Regression Analysis to Decide Whether the Ratings for a Education Program is Fair Enough

Ziyan Xia
Department of Statistics and Data Science, Carnegie Mellon University
zxia2@andrew.cmu.edu

Abstract:

It's always important to decide whether an evaluation experiment is fair before use the results to decide whether a education program is successful. In order to learn whether the experiment is fair, methods include making barplots, AVONA tests and Back-fit fixed effects and forward-fit random effects of an LMER model method. Eventually we find that it's fairer to use the reduced dataset, which is the data of 13 artifacts all seen by raters to do the evaluation. Also, rater 1 disagrees with rater 2 on ratings of Research Question. Semester seem an important factor affecting ratings. These thigs are worth considering when doing the evaluation of the program based on the ratings.

1. Introduction

It's always important for colleges to evaluate the quality of their education programs. Some colleges use the ratings of education-relevant statistics to decide whether an education program is successful. Dietrich College at Carnegie Mellon University is now in the process of implementing a new "General Education" program for undergraduates, which specifies a set of courses and experiences that all undergraduates must take. In order to determine whether this program is successful, the college hopes to rate student work performed in each of the "Gen Ed" courses each year. Recently the college has been experimenting with rating work in Freshman Statistics, using raters from across the college. For experiments like this, we always wonder whether it's truly fair to use the ratings from these raters based on these rubrics. To learn whether they are as follows:

1. **Do rater's ratings vary much?**
   Is the distribution of ratings for each rubric pretty much indistinguishable from the other rubrics, or are there rubrics that tend to get especially high or low ratings? Is the distribution of ratings given by each rater pretty much indistinguishable from the other raters, or are there raters that tend to give especially high or low ratings?

2. **Do rater's ratings reach a consensus?**
   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. **How do various factors affect ratings?**
   More generally, how are the various factors in this experiment (Rater, Semester, Sex, Repeated, Rubric) related to the ratings? Do the factors interact in any interesting ways?

4. **Other Interesting things about ratings**
   Is there anything else interesting to say about this data?

2. Data

In a recent rating work experiment, 91 project papers—referred to as "artifacts"—were randomly sampled from a Fall and Spring section of Fresh-man Statistics. Three raters from three different departments were asked to rate these artifacts on seven rubrics, as shown in Table 1.

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.

The rating scale for all rubrics is shown in Table 2.

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 raters did not know which class or which students produced the artifacts that they rated. Thirteen of the 91 artifacts were rated by all three raters; each of the remaining

78 artifacts were rated by only rater. The variables available for analysis are defined in Table 3. The file ratings.csv contains data organized exactly as in Table 3.

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 that we are using

3. Methods

To learn how the various factors in this experiment related to the rating and whether the rating depends largely on raters, there are four questions to answer. Before we answer these four questions, we create a subset of original dataset and this dataset contains the data of 13 artifacts seen by all 3 raters, we call it reduced dataset and call the original dataset the full dataset.
Our methods to answer these questions are as follows:

1. **Do rater's ratings vary much?**
   To answer this question, first we made barplots for the counts of ratings for each rubric both on the reduced dataset and full dataset. Besides, we also made barplots for the counts of ratings (with possibly NAs) for each rater both on the reduced dataset and full dataset.

2. **Do rater's ratings reach a consensus?**
   To answer this question, we fit seven random-intercept models, one for each rubric, and calculate the seven intraclass correlation (ICC) on both reduced dataset and full dataset to measure of agreement among the raters. Then we make a 2-way table of counts for the ratings of each pair of raters, on each rubric to determine who is agreeing with whom on each rubric.

3. **How do various factors affect ratings?**

To answer this question, we first add fixed effects to the seven rubric-specific models using just the data from the 13 common artifacts that are seen by all three raters using Back-fit fixed effects and forward-fit random effects of an LMER model method (fitLMER) and then redo the whole process on the full dataset after eliminating NAs in the full dataset. Then we add fixed effect and interactions for the "combined" model [See Technical Appendix, Page 20] using multiple ANOVA tests and add random effect using fitLMER.

4. **Other Interesting things about ratings**

   To further discover our data, we made barplots of counts of ratings for each rater during each semester separately.

4. Results

1. **Do rater's ratings vary much?**

   Figure 1 and Figure 2 are the barplots for the counts of ratings for each rubric both on the reduced dataset and full dataset.

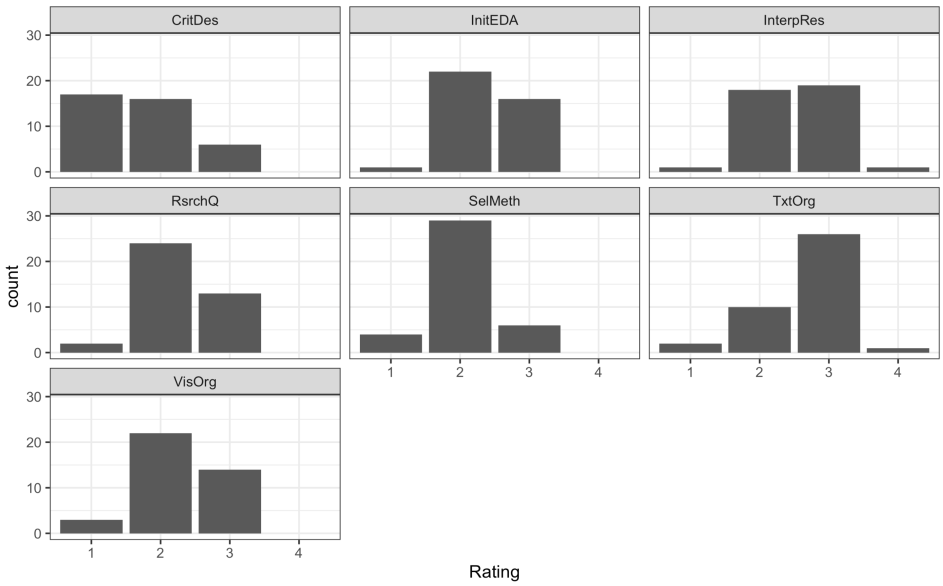


Figure 1: Barplots of ratings count on each rubric (reduced dataset)

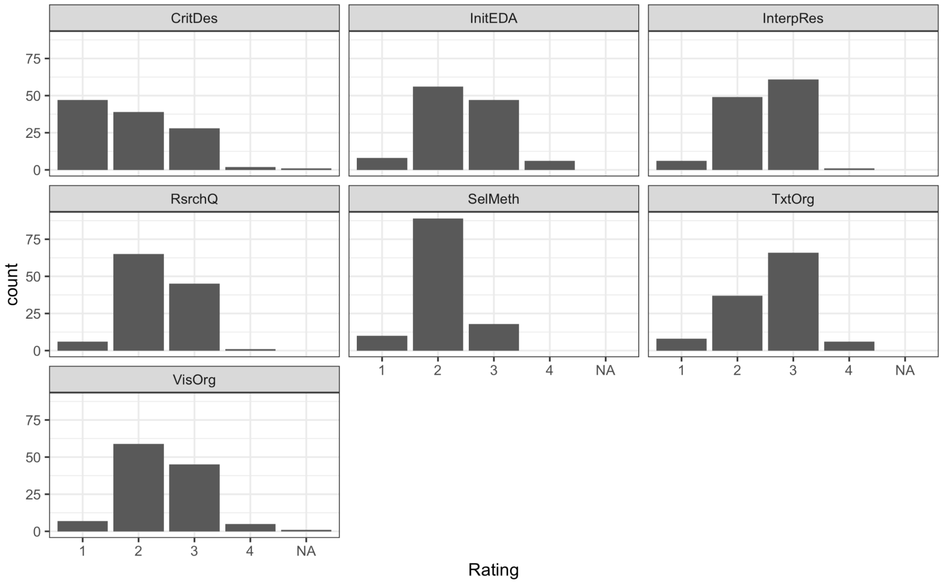


Figure 2: Barplots of ratings count on each rubric (full dataset)

After comparing Figure 1 and Figure 2, it is quite obvious that the distribution of ratings for some rubrics pretty much indistinguishable from the other rubrics on both dataset. Critique Design get especially low ratings. Interpret Results and Text Organization get especially low ratings. Except for the increase of NAs and rating value 4, the distribution of ratings for each rater on reduced dataset agrees with that on full dataset.

