Exploring ratings on undergraduates' project papers in a "General Education" program across different raters and different rubrics

Zixuan Jin
Department of Statistics and Data Science
Carnegie Mellon University
zixuanji@andrew.cmu.edu

Abstract

Dietrich College at Carnegie Mellon University has been experimenting with rating work in Freshman Statistics on different rubrics. We interested in analyzing the relationship among ratings, rubrics, raters, and some other factors related to this experiment. The experiment provided us with 91 project papers randomly sampled from Fall and Spring sections for Freshman Statistics and some personal characteristic of students writing these papers. We created bar plots and percentage tables to see the relationship among ratings, rubrics, and raters. We also calculated exact agreement rates between any raters on rating different rubrics. Then, we used AIC and BIC values to determine a final model predicting ratings using all other variables related to this experiment. Moreover, we used the same criterion to choose a final model predicting raters using all other variables. We found that the artifacts are not all of equal quality on each rubric and raters are not interpreting the evidence in the artifacts in the same way. These ratings might not correctly reflect the quality of this course. For future studies, the program might need to include more raters and change grading system to make raters more consistent in grading project papers.

1 Introduction

Dietrich College at Carnegie Mellon University is in the process of implementing a new "General Education" program for undergraduates. This program specifies a set of courses and experiences that all undergraduates must take. To determine whether the new program is successful, the college hopes to rate student work performed in each of the "General Education" courses each year. Recently the college has been experimenting with rating work in Freshman Statistics on different rubrics, using raters from across the college (Project 02 Description). In this report, we will investigate the relationship among ratings, rubrics, raters, and some other variables related to the experiment. Specifically, we will

1. Illustrate the relationship among ratings, rubrics, and raters
2. Identify the agreement or disagreement among raters for each rubric
3. Develop a regression model to predict scores from other factors in this experiment
4. Develop a regression model to predict raters from other factors in this experiment

2 Data

The data provides ratings on 91 project papers (referred to as "artifacts" in the following) randomly sampled from Fall and Spring sections of Freshman Statistics separately. Three raters from three different departments were asked to rate these artifacts on seven rubrics, as shown in

Table 1. The rating scale for all rubrics is shown in Table 2. 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.

Table 1: Rubrics for rating Freshman Statistics projects

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 2: Rating scale used for all rubrics

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

The data are given in the file ratings.csv in the project 02 folder for our course on Canvas. A detailed description of the variables in the dataset are shown in Table 3.

Table 3: Variable definitions for data in the file ratings.csv

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

The X variable here is the same as the row number in our data, and we did not include it in our data analyses. Since the Sample and Overlap variables are just identifiers for those artifacts, we did not include these two variables in our data analyses.

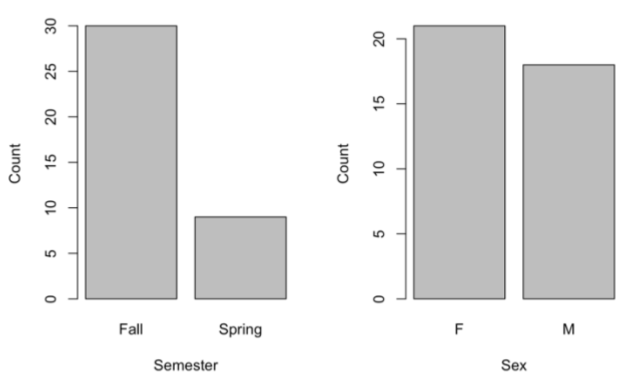
The file tall.csv provided in the project 02 folder contains the same data in the file ratings.csv but organized so that each row contains just one rating, in the column labelled Rating, and the rubric for that rating is listed in the column labelled Rubric. We used files ratings.csv and tall.csv interchangeably for our research questions. The summary statistics of ratings for seven rubrics for those 13 artifacts seen by all raters are given Table 4 below.

Table 4: Summary statistics for ratings for seven rubrics for those 13 artifacts

Variables	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
RsrchQ	1	2	2	2.282	3	3	0.560
CritDes	1	1	2	1.718	2	3	0.724
InitEDA	1	2	2	2.385	3	3	0.544
SelMeth	1	2	2	2.051	2	3	0.510
InterpRes	1	2	3	2.513	3	4	0.601
VisOrg	1	2	2	2.282	3	3	0.605
TxtOrg	1	2	3	2.667	3	4	0.621

Looking at the summary statistics above, we can tell that the mean rating of these rubrics is around 2. For rating on critique design, the mean is less than median, which might suggest that the distribution is right skewed. For rating on text organization and interpreting results, the mean is greater than median, which might suggest that the distributions are slightly skewed to the left. We will explore the distributions in the following section. However, there are no rubrics with a mean substantially smaller or larger than the median. The summary statistics for the dummy variables in the data are given in the Figure 1 below.

Figure 1: Histograms of dummy variables Semester and Sex for those 13 artifacts



In Figure 1, we see that most of artifacts are from Fall semester. Considering the sex or gender of students who created the artifact, we see similar number of female (sex 'F') students and male (sex 'M') students in our data for those 13 artifacts seen by all three raters.

3 Methods

3.1 The relationship among ratings, rubrics, and raters

To see if there is any rubric that tends to get especially high or low ratings and if there is any rater that tends to give especially high or low ratings, we provided bar plots and some numeric summaries such as counts, percentages, means, and SDs of the ratings of each rubric. We first analyzed the relationship among ratings, rubrics, and raters for those 13 artifacts seen by all three raters and then for all the 91 artifacts.

3.2 The agreement or disagreement among raters for each rubric

To see if raters agree on ratings for each rubric, we fitted a random-intercept model for each rubric (See equations 1 and 2 below) and calculated ICC as a measure of rater agreement which is going to be the correlation between ratings on any two different artifacts by the same rater (See equation 3 below).

$$y_i = \alpha_{j[i]} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2) \qquad (1)$$

$$\alpha_j = \beta_0 + \eta_j, \eta_j \sim N(0, \tau^2) \qquad (2)$$

$$ICC = \frac{\tau^2}{\tau^2 + \sigma^2} \qquad (3)$$

For our data, we ran the following model (See equation 4 below) for each rubric in R's model formula language to calculate ICC. A high value of ICC means a high correlation between any two raters, which implies a high agreement between any two raters for a specific rubric.

$$Rating \sim 1 + (1 \mid Artifact) \qquad (4)$$

To compute exact agreement rates between raters, we also constructed tables of counts cross classifying the ratings that each pair of raters gives. A high value directly indicates a high agreement between any two raters for a specific rubric. We calculated ICC and exact agreement rates for those 13 artifacts seen by all raters and for all artifacts separately.

3.3 Develop a regression model to predict scores from other factors in this experiment

3.3.1 Develop a regression model for each rubric

To explore how various factors in this experiment (Rater, Semester, Sex, Repeated, and Rubric) are related to ratings, we first added these factors as fixed effects to the equation (4) above for each rubric using those 13 artifacts seen by all three raters. For each rubric, we added one factor each time and used AIC, and BIC to determine a final model for each rubric. We re-ran this process for each rubric using all the data, considering any interaction and new random effects.

3.3.2 Develop a regression model examining interactions with rubrics

The models in 3.3.1 consider only one Rubric at a time, which did not allow us to examine the interactions with rubric directly. Then, we added fixed effects, interactions, and new random effects to the "combined" model (See equation 5) using all the data as follows:

$$Rating \sim 1 + (0 + Rubric \mid Artifact) \qquad (5)$$

We chose our final model based on F-tests, AIC, and BIC values.

3.4 Develop a regression model to predict raters from other factors in this experiment

To explore how various factors in this experiment (Semester, Rubric, Rating, and Sex) are related to raters, we developed a multi-nominal regression model as we have three raters in this experiment. Then, we chose our final model based on AIC values.

4 Results

4.1 The relationship among ratings, rubrics, and raters

4.1.1 The relationship among ratings, rubrics, and raters for 13 artifacts seen by all raters

Looking at Figure 2 and Table 5 below, we notice that most ratings are at 2 for most of the rubrics. Given the bar plots and percentage tables for rating on critique design, we see most counts are of ratings 1 and especially high percentages of ratings 1compared to other rubrics. Given the bar plots and percentages tables for rating on text organization, we see most counts are of ratings 3 and especially high percentages of ratings 3 compared to other rubrics. In this case, we believe rating on critique design tends to get especially low ratings and rating on text organization tends to get especially high ratings.

Figure 2: Bar plots of Ratings for each rubric for 13 artifacts seen by all raters

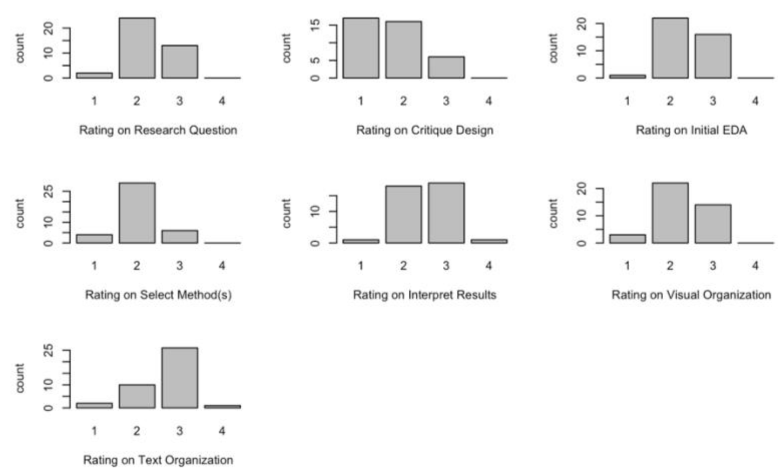
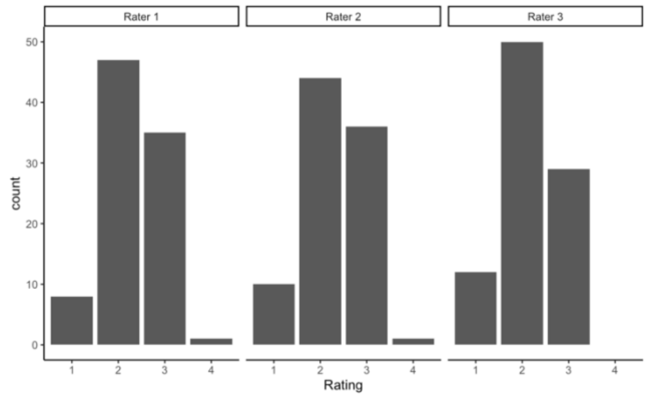


Table 5: Percentages of each rating score for each rubric for 13 artifacts seen by all raters

Rubrics	Rating = 1	Rating = 2	Rating = 3	Rating = 4
Rating on Research Question	5	62	33	0
Rating on Critique Design	44	41	15	0
Rating on Initial EDA	3	56	41	0
Rating on Select Method(s)	10	74	15	0
Rating on Interpret Results	3	46	49	3
Rating on Visual Organization	8	56	36	0
Rating on Text Organization	5	26	67	3

Given Figure 3 below, we can tell the patterns of ratings across three raters are comparable to each other. In this case, we believe that there is no rater giving especially high or low ratings for those 13 artifacts.

Figure 3: Distributions of ratings across three raters for 13 artifacts seen by all three raters



4.1.2 The relationship among ratings, rubrics, and raters for all artifacts

Looking at Figures 2 and 4, we notice that distributions of ratings for each rubric of those 13 artifacts seen by all three raters are indistinguishable from those distributions of ratings for each rubric of all artifacts. But the distribution of ratings on interpreting results for all artifacts is slightly different from that for those 13 artifacts. For all artifacts, the counts of rating 3 are higher for those 13 artifacts. Given Table 6, even we notice that there are some null values, we still find an especially high percentage of rating 1 for rating on critique design. However, we find that high percentages of rating 3 for rating on interpreting results and in text organization.

Figure 4: Bar plots of Ratings for each rubric for all artifacts

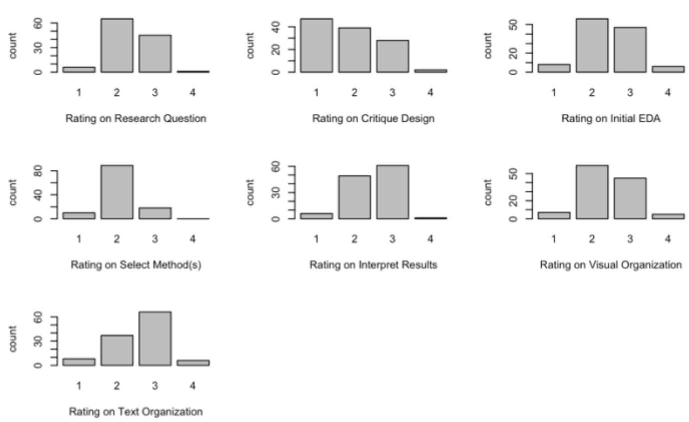
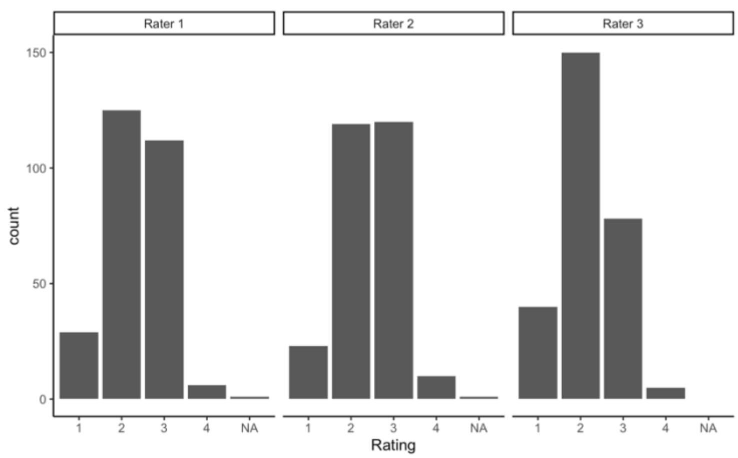


Table 6: Percentages of each rating score for each rubric for all artifacts

Rubrics	Rating = 1	Rating = 2	Rating = 3	Rating = 4	NA
Rating on Research Question	5	56	38	1	0
Rating on Critique Design	40	33	24	2	1
Rating on Initial EDA	7	48	40	5	0
Rating on Select Method(s)	9	76	15	0	0
Rating on Interpret Results	5	42	52	1	0
Rating on Visual Organization	6	50	38	4	1
Rating on Text Organization	7	32	56	5	0

Given Figures 3 and 5, we find the distributions of ratings for all artifacts are comparable for those artifacts seen by all raters. However, we notice that Rater 3 tends to give rating 2 more frequently compared to other raters. In general, we can tell that the ratings for those 13 artifacts are representative of the whole set of 91 artifacts.

