Breaking down the external factors that affect Students' Performance from New General Education Program

Wei-Yu Tseng, Department of Statistics and Data Science

weiyut@stat.cmu.edu

Abstract

This paper will investigate ratings students received in the new "General Education" program subcategorized by rubrics of their paper and raters that graded the paper. Using exploratory data analysis conducted by R programming language, we compare the ratings given by each rater, as well as ratings received by each rubric. We found that the distributions of ratings given by different rater may be different, so are the distributions each rubric received. We also found that each rater uses rubrics distinctively. However, the result of analysis had some discrepancy with our exploratory data analysis, and it also failed to include several important factors and. We probably need to dive deeper into the relationship between each rater and the ratings student received, as well as other important factors that are not captured by our analysis.

1 Introduction

Dietrich College at Carnegie Mellon University is running the experiment of a new "General Education" program which includes a combination of required courses and experiences for all undergraduates. To evaluate the effectiveness of this new program, the college hopes to rate student work performed in each of the "General Education" courses each year. The college asks raters from across the college to rate students' project papers on several rubrics. In this report, we will investigate whether the grading standards of each rater and rubrics are similar and discover the relationship between several factors and ratings. In particular, we will examine the following research topics:

1. Explore the distribution of ratings for each rubric and the distribution of ratings given by each rater and check whether are there particular rubrics or raters tend to perform different from others; and
2. Whether for each rubric, the raters generally agree on their ratings or not; and
3. How are various factors in the experiment related to the ratings; and
4. Some interesting discoveries from our experiment.

2 Data

This data came from a recent experiment sampled from a Fall and Spring section of Freshman Statistics at Carnegie Mellon University. The data provides the ratings information across 7 different rubrics on 91 project papers—referred to as "artifacts"— graded by three raters from three different departments, with 13 of these artifacts rated by all three raters and the remaining 78 artifacts graded by only rater, hence the data has $78+13\times3 = 117$ observations, but we have 3 NA's value in the dataset, we drop them to avoid any bias caused in our analysis. The raters did not know which class or which students produced the artifacts that they rated.

The description of each rubric is shown in Table1. The rating scale for all rubrics is shown in Table 2. The variables and information of our dataset is shown in Table 3.

Short Name	Full Name	Description
RsrchQ	Research Question	Given a scenario, the student generates, critiques, or evaluates a relevant empirical research question.
CritDes	Critique Design	Given an empirical research question, the student critiques or evaluates to what extent a study design convincingly answer that question
InitEDA	Initial EDA	Given a data set, the student appropriately describes the data and provides initial Exploratory Data Analysis.
SelMeth	Select Method(s)	Given a data set and a research question, the student selects appropriate method(s) to analyze the data
InterpRes	Interpret Results	The student appropriately interprets the results of the selected method(s).
VisOrg	Visual Organization	The student communicates in an organized, coherent, and effective fashion with visual elements (charts, graphs, tables, etc.).
TxtOrg	Text Organization	The student communicates in an organized, coherent, and effective fashion with text elements (words, sentences, paragraphs, section, and subsection titles, etc.).

Table 1: Rubrics for rating Freshman Statistics projects.

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

Table 2: Rating scale used for all rubrics.

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

Table 3: Variables in the dataset.

Table 4 displayed the summary statistics of ratings each rubric received. Table 5 showed the count of categorical variables.

Full Name	Count	mean	standard deviation	median	min	max	skewness
Research Question	114	2.35	0.59	2	1	4	-0.04
Critique Design	114	1.85	0.83	2	1	4	0.46
Initial EDA	114	2.44	0.70	2	1	4	0.07
Select Method(s)	114	2.05	0.48	2	1	3	0.16
Interpret Results	114	2.48	0.61	2	1	4	-0.51
Visual Organization	114	2.41	0.68	2	1	4	0.15
Text Organization	114	2.60	0.70	3	1	4	-0.50

Table 4: Summary statistics of ratings each rubric received.

Variable	Type 1 Count	Type 2 Count	Type 3 Count
Semester	Fall: 81	Spring: 33	N/A
Sex	Female: 62	Male: 52	N/A
Rater	Rater 1: 38	Rater 2: 38	Rater 3: 38

Table 5: Summary statistics of categorical variables.

3 Methods

1. Explore the distribution of ratings for each rubric and the distribution of ratings given by each rater and check whether are there particular rubrics or raters tend to perform different from others.

To discover whether exists any difference between ratings of each rubric (1) and rating standards of each rater (2), we used barplots and summary statistics to help us gain some insights from the data. We first perform our analysis on the small subset data, which includes only 13 artifacts that are rated by all three raters, and then expand the scale to the entire dataset. For (1), we plot one barplot for each rubric to count the ratings that particular rubric received (see A.1.1 and A.1.2), and we also use summary statistics (also see A.1.1 and A.1.2) of ratings each rubric received including mean, median, maximum, minimum, and even skewness to help us make the judgement. As for (2), we created barplots to count all ratings given by each rater (see A.2.1 and A.2.2) as well as summary statistics (also see A.2.1 and A.2.2) similar to one in (1).

2. Whether for each rubric, the raters generally agree on their ratings or not.

To illustrate whether for each rubric, the raters generally agree on their ratings or not, we used two approaches and compared the results of them. We focused mainly on the small subset data that contains13 artifacts that are rated by all three raters which allow us to compute agreement percentage between raters on the same item. First method is to calculate intraclass correlation, that is, the common correlation among the raters' ratings for each artifact. We performed this by fitting models for each rubric with random intercept with artifact as grouping variable. The second approach is to calculate the percent exact agreement. To do this, we made a 2-way table of counts for the ratings of each pair of raters, on each rubric, since there are three pairs of raters, each rubric will get three tables, and 21 tables in total. For each table, the percentage of observations on the main diagonal is the percent exact agreement between the two raters.

3. How are various factors in the experiment related to the ratings?

To find the relationships between several factors and ratings, we considered mixed effects regression. To be specific, we broke this process into four parts:

a. Adding fixed effects to the seven rubric-specific models we trained in Research Question 2 using subset data which includes only 13 rubrics rated by all 3 raters.

b. Repeat steps in (a) but fit the model using the entire dataset and compared them to part (a).

c. Adding interactions and random effects to the model in Research Question 2 using the entire dataset.

d. Combine the reasonable interactions, fixed effects and random effects into the initial model with rating as response and rubric as the random effect.

For this part we used built-in algorithm from R which compared BIC, AIC, and applied likelihood ratio test to help us select our model. These four parts gave us a model that could help us understand the relationships between the factors and ratings.

4. Some interesting discoveries from our experiment.

For this part, we consider some factors that are not included in our final model of Research Question 3 (part 3). We performed some exploratory data analysis on them and see if whether they violated or disagreed with our findings and conclusions in part 3. We included barplots as our main method of analysis.

4 Results

1. Explore the distribution of ratings for each rubric and the distribution of ratings given by each rater and check whether are there particular rubrics or raters tend to perform different from others.

Section 1 – Distribution of ratings for each rubric

From barplots(see **Figure 1** and **Figure 2** below) and summary statistics (see A.1) of the subset data and full data, we discovered that:

a. Critique Design received relatively low ratings, with many 1's, compared to any other rubric.
b. The distribution of ratings Select Method(s) received is centered on 2, with majority of values being 2.
c. Text Organization and Interpret Results received relatively more rating 3's than any other rubric.
d. Initial EDA, Research Question, and Visual Organization have relatively more rating 2's than any other rubric but Select Method(s).
e. Despite we expanded the analysis scale to full dataset, Select Method(s) still didn't receive any rating 4's.