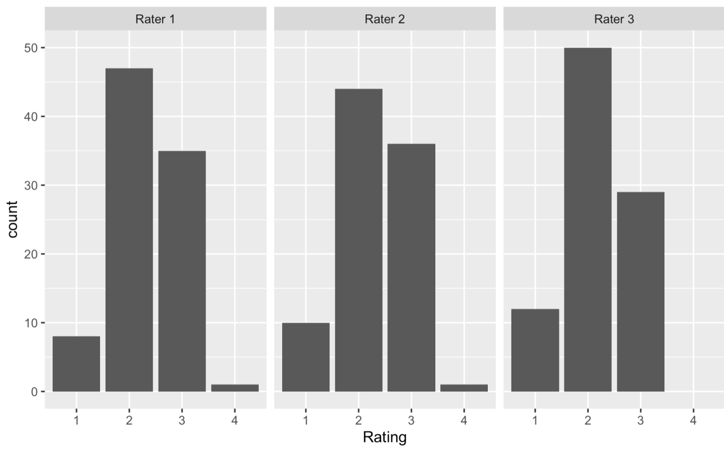


Figure 3: Barplots of ratings count for each rater (reduced dataset)

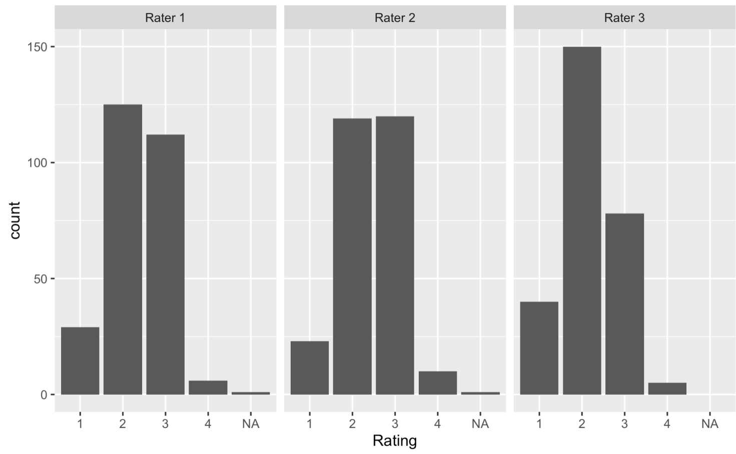


Figure 4: Barplots ratings count for each rater (full dataset)

Figure 3 and Figure 4 are the barplots for the counts of ratings for each rater both on the reduced dataset and full dataset. After comparing Figure 3 and Figure 4, it is quite obvious that the distribution of ratings given by each rater is not quite indistinguishable from the other raters. Except for the increase of NAs and rating value 4, the distribution of ratings for each rater on reduced dataset agrees with that on full dataset.

Therefore, the reduced dataset seems like a good representative of the full dataset here.

2. **Do rater's ratings reach a consensus?**

After calculating the intraclass correlation (ICC) on both reduced dataset and full dataset and calculating the agreement rate of the rubric for each two raters, we create a table called table 4 to compare them. As is shown in table 4, the column named "ICC.alldata" means the ICCs calculated from seven random-intercept models that are fitted on the full dataset and the column named "ICC.common" means the ICCs calculated from seven random-intercept models that are fitted on the reduced dataset. The column named "a12" means the agreement rate of rater 1 and 2 for the rubric. The column named "a23" means the agreement rate of rater 2 and 3 for the rubric. The column named "a13" means the agreement rate of rater 1 and 3 for the rubric.

The column "ICC.alldata" agrees with the column "ICC.common" while it is quite hard to see which agreement rate between two raters contributes most to the ICC calculated before.

	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

Table 4: ICCs and Raters Agreement Rate for each rubric

3. **How do various factors affect ratings?**
After adding fixed effects to the seven rubric-specific models using reduced dataset using Back-fit fixed effects and forward-fit random effects of an LMER model method, the final models we get is in Table 5. In Table 5, all the Rubric-specific models end up with formula "Rating (numeric) ~ (1 | Artifact)", which means for each specific, the model will give different overall mean based on different Artifact. (See Technical Appendix, Page 14) The final models in Table 5 are all random-intercept models.

Rubric	Final Models
CritDes	Rating (numeric) ~ (1 | Artifact)
InitEDA	Rating (numeric) ~ (1 | Artifact)
InterpRes	Rating (numeric) ~ (1 | Artifact)
RsrchQ	Rating (numeric) ~ (1 | Artifact)
SelMeth	Rating (numeric) ~ (1 | Artifact)
TxtOrg	Rating (numeric) ~ (1 | Artifact)
VisOrg	Rating (numeric) ~ (1 | Artifact)

Table 5: Final fixed effect on reduced dataset

After adding fixed effects to the seven rubric-specific models using full dataset using Back-fit fixed effects and forward-fit random effects of an LMER model method, the final models we get is in Table 6.

Rubric	Final Models
CritDes	Rating (numeric) ~ Rater (factor) + (1 | Artifact) -1
InitEDA	Rating (numeric) ~ (1 | Artifact)
InterpRes	Rating (numeric) ~ Rater (factor) + (1 | Artifact) -1
RsrchQ	Rating (numeric) ~ (1 | Artifact)
SelMeth	Rating (numeric) ~ Rater (factor) + Semester + (1 | Artifact)-1
TxtOrg	Rating (numeric) ~ (1 | Artifact)
VisOrg	Rating (numeric) ~ Rater (factor) + (1 | Artifact) -1

Table 6: Final fixed effect on full dataset

We see there are some differences among the models fitted on the full dataset: For rubrics InitEDA, RsrchQ and TxtOrg, the models are just the simple random-intercept models. However, for the other four rubrics, the models are a little more complex. For rubrics CritDes, InterpRes and VisOrg, compared to the simple random-intercept models, the models have one more fixed effect Rater. Also, rubric SelMeth has two more fixed effects Rater and Semester than random-intercept models.