Figure 5: Distributions of ratings across three raters for all artifacts



4.2 The agreement or disagreement among raters for each rubric

Given Table 7, we notice that ICC values for all artifacts are comparable to ICC values for those 13 artifacts. We do not find an especially high ICC value, which indicates no high correlation between any raters for each rubric. Given low ICC values for rating on research question, we believe there is low correlation between any two raters. Given the exact agreement rate of 0.85 between Rater 2 and Rater 3, we believe Rater 2 pretty much agree with Rater 3 for rating on initial EDA (exploratory data analysis). Looking at the exact agreement rate of 0.92 for rating on selecting method(s), we find that Rater 1 and Rater 2 agree with each other for the most time. Given exact agreement rates of 0.77, we believe Rater 1 and Rater 3 agree with each other for rating on research question and on visual organization. Moreover, we believe Rater 2 and Rater 3 agree with each other for rating on visual organization. In this case, we can tell raters agree on their ratings for rating on visual organization.

Table 7: ICC values and exact agreement rates between any two raters for each rubric

Rubrics	ICC (all artifacts)	ICC (13 artifacts)	a12	a23	a13
Rating on Research Question	0.21	0.19	0.38	0.54	0.77
Rating on Critique Design	0.67	0.57	0.54	0.69	0.62
Rating on Initial EDA	0.69	0.49	0.69	0.85	0.54
Rating on Select Method(s)	0.47	0.52	0.92	0.69	0.62
Rating on Interpret Results	0.22	0.23	0.62	0.62	0.54
Rating on Visual Organization	0.66	0.59	0.54	0.77	0.77
Rating on Text Organization	0.19	0.14	0.69	0.54	0.62

4.3 Develop a regression model to predict scores from other factors in this experiment

4.3.1 Develop a regression model for each rubric for 13 artifacts

Given AIC and BIC values, we do not need to add any fixed effect or interaction for predicting ratings on research question, initial EDA, and text organization (Appendix C, see pages 52 to 65). The model is the same as the equation 4 above with ratings treated as numeric values. By the same criterion, we should include Rater as a fixed effect in predicting ratings on critique design, interpreting results, and visual organization (Appendix C, see pages 52 to 65). Depending on AIC and BIC values, we need to include Sex and Semester in predicting ratings on selecting method(s) (Appendix C, see pages 58 to 61).
Then, we ran the same variable selection process as in 4.3.1 but with all the data available in this experiment this time and null values ignored. We ended up with an exact same model for each rubric as in 4.3.1 (Appendix C, see pages 65 to 71).

4.3.2 Develop a regression model examining interactions with rubric

Now we tried to add fixed effects, interactions, and new random effects to the combined model (see equation 5 above in 3.3.2) including interaction with rubric. Without considering new random effects, we used AIC and BIC values to choose our final model as follows:

$$Rating \sim Rater + Semester + Rubric + Rater \times Rubric + (0 + Rubric \mid Artifact) \quad (6)$$

Table 8 gives the full table of estimated coefficients and standard errors for the fixed effects for model (6).

Table 8: Estimated coefficients and standard errors for model (6)

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	1.73481	0.16228	10.690
Rater	0.10849	0.06712	1.616
SemesterS19	-0.17592	0.08328	-2.112
RubricInitEDA	0.84228	0.19681	4.280
RubricInterpRes	1.34106	0.19440	6.898
RubricRsrchQ	0.83297	0.18274	4.558
RubricSelMeth	0.57349	0.18852	3.042
RubricTxtOrg	1.18186	0.19763	5.980
RubricVisOrg	0.86947	0.19908	4.367
Rater:RubricInitEDA	-0.14832	0.08671	-1.711
Rater:RubricInterpRes	-0.37753	0.08559	-4.411
Rater:RubricRsrchQ	-0.18759	0.08146	-2.303
Rater:RubricSelMeth	-0.20764	0.08292	-2.504
Rater:RubricTxtOrg	-0.24519	0.08697	-2.819
Rater:RubricVisOrg	-0.17036	0.08740	-1.949

Given the facet graphs in Appendix C (see page 81 below), it does look as if the 3 raters have different ways of scoring the 7 rubrics, so the interaction we found in model 6 makes sense. Looking at the summary table above, we can tell Rater 1's ratings on interpreting results or text organization are 1 point greater than the ratings on critique design. For each rubric, we cannot tell the differences in ratings across raters. For rating on critique design, rater 2 generally gives 0.37 point higher than rater 1 does on average and rater 3 generally gives 0.21 point higher than rater 1 does. For rating on critique design, rater 3 generally gives 0.17 point lower than rater 2 does on average. For rating on initial EDA, rater 2 gives 0.06 point higher than rater 1 does on average and rater 3 gives 0.09 point less than rater 1 does on average. For rating on initial EDA, rater 3 gives 0.15 point higher than rater 2 does on average. For rating on interpreting results, rater 2 gives 0.17 point less than rater 1 does and rater 3 gives 0.54 point lower than rater 1 does on average. For rating on interpreting results, rater 2 generally gives 0.37 point higher than rater 3 does on average. For rating on research question, rater 2 gives 0.13 point less than rater 2 does and rater 3 gives 0.16 point less than rater 1 does on average. For rating on interpreting results, rater 2 generally gives 0.03 point higher than rater 2 does on average. For rating on selecting methods, rater 2 gives 0.03 point less than rater 1 does and rater 3 gives 0.2 point less than rater 1 does on average. For rating on selecting methods, rater 3 gives 0.17 point less than rater 2 does on average. For rating on text organization, rater 2 gives 0.21 point less than rater 1 does and rater 3 gives 0.28 point less than rater 1 does on average. For rating on text organization, rater 3 gives 0.07 point less than rater 2 does on average. For rating on visual organization, rater 2 gives 0.24 point less than rater 1 does and rater 3 gives 0.12 point less than rater 1 does on average. For rating on visual organization, rater 3 gives 0.12 point higher than rater 1 does on average. Moreover, ratings in Spring 2019 are 0.18 point less than ratings in fall on average, holding others constant.

Considering adding new random effects to the model (6) above, we used AIC and BIC values to choose our final model as follows:

$$Rating \sim Rater + Semester + Rubric + Rater \times Rubric + (0 + Rubric \mid Artifact) + (0 + Rater \mid Artifact) \qquad (7)$$

The model (7) includes an interaction between rater and artifact. This model allows that 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. As we mentioned above, the three raters have different ways of scoring the seven rubrics. The new random effect of rubric shows that 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. In all of this, the artifacts are not all of equal quality on each rubric, and so we should expect the average scores on each rubric to vary from one artifact to the next. The interaction between rater and rubric suggests that the raters are not all interpreting the rubrics in the same way. The interaction between raters and artifact suggests that the raters are not interpreting the evidence in the artifacts in the same way. These interactions suggest that perhaps the raters should be trained more, to make the raters' ratings more like each other.

4.4 Develop a regression model to predict raters from other factors in this experiment

We started with fitting the following full model as follows:

$$Rater \sim Semester + Rubric + Rating + Sex \qquad (8)$$

We treated Rating as a categorical in this model (8) above. Then we use AIC to choose our final mode (9) shown below.

$$Rater \sim Rating + Sex \qquad (9)$$

Table 9 gives the full table of estimated coefficients and standard errors for model (9).

Table 9: Estimated coefficients and standard errors for model (9)

	Dependent variable:	
	2	3
	(1)	(2)
Rating2	0.184	-0.138
	(0.307)	(0.273)
Rating3	0.302	-0.776***
	(0.309)	(0.288)
Rating4	0.738	-0.509
	(0.588)	(0.654)
SexF	3.688	-12.840
	(4.227)	(172.650)
SexM	3.896	-12.590
	(4.227)	(172.650)
Constant	-4.012	13.050
	(4.230)	(172.650)
Akaike Inf. Crit.	1,776.150	1,776.150
Note:	*p<0.1; **p<0.05; ***p<0.01	

Keeping all other variables constant, if the rating increases one unit, the rating is 0.669 times more likely to come from Rater 3 as compared to the Rater 1. The coefficient is statistically significant at 1% significance level. Keeping all other variables constant, if the rating increases one unit, the rating is 1.184 times more likely to come from Rater 2 as compared to the Rater 1. However, the coefficient is not statistically significant.

5 Discussion

Looking at bar plots and percentage tables, we found that rating on critique design tends to get especially low ratings and ratings on text organization tends to get especially high ratings. Moreover, we found that raters do not agree with each other on ratings seven rubrics. Given the section 4.3, we notice that the artifacts are not all of equal quality on each rubric and raters are not interpreting the evidence in the artifacts in the same way. Given the section 4.4, it seems that Rater 3 are more likely to give high ratings. In all of this, raters are not consistent in ratings all the artifacts. Hence the ratings might not correctly reflect the results of this program or this experiment, which might lead to an incorrect decision of implementing certain undergraduate course. Maybe raters should have rubrics on grading each topic in a paper to be more consistent in their ratings.

One of the limitations for this experiment is that we have only three raters. Three raters' ratings might not be representative, and they cannot truly reflect how students perform on those seven rubrics. In this case, it is better to include more raters in this experiment. Another limitation could the range of the rating. Currently, the ratings take 1, 2, 3, and 4 levels. It might be hard for raters to give ratings if they find an artifact is not worth of score 3 but is better than score 2. It would be better if they change their grading system.

For future studies, we could include more raters to eliminate the limitation mentioned above. In addition, the program manager needs to generate detailed guidelines of grading these rubrics above. For example, the program needs to list out where raters should deduct their points to make raters more consistent in grading and reflect the true ratings of the program.

References

Project 02 Description, *project-02.pdf*, available on 36617 course canvas page

Technical Appendices

Read the data

```
tall <- read.csv("/Users/rosahhh/Desktop/Fall 2021/36617 Applied Linear Model
/project02/tall.csv",
                 header = TRUE)
ratings <- read.csv("/Users/rosahhh/Desktop/Fall 2021/36617 Applied Linear Mo
del/project02/ratings.csv",
                    header = TRUE)
```

Appendix A

Make a subset of the data for just the 13 artifacts seen by all three raters and some EDA

```
ratings_sub <- ratings %>%
  filter(ratings$Repeated == 1)

summary(ratings_sub)

##        X              Rater       Sample           Overlap     Semester

##  Min.   :  1.00   Min.   :1   Min.   :  1.00   Min.   : 1   Length:39

##  1st Qu.: 23.50   1st Qu.:1   1st Qu.: 24.50   1st Qu.: 4   Class :charact
er
##  Median : 51.00   Median :2   Median : 52.00   Median : 7   Mode  :charact
er
##  Mean   : 53.46   Mean   :2   Mean   : 54.28   Mean   : 7

##  3rd Qu.: 81.50   3rd Qu.:3   3rd Qu.: 82.50   3rd Qu.:10

##  Max.   :109.00   Max.   :3   Max.   :110.00   Max.   :13

##      Sex                RsrchQ          CritDes          InitEDA
##  Length:39          Min.   :1.000   Min.   :1.000   Min.   :1.000
##  Class :character   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000
##  Mode  :character   Median :2.000   Median :2.000   Median :2.000
##                     Mean   :2.282   Mean   :1.718   Mean   :2.385
##                     3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000
##                     Max.   :3.000   Max.   :3.000   Max.   :3.000
##     SelMeth         InterpRes          VisOrg           TxtOrg
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :3.000   Median :2.000   Median :3.000
##  Mean   :2.051   Mean   :2.513   Mean   :2.282   Mean   :2.667
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :3.000   Max.   :4.000   Max.   :3.000   Max.   :4.000
##    Artifact            Repeated
##  Length:39          Min.   :1
##  Class :character   1st Qu.:1
```

```
##   Mode  :character   Median :1
##                      Mean   :1
##                      3rd Qu.:1
##                      Max.   :1
```

```
round(sd(ratings_sub$RsrchQ), 3)
```

```
## [1] 0.56
```

```
round(sd(ratings_sub$CritDes), 3)
```

```
## [1] 0.724
```

```
round(sd(ratings_sub$InitEDA), 3)
```

```
## [1] 0.544
```

```
round(sd(ratings_sub$SelMeth), 3)
```

```
## [1] 0.51
```

```
round(sd(ratings_sub$InterpRes), 3)
```

```
## [1] 0.601
```

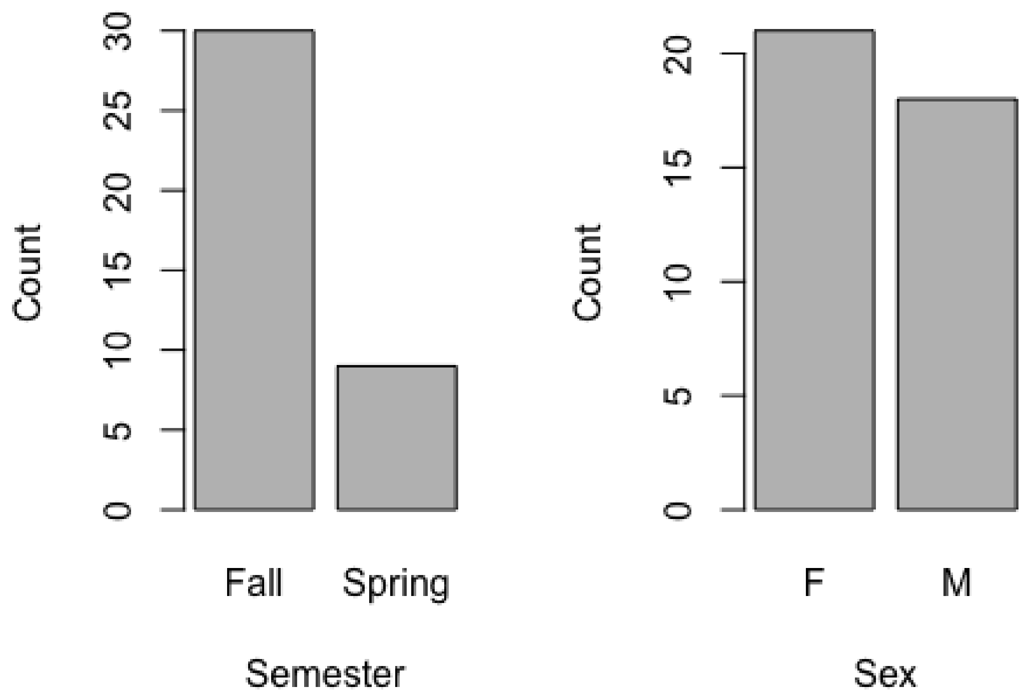
```
round(sd(ratings_sub$VisOrg), 3)
```

```
## [1] 0.605
```

```
round(sd(ratings_sub$TxtOrg), 3)
```

```
## [1] 0.621
```

```
par(mfrow=c(1,2))
ratings_sub$Semester <- as.factor(ratings_sub$Semester)
plot(ratings_sub$Semester, xlab="Semester", ylab="Count")
ratings_sub$Sex <- as.factor(ratings_sub$Sex)
plot(ratings_sub$Sex, xlab="Sex", ylab="Count")
```