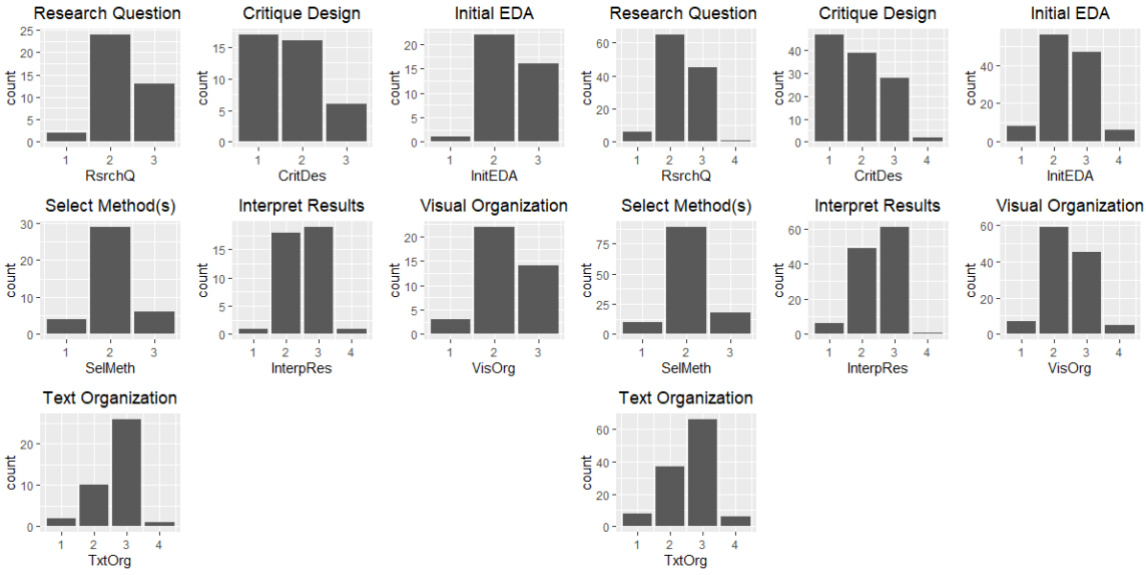


Figure 1: Barplots, Ratings each rubric received for subset data

Figure 2: Barplots, Ratings each rubric received for full data

Section 2 – Distribution of ratings given by each rater

From barplots (see **Figure 3** and **Figure 4** below) and summary statistics (see A.2) of the subset data and full data, we conclude that:

a. Rater 1 and Rater 1 both gave comparable amount of 2's and 3's.
b. Rater 3 gave significantly more 2's than 3's

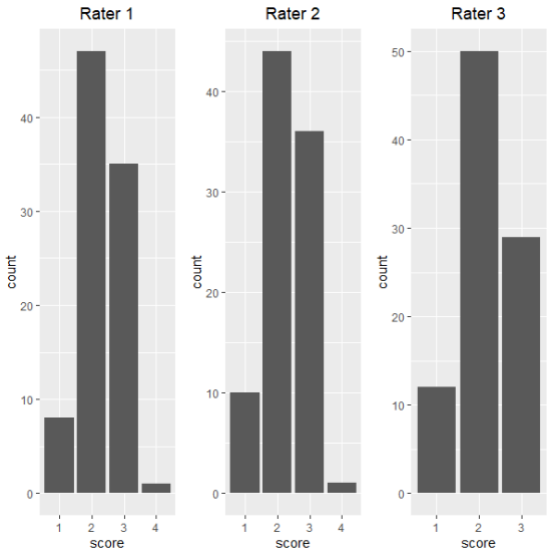


Figure 3: Barplots,
Ratings given by each rater for subset data

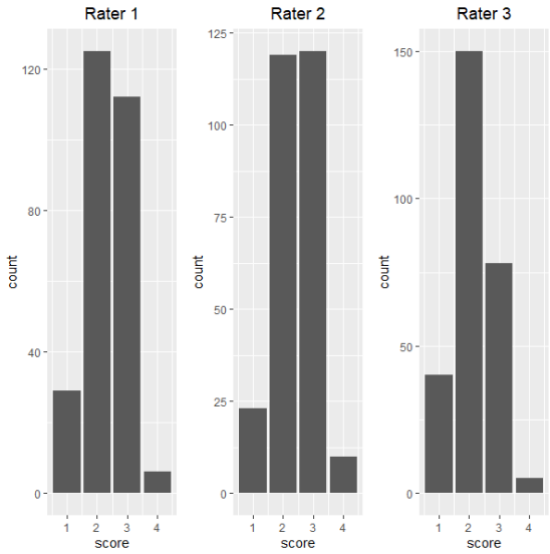


Figure 4: Barplots,
Ratings given by each rater for full data

2. Whether for each rubric, the raters generally agree on their ratings or not.

The following table (see **Table 1**) shows the ICC's we obtained from our mdoels (for details, see B) and the percent exact agreement of each rubric between each pair of raters.

Rubric	Intraclass Correlation (ICC)		Percent Exact Agreement		
	Subset Data	Full Data	Rater 1 & 2	Rater 1 & 3	Rater 2 &3
Research Question	0.19	0.21	0.38	0.77	0.54
Critique Design	0.57	0.67	0.54	0.62	0.69
Initial EDA	0.49	0.69	0.69	0.54	0.85
Select Method(s)	0.52	0.47	0.92	0.62	0.69
Interpret Results	0.23	0.22	0.62	0.54	0.62
Visual Organization	0.59	0.66	0.54	0.77	0.77
Text Organization	0.14	0.19	0.69	0.62	0.54

Table 1: ICC's and Percent Exact Agreement

From the table, we found that:

a. Research Question, Interpret Results, and Text Organization have relatively low ICC's compared to other rubrics.
b. While Interpret Results and Text Organization have low ICC, the percent exact agreements between any pair of raters are noticeable.
c. Although Select Method(s) has relatively high percent exact agreements, it's ICC isn't any prominent compared to others.

3. How are various factors in the experiment related to the ratings?

We gained some inspirations from adding fixed effects, random effects, and interactions to the model we obtained in Research Question 2 and used these findings to decide our final model along with AIC, BIC comparisons and likelihood ratio test (see C.4).

The final model we obtained has the following characteristics:

	Grouping Variable	Variable
Fixed Effect	N/A	(1) Semester
		(2) Rubric
		(3) Rater
		(4) Interaction between Rater and Rubric
Random Effect	Artifact	(5) Rubric
		(6) Rater

We can partition the model into 3 parts:

Part 1: Fixed effects: Rubric (2) + Rater (3) + Interaction between Rater and Rubric (4)

This part spoke that each Rater uses each Rubric in a way that is not like, or even parallel to, other rater's Rubric usage.

Part 2: Fixed effect: Rubric (2) + Random effect: Rubric grouped by Artifact (5)

This told us that not only does each rating each Rubric received varies by a random effect that depends on Artifact but also exists general difference in average score of each Rubric.

Part 3: Fixed effect: Rater (3) + Random effect: Rater grouped by Artifact (6)

This part suggested that each Rater's rating on each Artifact fluctuates by a small random effect that depends on the Artifact.

4. Some interesting discoveries from our experiment.

Our final model shown in the Result of Research Question 3 did not include predictors Semester and Sex, as random effects. Therefore, we performed some exploratory data analysis to see if the selection made by the final model is justifiable.

