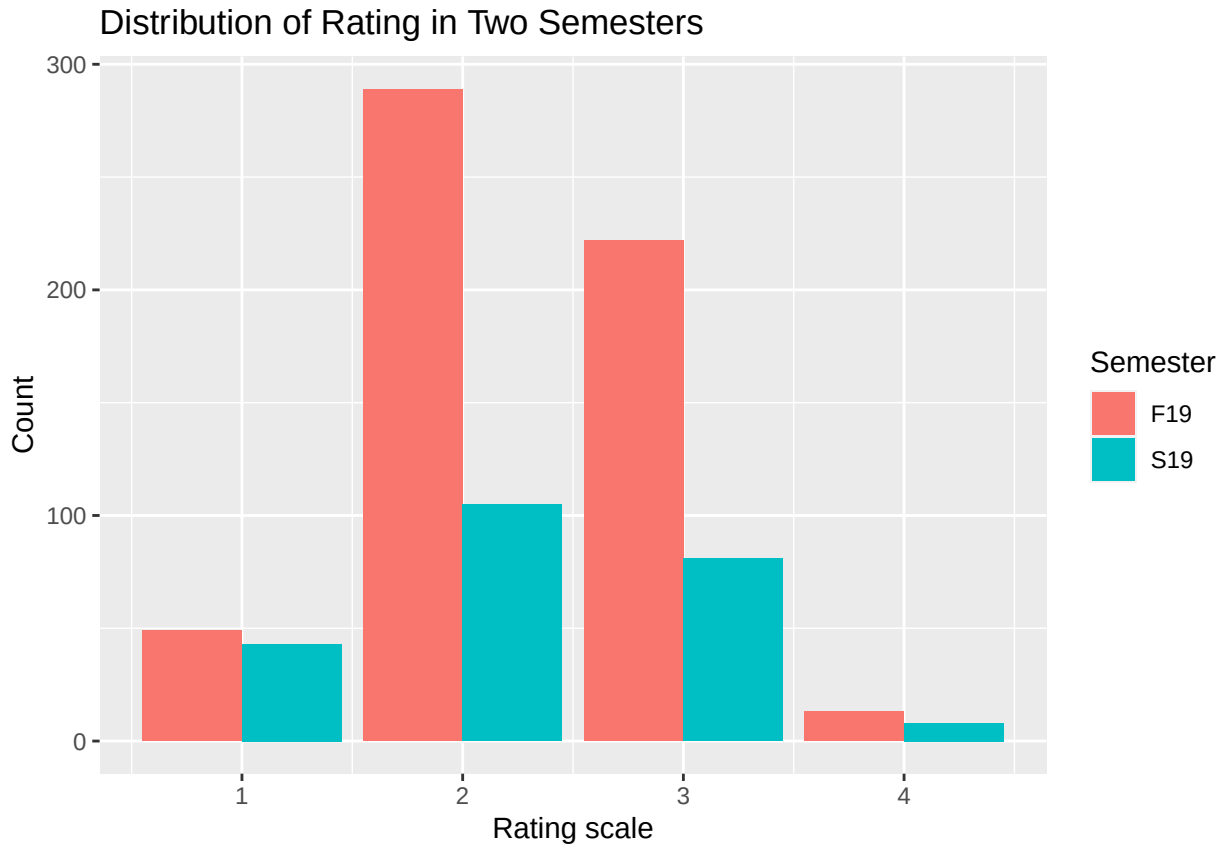
```
summary(Fall$Rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.347   3.000   4.000
```

```
summary(Spring$Rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.228   3.000   4.000
```

```
tall %>%
  group_by(Rating, Semester) %>%
  count(Rating) %>%
  ggplot()+
  geom_bar(aes(Rating, n, fill = Semester),
           position = position_dodge(),
           stat = "identity")+
  xlab("Rating scale")+
  ylab("Count")+
  ggtitle("Distribution of Rating in Two Semesters")
```



2. Sex

```
Male = tall[tall$Sex == 'M',]
Female = tall[tall$Sex == 'F',]
```

```
summary(Male$Rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    2.00    2.31    3.00    4.00
```

```
summary(Female$Rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.314   3.000   4.000
```

```
tall %>%
  group_by(Rating, Sex) %>%
  count(Rating) %>%
  ggplot()+
  geom_bar(aes(Rating, n, fill = Sex),
           position = position_dodge(),
           stat = "identity")+
  xlab("Rating scale")+
  ylab("Count")+
  ggtitle("Distribution of Rating Between Male and Female")
```