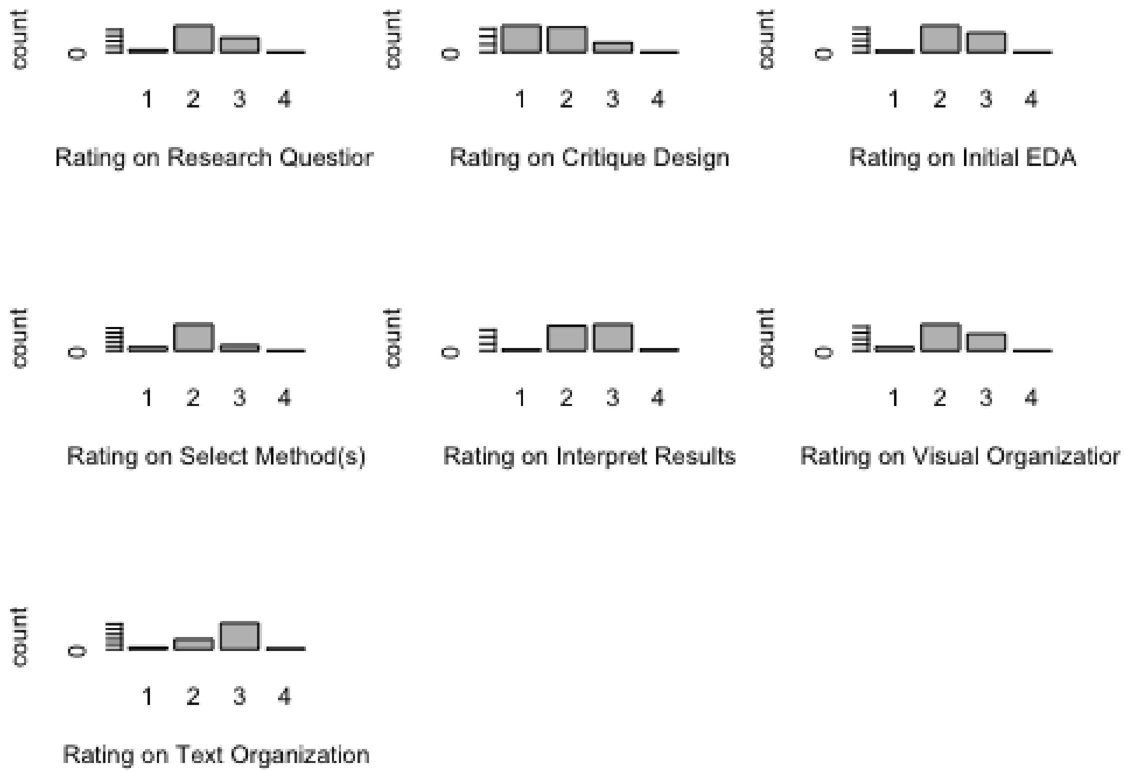


Make these rubric as categorical variables for those 13 artifacts

```
ratings_sub$RsrchQ <- factor(ratings_sub$RsrchQ, levels=1:4)
ratings_sub$CritDes <- factor(ratings_sub$CritDes, levels=1:4)
ratings_sub$InitEDA <- factor(ratings_sub$InitEDA, levels=1:4)
ratings_sub$SelMeth <- factor(ratings_sub$SelMeth, levels=1:4)
ratings_sub$InterpRes <- factor(ratings_sub$InterpRes, levels=1:4)
ratings_sub$VisOrg <- factor(ratings_sub$VisOrg, levels=1:4)
ratings_sub$TxtOrg <- factor(ratings_sub$TxtOrg, levels=1:4)
```

The distributions of ratings of each rubric

```
par(mfrow=c(3,3))
plot(ratings_sub$RsrchQ, xlab = "Rating on Research Question", ylab = "count"
)
plot(ratings_sub$CritDes, xlab = "Rating on Critique Design", ylab = "count")
plot(ratings_sub$InitEDA, xlab = "Rating on Initial EDA", ylab = "count")
plot(ratings_sub$SelMeth, xlab = "Rating on Select Method(s)", ylab = "count"
)
plot(ratings_sub$InterpRes, xlab = "Rating on Interpret Results", ylab = "cou
nt")
plot(ratings_sub$VisOrg, xlab = "Rating on Visual Organization", ylab = "coun
t")
plot(ratings_sub$TxtOrg, xlab = "Rating on Text Organization", ylab = "count"
)
```



Distributions of Ratings on each rubric

Calculating the percentage of ratings given each rubric

Frequency table and proportion table for each rubric of those 13 artifacts

```
RsrchQ <- table(ratings_sub$RsrchQ)
addmargins(RsrchQ)

##
##   1   2   3   4 Sum
##   2  24  13   0  39

round(prop.table(RsrchQ)*100,digits=0)

##
##  1  2  3  4
##  5 62 33  0

CritDes <- table(ratings_sub$CritDes)
addmargins(CritDes)

##
##   1   2   3   4 Sum
##  17  16   6   0  39
```

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

##
##  1  2  3  4
## 44 41 15  0

InitEDA <- table(ratings_sub$InitEDA)
addmargins(InitEDA)

##
##   1   2   3   4 Sum
##   1  22  16   0  39

round(prop.table(InitEDA)*100,digits=0)

##
##  1  2  3  4
##  3 56 41  0

SelMeth <- table(ratings_sub$SelMeth)
addmargins(SelMeth)

##
##   1   2   3   4 Sum
##   4  29   6   0  39

round(prop.table(SelMeth)*100,digits=0)

##
##  1  2  3  4
## 10 74 15  0

InterpRes <- table(ratings_sub$InterpRes)
addmargins(InterpRes)

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

round(prop.table(InterpRes)*100,digits=0)

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

VisOrg<- table(ratings_sub$VisOrg)
addmargins(VisOrg)

##
##   1   2   3   4 Sum
##   3  22  14   0  39

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

```
##
##   1  2  3  4
##   8 56 36  0
```

```r
TxtOrg <- table(ratings_sub$TxtOrg)
addmargins(TxtOrg)
```

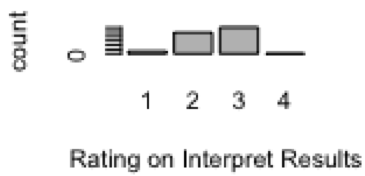
```
##
##    1   2   3   4 Sum
##    2  10  26   1  39
```

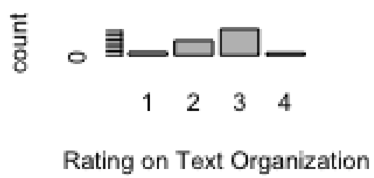
```r
round(prop.table(TxtOrg)*100,digits=0)
```

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

Working on the full data

```r
ratings$RsrchQ <- factor(ratings$RsrchQ, levels=1:4)
ratings$CritDes <- factor(ratings$CritDes, levels=1:4)
ratings$InitEDA <- factor(ratings$InitEDA, levels=1:4)
ratings$SelMeth <- factor(ratings$SelMeth, levels=1:4)
ratings$InterpRes <- factor(ratings$InterpRes, levels=1:4)
ratings$VisOrg <- factor(ratings$VisOrg, levels=1:4)
ratings$TxtOrg <- factor(ratings$TxtOrg, levels=1:4)

par(mfrow=c(3,3))
plot(ratings$RsrchQ, xlab = "Rating on Research Question", ylab = "count")
plot(ratings$CritDes, xlab = "Rating on Critique Design", ylab = "count")
plot(ratings$InitEDA, xlab = "Rating on Initial EDA", ylab = "count")
plot(ratings$SelMeth, xlab = "Rating on Select Method(s)", ylab = "count")
plot(ratings$InterpRes, xlab = "Rating on Interpret Results", ylab = "count")
plot(ratings$VisOrg, xlab = "Rating on Visual Organization", ylab = "count")
plot(ratings$TxtOrg, xlab = "Rating on Text Organization", ylab = "count")
```






```
RsrchQ_full <- table(ratings$RsrchQ, useNA = "always")
addmargins(RsrchQ_full)
```

```
##
##    1    2    3    4 <NA>  Sum
##    6   65   45    1    0  117
```

```
round(prop.table(RsrchQ_full)*100,digits=0)
```

```
##
##    1    2    3    4 <NA>
##    5   56   38    1    0
```

```
CritDes_full <- table(ratings$CritDes, useNA = "always")
addmargins(CritDes_full)
```

```
##
##    1    2    3    4 <NA>  Sum
##   47   39   28    2    1  117
```

```
round(prop.table(CritDes_full)*100,digits=0)
```

```
##
##    1    2    3    4 <NA>
##   40   33   24    2    1
```

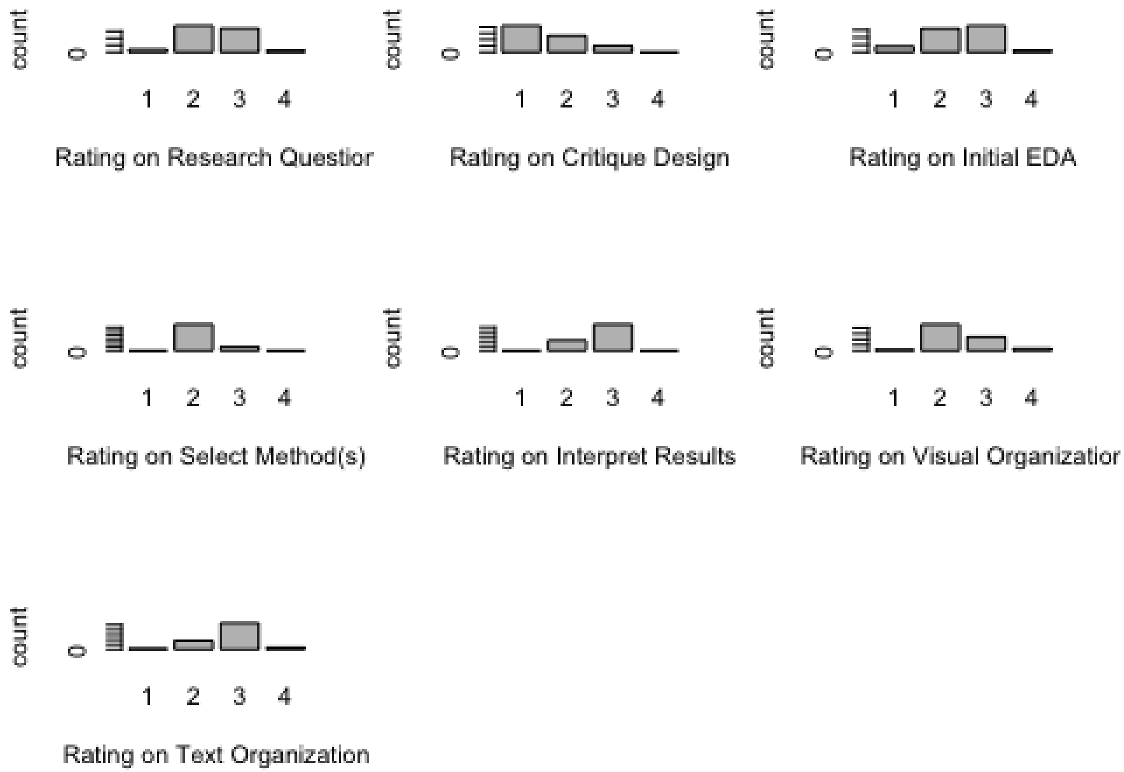
```r
InitEDA_full <- table(ratings$InitEDA, useNA = "always")
addmargins(InitEDA_full)

##
##    1    2    3    4 <NA>  Sum
##    8   56   47    6    0  117

round(prop.table(InitEDA_full)*100,digits=0)

##
##    1    2    3    4 <NA>
##    7   48   40    5    0

SelMeth_full <- table(ratings$SelMeth, useNA = "always")
addmargins(SelMeth_full)

##
##    1    2    3    4 <NA>  Sum
##   10   89   18    0    0  117

round(prop.table(SelMeth_full)*100,digits=0)

##
##    1    2    3    4 <NA>
##    9   76   15    0    0

InterpRes_full <- table(ratings$InterpRes, useNA = "always")
addmargins(InterpRes_full)

##
##    1    2    3    4 <NA>  Sum
##    6   49   61    1    0  117

round(prop.table(InterpRes_full)*100,digits=0)

##
##    1    2    3    4 <NA>
##    5   42   52    1    0

VisOrg_full <- table(ratings$VisOrg, useNA = "always")
addmargins(VisOrg_full)

##
##    1    2    3    4 <NA>  Sum
##    7   59   45    5    1  117

round(prop.table(VisOrg_full)*100,digits=0)

##
##    1    2    3    4 <NA>
##    6   50   38    4    1
```

```
TxtOrg_full <- table(ratings$TxtOrg, useNA = "always")
addmargins(TxtOrg_full)

##
##    1    2    3    4 <NA>  Sum
##    8   37   66    6    0  117

round(prop.table(TxtOrg_full)*100,digits=0)

##
##    1    2    3    4 <NA>
##    7   32   56    5    0
```