Section 1 – Variable Sex

Figure 5 displayed the ratings received by each rubric separated by Sex (Male or Female).

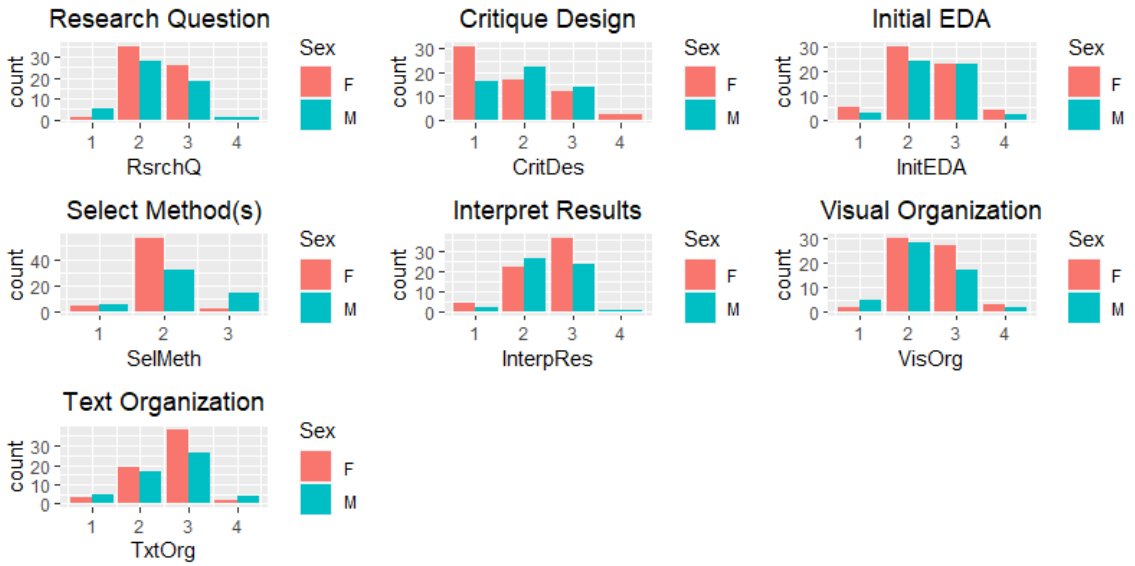


Figure 5: Ratings received by each rubric grouped by Sex.

From Figure 5, we could easily see the difference between each gender (Female or Male) across every rubric, this told us gender may be a possible random effect with rubrics as grouping variable.

Section 2 – Variable Semester

Figure 6 illustrated the ratings received by each rubric separated by Semester (Spring or Fall).

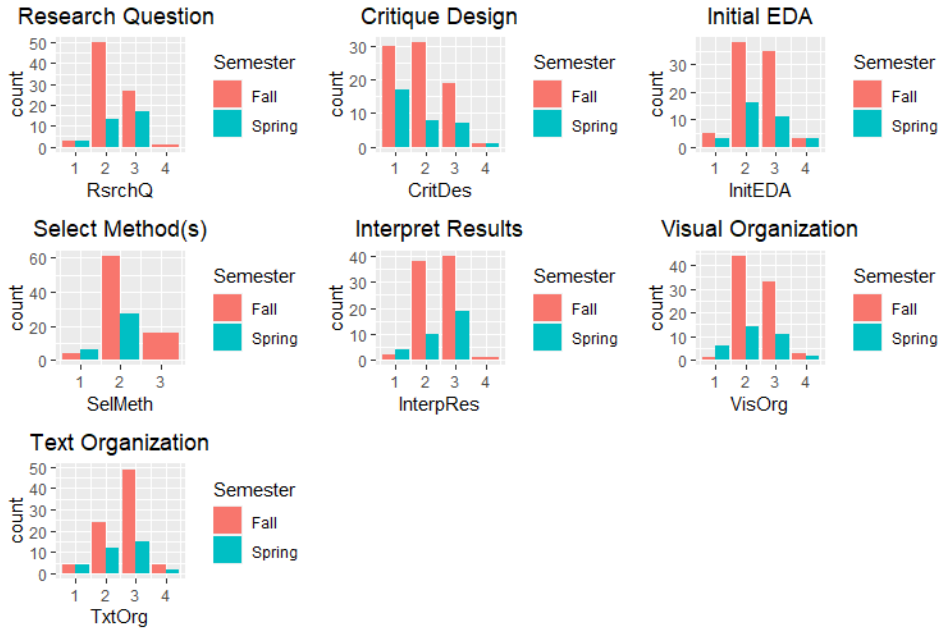


Figure 6: Ratings received by each rubric grouped by Semester.

Similar to what we've found in section 1 (variable Sex), the difference in patterns of barplots could be easily told between Semesters (Spring and Fall), which suggested that Semester is also likely a random effect with rubrics as grouping variable.

5 Discussion

1. Explore the distribution of ratings for each rubric and the distribution of ratings given by each rater and check whether are there particular rubrics or raters tend to perform different from others.

Section 1 – Distribution of ratings for each rubric

From the arguments we made in Result of the Research Question, we could easily tell the distribution of each rubric are quite different. Critique Design is the particular rubric that had distinct patterns. Most students/ artifacts received relatively low ratings in this rubric. This is reasonable since Critique Design seems to be the toughest section among all rubrics, which required experience to hone and master.

Section 2 – Distribution of ratings given by each rater

Based on our graphs, summary statistics (see A.2) and findings in Result, the distribution of ratings given by each rater appears to similar, although Rater 3 seemed to have a different (stricter) rating standard than the other two Raters.

2. Whether for each rubric, the raters generally agree on their ratings or not.

From our result, ICC's and Percent Exact agreement seemed to not agree on each other.

In fact, Percent exact agreement measures the proportion of time two raters have the same rating on each artifact, however, ICC's measure the relationship between raters in a more general way, they measure correlation between raters in that particular rubric. That said, for example, if one rater gave every artifact 2 points in rating while the other rater gave every artifact 3 in rating, we are going to get an ICC equals 1, because whenever one rater gave 2 points in rating, the other one always gave 3 points in rating, however, in this example their percent exact agreement is going to be 0. This explains why for some rubrics the ICC's and percent agreement don't really give comparable answers.

At this point, Percent exact agreement is probably a better measurement over ICC.

The mean of the percent exact agreement across all rubrics of each pair of raters are shown below:

Pair	Rater 1 & Rater 2	Rater 1 & Rater3	Rater 2 & Rater 3
Average Percent Exact Agreement	0.6264	0.6374	0.6703

In general, all three raters seemed to agree in most cases, although Rater 2 and Rater 3 appeared to have slightly more chances to reach consensus in most cases.

3. How are various factors in the experiment related to the ratings?

Our final model displayed the characteristics that matched our exploratory data analysis in discussed in Result of Research Question 1, where both Rubric and Rater played a crucial role in predicting the ratings received by artifacts. Rater and Rubric both presented fixed effect and random effect in the model, as well as the interaction between two variables as a fixed effect. From the barplots of the ratings each rubric received (see **Figure 1 & 2**), we could see that the ratings each rubric received not only followed different distribution, but also seemed to have different mean, which reflected the random effect and fixed effect respectively.

What we need to pay more attention to are the interaction between Rater and Rubric, and the random effect of Rater grouped by Artifact. The former suggested that the Raters are not all interpreting the Rubrics in the same way, while the latter informed that the Raters are not interpreting the evidence in the artifacts in the same way. Since in our barplots (see **Figure 3 & 4**) and argument from Research Question 1, there is not much difference across ratings given by each Rater, these interactions suggest that perhaps the raters should be trained more, to make the raters' ratings more similar to each other.