After multiple ANOVA tests, we are able to select fixed effects Rater, Semester, Rubric, Repeated and interactions Rater * Rubric. After fitLMER, we are able to select random effects Rater, Rubric. The final model's output is in Figure 5.

```
lmer(formula = Rating ~ 1 + Rater + Semester + Rubric + Repeated +
    Rater * Rubric + (0 + Rater + Rubric | Artifact), data = tall)
                      coef.est coef.se
(Intercept)            1.80     0.17
Rater                  0.08     0.07
SemesterS19           -0.13     0.08
RubricInitEDA          0.83     0.19
RubricInterpRes        1.30     0.19
RubricRsrchQ           0.81     0.18
RubricSelMeth          0.51     0.19
RubricTxtOrg           1.15     0.19
RubricVisOrg           0.84     0.19
Repeated              -0.07     0.09
Rater:RubricInitEDA   -0.15     0.08
Rater:RubricInterpRes -0.36     0.08
Rater:RubricRsrchQ    -0.18     0.08
Rater:RubricSelMeth   -0.18     0.08
Rater:RubricTxtOrg    -0.23     0.08
Rater:RubricVisOrg    -0.16     0.08
```

Figure 5: The output of the final "combined" model

4. Other Interesting things about ratings

The barplots of counts of ratings for each rater during each semester separately on reduced dataset are in Figure 6. Figure 6 shows for each semester the raters ratings will vary a lot.

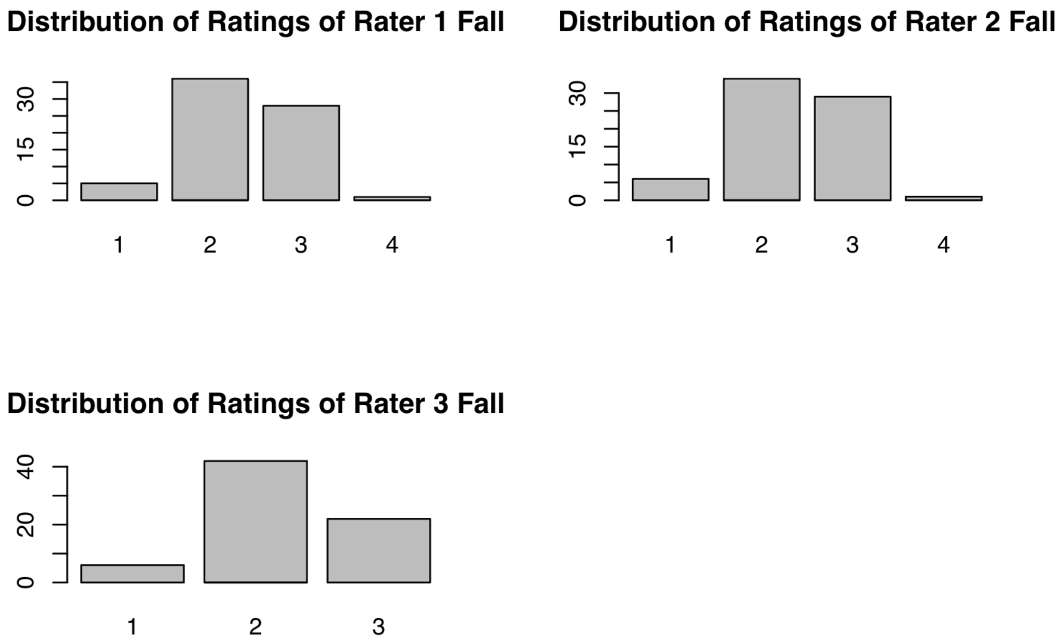


Figure 6: The barplots of counts of ratings for each rater during each semester

5. Discussion

The ratings for each rubrics vary a lot while each rater's ratings don't vary a lot. The fact that the distribution of ratings for some rubrics indistinguishable from the other rubrics on both dataset indicates the program may be considered as successful on some rubrics but fail on others.
The ICC on both reduced dataset and full dataset indicates are low for most of the rubrics, meaning the intraclass correlation between different raters is quite low. It is worth notifying that rater 1 and rater 2 quite disagree on rubric Research Question while rater 3's ratings quite agree with other raters based on the agreement rate.
It's quite interesting that for seven rubric-specific models, if we apply fitLMER method on them, the final selection of fixed effects that should be added to the models is quite different for the reduced dataset and the full dataset. In the full dataset, the fixed effect Rater is added for some rubrics. In the "combined" model fitting process, we find the interaction between rater and rubric is quite significant. It makes sense cause not all artifacts were seen by all raters. From the barplots we made to answer question 1, we can see that the reduced dataset is actually a good representative of the full dataset, considering using the full dataset there will be interactions between rubric and rater, it's better to use the reduced dataset to do the analysis.
In the full dataset, the fixed effect Semester is also added for one rubric and the fixed effect Semester is added in the "combined" model. The barplots of counts of ratings for each rater during each semester show which semester does have effect on the ratings distribution.

References

Dietrich College General Education Program, Dietrich College of Humanities and Social Sciences, Carnegie Mellon University

Technical Appendix for Project 2

Ziyan Xia

11/28/2021

```
tall<-read.csv("/Users/ceciliaxia/Desktop/tall.csv")
rating<-read.csv("/Users/ceciliaxia/Desktop/ratings.csv")
subset_rating<-rating[grep("0",rating$Artifact,fixed=TRUE),]
subset_tall<-tall[grep("0",tall$Artifact,fixed=TRUE),]
```

```
library(LMERConvenienceFunctions)
```

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

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

```
library(RLRsim)
library(scales)
library(performance)
library(lme4)
library(arm)
```

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

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

```
## Working directory is /Users/ceciliaxia/Desktop
```

```
##
## Attaching package: 'arm'
```

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

```
## The following object is masked from 'package:scales':
##
##     rescale
```

```
library(lme4)
library(ggplot2)
library(plyr)
library(LMERConvenienceFunctions)
```

1. Part A: EDA on subset datasets results

```
par(mfrow=c(3,3))
with(subset_rating,{
  barplot(table(RsrchQ),main=" Rating on Research Question")
  barplot(table(CritDes),main=" Rating on Critique Design")
  barplot(table(InitEDA),main=" Rating on Initial EDA")
  barplot(table(SelMeth),main=" Rating on Select Method(s)")
```

```
  barplot(table(InterpRes),main=" Rating on Interpret Results")
  barplot(table(VisOrg),main=" Rating on Visual Organization")
  barplot(table(TxtOrg),main=" Rating on Text Organization")
})


with(subset_rating,table(RsrchQ))
```

```
## RsrchQ
##  1  2  3
##  2 24 13
```

```
with(subset_rating, table(CritDes))
```

```
## CritDes
##  1  2  3
## 17 16  6
```

```
with(subset_rating, table(InitEDA))
```

```
## InitEDA
##  1  2  3
##  1 22 16
```

```
with(subset_rating, table(SelMeth))
```

```
## SelMeth
##  1  2  3
##  4 29  6
```

```
with(subset_rating, table(InterpRes))
```

```
## InterpRes
##  1  2  3  4
##  1 18 19  1
```

```
with(subset_rating, table(VisOrg))
```

```
## VisOrg
##  1  2  3
##  3 22 14
```