Rater 1

The distributions of how Rater 1 rates on different rubrics

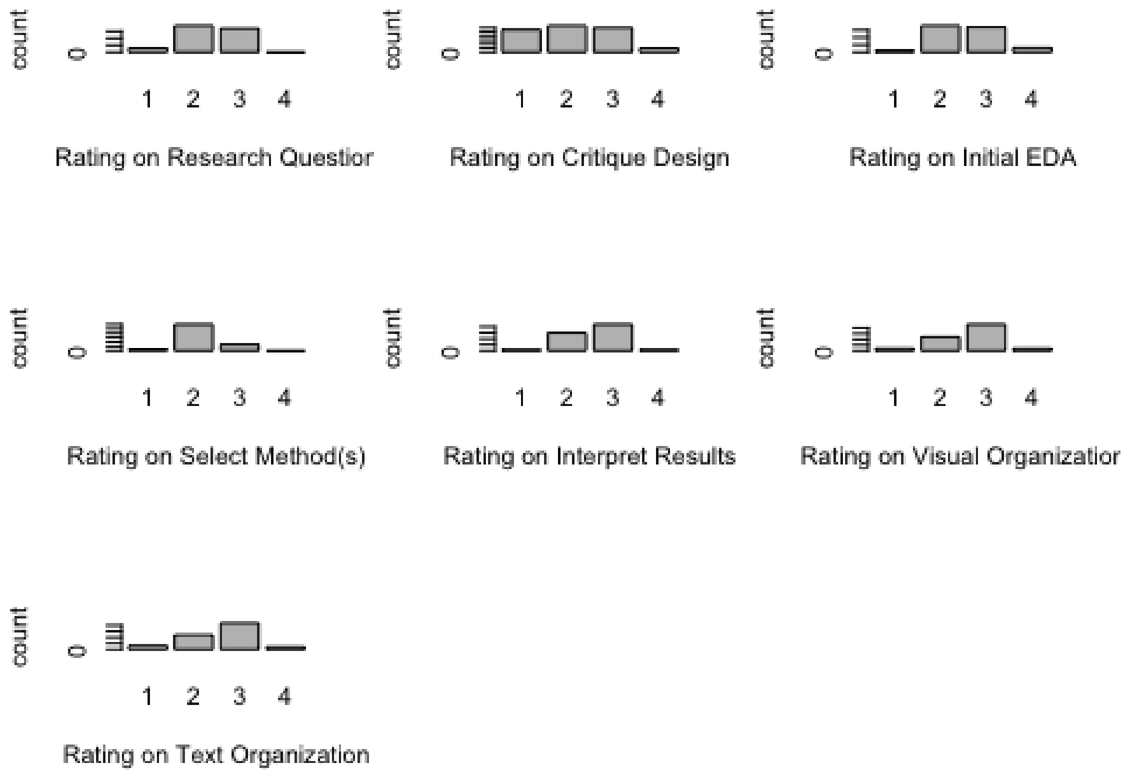
```
ratings_1 <- ratings %>%
  filter(ratings$Rater == 1)

par(mfrow=c(3,3))
plot(ratings_1$RsrchQ, xlab = "Rating on Research Question", ylab = "count")
plot(ratings_1$CritDes, xlab = "Rating on Critique Design", ylab = "count")
plot(ratings_1$InitEDA, xlab = "Rating on Initial EDA", ylab = "count")
plot(ratings_1$SelMeth, xlab = "Rating on Select Method(s)", ylab = "count")
plot(ratings_1$InterpRes, xlab = "Rating on Interpret Results", ylab = "count
")
plot(ratings_1$VisOrg, xlab = "Rating on Visual Organization", ylab = "count"
)
plot(ratings_1$TxtOrg, xlab = "Rating on Text Organization", ylab = "count")
```



Rater 2

The distributions of how Rater 2 rates on different rubrics

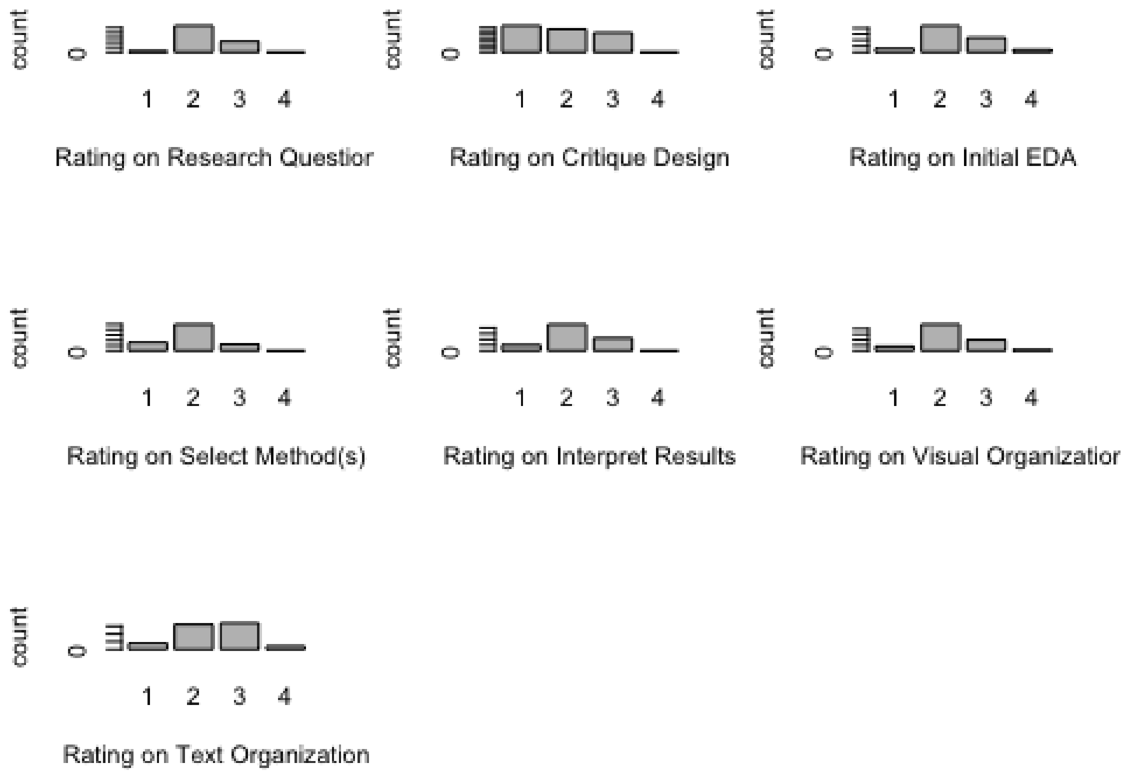
```
ratings_2 <- ratings %>%
  filter(ratings$Rater == 2)

par(mfrow=c(3,3))
plot(ratings_2$RsrchQ, xlab = "Rating on Research Question", ylab = "count")
plot(ratings_2$CritDes, xlab = "Rating on Critique Design", ylab = "count")
plot(ratings_2$InitEDA, xlab = "Rating on Initial EDA", ylab = "count")
plot(ratings_2$SelMeth, xlab = "Rating on Select Method(s)", ylab = "count")
plot(ratings_2$InterpRes, xlab = "Rating on Interpret Results", ylab = "count
")
plot(ratings_2$VisOrg, xlab = "Rating on Visual Organization", ylab = "count"
)
plot(ratings_2$TxtOrg, xlab = "Rating on Text Organization", ylab = "count")
```



Rater 3

The distributions of how Rater 3 rates on different rubrics

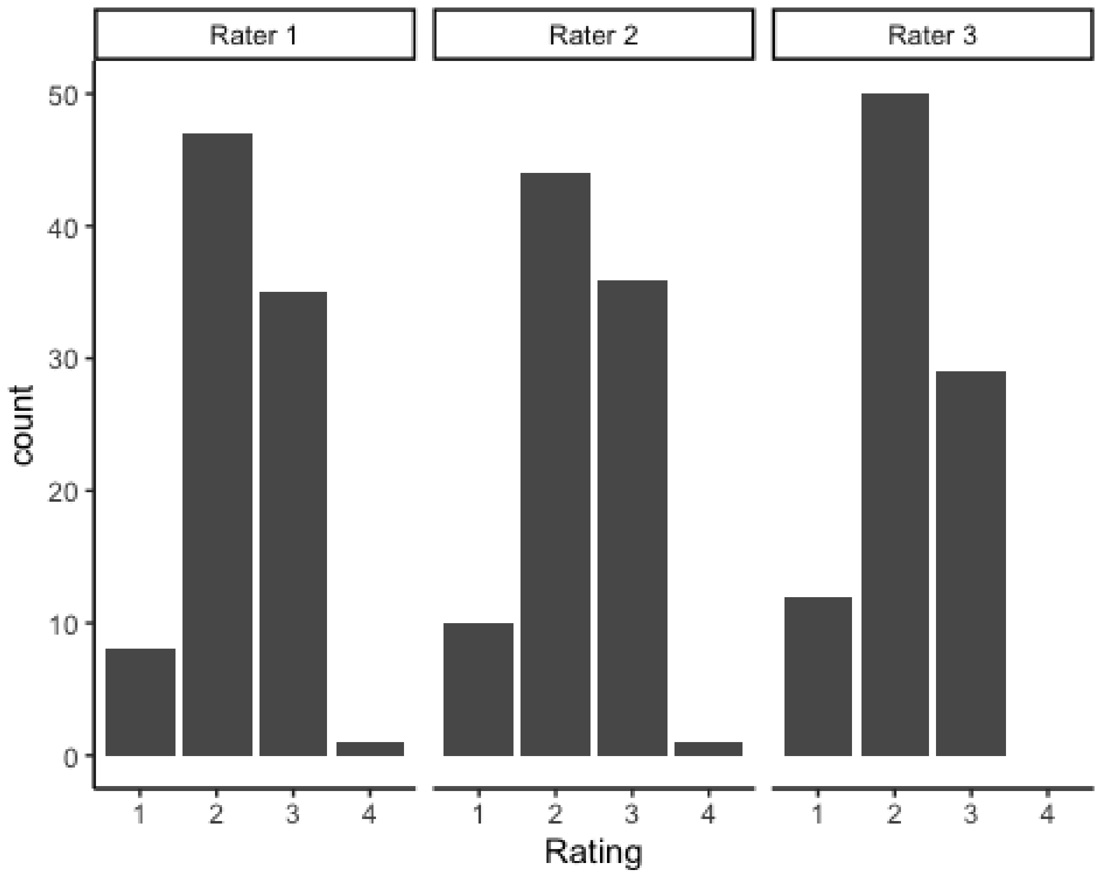
```
ratings_3 <- ratings %>%
  filter(ratings$Rater == 3)

par(mfrow=c(3,3))
plot(ratings_3$RsrchQ, xlab = "Rating on Research Question", ylab = "count")
plot(ratings_3$CritDes, xlab = "Rating on Critique Design", ylab = "count")
plot(ratings_3$InitEDA, xlab = "Rating on Initial EDA", ylab = "count")
plot(ratings_3$SelMeth, xlab = "Rating on Select Method(s)", ylab = "count")
plot(ratings_3$InterpRes, xlab = "Rating on Interpret Results", ylab = "count
")
plot(ratings_3$VisOrg, xlab = "Rating on Visual Organization", ylab = "count"
)
plot(ratings_3$TxtOrg, xlab = "Rating on Text Organization", ylab = "count")
```



Compare distributions across Raters using tall dataset

First, we focus on those 13 artifacts

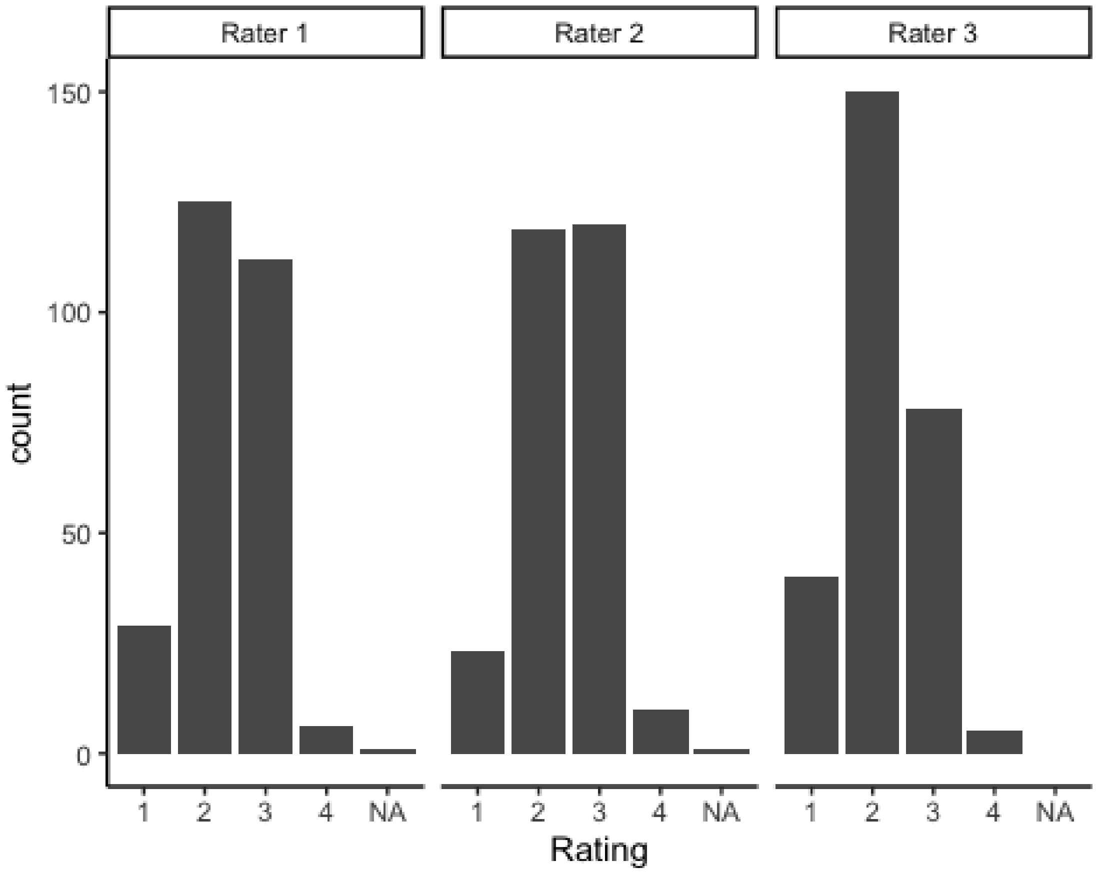
```
tall$Rating <- factor(tall$Rating,levels=1:4)
tall$Sex[nchar(tall$Sex)==0] <- "--"
tall_sub <- tall[grep("O",tall$Artifact),]

## Needed to make the title of each facet more human-readable...
rater.name <- function(x) { paste("Rater",x) }
## Barplots for reduced data...
g <- ggplot(tall_sub, aes(x = Rating)) +
  facet_wrap( ~ Rater, labeller=labeller(Rater=rater.name)) +
  geom_bar() + theme_classic()

g
```




```
## Corresponding table of counts...
tmp <- data.frame(lapply(split(tall_sub$Rating,tall_sub$Rater),summary))
row.names(tmp) <- paste("Rating",1:4)
names(tmp) <- paste("Rater",1:3)

tmp

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

Working on full data

```
## Barplots for full data...
g <- ggplot(tall,aes(x = Rating)) +
  facet_wrap( ~ Rater, labeller=labeller(Rater=rater.name)) +
  geom_bar() + theme_classic()
g
```




```
## Corresponding table of counts...
tmp0 <- lapply(split(tall$Rating,tall$Rater),summary)
tmp <- data.frame(matrix(0,nrow=5,ncol=3)) ## three raters...
names(tmp) <- names(tmp0)
row.names(tmp) <- c(paste("Rating",1:4),"<NA>")
for (i in names(tmp0)) {
  tmp[,i] <- tmp[,i] + c(tmp0[[i]],0)[1:5]
}
names(tmp) <- paste("Rater",1:3)
tmp

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

tall[apply(tall,1,function(x){any(is.na(x))}),]

##        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=="--",]

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

Appendix B

Investigate the rating consistency over these three raters on those 13 artifacts

Calculating ICC's as a measure of rater agreement

```
names(tall)

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

Group the ratings

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

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

dim(common)

## [1] 273   8
```