4. Some interesting discoveries from our experiment.

As we presented in Result, both predictors seem to have different distributions across rubrics, which means they should have random effects grouped by different rubrics. We can also tell the fixed effect of Semester in our final model from the barplots, where in general different semesters appear to have different mean of

ratings (Artifacts in Fall semester always has higher ratings than in Spring). However, since our final model does not include these two variables as random effects, we may need to carefully consider them in the model of the future analysis.

References

Sheather, S.J. (2009), *A Modern Approach to Regression with R*. New York: Springer Science + Business Media LLC.

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

Contents

Techincal Appendix

A

A.1 Is the distribution of ratings for each rubrics pretty much indistinguishable from the other rubrics, or are there rubrics that tend to get especially high or low ratings?

A.1.1 Subset of the data for the 13 artifacts seen by all three raters

Summary Statistics

```
describe(sp_ratings[,7:13]) %>% as.data.frame() %>%
  dplyr::select(-vars, -trimmed, -mad, -se, -vars, -kurtosis ) %>%
  round(2) %>% kbl(booktabs=T,caption="Summary Statistics for Rubric scores subset") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 1: Summary Statistics for Rubric scores subset

	n	mean	sd	median	min	max	range	skew
RsrchQ	39	2.28	0.56	2	1	3	2	0.01
CritDes	39	1.72	0.72	2	1	3	2	0.45
InitEDA	39	2.38	0.54	2	1	3	2	-0.03
SelMeth	39	2.05	0.51	2	1	3	2	0.09
InterpRes	39	2.51	0.60	3	1	4	3	-0.05
VisOrg	39	2.28	0.60	2	1	3	2	-0.19
TxtOrg	39	2.67	0.62	3	1	4	3	-0.95

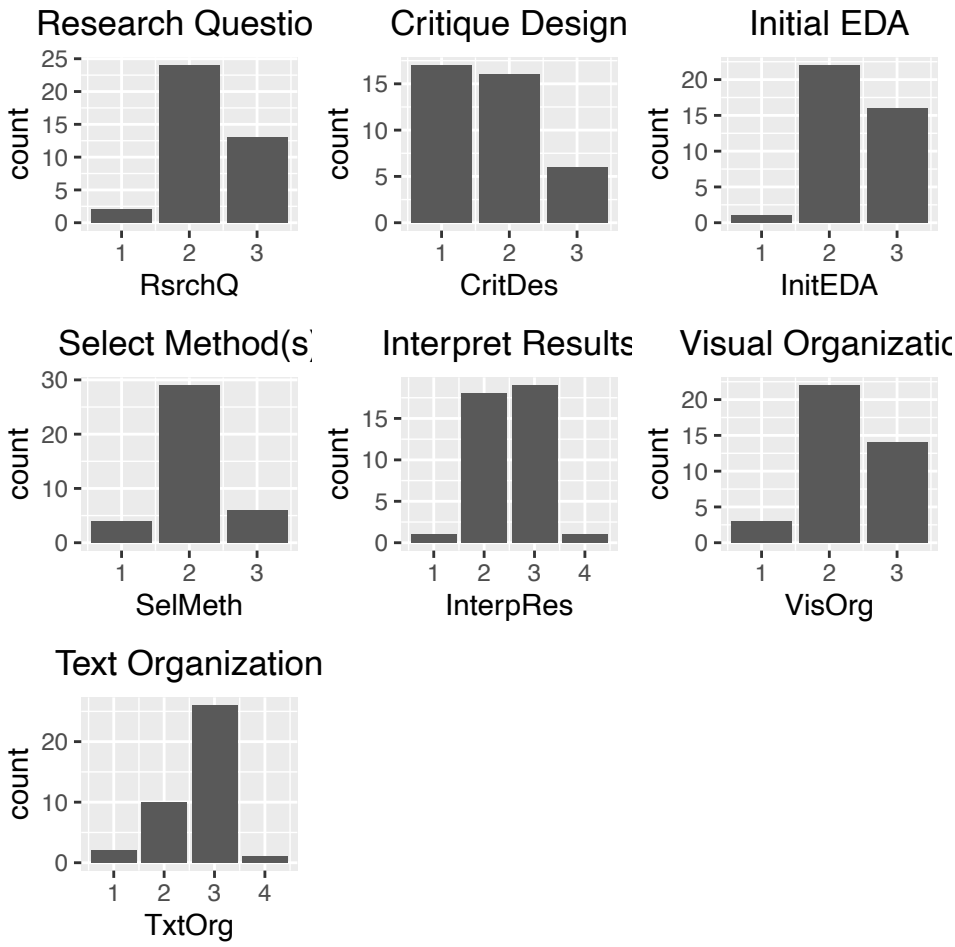
Raters seem to be quite neutral across each rubric, there isn't any significant difference between rubrics except for Text Organization and Critique Design, which have large negative skewness and large positive skewness respectively. Also, raters only gave 4s in Interpret Results and Text Organization.

Barplots

```
Rsr_g = ggplot(sp_ratings, aes(x = RsrchQ)) +
  geom_bar() + labs(title = 'Research Question ') +
  theme(plot.title = element_text(hjust = 0.5))
Cri_g = ggplot(sp_ratings, aes(x = CritDes)) +
  geom_bar() + labs(title = 'Critique Design') +
  theme(plot.title = element_text(hjust = 0.5))
Ini_g = ggplot(sp_ratings, aes(x = InitEDA)) +
  geom_bar() + labs(title = 'Initial EDA') +
  theme(plot.title = element_text(hjust = 0.5))
Sel_g = ggplot(sp_ratings, aes(x = SelMeth)) +
  geom_bar() + labs(title = 'Select Method(s)') +
  theme(plot.title = element_text(hjust = 0.5))
Int_g = ggplot(sp_ratings, aes(x = InterpRes)) +
  geom_bar() + labs(title = 'Interpret Results') +
```

```
  theme(plot.title = element_text(hjust = 0.5))
Vis_g = ggplot(sp_ratings, aes(x = VisOrg)) +
  geom_bar() + labs(title = 'Visual Organization') +
  theme(plot.title = element_text(hjust = 0.5))
Txt_g = ggplot(sp_ratings, aes(x = TxtOrg)) +
  geom_bar() + labs(title = 'Text Organization') +
  theme(plot.title = element_text(hjust = 0.5))


fig1 = grid.arrange(Rsr_g, Cri_g, Ini_g, Sel_g, Int_g, Vis_g, Txt_g, ncol = 3)
```



Quite similar as what we've seen in summary statistics of these rubrics, most rubrics have comparable amount of 2's and 3's for scores, while only Critique Design has many scores of 1's, and most artifacts get 3 points in Text Organization.

Another interesting finding is that in summary statistics we couldn't really tell Select Method(s)' distribution because almost every rubric is centered around 2. However, the scores these artifacts received are not only just centered around 2, but they are also mostly 2's.

A.1.2 Entire dataset

Summary Statistics

```
describe(ratings[,7:12]) %>% as.data.frame() %>%
  dplyr::select(-vars, -trimmed, -mad, -se, -vars,-kurtosis ) %>%
  round(2) %>% kbl(booktabs=T,caption="Summary Statistics for Rubric scores") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 2: Summary Statistics for Rubric scores