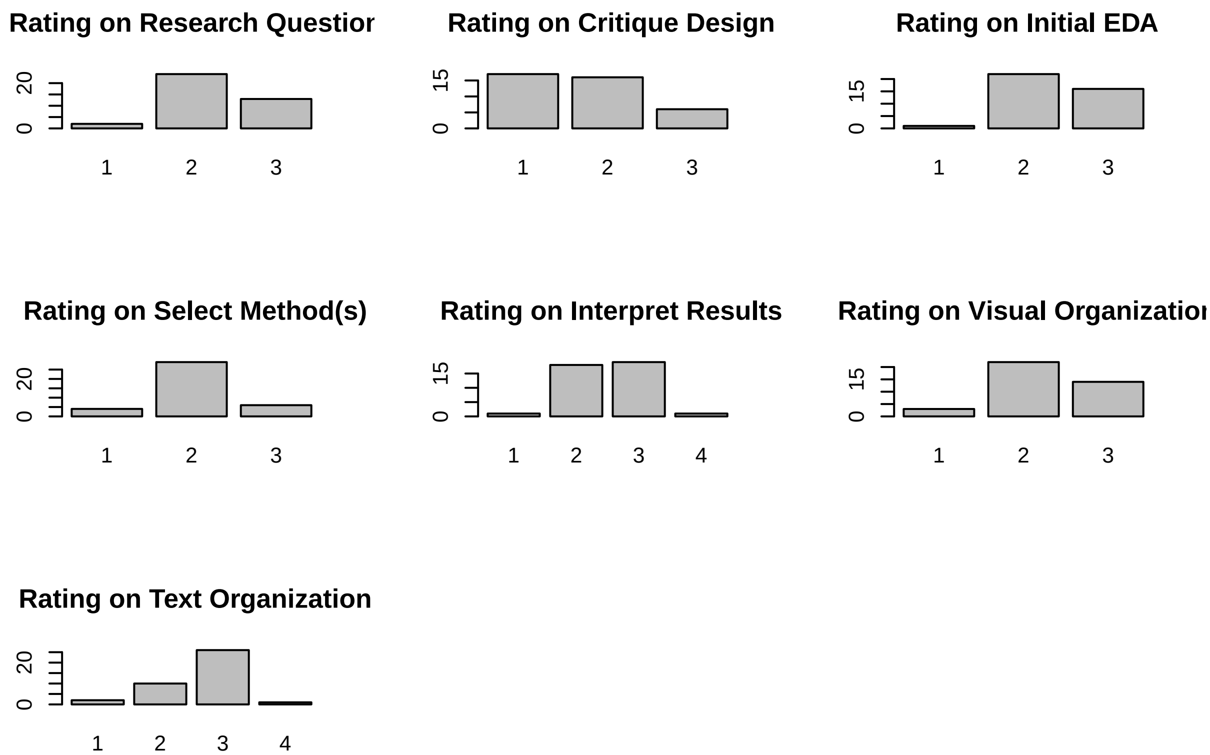
```
with(subset_rating, table(TxtOrg))
```

```
## TxtOrg
##  1  2  3  4
##  2 10 26  1
```

```
summary(subset_rating[,7:13])
```

```
##      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
```

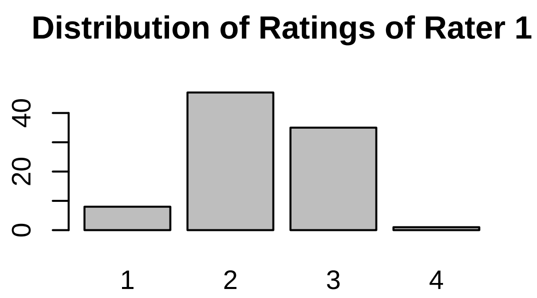
```
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000   Median :2.000   Median :3.000
##  Mean   :2.513   Mean   :2.282   Mean   :2.667
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :4.000   Max.   :3.000   Max.   :4.000
```

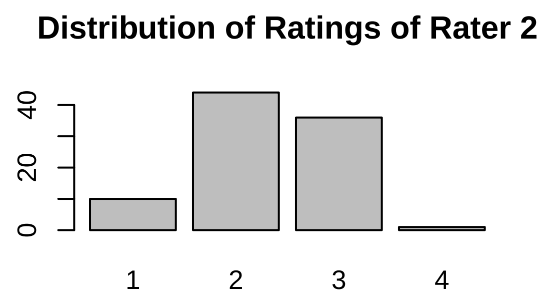


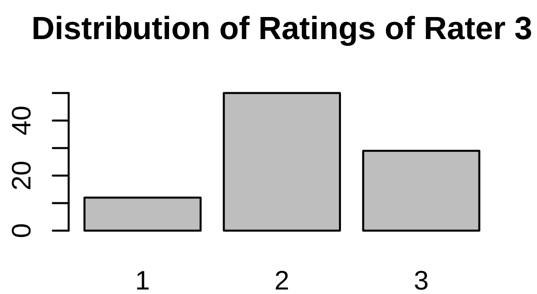
From the barplots and counts of ratings for each rubrics, it is quite obvious that the distribution of ratings for some rubrics pretty much indistinguishable from the other rubrics. Critique Design get especially low ratings. Interpret Results and Text Organization get especially low ratings.

```
par(mfrow=c(2,2))
barplot(table(subset_tall[which(subset_tall$Rater==1),]$Rating),main="Distribution of Ratings of Rater
barplot(table(subset_tall[which(subset_tall$Rater==2),]$Rating),main="Distribution of Ratings of Rater
barplot(table(subset_tall[which(subset_tall$Rater==3),]$Rating),main="Distribution of Ratings of Rater
tmp1<-data.frame(r1=subset_tall[which(subset_tall$Rater==1),]$Rating,r2=subset_tall[which(subset_tall$Ra
summary(tmp1)
```

```
##        r1              r2              r3
##  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 :2.000
##  Mean   :2.319   Mean   :2.308   Mean   :2.187
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :4.000   Max.   :4.000   Max.   :3.000
```







From the barplots and counts of ratings for each rubrics, it is quite obvious that the distribution of ratings given by each rater is not quite indistinguishable from the other raters.

Part B: EDA on full datset results

```
par(mfrow=c(2,4))
with(rating,{
  barplot(table(RsrchQ),main=" Rating on Research Question")
  barplot(table(CritDes),main=" Rating on Critique Design")
  barplot(table(InitEDA),main=" Rating on Initial EDA")
  barplot(table(SelMeth),main=" Rating on Select Method(s)")
  barplot(table(InterpRes),main=" Rating on Interpret Results")
  barplot(table(VisOrg),main=" Rating on Visual Organization")
  barplot(table(TxtOrg),main=" Rating on Text Organization")
})


with(rating,table(RsrchQ))
```

```
## RsrchQ
##   1  2  3  4
##   6 65 45  1
```

```
with(rating, table(CritDes))
```

```
## CritDes
##   1  2  3  4
## 47 39 28  2
```

```
with(rating, table(InitEDA))
```

```
## InitEDA
##   1  2  3  4
##   8 56 47  6
```

```r
with(rating, table(SelMeth))
```

```
## SelMeth
##  1  2  3
## 10 89 18
```

```r
with(rating, table(InterpRes))
```

```
## InterpRes
##  1  2  3  4
##  6 49 61  1
```

```r
with(rating, table(VisOrg))
```

```
## VisOrg
##  1  2  3  4
##  7 59 45  5
```