ICC calculations on each rubric for the 13 artifacts that all three raters saw

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

Ratings on Rating on Research Question

```
RsrchQ.ratings <- common[common$Rubric=="RsrchQ",]
RsrchQ_1 <- lmer(as.numeric(Rating) ~ 1 + (1|Rater), data=RsrchQ.ratings)

## boundary (singular) fit: see ?isSingular

summary(RsrchQ_1)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Rater)
```

```
##    Data: RsrchQ.ratings
##
## REML criterion at convergence: 67.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2912 -0.5041 -0.5041  1.2831  1.2831
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Rater    (Intercept) 0.0000   0.0000
##  Residual             0.3131   0.5595
## Number of obs: 39, groups:  Rater, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.2820     0.0896   25.47
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

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

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

```
RsrchQ_2 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=RsrchQ.ratings)
summary(RsrchQ_2)
```

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

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

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

Rating on Critique Design

```
CritDes.ratings <- common[common$Rubric=="CritDes",]
CritDes_1 <- lmer(as.numeric(Rating) ~ 1 + (1|Rater), data=CritDes.ratings)

## boundary (singular) fit: see ?isSingular

summary(CritDes_1)

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

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

## [1] 0

CritDes_2 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=CritDes.ratings
)
summary(CritDes_2)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Artifact)
##    Data: CritDes.ratings
##
## REML criterion at convergence: 75.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9647 -0.4386 -0.2978  0.5318  2.1987
##
```

```
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3091   0.5560
##  Residual             0.2308   0.4804
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   1.7179     0.1723   9.969

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

## [1] 0.5725134
```

Rating on Initial EDA

```
InitEDA.ratings <- common[common$Rubric=="InitEDA",]
InitEDA_1 <- lmer(as.numeric(Rating) ~ 1 + (1|Rater), data=InitEDA.ratings)
summary(InitEDA_1)

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

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

## [1] 0.003333356

InitEDA_2 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=InitEDA.ratings
)
summary(InitEDA_2)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Artifact)
```

```
##    Data: InitEDA.ratings
##
## REML criterion at convergence: 56.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1670 -0.2504 -0.2504  0.4006  1.6663
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.1496   0.3867
##  Residual             0.1538   0.3922
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.3846     0.1243   19.18
```

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

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

Rating on Select Method(s)

```
SelMeth.ratings <- common[common$Rubric=="SelMeth",]
SelMeth_1 <- lmer(as.numeric(Rating) ~ 1 + (1|Rater), data=SelMeth.ratings)
```

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

```
summary(SelMeth_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Rater)
##    Data: SelMeth.ratings
##
## REML criterion at convergence: 60.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0599 -0.1005 -0.1005 -0.1005  1.8590
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Rater    (Intercept) 0.0000   0.0000
##  Residual             0.2605   0.5104
## Number of obs: 39, groups:  Rater, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.05128    0.08172    25.1
```

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

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

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

```
SelMeth_2 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=SelMeth.ratings
)
summary(SelMeth_2)
```

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

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

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

Rating on Interpret Results

```
InterpRes.ratings <- common[common$Rubric=="InterpRes",]
InterpRes_1 <- lmer(as.numeric(Rating) ~ 1 + (1|Rater), data=InterpRes.rating
s)
summary(InterpRes_1)
```

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

```
## -2.4822 -0.8773  0.7917  0.7917  2.4608
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Rater    (Intercept) 0.003945 0.06281
##  Residual             0.358974 0.59914
## Number of obs: 39, groups:  Rater, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.5128     0.1026    24.5
```

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

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

```
InterpRes_2 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=InterpRes.rat
ings)
summary(InterpRes_2)
```

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

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

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

Rating on Visual Organization

```
VisOrg.ratings <- common[common$Rubric=="VisOrg",]
VisOrg_1 <- lmer(as.numeric(Rating) ~ 1 + (1|Rater), data=VisOrg.ratings)
```

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

```
summary(VisOrg_1)

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

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

## [1] 0

VisOrg_2 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=VisOrg.ratings)
summary(VisOrg_2)

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

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

## [1] 0.5924748
```

Rating on Text Organization

```
TxtOrg.ratings <- common[common$Rubric=="TxtOrg",]
TxtOrg_1 <- lmer(as.numeric(Rating) ~ 1 + (1|Rater), data=TxtOrg.ratings)

## boundary (singular) fit: see ?isSingular

summary(TxtOrg_1)

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

TxtOrg_ICC_1 <- (0.000)/(0.000+0.386)
TxtOrg_ICC_1

## [1] 0

TxtOrg_2 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=TxtOrg.ratings)
summary(TxtOrg_2)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Artifact)
##    Data: TxtOrg.ratings
##
## REML criterion at convergence: 74.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6943 -0.7698  0.3849  0.3849  2.5019
```

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

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

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

Calculations of exact agreement between any two raters and on each rubric

Cross-classifying the ratings that each pair of raters gives on the subset of 13 artifacts seen by every rater

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

Rating on Research Question

```
raters_1_and_2_on_CritDes <- data.frame(r1=repeated$RsrchQ[repeated$Rater==1]
,
                                        r2=repeated$RsrchQ[repeated$Rater==2],
                                        a1=repeated$Artifact[repeated$Rater==1
],
                                        a2=repeated$Artifact[repeated$Rater==2
])
```

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

```
(t12 <- table(r1,r2))
```

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

```
(4+1)/(1+4+3+1+3+1)
```

```
## [1] 0.3846154
```

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

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

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

(t13 <- table(r1,r3))

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

(7+3)/(7+3+2+3)

## [1] 0.6666667

raters_2_and_3_on_RsrchQ <- data.frame(r2=repeated$RsrchQ[repeated$Rater==2],
                                       r3=repeated$RsrchQ[repeated$Rater==3],
                                       a2=repeated$Artifact[repeated$Rater==2
],
                                       a3=repeated$Artifact[repeated$Rater==3
])

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

(t23 <- table(r2,r3))

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

(5+2)/(2+5+2+2+2)

## [1] 0.5384615
```

Rating on Critique Design

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

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

(t12 <- table(r1,r2))

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

(3+3+1)/(3+2+1+2+3+1+1)

## [1] 0.5384615

raters_1_and_3_on_CritDes <- data.frame(r1=repeated$CritDes[repeated$Rater==1
],
                                       r3=repeated$CritDes[repeated$Rater==3]
,
                                       a1=repeated$Artifact[repeated$Rater==1
],
                                       a3=repeated$Artifact[repeated$Rater==3
])

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

(t13 <- table(r1,r3))

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

(4+3+1)/(4+2+2+3+1+1)

## [1] 0.6153846

raters_2_and_3_on_CritDes <- data.frame(r2=repeated$CritDes[repeated$Rater==2
],
                                       r3=repeated$CritDes[repeated$Rater==3]
,
                                       a2=repeated$Artifact[repeated$Rater==2
],
                                       a3=repeated$Artifact[repeated$Rater==3
])

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

```
(t23 <- table(r2,r3))

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

(5+3+1)/(5+1+3+1+2+1)

## [1] 0.6923077
```

Rating on Rating on Initial EDA

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

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

(t12 <- table(r1,r2))

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

(4+5)/(1+4+3+5)

## [1] 0.6923077

raters_1_and_3_on_InitEDA <- data.frame(r1=repeated$InitEDA[repeated$Rater==1
],
                                       r3=repeated$InitEDA[repeated$Rater==3]
,
                                       a1=repeated$Artifact[repeated$Rater==1
],
                                       a3=repeated$Artifact[repeated$Rater==3
])

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

(t13 <- table(r1,r3))
```

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

```
(4+3)/(1+4+5+3)
```

```
## [1] 0.5384615
```

```
raters_2_and_3_on_InitEDA <- data.frame(r2=repeated$InitEDA[repeated$Rater==2
],
                                        r3=repeated$InitEDA[repeated$Rater==3]
,
                                        a2=repeated$Artifact[repeated$Rater==2
],
                                        a3=repeated$Artifact[repeated$Rater==3
])

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

(t23 <- table(r2,r3))
```

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

```
(8+3)/(8+2+3)
```

```
## [1] 0.8461538
```

Rating on Select Method(s)

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

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

(t12 <- table(r1,r2))
```

```
##    r2
## r1   1  2  3  4
##   1  0  0  0  0
##   2  1 10  0  0
##   3  0  0  2  0
##   4  0  0  0  0
```

```
(10+2)/(1+10+2)
```

```
## [1] 0.9230769
```

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

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

```
(t13 <- table(r1,r3))
```

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

```
(7+1)/(3+7+1+1+1)
```

```
## [1] 0.6153846
```

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

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

```
(t23 <- table(r2,r3))
```

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

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

(1+7+1)/(1+2+7+1+1+1)

## [1] 0.6923077
```

Rating on Interpret Results

```
raters_1_and_2_on_InterpRes <- data.frame(r1=repeated$InterpRes[repeated$Rate
r==1],
                                          r2=repeated$InterpRes[repeated$Rater==
2],
                                          a1=repeated$Artifact[repeated$Rater==1
],
                                          a2=repeated$Artifact[repeated$Rater==2
])

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

(t12 <- table(r1,r2))

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

(3+5)/(3+1+1+3+5)

## [1] 0.6153846

raters_1_and_3_on_InterpRes <- data.frame(r1=repeated$InterpRes[repeated$Rate
r==1],
                                          r3=repeated$InterpRes[repeated$Rater==
3],
                                          a1=repeated$Artifact[repeated$Rater==1
],
                                          a3=repeated$Artifact[repeated$Rater==3
])

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

(t13 <- table(r1,r3))

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

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

```
(3+4)/(1+3+1+4+4)
```

```
## [1] 0.5384615
```

```
raters_2_and_3_on_InterpRes <- data.frame(r2=repeated$InterpRes[repeated$Rate
r==2],
                                          r3=repeated$InterpRes[repeated$Rater==
3],
                                          a2=repeated$Artifact[repeated$Rater==2
],
                                          a3=repeated$Artifact[repeated$Rater==3
])
```

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

```
(t23 <- table(r2,r3))
```

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

```
(4+4)/(1+4+1+2+4+1)
```

```
## [1] 0.6153846
```

Rating on Visual Organization

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

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

```
(t12 <- table(r1,r2))
```

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

```
(1+4+2)/(1+4+5+1+2)

## [1] 0.5384615

raters_1_and_3_on_VisOrg <- data.frame(r1=repeated$VisOrg[repeated$Rater==1],
                                       r3=repeated$VisOrg[repeated$Rater==3],
                                       a1=repeated$Artifact[repeated$Rater==1
],
                                       a3=repeated$Artifact[repeated$Rater==3
])

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

(t13 <- table(r1,r3))

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

(1+7+2)/(1+7+2+1+2)

## [1] 0.7692308

raters_2_and_3_on_VisOrg <- data.frame(r2=repeated$VisOrg[repeated$Rater==2],
                                       r3=repeated$VisOrg[repeated$Rater==3],
                                       a2=repeated$Artifact[repeated$Rater==2
],
                                       a3=repeated$Artifact[repeated$Rater==3
])

r2 <- factor(raters_2_and_3_on_VisOrg$r2,levels=1:4)
r3 <- factor(raters_2_and_3_on_VisOrg$r3,levels=1:4)

(t23 <- table(r2,r3))

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

(1+5+4)/(1+5+3+4)

## [1] 0.7692308
```

Rating on Text Organization

```
raters_1_and_2_on_TxtOrg <- data.frame(r1=repeated$TxtOrg[repeated$Rater==1],
                                       r2=repeated$TxtOrg[repeated$Rater==2],
```

```
                                                      a1=repeated$Artifact[repeated$Rater==1
],
                                                      a2=repeated$Artifact[repeated$Rater==2
])

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

(t12 <- table(r1,r2))

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

(2+7)/(2+2+1+7+1)

## [1] 0.6923077

raters_1_and_3_on_TxtOrg <- data.frame(r1=repeated$TxtOrg[repeated$Rater==1],
                                       r3=repeated$TxtOrg[repeated$Rater==3],
                                       a1=repeated$Artifact[repeated$Rater==1
],
                                       a3=repeated$Artifact[repeated$Rater==3
])

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

(t13 <- table(r1,r3))

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

(1+7)/(1+1+2+1+7+1)

## [1] 0.6153846

raters_2_and_3_on_TxtOrg <- data.frame(r2=repeated$TxtOrg[repeated$Rater==2],
                                       r3=repeated$TxtOrg[repeated$Rater==3],
                                       a2=repeated$Artifact[repeated$Rater==2
],
                                       a3=repeated$Artifact[repeated$Rater==3
])

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

```
(t23 <- table(r2,r3))

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

(7)/(1+1+2+2+7)

## [1] 0.5384615

Rubric.names <- sort(unique(tall$Rubric))
ICC.vec <- NULL
for (i in Rubric.names) {
  tmp <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=tall_sub[tall_sub$R
ubric==i,])
  sig2 <- summary(tmp)$sigma^2
  tau2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
  ICC <- tau2 / (tau2 + sig2)
  ICC.vec <- c(ICC.vec,ICC)
}
names(ICC.vec) <- Rubric.names
agreement.results <- cbind(ICC.common=ICC.vec,"    a12"=0,a23=0,a13=0)
agreement.tables <- as.list(rep(NA,7))
names(agreement.tables) <- Rubric.names
for (i in Rubric.names) {
  r12 <- data.frame(r1=factor(ratings_sub[ratings_sub$Rater==1,i],levels=1:4)
,
                    r2=factor(ratings_sub[ratings_sub$Rater==2,i],levels=1:4)
,
                    a1=ratings_sub[ratings_sub$Rater==1,"Artifact"],
                    a2=ratings_sub[ratings_sub$Rater==2,"Artifact"])
  if(any(r12[,3]!=r12[,4])) { stop(paste("Rater 1-2 Artifact mismatch on rubr
ic",i)) }
  a12 <- mean(r12[,1]==r12[,2])
  r12 <- table(r12[,1:2]) ## print this to see how much agreement there is am
ong raters 1-2
  r23 <- data.frame(r2=factor(ratings_sub[ratings_sub$Rater==2,i],levels=1:4)
,
                    r3=factor(ratings_sub[ratings_sub$Rater==3,i],levels=1:4)
,
                    a2=ratings_sub[ratings_sub$Rater==2,"Artifact"],
                    a3=ratings_sub[ratings_sub$Rater==3,"Artifact"])
  if(any(r23[,3]!=r23[,4])) { stop(paste("Rater 2-3 Artifact mismatch on rubr
ic",i)) }
  a23 <- mean(r23[,1]==r23[,2])
  r23 <- table(r23[,1:2]) ## print this to see how much agreement there is am
ong raters 2-3
  r13 <- data.frame(r1=factor(ratings_sub[ratings_sub$Rater==1,i],levels=1:4)
```

```
,
                         r3=factor(ratings_sub[ratings_sub$Rater==3,i],levels=1:4)
,
                         a1=ratings_sub[ratings_sub$Rater==1,"Artifact"],
                         a3=ratings_sub[ratings_sub$Rater==3,"Artifact"])
  if(any(r13[,3]!=r13[,4])) { stop(paste("Rater 1-3 Artifact mismatch on rubr
ic",i)) }
  a13 <- mean(r13[,1]==r13[,2])
  r13 <- table(r13[,1:2]) ## print this to see how much agreement there is am
ong raters 1-3
  agreement.results[i,2:4] <- c(a12,a23,a13)

  agreement.tables[[i]] <- list(r12,r23,r13)
}
round(agreement.results,2)