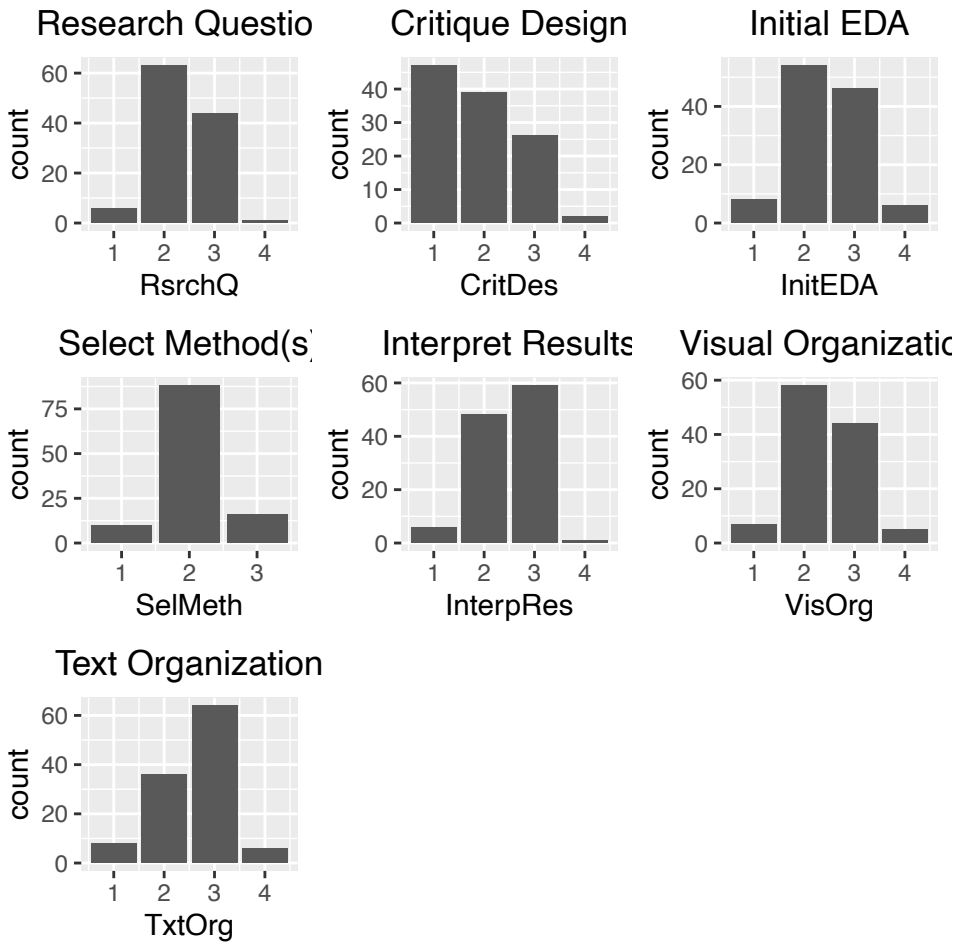
	n	mean	sd	median	min	max	range	skew
RsrchQ	114	2.35	0.59	2	1	4	3	-0.04
CritDes	114	1.85	0.83	2	1	4	3	0.46
InitEDA	114	2.44	0.70	2	1	4	3	0.07
SelMeth	114	2.05	0.48	2	1	3	2	0.16
InterpRes	114	2.48	0.61	3	1	4	3	-0.51
VisOrg	114	2.41	0.68	2	1	4	3	0.15

Now we look at the full dataset, there is no much difference from 13-repeated subset, except for Interpret Results whose skewness plummeted to -0.52 compared to its -0.05 in the subset.

Barplots

```
Rsr_g = ggplot(ratings, aes(x = RsrchQ)) +
  geom_bar() + labs(title = 'Research Question ') +
  theme(plot.title = element_text(hjust = 0.5))
Cri_g = ggplot(ratings, aes(x = CritDes)) +
  geom_bar() + labs(title = 'Critique Design') +
  theme(plot.title = element_text(hjust = 0.5))
Ini_g = ggplot(ratings, aes(x = InitEDA)) +
  geom_bar() + labs(title = 'Initial EDA') +
  theme(plot.title = element_text(hjust = 0.5))
Sel_g = ggplot(ratings, aes(x = SelMeth)) +
  geom_bar() + labs(title = 'Select Method(s)') +
  theme(plot.title = element_text(hjust = 0.5))
Int_g = ggplot(ratings, aes(x = InterpRes)) +
  geom_bar() + labs(title = 'Interpret Results') +
  theme(plot.title = element_text(hjust = 0.5))
Vis_g = ggplot(ratings, aes(x = VisOrg)) +
  geom_bar() + labs(title = 'Visual Organization') +
  theme(plot.title = element_text(hjust = 0.5))
Txt_g = ggplot(ratings, aes(x = TxtOrg)) +
  geom_bar() + labs(title = 'Text Organization') +
  theme(plot.title = element_text(hjust = 0.5))


fig2 = grid.arrange(Rsr_g, Cri_g, Ini_g, Sel_g, Int_g, Vis_g, Txt_g, ncol = 3)
```



Similar to what we've seen in A.1.2, the barplots of 13-repeated subset, Critique Design still has many 1's, the distribution of Select Method(s) is still close to symmetric, centered on 2's with majority being 2. While the skewness of Interpret Results changes, the new distribution still looks alike to the old one, a large proportion of data are still around 2's and 3's, although it has more 2's and 1's compared to the small subset.

A.2 Is the distribution of ratings given by each rater pretty much indistinguishable from the other raters, or are there raters that tend to give especially high or low ratings?

A.2.1 Subset of the data for the 13 artifacts seen by all three raters

Summary Statistics

```
sp_long_2 <- sp_ratings %>%
  pivot_longer(cols = c(RsrchQ, CritDes, InitEDA, SelMeth, InterpRes, VisOrg, TxtOrg),
               names_to = "rubric", values_to = 'score')


r1_s = sp_long_2 %>% filter(Rater == 1)
r2_s = sp_long_2 %>% filter(Rater == 2)
r3_s = sp_long_2 %>% filter(Rater == 3)
```

```r
describe(r1_s$score) %>% as.data.frame() %>%
  dplyr::select(-vars, -trimmed, -mad, -se, -vars ,-kurtosis) %>%
  round(2) %>% kbl(booktabs=T,caption="Summary Statistics of Rater 1's Rubric scores ") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 3: Summary Statistics of Rater 1's Rubric scores

	n	mean	sd	median	min	max	range	skew
X1	91	2.32	0.65	2	1	4	3	-0.16