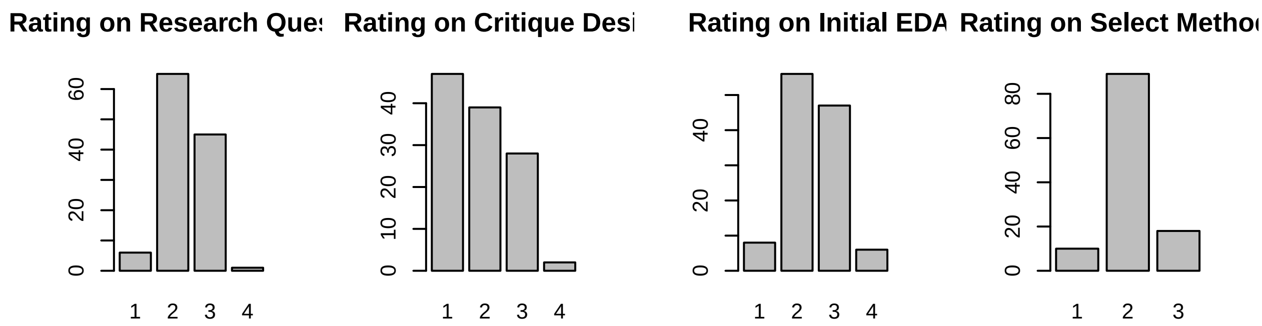
```r
with(rating, table(TxtOrg))
```

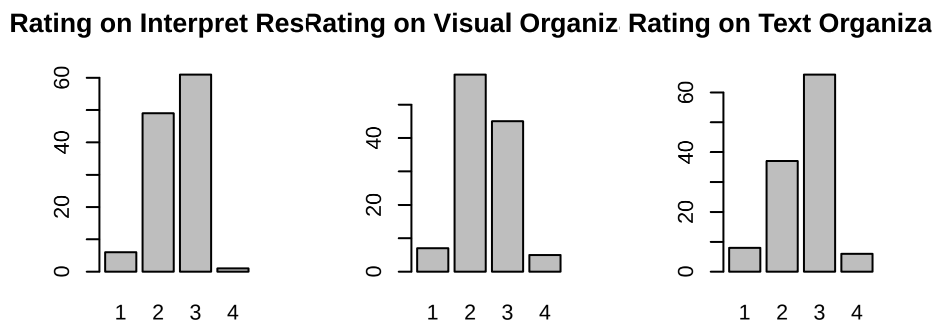
```
## TxtOrg
##  1  2  3  4
##  8 37 66  6
```

```r
summary(rating[,7:13])
```

```
##      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
```

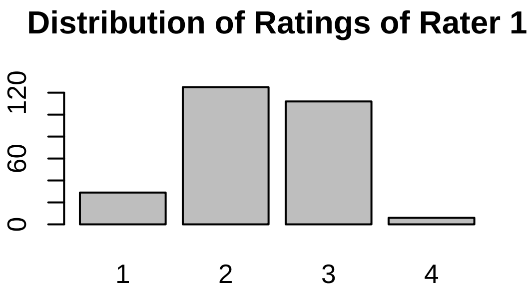
```r
par(mfrow=c(2,2))
```

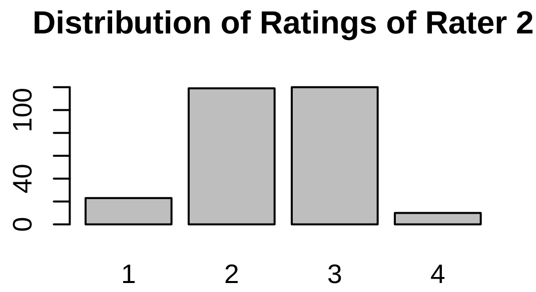


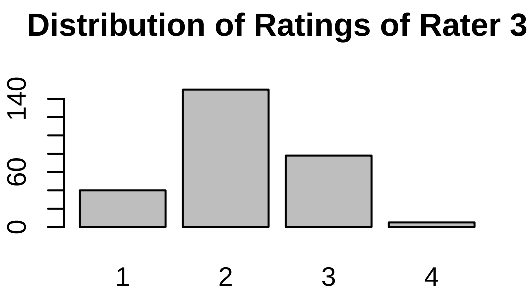



```
barplot(table(tall[which(tall$Rater==1),]$Rating),main="Distribution of Ratings of Rater 1")
barplot(table(tall[which(tall$Rater==2),]$Rating),main="Distribution of Ratings of Rater 2")
barplot(table(tall[which(tall$Rater==3),]$Rating),main="Distribution of Ratings of Rater 3")
tmp1<-data.frame(r1=tall[which(tall$Rater==1),]$Rating,r2=tall[which(tall$Rater==2),]$Rating,r3=tall[whi
summary(tmp1)
```

```
##        r1              r2              r3       
##  Min.   :1.000   Min.   :1.00   Min.   :1.000  
##  1st Qu.:2.000   1st Qu.:2.00   1st Qu.:2.000  
##  Median :2.000   Median :2.00   Median :2.000  
##  Mean   :2.349   Mean   :2.43   Mean   :2.176  
##  3rd Qu.:3.000   3rd Qu.:3.00   3rd Qu.:3.000  
##  Max.   :4.000   Max.   :4.00   Max.   :4.000  
##  NA's   :1       NA's   :1
```







Comparing the EDA results of full dataset with subset dataset, it seems thirteen artifacts are representative of the whole set of 91 artifacts.

Part C: ICC and agreement rate on subset data

```
subset_icc<-rep(0,7)
for(i in 7:13){
  model<-lmer(subset_rating[,i]~1+(1|Artifact),data=subset_rating)
  j=i-6
 subset_icc[j]<-unlist(icc(model)[[1]])
}
```

```
repeated <- subset_rating[subset_rating$Repeated==1,]
store_rate1<-as.data.frame(matrix(rep(0,n=3*7),nrow=7,ncol=3))
colnames(store_rate1)<-c("rate_1_and_2","rater_2_and_3","rater_1_and_3")
rownames(store_rate1)<-colnames(rating)[7:13]
for(i in 7:13){
k=i-6
raters_1_and_2_on_RsrchQ <-(
data.frame(r1=repeated[,i][repeated$Rater==1],
           r2=repeated[,i][repeated$Rater==2],
           a1=repeated$Artifact[repeated$Rater==1],
           a2=repeated$Artifact[repeated$Rater==2]
))
r1 <- factor(raters_1_and_2_on_RsrchQ$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_RsrchQ$r2,levels=1:4)
(t12 <- table(r1,r2))
store_rate1[k,1]<-sum(diag(t12))/sum(t12)


raters_2_and_3_on_RsrchQ <-(
data.frame(r1=repeated[,i][repeated$Rater==2],
```

```
           r2=repeated[,i][repeated$Rater==3],
           a1=repeated$Artifact[repeated$Rater==2],
           a2=repeated$Artifact[repeated$Rater==3]
))
r1 <- factor(raters_2_and_3_on_RsrchQ$r1,levels=1:4)
r2 <- factor(raters_2_and_3_on_RsrchQ$r2,levels=1:4)
(t23 <- table(r1,r2))
store_rate1[k,2]<-sum(diag(t23))/sum(t23)

raters_1_and_3_on_RsrchQ <-(
data.frame(r1=repeated[,i][repeated$Rater==1],
           r2=repeated[,i][repeated$Rater==3],
           a1=repeated$Artifact[repeated$Rater==1],
           a2=repeated$Artifact[repeated$Rater==3]
))
r1 <- factor(raters_1_and_3_on_RsrchQ$r1,levels=1:4)
r2 <- factor(raters_1_and_3_on_RsrchQ$r2,levels=1:4)
(t13 <- table(r1,r2))
store_rate1[k,3]<-sum(diag(t13))/sum(t13)
}

data.frame(store_rate1,subset_icc)
```

```
##           rate_1_and_2 rater_2_and_3 rater_1_and_3 subset_icc
## RsrchQ       0.3846154     0.5384615     0.7692308  0.1891892
## CritDes      0.5384615     0.6923077     0.6153846  0.5725594
## InitEDA      0.6923077     0.8461538     0.5384615  0.4929577
## SelMeth      0.9230769     0.6923077     0.6153846  0.5212766
## InterpRes    0.6153846     0.6153846     0.5384615  0.2295720
## VisOrg       0.5384615     0.7692308     0.7692308  0.5924529
## TxtOrg       0.6923077     0.5384615     0.6153846  0.1428571
```