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

Working on full data

```
ratings$Rater <- as.factor(ratings$Rater)
tall$Artifact <- as.factor(tall$Artifact)
tall$Rater <- as.factor(tall$Rater)
tall$Sex <- as.factor(tall$Sex)
tall$Semester <- as.factor(tall$Semester)
tall$Rubric <- as.factor(tall$Rubric)
tall$Repeated <- as.factor(tall$Repeated)
```

Ratings on Rating on Research Question

```
RsrchQ.ratings_full <- tall[tall$Rubric=="RsrchQ",]
RsrchQ_3 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=RsrchQ.ratings_f
ull)
summary(RsrchQ_3)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Artifact)
##    Data: RsrchQ.ratings_full
##
## REML criterion at convergence: 211.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2748 -0.5365 -0.3780  0.9626  2.4617
##
```

```
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.07372  0.2715
##  Residual             0.27797  0.5272
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.35790    0.05774   40.84

RsrchQ_ICC_3 <- (0.07372)/(0.07372+0.27797)
RsrchQ_ICC_3

## [1] 0.2096164

RsrchQ_ICC_2

## [1] 0.1891918
```

Rating on Critique Design

```
CritDes.ratings_full <- tall[tall$Rubric=="CritDes",]
CritDes_3 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=CritDes.ratings
_full)
summary(CritDes_3)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Artifact)
##    Data: CritDes.ratings_full
##
## REML criterion at convergence: 277.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.01042 -0.60409  0.04407  0.72769  2.06310
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.4963   0.7045
##  Residual             0.2411   0.4910
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.90720    0.08874   21.49

CritDes_ICC_3 <- (0.4963)/(0.4963+0.2411)
CritDes_ICC_3

## [1] 0.6730404

CritDes_ICC_2
```

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

Rating on Initial EDA

```
InitEDA.ratings_full <- tall[tall$Rubric=="InitEDA",]
InitEDA_3 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=InitEDA.ratings
_full)
summary(InitEDA_3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Artifact)
##    Data: InitEDA.ratings_full
##
## REML criterion at convergence: 240.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8923 -0.3451 -0.1454  0.4250  1.6015
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3628   0.6023
##  Residual             0.1655   0.4068
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.44815    0.07479   32.73
```

```
InitEDA_ICC_3 <- (0.3628)/(0.3628+0.1655)
InitEDA_ICC_3
```

```
## [1] 0.686731
```

```
InitEDA_ICC_2
```

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

Rating on Select Method(s)

```
SelMeth.ratings_full <- tall[tall$Rubric=="SelMeth",]
SelMeth_3 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=SelMeth.ratings
_full)
summary(SelMeth_3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Artifact)
##    Data: SelMeth.ratings_full
##
## REML criterion at convergence: 157.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.2057 -0.1075 -0.1075 -0.0553  2.0951
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.1108   0.3329
##  Residual             0.1240   0.3521
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.07168    0.04893   42.34
```

```
SelMeth_ICC_3 <- (0.1108)/(0.1108+0.1240)
SelMeth_ICC_3
```

```
## [1] 0.471891
```

```
SelMeth_ICC_2
```

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

Rating on Interpret Results

```
InterpRes.ratings_full <- tall[tall$Rubric=="InterpRes",]
InterpRes_3 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=InterpRes.rat
ings_full)
summary(InterpRes_3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Artifact)
##    Data: InterpRes.ratings_full
##
## REML criterion at convergence: 217.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1448 -0.6998  0.5175  0.7452  2.6532
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.08219  0.2867
##  Residual             0.29136  0.5398
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.48427    0.05962   41.67
```

```
InterpRes_ICC_3 <- (0.08219)/(0.08219+0.29136)
InterpRes_ICC_3
```

```
## [1] 0.2200241
```

```
InterpRes_ICC_2

## [1] 0.2295821
```

Rating on Visual Organization

```
VisOrg.ratings_full <- tall[tall$Rubric=="VisOrg",]
VisOrg_3 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=VisOrg.ratings_f
ull)
summary(VisOrg_3)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Artifact)
##    Data: VisOrg.ratings_full
##
## REML criterion at convergence: 226.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5918 -0.3789 -0.1632  0.4726  1.6322
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3092   0.5561
##  Residual             0.1588   0.3985
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.44497    0.07063   34.62

VisOrg_ICC_3 <- (0.3092)/(0.3092+0.1588)
VisOrg_ICC_3

## [1] 0.6606838

VisOrg_ICC_2

## [1] 0.5924748
```

Rating on Text Organization

```
TxtOrg.ratings_full <- tall[tall$Rubric=="TxtOrg",]
TxtOrg_3 <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=TxtOrg.ratings_f
ull)
summary(TxtOrg_3)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Artifact)
##    Data: TxtOrg.ratings_full
##
## REML criterion at convergence: 249
##
## Scaled residuals:
```

```
##      Min       1Q  Median      3Q     Max
## -2.3638 -0.7641  0.3836  0.5278  2.4094
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.09145  0.3024
##  Residual             0.39503  0.6285
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.59144    0.06764   38.31
```

```
TxtOrg_ICC_3 <- (0.09145)/(0.09145+0.39503)
TxtOrg_ICC_3
```

```
## [1] 0.1879831
```

```
TxtOrg_ICC_2
```

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

```
if (F) { print(agreement.tables) }
ICC.vec <- NULL
for (i in Rubric.names) {
  tmp <- lmer(as.numeric(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 <- c(ICC.vec,ICC)
}
names(ICC.vec) <- Rubric.names
agreement.results <- cbind(ICC.alldata=ICC.vec,agreement.results)
round(agreement.results,2)
```

```
##           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
```

Appendix C

Exploring the relationship of various factors in this experiment (Rater, Semester, Sex, Repeated, Rubric) and the ratings

Rating on Research Question Variable Selection

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

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

## Data: RsrchQ.ratings_full
## Models:
## RsrchQ_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## RsrchQ_3_1: as.numeric(Rating) ~ (1 | Artifact) + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## RsrchQ_3      3 213.19 221.48 -103.60   207.19
## RsrchQ_3_1    5 215.37 229.18 -102.68   205.37 1.8253  2     0.4015

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

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

## Data: RsrchQ.ratings_full
## Models:
## RsrchQ_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## RsrchQ_3_2: as.numeric(Rating) ~ (1 | Artifact) + Semester
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## RsrchQ_3      3 213.19 221.48 -103.60   207.19
## RsrchQ_3_2    4 214.57 225.62 -103.28   206.57 0.6253  1     0.4291

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

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

## Data: RsrchQ.ratings_full
## Models:
## RsrchQ_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## RsrchQ_3_3: as.numeric(Rating) ~ (1 | Artifact) + Rater
##            npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
## RsrchQ_3      3 213.19 221.48 -103.6   207.19
## RsrchQ_3_3    5 215.39 229.20 -102.7   205.39 1.8013  2     0.4063

summary(RsrchQ_3)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (1 | Artifact)
##    Data: RsrchQ.ratings_full
##
## REML criterion at convergence: 211.1
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2748 -0.5365 -0.3780  0.9626  2.4617
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.07372  0.2715
##  Residual             0.27797  0.5272
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.35790    0.05774   40.84
```

Rating on Critique Design Variable Selection

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

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

## Data: CritDes.ratings_full
## Models:
## CritDes_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## CritDes_3_1: as.numeric(Rating) ~ (1 | Artifact) + Sex
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## CritDes_3      3 280.86 289.12 -137.43   274.86
## CritDes_3_1    5 282.65 296.42 -136.33   272.65 2.2017  2     0.3326

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

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

## Data: CritDes.ratings_full
## Models:
## CritDes_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## CritDes_3_2: as.numeric(Rating) ~ (1 | Artifact) + Semester
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## CritDes_3      3 280.86 289.12 -137.43   274.86
## CritDes_3_2    4 282.58 293.60 -137.29   274.58 0.2751  1     0.5999

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

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

## Data: CritDes.ratings_full
## Models:
## CritDes_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## CritDes_3_3: as.numeric(Rating) ~ (1 | Artifact) + Rater
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
```

```
## CritDes_3      3 280.86 289.12 -137.43   274.86
## CritDes_3_3    5 276.86 290.62 -133.43   266.86 7.9996  2    0.01832 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

CritDes_3_4 <- update(CritDes_3, . ~ . + Sex + Rater)
anova(CritDes_3, CritDes_3_3, CritDes_3_4)

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

## Data: CritDes.ratings_full
## Models:
## CritDes_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## CritDes_3_3: as.numeric(Rating) ~ (1 | Artifact) + Rater
## CritDes_3_4: as.numeric(Rating) ~ (1 | Artifact) + Sex + Rater
##              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## CritDes_3      3 280.86 289.12 -137.43   274.86
## CritDes_3_3    5 276.86 290.62 -133.43   266.86 7.9996  2    0.01832 *
## CritDes_3_4    7 278.65 297.93 -132.33   264.65 2.2062  2    0.33185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

CritDes_3_5 <- update(CritDes_3, . ~ . + Sex * Rater)

## fixed-effect model matrix is rank deficient so dropping 2 columns / coeffi
cients

anova(CritDes_3, CritDes_3_3, CritDes_3_4, CritDes_3_5)

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

## Data: CritDes.ratings_full
## Models:
## CritDes_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## CritDes_3_3: as.numeric(Rating) ~ (1 | Artifact) + Rater
## CritDes_3_4: as.numeric(Rating) ~ (1 | Artifact) + Sex + Rater
## CritDes_3_5: as.numeric(Rating) ~ (1 | Artifact) + Sex + Rater + Sex:Rater
##              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## CritDes_3      3 280.86 289.12 -137.43   274.86
## CritDes_3_3    5 276.86 290.62 -133.43   266.86 7.9996  2    0.01832 *
## CritDes_3_4    7 278.65 297.93 -132.33   264.65 2.2062  2    0.33185
## CritDes_3_5    9 282.04 306.82 -132.02   264.04 0.6150  2    0.73529
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

CritDes_3_6 <- update(CritDes_3, . ~ . + Semester + Rater)
anova(CritDes_3, CritDes_3_3, CritDes_3_6)

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

## Data: CritDes.ratings_full
## Models:
```

```
## CritDes_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## CritDes_3_3: as.numeric(Rating) ~ (1 | Artifact) + Rater
## CritDes_3_6: as.numeric(Rating) ~ (1 | Artifact) + Semester + Rater
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## CritDes_3      3 280.86 289.12 -137.43   274.86
## CritDes_3_3    5 276.86 290.62 -133.43   266.86 7.9996  2    0.01832 *
## CritDes_3_6    6 278.65 295.18 -133.33   266.65 0.2038  1    0.65164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
CritDes_3_7 <- update(CritDes_3, . ~ . + Semester * Rater)
anova(CritDes_3, CritDes_3_3, CritDes_3_6, CritDes_3_7)
```

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

```
## Data: CritDes.ratings_full
## Models:
## CritDes_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## CritDes_3_3: as.numeric(Rating) ~ (1 | Artifact) + Rater
## CritDes_3_6: as.numeric(Rating) ~ (1 | Artifact) + Semester + Rater
## CritDes_3_7: as.numeric(Rating) ~ (1 | Artifact) + Semester + Rater + Seme
ster:Rater
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## CritDes_3      3 280.86 289.12 -137.43   274.86
## CritDes_3_3    5 276.86 290.62 -133.43   266.86 7.9996  2    0.01832 *
## CritDes_3_6    6 278.65 295.18 -133.33   266.65 0.2038  1    0.65164
## CritDes_3_7    8 280.32 302.35 -132.16   264.32 2.3353  2    0.31110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
formula(CritDes_3_3)
```

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

```
summary(CritDes_3_3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (1 | Artifact) + Rater
##    Data: CritDes.ratings_full
##
## REML criterion at convergence: 274.2
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.54697 -0.50107 -0.08068  0.63782  1.61697
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.4401   0.6634
##  Residual             0.2475   0.4975
## Number of obs: 116, groups:  Artifact, 90
```

```
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   1.6926     0.1210  13.994
## Rater2        0.4259     0.1494   2.850
## Rater3        0.2218     0.1488   1.491
##
## Correlation of Fixed Effects:
##        (Intr) Rater2
## Rater2 -0.609
## Rater3 -0.615  0.498
```

```
qt(p=.05/2, df=116-3, lower.tail=FALSE)
```

```
## [1] 1.98118
```

```
CritDes_ICC_4 <- (0.4401)/(0.4401+0.2475)
CritDes_ICC_4
```

```
## [1] 0.6400524
```

```
CritDes_ICC_3
```

```
## [1] 0.6730404
```

Rating on Initial EDA Variable Selection

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

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

```
## Data: InitEDA.ratings_full
## Models:
## InitEDA_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## InitEDA_3_1: as.numeric(Rating) ~ (1 | Artifact) + Sex
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## InitEDA_3      3 243.42 251.71 -118.71   237.42
## InitEDA_3_1    5 246.75 260.56 -118.38   236.75 0.6718  2     0.7147
```

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

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

```
## Data: InitEDA.ratings_full
## Models:
## InitEDA_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## InitEDA_3_2: as.numeric(Rating) ~ (1 | Artifact) + Semester
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## InitEDA_3      3 243.42 251.71 -118.71   237.42
## InitEDA_3_2    4 245.38 256.43 -118.69   237.38 0.0391  1     0.8432
```

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

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

## Data: InitEDA.ratings_full
## Models:
## InitEDA_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## InitEDA_3_3: as.numeric(Rating) ~ (1 | Artifact) + Rater
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## InitEDA_3      3 243.42 251.71 -118.71   237.42
## InitEDA_3_3    5 244.28 258.09 -117.14   234.28 3.1408  2      0.208
```

Rating on Select Method(s) Variable Selection

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

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

## Data: SelMeth.ratings_full
## Models:
## SelMeth_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## SelMeth_3_1: as.numeric(Rating) ~ (1 | Artifact) + Sex
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## SelMeth_3      3 159.53 167.82 -76.768   153.53
## SelMeth_3_1    5 155.32 169.13 -72.660   145.32 8.2155  2    0.01644 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

SelMeth_3_2 <- update(SelMeth_3_1, . ~ . + Semester)
anova(SelMeth_3, SelMeth_3_1, SelMeth_3_2)

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

## Data: SelMeth.ratings_full
## Models:
## SelMeth_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## SelMeth_3_1: as.numeric(Rating) ~ (1 | Artifact) + Sex
## SelMeth_3_2: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## SelMeth_3      3 159.53 167.82 -76.768   153.53
## SelMeth_3_1    5 155.32 169.13 -72.660   145.32 8.2155  2   0.016445 *
## SelMeth_3_2    6 147.94 164.51 -67.968   135.94 9.3837  1   0.002189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

SelMeth_3_3 <- update(SelMeth_3, . ~ . + Sex*Semester)

## fixed-effect model matrix is rank deficient so dropping 1 column / coeffic
ient

anova(SelMeth_3, SelMeth_3_2, SelMeth_3_3)
```