```r
describe(r2_s$score) %>% as.data.frame() %>%
  dplyr::select(-vars, -trimmed, -mad, -se, -vars ,-kurtosis) %>%
  round(2) %>% kbl(booktabs=T,caption="Summary Statistics of Rater 2's Rubric scores ") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 4: Summary Statistics of Rater 2's Rubric scores

	n	mean	sd	median	min	max	range	skew
X1	91	2.31	0.68	2	1	4	3	-0.24

```r
describe(r3_s$score) %>% as.data.frame() %>%
  dplyr::select(-vars, -trimmed, -mad, -se, -vars ,-kurtosis) %>%
  round(2) %>% kbl(booktabs=T,caption="Summary Statistics of Rater 3's Rubric scores ") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 5: Summary Statistics of Rater 3's Rubric scores

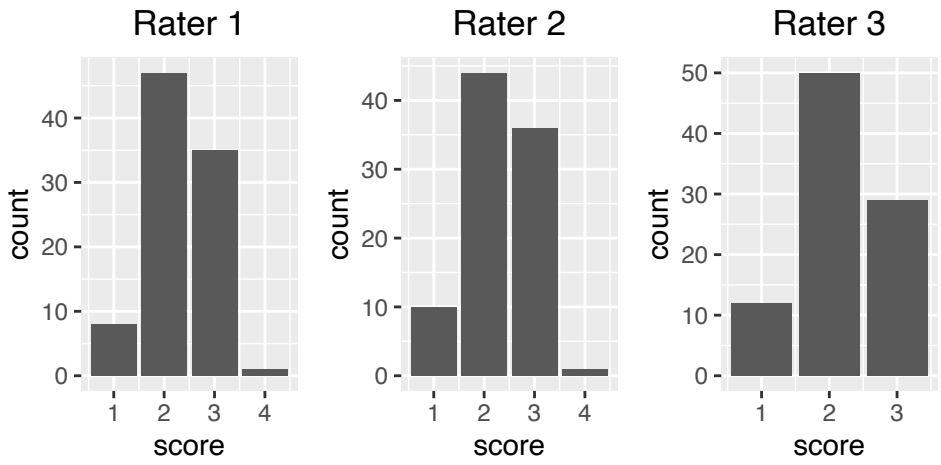
	n	mean	sd	median	min	max	range	skew
X1	91	2.19	0.65	2	1	3	2	-0.19

The summary statistics shows there is no significant difference between 3 raters, they share similar means, standard deviations,medians and even skenesses. However, in this small subset, rater 3 didn't give out any 4's.

Barplots

```r
r1_g = ggplot(r1_s, aes(x = score)) +
  geom_bar() + labs(title = 'Rater 1') + theme(plot.title = element_text(hjust = 0.5))
r2_g = ggplot(r2_s, aes(x = score)) +
  geom_bar() + labs(title = 'Rater 2') + theme(plot.title = element_text(hjust = 0.5))
r3_g = ggplot(r3_s, aes(x = score)) +
  geom_bar() + labs(title = 'Rater 3') + theme(plot.title = element_text(hjust = 0.5))

fig3 = grid.arrange(r1_g, r2_g, r3_g, ncol = 3)
```



There is no noticeable difference between 3 raters.

A.2.2 Entire dataset

Summary Statistics

```
long_2 <- ratings %>%
  pivot_longer(cols = c(RsrchQ, CritDes, InitEDA, SelMeth, InterpRes, VisOrg, TxtOrg),
               names_to = "rubric", values_to = 'score')


r1 = long_2 %>% filter(Rater == 1)
r2 = long_2 %>% filter(Rater == 2)
r3 = long_2 %>% filter(Rater == 3)

describe(r1$score) %>% as.data.frame() %>%
  dplyr::select(-vars, -trimmed, -mad, -se, -vars ,-kurtosis) %>%
  round(2) %>% kbl(booktabs=T,caption="Summary Statistics of Rater 1's Rubric scores ") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 6: Summary Statistics of Rater 1's Rubric scores

	n	mean	sd	median	min	max	range	skew
X1	266	2.35	0.7	2	1	4	3	-0.2

```
describe(r2$score) %>% as.data.frame() %>%
  dplyr::select(-vars, -trimmed, -mad, -se, -vars ,-kurtosis) %>%
  round(2) %>% kbl(booktabs=T,caption="Summary Statistics of Rater 2's Rubric scores ") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 7: Summary Statistics of Rater 2's Rubric scores

	n	mean	sd	median	min	max	range	skew
X1	266	2.44	0.7	2	1	4	3	-0.19

```r
describe(r3$score) %>% as.data.frame() %>%
  dplyr::select(-vars, -trimmed, -mad, -se, -vars ,-kurtosis) %>%
  round(2) %>% kbl(booktabs=T,caption="Summary Statistics of Rater 3's Rubric scores ") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 8: Summary Statistics of Rater 3's Rubric scores

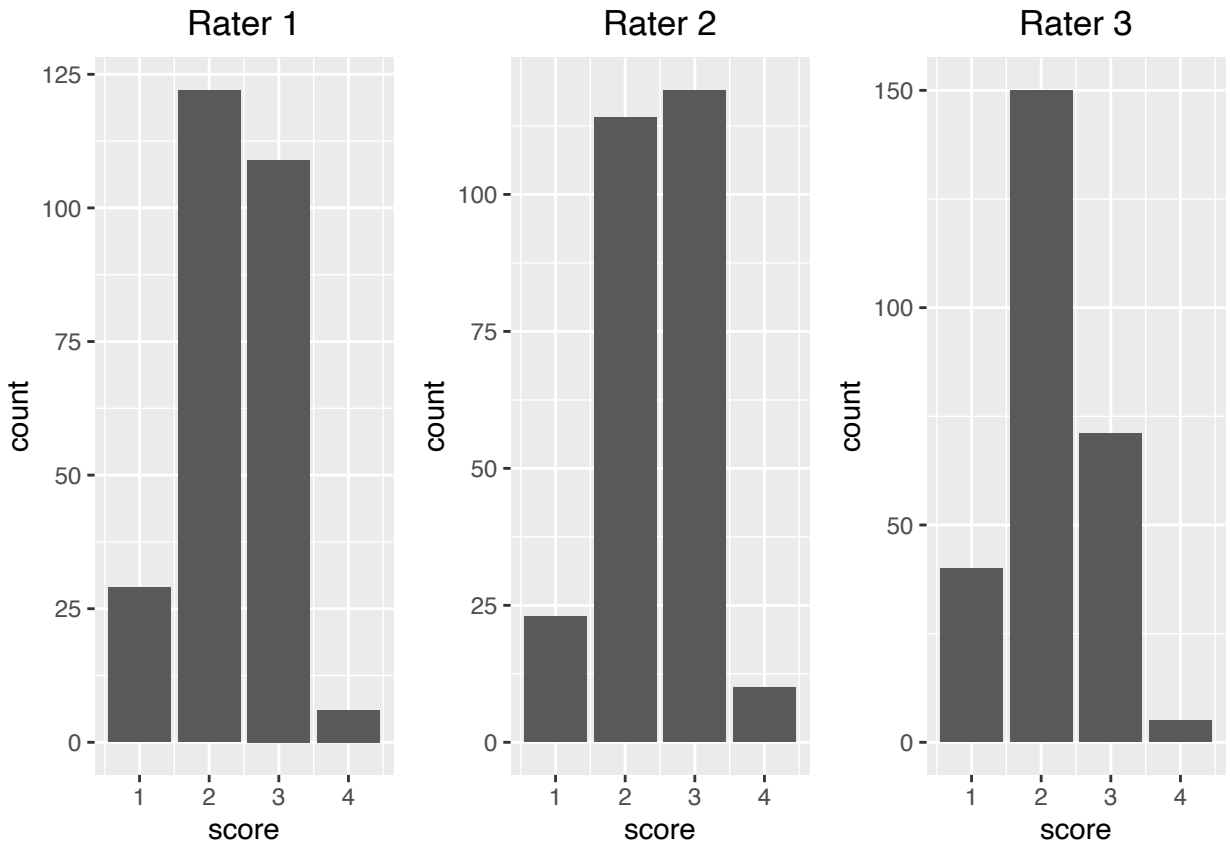
	n	mean	sd	median	min	max	range	skew
X1	266	2.15	0.69	2	1	4	3	0.14

The full dataset delivers the similar information as the subset dataset, the only difference is that now all raters gave 4's in rating.