Part D: ICC on full data

```
full_icc<-rep(0,7)
for(i in 7:13){
  model<-lmer(rating[,i]~1+(1|Artifact),data=rating)
  j=i-6
 full_icc[j]<-unlist(icc(model)[[1]])
}
```

We should redo the percent exact agreement calculations because the when select records that repeated is 1, we also selected the 13 Artifacts record. Therefore for this procedure, the subset dataset and the full dataset will have exact same agreement calculations.

```
data.frame(store_rate1,subset_icc,full_icc)
```

```
##           rate_1_and_2 rater_2_and_3 rater_1_and_3 subset_icc  full_icc
## RsrchQ       0.3846154     0.5384615     0.7692308  0.1891892 0.2096214
## CritDes      0.5384615     0.6923077     0.6153846  0.5725594 0.6730647
## InitEDA      0.6923077     0.8461538     0.5384615  0.4929577 0.6867210
## SelMeth      0.9230769     0.6923077     0.6153846  0.5212766 0.4719014
## InterpRes    0.6153846     0.6153846     0.5384615  0.2295720 0.2200285
## VisOrg       0.5384615     0.7692308     0.7692308  0.5924529 0.6607372
## TxtOrg       0.6923077     0.5384615     0.6153846  0.1428571 0.1879927
```

ICC is the correlation between any two rater's ratings on the same artifact. If the raters are consistent with one another in how they rate, we would expect this correlation to be higher. This between-raters correlation does tell us something useful about rater agreement: raters agree more when their correlations are higher.

The seven ICC's for the full data set agree with the seven ICC's for the subset corresponding to the 13 artifacts that all three raters saw.

For each rubric, the raters generally agree on their scores.

Part E: fit the best Rubric-specific models

```r
tall <- read.csv("/Users/ceciliaxia/Desktop/tall.csv",header=T)
ratings <-  read.csv("/Users/ceciliaxia/Desktop/ratings.csv",header=T)
tall$Rating <- factor(tall$Rating,levels=1:4)
for (i in unique(tall$Rubric)) {
  ratings[,i] <- factor(ratings[,i],levels=1:4)
}
tall$Sex[nchar(tall$Sex)==0] <- "--"

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

Rubric.names <- sort(unique(tall$Rubric))
tmp <- lmer(as.numeric(Rating) ~ -1 + as.factor(Rater) +
              Semester + Sex + (1|Artifact),
            data=tall.13[tall.13$Rubric=="RsrchQ",],REML=FALSE)
tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)
```

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

```r
formula(tmp.back_elim)
```

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

```r
tmp.int_only <- update(tmp.back_elim, . ~ . + 1 - as.factor(Rater))
anova(tmp.int_only,tmp.back_elim)
```

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

```r
anova(tmp.int_only,tmp.back_elim)$"Pr(>Chisq)"[2]
```

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

```r
Rubric.names <- sort(unique(tall$Rubric))
model.formula.13 <- as.list(rep(NA,7))
names(model.formula.13) <- Rubric.names
for (i in Rubric.names) {

  ## fit each base model
  rubric.data <- tall.13[tall.13$Rubric==i,]
  tmp <- lmer(as.numeric(Rating) ~ -1 + as.factor(Rater) +
              Semester + Sex + (1|Artifact),
            data=rubric.data,REML=FALSE)

  ## do backwards elimination
  tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)

  ## check to see if the raters are significantly different from one another
  tmp.single_intercept <- update(tmp.back_elim, . ~ . + 1 - as.factor(Rater))
  pval <- anova(tmp.single_intercept,tmp.back_elim)$"Pr(>Chisq)"[2]

  ## choose the best model
  if (pval<=0.05) {
    tmp_final <- tmp.back_elim
  } else {
    tmp_final <- tmp.single_intercept
  }

  ## and add to list...
  model.formula.13[[i]] <- formula(tmp_final)

}
```

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

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

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

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

## Note: Now the missing ratings become important.  We want to use the same data
## set for every model fit and model comparison.  I am going to eliminate by
## hand the two observations with missing data, and only do fitting and comparison
## on this "slightly" reduced data set.

tall[c(161,684),] ## just to check that these are the rows with missing ratings...
```

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

```
tall.nonmissing <- tall[-c(161,684),]  ## now delete them...

## I can't think of a good justification for imputing the "Sex" of the student who
## didn't report this to either M or F, and leaving it as "--" makes the models
## harder to interpret.  So I will eliminate that person from the data set also...
```

```
tall.nonmissing[tall.nonmissing$Sex=="--",]  ## check which rows will be eliminated
```

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

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

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

## There will be a lot of output from fitLMER.fnc() here... Sorry!

for (i in Rubric.names) {

  ## fit each base model
  rubric.data <- tall.nonmissing[tall.nonmissing$Rubric==i,]
  tmp <- lmer(as.numeric(Rating) ~ -1 + as.factor(Rater) +
              Semester + Sex + (1|Artifact),
            data=rubric.data,REML=FALSE)

  ## do backwards elimination
  tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)

  ## check to see if the raters are significantly different from one another
  tmp.single_intercept <- update(tmp.back_elim, . ~ . + 1 - as.factor(Rater))
  pval <- anova(tmp.single_intercept,tmp.back_elim)$"Pr(>Chisq)"[2]

  ## choose the best model
  if (pval<=0.05) {
    tmp_final <- tmp.back_elim
  } else {
    tmp_final <- tmp.single_intercept
  }

  ## and add to list...
  model.formula.alldata[[i]] <- formula(tmp_final)

}
```

```
## 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" = 0.608 >= 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 "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 "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" = 0.5041 >= 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)

## 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)
```

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

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

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

Part F: fit the best combined model

```
tall<-read.csv("/Users/ceciliaxia/Desktop/tall.csv")
rating<-read.csv("/Users/ceciliaxia/Desktop/ratings.csv")
subset_rating<-rating[grep("O",rating$Artifact,fixed=TRUE),]
subset_tall<-tall[grep("O",tall$Artifact,fixed=TRUE),]
```

```
lmer.1<-lmer(Rating~(0+Rubric|Artifact),data=tall)
lmer.2 <- update(lmer.1, . ~ . +Rubric)
```

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

```
anova(lmer.1,lmer.2)
```