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

## Data: SelMeth.ratings_full
## Models:
## SelMeth_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## SelMeth_3_2: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester
## SelMeth_3_3: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester + Sex:Se
mester
##             npar    AIC    BIC  logLik deviance   Chisq Df Pr(>Chisq)
## SelMeth_3      3 159.53 167.82 -76.768   153.53
## SelMeth_3_2    6 147.94 164.51 -67.968   135.94 17.5992  3   0.000532 ***
## SelMeth_3_3    7 145.28 164.62 -65.641   131.28  4.6538  1   0.030984 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

SelMeth_3_4 <- update(SelMeth_3_2, . ~ . + Rater)
anova(SelMeth_3, SelMeth_3_2, SelMeth_3_3, SelMeth_3_4)

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

## Data: SelMeth.ratings_full
## Models:
## SelMeth_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## SelMeth_3_2: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester
## SelMeth_3_3: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester + Sex:Se
mester
## SelMeth_3_4: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester + Rater
##             npar    AIC    BIC  logLik deviance   Chisq Df Pr(>Chisq)
## SelMeth_3      3 159.53 167.82 -76.768   153.53
## SelMeth_3_2    6 147.94 164.51 -67.968   135.94 17.5992  3   0.000532 ***
## SelMeth_3_3    7 145.28 164.62 -65.641   131.28  4.6538  1   0.030984 *
## SelMeth_3_4    8 144.52 166.62 -64.260   128.52  2.7615  1   0.096557 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

SelMeth_3_5 <- update(SelMeth_3_4, . ~ . + Sex*Rater)

## fixed-effect model matrix is rank deficient so dropping 2 columns / coeffi
cients

anova(SelMeth_3, SelMeth_3_4, SelMeth_3_2, SelMeth_3_5)

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

## Data: SelMeth.ratings_full
## Models:
## SelMeth_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## SelMeth_3_2: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester
## SelMeth_3_4: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester + Rater
## SelMeth_3_5: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester + Rater
+ Sex:Rater
##             npar    AIC    BIC  logLik deviance   Chisq Df Pr(>Chisq)
```

```
## SelMeth_3      3 159.53 167.82 -76.768   153.53
## SelMeth_3_2    6 147.94 164.51 -67.968   135.94 17.5992  3   0.000532 ***
## SelMeth_3_4    8 144.52 166.62 -64.260   128.52  7.4154  2   0.024534 *
## SelMeth_3_5   10 146.50 174.12 -63.248   126.50  2.0238  2   0.363524
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

SelMeth_3_6 <- update(SelMeth_3_4, . ~ . + Semester*Rater)
anova(SelMeth_3, SelMeth_3_2, SelMeth_3_4, SelMeth_3_6)

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

## Data: SelMeth.ratings_full
## Models:
## SelMeth_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## SelMeth_3_2: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester
## SelMeth_3_4: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester + Rater
## SelMeth_3_6: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester + Rater
+ Semester:Rater
##             npar    AIC    BIC  logLik deviance   Chisq Df Pr(>Chisq)
## SelMeth_3      3 159.53 167.82 -76.768   153.53
## SelMeth_3_2    6 147.94 164.51 -67.968   135.94 17.5992  3   0.000532 ***
## SelMeth_3_4    8 144.52 166.62 -64.260   128.52  7.4154  2   0.024534 *
## SelMeth_3_6   10 145.77 173.40 -62.887   125.77  2.7467  2   0.253259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(SelMeth_3_2)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (1 | Artifact) + Sex + Semester
##    Data: SelMeth.ratings_full
##
## REML criterion at convergence: 145.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.45094 -0.36643 -0.09728  0.34827  2.15811
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.08971  0.2995
##  Residual             0.11560  0.3400
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   3.0000     0.4531   6.621
## SexF         -0.8899     0.4594  -1.937
## SexM         -0.7787     0.4585  -1.698
## SemesterS19  -0.3204     0.1043  -3.072
```

```
##
## Correlation of Fixed Effects:
##              (Intr) SexF   SexM
## SexF         -0.986
## SexM         -0.988  0.978
## SemesterS19   0.000 -0.097 -0.034

qt(p=.05/2, df=117-7, lower.tail=FALSE)

## [1] 1.981765

SelMeth_ICC_4 <- (0.0850)/(0.0850+0.1057)
SelMeth_ICC_4

## [1] 0.4457263

SelMeth_ICC_3

## [1] 0.471891
```

Rating on Interpret Results Variable Selection

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

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

## Data: InterpRes.ratings_full
## Models:
## InterpRes_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## InterpRes_3_1: as.numeric(Rating) ~ (1 | Artifact) + Sex
##               npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## InterpRes_3      3 220.09 228.38 -107.05   214.09
## InterpRes_3_1    5 223.14 236.95 -106.57   213.14 0.9519  2     0.6213

InterpRes_3_2 <- update(InterpRes_3, . ~ . + Semester)
anova(InterpRes_3, InterpRes_3_2)

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

## Data: InterpRes.ratings_full
## Models:
## InterpRes_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## InterpRes_3_2: as.numeric(Rating) ~ (1 | Artifact) + Semester
##               npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## InterpRes_3      3 220.09 228.38 -107.05   214.09
## InterpRes_3_2    4 221.76 232.81 -106.88   213.76 0.3386  1     0.5606

InterpRes_3_3 <- update(InterpRes_3, . ~ . + Rater)
anova(InterpRes_3, InterpRes_3_3)

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

```
## Data: InterpRes.ratings_full
## Models:
## InterpRes_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## InterpRes_3_3: as.numeric(Rating) ~ (1 | Artifact) + Rater
##               npar    AIC    BIC   logLik deviance  Chisq Df Pr(>Chisq)

## InterpRes_3      3 220.09 228.38 -107.048   214.09

## InterpRes_3_3    5 203.66 217.47  -96.831   193.66 20.433  2  3.657e-05 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(InterpRes_3_3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (1 | Artifact) + Rater
##    Data: InterpRes.ratings_full
##
## REML criterion at convergence: 202.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5101 -0.7484  0.3763  0.6532  2.6479
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.06471  0.2544
##  Residual             0.25381  0.5038
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.70517    0.08961  30.187
## Rater2      -0.11816    0.12270  -0.963
## Rater3      -0.54402    0.12270  -4.434
##
## Correlation of Fixed Effects:
##        (Intr) Rater2
## Rater2 -0.685
## Rater3 -0.685  0.500
```

```
qt(p=.05/2, df=117-3, lower.tail=FALSE)
```

```
## [1] 1.980992
```

```
InterpRes_ICC_4 <- (0.06471)/(0.06471+0.25381)
InterpRes_ICC_4
```

```
## [1] 0.2031584
```

```
InterpRes_ICC_3

## [1] 0.2200241
```

Rating on Visual Organization Variable Selection

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

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

## Data: VisOrg.ratings_full
## Models:
## VisOrg_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## VisOrg_3_1: as.numeric(Rating) ~ (1 | Artifact) + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## VisOrg_3      3 228.95 237.21 -111.47   222.95
## VisOrg_3_1    5 231.47 245.23 -110.73   221.47 1.4831  2     0.4764

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

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

## Data: VisOrg.ratings_full
## Models:
## VisOrg_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## VisOrg_3_2: as.numeric(Rating) ~ (1 | Artifact) + Semester
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## VisOrg_3      3 228.95 237.21 -111.47   222.95
## VisOrg_3_2    4 229.33 240.34 -110.67   221.33 1.6196  1     0.2031

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

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

## Data: VisOrg.ratings_full
## Models:
## VisOrg_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## VisOrg_3_3: as.numeric(Rating) ~ (1 | Artifact) + Rater
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## VisOrg_3      3 228.95 237.21 -111.47   222.95
## VisOrg_3_3    5 222.97 236.74 -106.48   212.97 9.9784  2   0.006811 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(VisOrg_3_3)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (1 | Artifact) + Rater
##    Data: VisOrg.ratings_full
##
```

```
## REML criterion at convergence: 221.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5008 -0.3334 -0.2599  0.4108  1.8726
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.2937   0.5420
##  Residual             0.1454   0.3813
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.38148    0.09652  24.673
## Rater2       0.27121    0.11645   2.329
## Rater3      -0.08213    0.11645  -0.705
##
## Correlation of Fixed Effects:
##        (Intr) Rater2
## Rater2 -0.611
## Rater3 -0.611  0.504
```

```
qt(p=.05/2, df=116-3, lower.tail=FALSE)
```

```
## [1] 1.98118
```

```
VisOrg_ICC_4 <- (0.2937)/(0.2937+0.1454)
VisOrg_ICC_4
```

```
## [1] 0.6688681
```

```
VisOrg_ICC_3
```

```
## [1] 0.6606838
```

Rating on Text Organization Variable Selection

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

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

```
## Data: TxtOrg.ratings_full
## Models:
## TxtOrg_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## TxtOrg_3_1: as.numeric(Rating) ~ (1 | Artifact) + Sex
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## TxtOrg_3      3 251.45 259.74 -122.73   245.45
## TxtOrg_3_1    5 254.99 268.80 -122.50   244.99 0.4621  2     0.7937
```

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

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

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

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

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

## Data: TxtOrg.ratings_full
## Models:
## TxtOrg_3: as.numeric(Rating) ~ 1 + (1 | Artifact)
## TxtOrg_3_3: as.numeric(Rating) ~ (1 | Artifact) + Rater
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## TxtOrg_3      3 251.45 259.74 -122.73   245.45
## TxtOrg_3_3    5 250.86 264.68 -120.43   240.86 4.5892  2     0.1008
```

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

```
Rubric.names <- sort(unique(tall$Rubric))
tall[c(161,684),] ## just to check that these are the rows with missing ratin
gs...

##       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...
tall.nonmissing[tall.nonmissing$Sex=="--",]

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

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 = FALS
E)
  ## 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 empty, which means you will not be forward-fitting
 the random effect structure of your model. You could just as well run functi
on "bfFixefLMER_F.fnc" or "bfFixefLMER_t.fnc".
## 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 empty, which means you will not be forward-fitting
 the random effect structure of your model. You could just as well run functi
on "bfFixefLMER_F.fnc" or "bfFixefLMER_t.fnc".
## 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 empty, which means you will not be forward-fitting
 the random effect structure of your model. You could just as well run functi
on "bfFixefLMER_F.fnc" or "bfFixefLMER_t.fnc".
## 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 empty, which means you will not be forward-fitting
 the random effect structure of your model. You could just as well run functi
on "bfFixefLMER_F.fnc" or "bfFixefLMER_t.fnc".
## 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 empty, which means you will not be forward-fitting
 the random effect structure of your model. You could just as well run functi
on "bfFixefLMER_F.fnc" or "bfFixefLMER_t.fnc".
## 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 empty, which means you will not be forward-fitting
 the random effect structure of your model. You could just as well run functi
on "bfFixefLMER_F.fnc" or "bfFixefLMER_t.fnc".
## 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 empty, which means you will not be forward-fitting
 the random effect structure of your model. You could just as well run functi
on "bfFixefLMER_F.fnc" or "bfFixefLMER_t.fnc".
## 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
```

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

Examing interactions with Rubric

```
model.01 <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact), data=tall.nonm
issing)

## boundary (singular) fit: see ?isSingular

summary(model.01)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (0 + Rubric | Artifact)
##    Data: tall.nonmissing
##
## 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
```

Adding all factors available in the dataset

```
model.full <- update(model.01, . ~ . + Rater + Semester + Sex + Repeated + Ru
bric)
summary(model.full)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (0 + Rubric | Artifact) + Rater + Semester +

##     Sex + Repeated + Rubric
##    Data: tall.nonmissing
##
## 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
## Repeated1       -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   Reptd1 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
## Repeated1   -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
## Repeated1
## RubrcIntEDA
## RbrcIntrpRs
## RubrcRsrchQ
## RubricSlMth
## RubrcTxtOrg  0.725
## RubricVsOrg  0.680  0.750
```

Using fnc to choose factor

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

## Warning in fitLMER.fnc(model.full, log.file.name = FALSE): Argument "ran.e
ffects" is empty, which means you will not be forward-fitting the random effe
ct structure of your model. You could just as well run function "bfFixefLMER_
F.fnc" or "bfFixefLMER_t.fnc".
## 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(model.back_elim)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (0 + Rubric | Artifact) + Rater + Semester +

##     Rubric
##    Data: tall.nonmissing
##
## 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
```

Considering all interactions among factors available

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

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

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

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

```
summary(model.inter.u)

## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (0 + Rubric | Artifact) + Rater + Semester +

##     Rubric + Rater:Semester + Rater:Rubric + Semester:Rubric +
##     Rater:Semester:Rubric
##    Data: tall.nonmissing
## 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
```

Using fnc function to choose out final model

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

## Warning in fitLMER.fnc(model.inter.u, log.file.name = FALSE): Argument "ra
n.effects" is empty, which means you will not be forward-fitting the random e
ffect structure of your model. You could just as well run function "bfFixefLM
ER_F.fnc" or "bfFixefLMER_t.fnc".
## 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
```

Initial interaction model

```
formula(model.inter.u)