Barplots

```r
r1_g = ggplot(r1, aes(x = score)) +
  geom_bar() + labs(title = 'Rater 1') + theme(plot.title = element_text(hjust = 0.5))
r2_g = ggplot(r2, aes(x = score)) +
  geom_bar() + labs(title = 'Rater 2') + theme(plot.title = element_text(hjust = 0.5))
r3_g = ggplot(r3, aes(x = score)) +
  geom_bar() + labs(title = 'Rater 3') + theme(plot.title = element_text(hjust = 0.5))

fig4 = grid.arrange(r1_g, r2_g, r3_g, ncol = 3)
```



The barplots of the full dataset give a slightly different answer compared to the barplots of subset data, Rater 1 and Rater 2 both gave comparable amount of 2's and 3's while Rater 3 gave significantly more 2's than 3's.

B

B.1 For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees? Or do they all disagree?

```
sp_ratings$Artifact <- sp_ratings$Artifact %>% as.factor()
lmer1 <- lmer(RsrchQ ~ (1|Artifact) + 1 ,data=sp_ratings)
ICC1 = 0.05983/(0.05983 + 0.25641)
lmer2 <- lmer(CritDes ~ (1|Artifact) + 1 ,data=sp_ratings)
ICC2 <- 0.3091/(0.3091 + 0.2308)
lmer3 <- lmer(InitEDA ~ (1|Artifact) + 1 ,data=sp_ratings)
ICC3 <- 0.1496/(0.1496 + 0.1538)
lmer4 <- lmer(SelMeth ~ (1|Artifact) + 1 ,data=sp_ratings)
ICC4 <- 0.1396/(0.1396 + 0.1282)
lmer5 <- lmer(InterpRes ~ (1|Artifact) + 1 ,data=sp_ratings)
ICC5 <- 0.08405/(0.08405 + 0.28205)
lmer6 <- lmer(VisOrg ~ (1|Artifact) + 1 ,data=sp_ratings)
ICC6 <- 0.2236/(0.2236 + 0.1538)
lmer7 <- lmer(TxtOrg ~ (1|Artifact) + 1 ,data=sp_ratings)
ICC7 <- 0.05556/(0.05556 + 0.33333)
ICC_sub  = c(ICC1, ICC2, ICC3, ICC4, ICC5, ICC6, ICC7)
```

```
ratings$Artifact <- ratings$Artifact %>% as.factor()
lmer1.l <- lmer(RsrchQ ~ (1|Artifact) + 1 ,data=ratings)
ICC1_l = 0.07372/(0.07372 + 0.27797)
lmer2.l <- lmer(CritDes ~ (1|Artifact) + 1 ,data=ratings)
ICC2_l = 0.4963/(0.4963 + 0.2411)
lmer3.l <- lmer(InitEDA ~ (1|Artifact) + 1 ,data=ratings)
ICC3_l = 0.3628/(0.3628 + 0.1655)
lmer4.l <- lmer(SelMeth ~ (1|Artifact) + 1 ,data=ratings)
ICC4_l = 0.1108/(0.1108 + 0.1240)
lmer5.l <- lmer(InterpRes ~ (1|Artifact) + 1 ,data=ratings)
ICC5_l = 0.08219/(0.08219 + 0.29136)
lmer6.l <- lmer(VisOrg ~ (1|Artifact) + 1 ,data=ratings)
ICC6_l = 0.3092/(0.3092+0.1588)
lmer7.l <- lmer(TxtOrg ~ (1|Artifact) + 1 ,data=ratings)
ICC7_l = 0.09145/(0.09145 + 0.39503)

ICC_full  = c(ICC1_l, ICC2_l, ICC3_l, ICC4_l, ICC5_l, ICC6_l, ICC7_l)
```

```
rubrics = r1$rubric %>% unique()
ICC = data.frame(Rubric = rubrics, ICC_sub = ICC_sub,  ICC_full = ICC_full)
ICC %>% kbl(booktabs=T,caption="ICCs for each Rubric scores ") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 9: ICCs for each Rubric scores

Rubric	ICC_sub	ICC_full
RsrchQ	0.1891918	0.2096164
CritDes	0.5725134	0.6730404
InitEDA	0.4930784	0.6867310
SelMeth	0.5212845	0.4718910
InterpRes	0.2295821	0.2200241
VisOrg	0.5924748	0.6606838
TxtOrg	0.1428682	0.1879831

For both subset data and full dataset, the ICC's are higher in rubrics Critique Design, Initial EDA, Select Method(s) and Interpret Results, Visual Organization.

```
r1_s$score = factor(r1_s$score, levels = 1:4)
r2_s$score = factor(r2_s$score, levels = 1:4)
r3_s$score = factor(r3_s$score, levels = 1:4)
per_exp_agree_r1r2 = NULL
per_exp_agree_r1r3 = NULL
per_exp_agree_r2r3 = NULL
for (i in 1:7) {
  t_r1r2 = table(filter(r1_s, rubric == rubrics[i])$score,
                 filter(r2_s, rubric == rubrics[i])$score)
  per_exp_agree_r1r2[i] = sum(diag(t_r1r2))/sum(t_r1r2)
  t_r1r3 = table(filter(r1_s, rubric == rubrics[i])$score,
                 filter(r3_s, rubric == rubrics[i])$score)
  per_exp_agree_r1r3[i] = sum(diag(t_r1r3))/sum(t_r1r3)
  t_r2r3 = table(filter(r2_s, rubric == rubrics[i])$score,
                 filter(r3_s, rubric == rubrics[i])$score)
  per_exp_agree_r2r3[i] = sum(diag(t_r2r3))/sum(t_r2r3)
}
`row.names<-`(cbind(per_exp_agree_r1r2,
                    per_exp_agree_r1r3,
                    per_exp_agree_r2r3),
              rubrics) %>%
  kbl(booktabs=T,caption="ICCs for each Rubric scores ") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 10: ICCs for each Rubric scores

	per_exp_agree_r1r2	per_exp_agree_r1r3	per_exp_agree_r2r3
RsrchQ	0.3846154	0.7692308	0.5384615
CritDes	0.5384615	0.6153846	0.6923077
InitEDA	0.6923077	0.5384615	0.8461538
SelMeth	0.9230769	0.6153846	0.6923077
InterpRes	0.6153846	0.5384615	0.6153846
VisOrg	0.5384615	0.7692308	0.7692308
TxtOrg	0.6923077	0.6153846	0.5384615

Percent exact agreement measures the proportion of time two raters have the same rating on each artifact, however, ICC's measure the relationship between raters in a more general way, they measure correlation between raters in that particular rubric. That said, for example, if one rater gave every artifact 2 points in rating while the other rater gave every artifact 3 in rating, we are going to get an ICC equals 1, because whenever one rater gave 2 points in rating, the other one always gave 3 points in rating, however, in this example their percent exact agreement is going to be 0. This explains why for some rubrics the ICC's and percent agreement don't really give comparable answers.