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

## Data: tall
## Models:
## lmer.1: Rating ~ (0 + Rubric | Artifact)
## lmer.2: Rating ~ (0 + Rubric | Artifact) + Rubric
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.1   30 1537.2 1678.3 -738.58   1477.2
## lmer.2   36 1485.0 1654.4 -706.51   1413.0 64.134  6  6.481e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lmer.3 <- update(lmer.2, . ~ . + Semester)
anova(lmer.2,lmer.3)
```

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

## Data: tall
## Models:
## lmer.2: Rating ~ (0 + Rubric | Artifact) + Rubric
## lmer.3: Rating ~ (0 + Rubric | Artifact) + Rubric + Semester
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.2   36 1485.0 1654.4 -706.51   1413.0
## lmer.3   37 1483.1 1657.2 -704.57   1409.1 3.8888  1    0.04861 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lmer.4 <- update(lmer.3, . ~ . + Sex)
```

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

```
anova(lmer.3,lmer.4)
```

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

## Data: tall
## Models:
```

```
## lmer.3: Rating ~ (0 + Rubric | Artifact) + Rubric + Semester
## lmer.4: Rating ~ (0 + Rubric | Artifact) + Rubric + Semester + Sex
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.3   37 1483.1 1657.2 -704.57   1409.1
## lmer.4   39 1483.9 1667.4 -702.93   1405.9 3.2665  2     0.1953
```

```
lmer.5 <- update(lmer.3, . ~ . +Rater)
anova(lmer.3,lmer.5)
```

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

```
## Data: tall
## Models:
## lmer.3: Rating ~ (0 + Rubric | Artifact) + Rubric + Semester
## lmer.5: Rating ~ (0 + Rubric | Artifact) + Rubric + Semester + Rater
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.3   37 1483.1 1657.2 -704.57   1409.1
## lmer.5   38 1476.2 1655.0 -700.09   1400.2 8.9478  1   0.002778 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lmer.6 <- update(lmer.5, . ~ . +Repeated)
```

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

```
anova(lmer.3,lmer.6)
```

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

```
## Data: tall
## Models:
## lmer.3: Rating ~ (0 + Rubric | Artifact) + Rubric + Semester
## lmer.6: Rating ~ (0 + Rubric | Artifact) + Rubric + Semester + Rater + Repeated
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.3   37 1483.1 1657.2 -704.57   1409.1
## lmer.6   39 1477.6 1661.1 -699.81   1399.6 9.5169  2   0.008579 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
display(final_fixed<-lmer.6)
```

```
## lmer(formula = Rating ~ (0 + Rubric | Artifact) + Rubric + Semester +
##     Rater + Repeated, data = tall)
##                 coef.est coef.se
## (Intercept)      2.15     0.11
## RubricInitEDA    0.54     0.09
## RubricInterpRes  0.58     0.10
## RubricRsrchQ     0.46     0.09
## RubricSelMeth    0.16     0.09
## RubricTxtOrg     0.69     0.10
## RubricVisOrg     0.52     0.10
## SemesterS19     -0.19     0.09
## Rater           -0.08     0.03
## Repeated        -0.08     0.10
##
## Error terms:
##  Groups   Name            Std.Dev. Corr
##  Artifact RubricCritDes   0.76
```

```
##            RubricInitEDA   0.60      0.49
##            RubricInterpRes 0.42      0.27 0.76
##            RubricRsrchQ    0.42      0.61 0.46 0.72
##            RubricSelMeth   0.27      0.45 0.63 0.76 0.45
##            RubricTxtOrg    0.51      0.36 0.63 0.71 0.57 0.68
##            RubricVisOrg    0.52      0.38 0.75 0.70 0.54 0.45 0.77
##  Residual                  0.43
## ---
## number of obs: 817, groups: Artifact, 91
## AIC = 1515.6, DIC = 1361.7
## deviance = 1399.6
```

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

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

```r
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 = 0.0005783844
##  adding (Rater|Artifact) to model
## evaluating addition of (Semester|Artifact) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 0.9224995
##  not adding (Semester|Artifact) to model
## evaluating addition of (Sex|Artifact) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 0.4953412
```

```
##  not adding (Sex|Artifact) to model
## evaluating addition of (Repeated|Artifact) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 0.90132
##  not adding (Repeated|Artifact) to model
## ======================================================
## ===            re-backfitting fixed effects        ===
## ======================================================
## setting REML to FALSE

## boundary (singular) fit: see ?isSingular

## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.0694 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     BIC simple = 1659; BIC complex = 1658; decrease = 1 < 5
##     removing term
## resetting REML to TRUE

## boundary (singular) fit: see ?isSingular

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

Above we decide which random effect is significant and should be added to the model. The significant random effect is (Rater|Artifact).

```
model1<-lmer(Rating ~ 1 + Rater + Semester + Rubric+ Repeated + (0+Rater+Rubric|Artifact), data = tall)
```

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

```
model2<-lmer(Rating ~ 1 + Rater + Semester + Rubric+ Repeated + Rater*Semester+(0+Rater+Rubric|Artifact
```

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

```
anova(model1,model2)
```

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

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

## Warning in optwrap(optimizer, devfun, x@theta, lower = x@lower, calc.derivs =
## TRUE, : convergence code 1 from bobyqa: bobyqa -- maximum number of function
## evaluations exceeded

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

## Warning in optwrap(optimizer, devfun, x@theta, lower = x@lower, calc.derivs =
## TRUE, : convergence code 1 from bobyqa: bobyqa -- maximum number of function
## evaluations exceeded

## Data: tall
## Models:
## model1: Rating ~ 1 + Rater + Semester + Rubric + Repeated + (0 + Rater + Rubric | Artifact)
```

```
## model2: Rating ~ 1 + Rater + Semester + Rubric + Repeated + Rater * Semester + (0 + Rater + Rubric |
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model1   47 1469.5 1690.7 -687.77   1375.5
## model2   48 1471.0 1696.9 -687.52   1375.0 0.5047  1     0.4774
```

```
model3<-lmer(Rating ~ 1 + Rater + Semester + Rubric+ Repeated + Rubric*Semester+(0+Rater+Rubric|Artifac
```

```
## 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(model1,model3)
```

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

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

## Warning in optwrap(optimizer, devfun, x@theta, lower = x@lower, calc.derivs =
## TRUE, : convergence code 1 from bobyqa: bobyqa -- maximum number of function
## evaluations exceeded

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

## Data: tall
## Models:
## model1: Rating ~ 1 + Rater + Semester + Rubric + Repeated + (0 + Rater + Rubric | Artifact)
## model3: Rating ~ 1 + Rater + Semester + Rubric + Repeated + Rubric * Semester + (0 + Rater + Rubric
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model1   47 1469.5 1690.7 -687.77   1375.5
## model3   53 1470.7 1720.1 -682.37   1364.7 10.808  6    0.09451 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model4<-lmer(Rating ~ 1 + Rater + Semester + Rubric+ Repeated + Rubric*Repeated+(0+Rater+Rubric|Artifac
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00901118 (tol = 0.002, component 1)
```

```
anova(model1,model4)
```

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

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

## Warning in optwrap(optimizer, devfun, x@theta, lower = x@lower, calc.derivs =
## TRUE, : convergence code 1 from bobyqa: bobyqa -- maximum number of function
## evaluations exceeded