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

Final interaction model

```
formula(model.inter_elim)

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

Final model without interaction

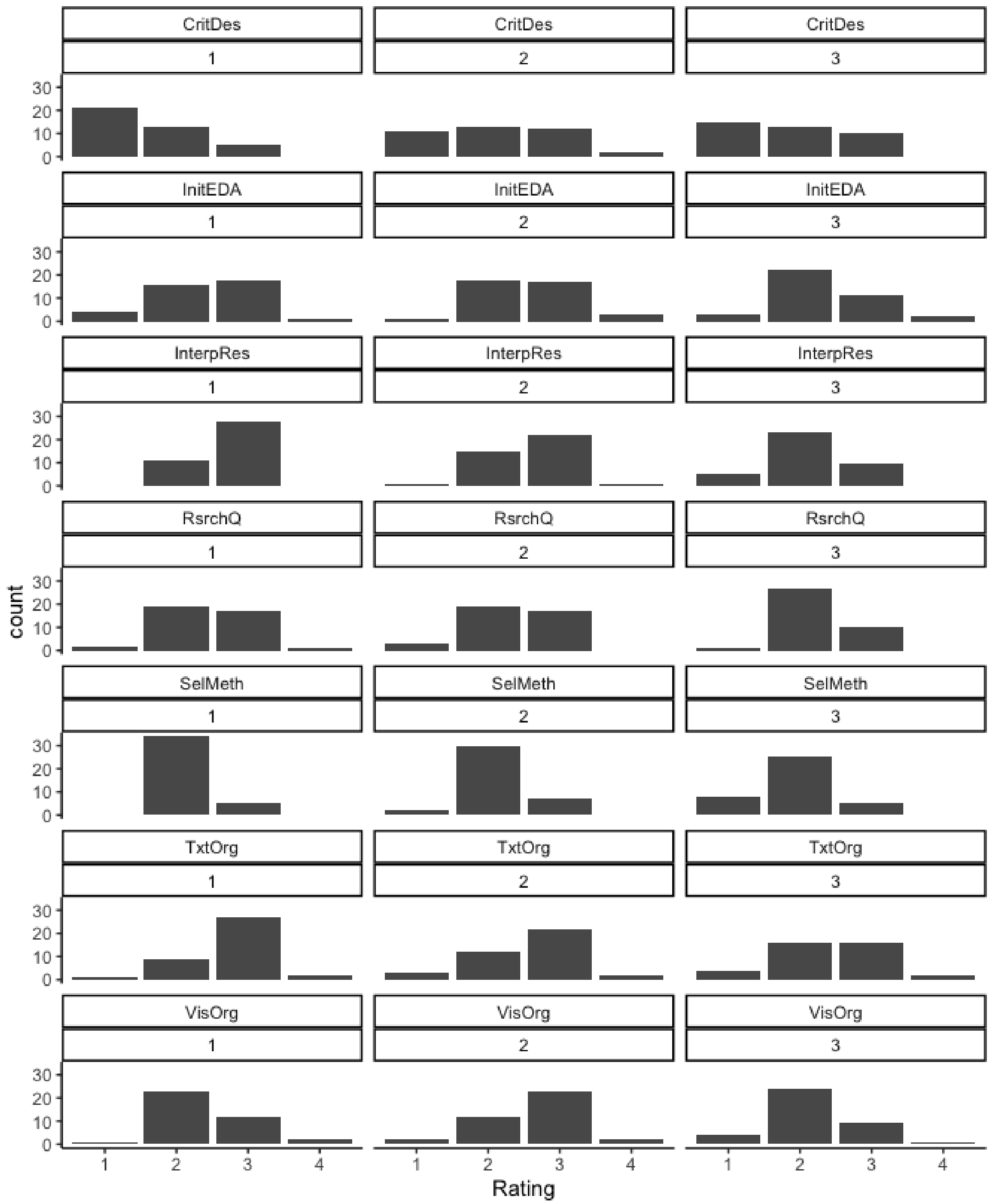
```
formula(model.back_elim)

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

summary(model.inter.u)$varcor

##  Groups   Name              Std.Dev. Corr
##  Artifact RubricCritDes     0.69678
##           RubricInitEDA     0.59378  0.416
##           RubricInterpRes   0.38235  0.324 0.800
##           RubricRsrchQ      0.40551  0.655 0.430 0.723
##           RubricSelMeth     0.25094  0.446 0.639 0.784 0.488
##           RubricTxtOrg      0.50439  0.436 0.649 0.667 0.604 0.622
##           RubricVisOrg      0.50524  0.349 0.727 0.675 0.567 0.346 0.757
##  Residual                   0.43404

summary(model.inter_elim)$varcor

##  Groups   Name              Std.Dev. Corr
##  Artifact RubricCritDes     0.70956
##           RubricInitEDA     0.59565  0.445
##           RubricInterpRes   0.38977  0.354 0.815
##           RubricRsrchQ      0.42371  0.631 0.440 0.716
##           RubricSelMeth     0.25937  0.424 0.601 0.737 0.364
##           RubricTxtOrg      0.51058  0.417 0.637 0.675 0.547 0.636
##           RubricVisOrg      0.50489  0.339 0.715 0.677 0.512 0.376 0.772
##  Residual                   0.43034

summary(model.back_elim)$varcor

##  Groups   Name              Std.Dev. Corr
##  Artifact RubricCritDes     0.74495
##           RubricInitEDA     0.59215  0.467
##           RubricInterpRes   0.41100  0.230 0.749
##           RubricRsrchQ      0.40960  0.588 0.436 0.704
##           RubricSelMeth     0.25493  0.399 0.603 0.736 0.397
##           RubricTxtOrg      0.50612  0.335 0.614 0.691 0.551 0.656
##           RubricVisOrg      0.50886  0.350 0.731 0.679 0.516 0.414 0.752
##  Residual                   0.43513

anova(model.back_elim, model.inter_elim, model.inter.u)

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

## Data: tall.nonmissing
## Models:
## model.back_elim: as.numeric(Rating) ~ (0 + Rubric | Artifact) + Rater + Se
mester + Rubric
## model.inter_elim: as.numeric(Rating) ~ (0 + Rubric | Artifact) + Rater + S
emester + Rubric + Rater:Rubric
```

```
## model.inter.u: as.numeric(Rating) ~ (0 + Rubric | Artifact) + Rater + Seme
ster + Rubric + Rater:Semester + Rater:Rubric + Semester:Rubric + Rater:Semes
ter:Rubric
##                  npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)

## model.back_elim    39 1464.0 1647.2 -693.02   1386.0

## model.inter_elim   51 1454.5 1694.1 -676.26   1352.5 33.526 12   0.000801
***
## model.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
```

```
summary(model.inter_elim )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (0 + Rubric | Artifact) + Rater + Semester +

##     Rubric + Rater:Rubric
##    Data: tall.nonmissing
## 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

qt(p=.05/2, df=810-22, lower.tail=FALSE)

## [1] 1.962979

g <- ggplot(tall.nonmissing, aes(x=Rating)) +
  geom_bar() +
  facet_wrap( ~ Rubric + Rater, nrow=7) + theme_classic()

g
```



Random Effects

The fixed-effects terms we have to work with are: Rater, Semester, Rater:Rubric

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

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkCon
v, :
## Model failed to converge with max|grad| = 0.00347545 (tol = 0.002, compone
nt 1)

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

## boundary (singular) fit: see ?isSingular

anova(m0,mA.u)

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

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

## Data: tall.nonmissing
## Models:
## m0: as.numeric(Rating) ~ (0 + Rubric | Artifact) + Rater + Semester + Rubr
ic + Rater:Rubric
## mA.u: as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + Rater | Artifact
) + Rater + Semester + Rubric + Rater:Rubric
##      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m0     51 1454.5 1694.1 -676.26   1352.5
## mA.u   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

m0 <- model.inter_elim
mA.1 <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) +
             (0 + Semester | Artifact) + Rater +
             Semester + Rubric + Rater:Rubric, data=tall.nonmissing)

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

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkCon
v, :
## Model failed to converge: degenerate Hessian with 1 negative eigenvalues

ss_1 <- getME(mA.1, c("theta","fixef"))
mA_1.u<- update(mA.1, start=ss_1,
             control=lmerControl(optimizer="bobyqa",
                                 optCtrl=list(maxfun=2e5)))

## boundary (singular) fit: see ?isSingular

anova(m0,mA_1.u)
```

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

## Data: tall.nonmissing
## Models:
## m0: as.numeric(Rating) ~ (0 + Rubric | Artifact) + Rater + Semester + Rubr
ic + Rater:Rubric
## mA_1.u: as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + Semester | Art
ifact) + Rater + Semester + Rubric + Rater:Rubric
##         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m0        51 1454.5 1694.1 -676.26   1352.5
## mA_1.u    54 1458.4 1712.0 -675.18   1350.4 2.1534  3     0.5412

m0 <- model.inter_elim
#mA.2 <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) +
             #(0 + Rater | Artifact) +
             #(0 + Rater:Rubric | Artifact) + Rater +
             #Semester + Rubric + Rater:Rubric, data=tall.nonmissing)
#ss_2 <- getME(mA.2, c("theta","fixef"))
#mA_2.u<- update(mA.2, start=ss_2,
             #control=lmerControl(optimizer="bobyqa",
                                  #optCtrl=list(maxfun=2e5)))
#anova(m0,mA_2.u)

m.final <- mA.u

formula(m.final)

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

summary(m.final)$varcor

##  Groups     Name            Std.Dev. Corr

##  Artifact   RubricCritDes   0.70459

##             RubricInitEDA   0.56379   0.318

##             RubricInterpRes 0.31947   0.142  0.674

##             RubricRsrchQ    0.42309   0.500  0.194  0.538

##             RubricSelMeth   0.19556   0.145  0.226  0.376 -0.240

##             RubricTxtOrg    0.50025   0.268  0.437  0.364  0.305  0.213

##             RubricVisOrg    0.48200   0.175  0.504  0.445  0.276 -0.161  0
.537
##  Artifact.1 Rater1          0.11319

##             Rater2          0.33428  -0.486
```

```
##              Rater3          0.30681   0.332  0.663

##  Residual                    0.36700
```

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

```
##                           Estimate Std. Error    t value
## (Intercept)              1.7575529 0.11404064 15.4116374
## Rater2                   0.3660533 0.13918297  2.6300146
## Rater3                   0.1959108 0.12966717  1.5108747
## SemesterS19             -0.1591787 0.07647482 -2.0814529
## RubricInitEDA            0.7394943 0.12995953  5.6901889
## RubricInterpRes          0.9915204 0.12770698  7.7640272
## RubricRsrchQ             0.7261884 0.11793010  6.1577864
## RubricSelMeth            0.4106835 0.12470431  3.2932583
## RubricTxtOrg             1.0157846 0.12999569  7.8139862
## RubricVisOrg             0.6542579 0.13352924  4.8997349
## Rater2:RubricInitEDA    -0.2998068 0.15609073 -1.9207213
## Rater3:RubricInitEDA    -0.2947320 0.15635198 -1.8850546
## Rater2:RubricInterpRes -0.5132359 0.15348542 -3.3438739
## Rater3:RubricInterpRes -0.7148521 0.15364022 -4.6527671
## Rater2:RubricRsrchQ     -0.4874141 0.14722216 -3.3107390
## Rater3:RubricRsrchQ     -0.3223831 0.14726586 -2.1891228
## Rater2:RubricSelMeth    -0.3863783 0.15030982 -2.5705457
## Rater3:RubricSelMeth    -0.3871620 0.14961502 -2.5877213
## Rater2:RubricTxtOrg     -0.5510472 0.15646169 -3.5219304
## Rater3:RubricTxtOrg     -0.4448976 0.15673248 -2.8385794
## Rater2:RubricVisOrg     -0.1049082 0.15861058 -0.6614199
## Rater3:RubricVisOrg     -0.2752251 0.15884843 -1.7326269
```

Appendix D

Can we use data available to predict the rater?

```
library(nnet)
new_tall <- tall
new_tall <- new_tall[complete.cases(new_tall), ]
full.model <- multinom(Rater ~ Semester + Rubric + Rating + Sex, data = new_t
all)
```

```
## # weights:  42 (26 variable)
## initial  value 897.566240
## iter  10 value 880.496833
## iter  20 value 875.048358
## iter  30 value 874.692554
## iter  40 value 874.662278
## final  value 874.662238
## converged
```

```
step.model <- full.model %>% stepAIC(trace = FALSE)
```

```
## # weights:  39 (24 variable)
## initial  value 897.566240
## iter  10 value 880.204426
## iter  20 value 875.088503
## iter  30 value 874.761581
## final  value 874.754033
## converged
## # weights:  24 (14 variable)
## initial  value 897.566240
## iter  10 value 879.134481
## iter  20 value 876.006894
## iter  30 value 875.993851
## iter  30 value 875.993851
## final  value 875.993851
## converged
## # weights:  33 (20 variable)
## initial  value 897.566240
## iter  10 value 890.028630
## iter  20 value 888.509167
## iter  30 value 888.505180
## final  value 888.505128
## converged
## # weights:  36 (22 variable)
## initial  value 897.566240
## iter  10 value 886.191589
## iter  20 value 885.586337
## final  value 885.583621
## converged
## # weights:  24 (14 variable)
## initial  value 897.566240
## iter  10 value 879.134481
## iter  20 value 876.006894
## iter  30 value 875.993851
## iter  30 value 875.993851
## final  value 875.993851
## converged
## # weights:  21 (12 variable)
## initial  value 897.566240
## iter  10 value 878.113201
## iter  20 value 876.076980
## final  value 876.074906
## converged
## # weights:  15 (8 variable)
## initial  value 897.566240
## iter  10 value 888.655863
## final  value 888.522335
## converged
## # weights:  18 (10 variable)
## initial  value 897.566240
## iter  10 value 886.736230
```

```
## final  value 886.717319
## converged
## # weights:  21 (12 variable)
## initial  value 897.566240
## iter  10 value 878.113201
## iter  20 value 876.076980
## final  value 876.074906
## converged
## # weights:  12 (6 variable)
## initial  value 897.566240
## iter  10 value 888.687950
## final  value 888.631020
## converged
## # weights:  15 (8 variable)
## initial  value 897.566240
## iter  10 value 887.181928
## final  value 887.168091
## converged
```

```
coef(step.model)
```

```
##   (Intercept)    Rating2    Rating3    Rating4       SexF       SexM
## 2   -4.011588  0.1839819  0.3020539  0.7384226   3.687653   3.895948
## 3   13.049691 -0.1376085 -0.7758829 -0.5089739 -12.840377 -12.589546
```

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary St
atistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(step.model, type="text")
```

```
##
## ==============================================
##                       Dependent variable:
##                   ----------------------------
##                         2              3
##                        (1)            (2)
## ----------------------------------------------
## Rating2               0.184         -0.138
##                      (0.307)        (0.273)
##
## Rating3               0.302        -0.776***
##                      (0.309)        (0.288)
##
## Rating4               0.738         -0.509
##                      (0.588)        (0.654)
```

```
##
## SexF                    3.688         -12.840
##                        (4.227)       (172.650)
##
## SexM                    3.896         -12.590
##                        (4.227)       (172.650)
##
## Constant               -4.012         13.050
##                        (4.230)       (172.650)
##
## ------------------------------------------------
## Akaike Inf. Crit.    1,776.150      1,776.150
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
step.model.ratio = exp(coef(step.model))
stargazer(step.model, type="text", coef=list(step.model.ratio), p.auto=FALSE)
```

```
##
## ================================================
##                      Dependent variable:
##                  ----------------------------
##                         2              3
##                        (1)            (2)
## ------------------------------------------------
## Rating2                1.202          0.871
##                       (0.307)        (0.273)
##
## Rating3                1.353        0.460***
##                       (0.309)        (0.288)
##
## Rating4                2.093          0.601
##                       (0.588)        (0.654)
##
## SexF                   39.951        0.00000
##                       (4.227)       (172.650)
##
## SexM                   49.203        0.00000
##                       (4.227)       (172.650)
##
## Constant               0.018       464,952.700
##                       (4.230)       (172.650)
##
## ------------------------------------------------
## Akaike Inf. Crit.    1,776.150      1,776.150
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```