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

## Data: tall
## Models:
## model1: Rating ~ 1 + Rater + Semester + Rubric + Repeated + (0 + Rater + Rubric | Artifact)
## model4: Rating ~ 1 + Rater + Semester + Rubric + Repeated + Rubric * Repeated + (0 + Rater + Rubric
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
```

```
## model1   47 1469.5 1690.7 -687.77   1375.5
## model4   53 1477.1 1726.5 -685.54   1371.1 4.4525  6     0.6157
```

```r
model5<-lmer(Rating ~ 1 + Rater + Semester + Rubric+ Repeated + Rater*Repeated+(0+Rater+Rubric|Artifact
```

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

```r
anova(model1,model5)
```

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

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

## Warning in optwrap(optimizer, devfun, x@theta, lower = x@lower, calc.derivs =
## TRUE, : convergence code 1 from bobyqa: bobyqa -- maximum number of function
## evaluations exceeded

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

## Warning in optwrap(optimizer, devfun, x@theta, lower = x@lower, calc.derivs =
## TRUE, : convergence code 1 from bobyqa: bobyqa -- maximum number of function
## evaluations exceeded

## Data: tall
## Models:
## model1: Rating ~ 1 + Rater + Semester + Rubric + Repeated + (0 + Rater + Rubric | Artifact)
## model5: Rating ~ 1 + Rater + Semester + Rubric + Repeated + Rater * Repeated + (0 + Rater + Rubric |
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model1   47 1469.5 1690.7 -687.77   1375.5
## model5   48 1471.5 1697.3 -687.73   1375.5 0.0823  1     0.7742
```

```r
model6<-lmer(Rating ~ 1 + Rater + Semester + Rubric+ Repeated + Rater*Rubric+(0+Rater+Rubric|Artifact),
```

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

```r
anova(model1,model6)
```

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

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

## Warning in commonArgs(par, fn, control, environment()): convergence code 1 from
## bobyqa: bobyqa -- maximum number of function evaluations exceeded

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

## Data: tall
## Models:
## model1: Rating ~ 1 + Rater + Semester + Rubric + Repeated + (0 + Rater + Rubric | Artifact)
## model6: Rating ~ 1 + Rater + Semester + Rubric + Repeated + Rater * Rubric + (0 + Rater + Rubric | A
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model1   47 1469.5 1690.7 -687.77   1375.5
## model6   53 1461.3 1710.7 -677.66   1355.3 20.214  6   0.002537 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model7<-lmer(Rating ~ 1 + Rater + Semester + Rubric+ Repeated + Semester*Repeated+(0+Rater+Rubric|Artifa
```

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

```r
anova(model1,model7)
```

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

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

## Warning in optwrap(optimizer, devfun, x@theta, lower = x@lower, calc.derivs =
## TRUE, : convergence code 1 from bobyqa: bobyqa -- maximum number of function
## evaluations exceeded

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

## Warning in optwrap(optimizer, devfun, x@theta, lower = x@lower, calc.derivs =
## TRUE, : convergence code 1 from bobyqa: bobyqa -- maximum number of function
## evaluations exceeded

## Data: tall
## Models:
## model1: Rating ~ 1 + Rater + Semester + Rubric + Repeated + (0 + Rater + Rubric | Artifact)
## model7: Rating ~ 1 + Rater + Semester + Rubric + Repeated + Semester * Repeated + (0 + Rater + Rubri
##        npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## model1   47 1469.5 1690.7 -687.77   1375.5
## model7   48 1471.8 1697.7 -687.92   1375.8     0  1          1
```

The final model is:

```r
final_11<-lmer(Rating ~ 1 + Rater + Semester + Rubric + Repeated + Rater * Rubric + (0 + Rater + Rubric
```

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

```r
display(final_11)
```

```
## lmer(formula = Rating ~ 1 + Rater + Semester + Rubric + Repeated +
##     Rater * Rubric + (0 + Rater + Rubric | Artifact), data = tall)
##                        coef.est coef.se
## (Intercept)             1.80     0.17
## Rater                   0.08     0.07
## SemesterS19            -0.13     0.08
## RubricInitEDA           0.83     0.19
## RubricInterpRes         1.30     0.19
## RubricRsrchQ            0.81     0.18
## RubricSelMeth           0.51     0.19
## RubricTxtOrg            1.15     0.19
## RubricVisOrg            0.84     0.19
## Repeated               -0.07     0.09
## Rater:RubricInitEDA    -0.15     0.08
## Rater:RubricInterpRes  -0.36     0.08
## Rater:RubricRsrchQ     -0.18     0.08
## Rater:RubricSelMeth    -0.18     0.08
## Rater:RubricTxtOrg     -0.23     0.08
## Rater:RubricVisOrg     -0.16     0.08
##
## Error terms:
```

```
##  Groups   Name             Std.Dev. Corr
##  Artifact Rater            0.17
##           RubricCritDes    0.78     -0.36
##           RubricInitEDA    0.54     -0.13  0.47
##           RubricInterpRes  0.34     -0.37  0.44  0.73
##           RubricRsrchQ     0.53     -0.65  0.67  0.41  0.81
##           RubricSelMeth    0.15     -0.53  0.71  0.36  0.62  0.70
##           RubricTxtOrg     0.41     -0.10  0.39  0.45  0.40  0.51  0.07
##           RubricVisOrg     0.48     -0.24  0.41  0.62  0.58  0.52  0.03  0.59
##  Residual                  0.41
```

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

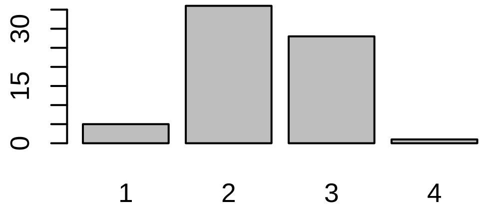
```
## ---
## number of obs: 817, groups: Artifact, 91
## AIC = 1521.9, DIC = 1294.7
## deviance = 1355.3
```

Part G: interesting things about the data

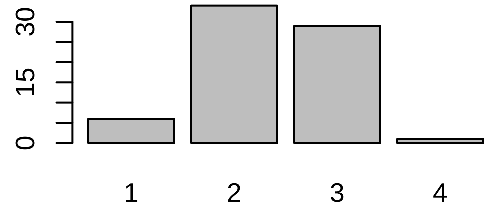
```
par(mfrow=c(2,2))
dat1<-subset_tall[which(subset_tall$Semester=="F19"),]
barplot(table(dat1[which(dat1$Rater==1),]$Rating),main="Distribution of Ratings of Rater 1 Fall")
barplot(table(dat1[which(dat1$Rater==2),]$Rating),main="Distribution of Ratings of Rater 2 Fall")
barplot(table(dat1[which(dat1$Rater==3),]$Rating),main="Distribution of Ratings of Rater 3 Fall")
tmp1<-data.frame(r1=dat1[which(dat1$Rater==1),]$Rating,r2=dat1[which(dat1$Rater==2),]$Rating,r3=dat1[wh
summary(tmp1)
```

```
##        r1              r2              r3
##  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 :2.000
##  Mean   :2.357   Mean   :2.357   Mean   :2.229
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :4.000   Max.   :4.000   Max.   :3.000
```

Distribution of Ratings of Rater 1 Fall



Distribution of Ratings of Rater 2 Fall



Distribution of Ratings of Rater 3 Fall