At this point, Percent exact agreement is probably a better measurement.

```
Q3 = data.frame(Rubric = rubrics,
                ICC_subset = ICC_sub %>% round(2),
                ICC_full = ICC_full %>% round(2),
                R12 = per_exp_agree_r1r2 %>% round(2),
                R13 = per_exp_agree_r1r3 %>% round(2),
                R23 = per_exp_agree_r2r3 %>% round(2))
Q3%>% kbl(booktabs=T,caption=" ") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 11:

Rubric	ICC_subset	ICC_full	R12	R13	R23
RsrchQ	0.19	0.21	0.38	0.77	0.54
CritDes	0.57	0.67	0.54	0.62	0.69
InitEDA	0.49	0.69	0.69	0.54	0.85
SelMeth	0.52	0.47	0.92	0.62	0.69
InterpRes	0.23	0.22	0.62	0.54	0.62
VisOrg	0.59	0.66	0.54	0.77	0.77
TxtOrg	0.14	0.19	0.69	0.62	0.54

C

C.1 Adding fixed effects to the seven rubric-specific models using just the data from the 13 common artifacts that al three raters saw

```
tall.13 <- tall[grep("O",tall$Artifact),]
Rubric.names <- sort(unique(tall$Rubric))
model.formula.13 <- as.list(rep(NA,7))
names(model.formula.13) <- Rubric.names
for (i in Rubric.names) {
## fit each base model
  rubric.data <- tall.13[tall.13$Rubric==i,]
  tmp <- lmer(as.numeric(Rating) ~ -1 + as.factor(Rater) + Semester +
                Sex + (1|Artifact),data=rubric.data,REML=FALSE)
## do backwards elimination
  tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)
## check to see if the raters are significantly different from one another
  tmp.single_intercept <- update(tmp.back_elim, . ~ . + 1 - as.factor(Rater))
  pval <- anova(tmp.single_intercept,tmp.back_elim)$"Pr(>Chisq)"[2]
```

```
## choose the best model
  if (pval<=0.05) {
    tmp_final <- tmp.back_elim
    }
  else {
          tmp_final <- tmp.single_intercept
    }
## and add to list...
  model.formula.13[[i]] <- formula(tmp_final)
}
```

```
model.formula.13
```

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

It looks like we don't need to add any fixed effects or interactions to the models for each rubric, using only the data reduced to the 13 common rubrics.

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

```
tall.nonmissing <- tall[-c(161,684),]
tall.nonmissing <- tall.nonmissing[tall.nonmissing$Sex!="--",]
model.formula.alldata <- as.list(rep(NA,7))
names(model.formula.alldata) <- Rubric.names
for (i in Rubric.names) {
## fit each base model
  rubric.data <- tall.nonmissing[tall.nonmissing$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 = FALSE)
```

```
## check to see if the raters are significantly different from one another
  tmp.single_intercept <- update(tmp.back_elim, . ~ . + 1 - as.factor(Rater))
  pval <- anova(tmp.single_intercept,tmp.back_elim)$"Pr(>Chisq)"[2]
## choose the best model
if (pval<=0.05) {
    tmp_final <- tmp.back_elim
  }
  else {
    tmp_final <- tmp.single_intercept
  }
## and add to list...
    model.formula.alldata[[i]] <- formula(tmp_final)
}
```

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

It appears to be that rater might be an important fixed effect given the model trained on full dataset.

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

Now we see there are some differences among the models: For InitEDA, RsrchQ and SelMeth, the models are just the simple random-intercept models. For the other four, the models are a little more complex. We should examine each of these 4 models to see (a) if the fixed effects make sense to us; and (2) if there are any interactions or additional random effects to consider.

C.3.1 Critique Design

```
fla <- formula(model.formula.alldata[["CritDes"]])
tmp <- lmer(fla,data=tall.nonmissing[tall.nonmissing$Rubric=="CritDes",])
round(summary(tmp)$coef,2)
```

```
##                   Estimate Std. Error t value
## as.factor(Rater)1     1.69       0.12   13.99
## as.factor(Rater)2     2.12       0.12   17.34
## as.factor(Rater)3     1.91       0.12   15.83
```

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

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

```
## Data: tall.nonmissing[tall.nonmissing$Rubric == "CritDes", ]
## Models:
## tmp.single_intercept: as.numeric(Rating) ~ (1 | Artifact)
## tmp: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##                       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    3 280.86 289.12 -137.43   274.86
## tmp                     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
```

The ANOVA stated that we really need "Rater" as a factor.

```
m0 <- tmp ## Null hypothesis
mA <- update(m0, . ~ . + (as.factor(Rater)|Artifact))
```

```
## Error: number of observations (=116) <= number of random effects (=270) for term (as.factor(Rater) |
```

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

```
## Error in update(mA, . ~ . - (1 | Artifact)): object 'mA' not found
```

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

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

```
summary(tmp)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "CritDes", ]
##
## 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
## as.factor(Rater)1   1.6926     0.1210   13.99
## as.factor(Rater)2   2.1184     0.1222   17.34
## as.factor(Rater)3   1.9144     0.1210   15.83
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2
## as.fctr(R)2 0.245
## as.fctr(R)3 0.243  0.245
```

The random effect here are not needed.

C.3.2 Interpret Result

```
fla <- formula(model.formula.alldata[["InterpRes"]])
tmp <- lmer(fla,data=tall.nonmissing[tall.nonmissing$Rubric=="InterpRes",])
round(summary(tmp)$coef,2)
```

```
##                   Estimate Std. Error t value
## as.factor(Rater)1     2.71       0.09   30.19
## as.factor(Rater)2     2.59       0.09   28.87
## as.factor(Rater)3     2.16       0.09   24.12
```

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

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

## Data: tall.nonmissing[tall.nonmissing$Rubric == "InterpRes", ]
## Models:
## tmp.single_intercept: as.numeric(Rating) ~ (1 | Artifact)
## tmp: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##                      npar    AIC    BIC   logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    3 220.09 228.38 -107.048   214.09
## tmp                     5 203.66 217.47  -96.831   193.66 20.433  2  3.657e-05
##
## tmp.single_intercept
## tmp                  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA stated that we really need "Rater" as a factor.

```
m0 <- tmp ## Null hypothesis
mA <- update(m0, . ~ . + (as.factor(Rater)|Artifact))
```

```
## Error: number of observations (=117) <= number of random effects (=273) for term (as.factor(Rater) |
```

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

```
## Error in update(mA, . ~ . - (1 | Artifact)): object 'mA' not found
```

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

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

```
summary(tmp)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "InterpRes", ]
##
## 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
## as.factor(Rater)1  2.70517    0.08961   30.19
## as.factor(Rater)2  2.58701    0.08961   28.87
## as.factor(Rater)3  2.16116    0.08961   24.12
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2
## as.fctr(R)2 0.063
## as.fctr(R)3 0.063  0.063
```

The random effect here are not needed.

C.3.3 Select Method(s)

```
fla <- formula(model.formula.alldata[["SelMeth"]])
tmp <- lmer(fla,data=tall.nonmissing[tall.nonmissing$Rubric=="SelMeth",])
round(summary(tmp)$coef,2) ## fixed effects and their t-values
```

```
##                   Estimate Std. Error t value
## as.factor(Rater)1     3.22       0.45    7.11
## as.factor(Rater)2     3.19       0.45    7.05
## as.factor(Rater)3     3.00       0.44    6.75
## SemesterS19          -0.32       0.10   -3.12
## SexF                 -1.04       0.45   -2.28
## SexM                 -0.91       0.45   -2.02
```

The ANOVA stated that we really need "Rater" as a factor.

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

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

## Data: tall.nonmissing[tall.nonmissing$Rubric == "SelMeth", ]
## Models:
## tmp.single_intercept: as.numeric(Rating) ~ Semester + Sex + (1 | Artifact)
## tmp: as.numeric(Rating) ~ as.factor(Rater) + Semester + Sex + (1 |
## tmp:     Artifact) - 1
##                      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    6 147.94 164.51 -67.968   135.94
## tmp                     8 144.52 166.62 -64.260   128.52 7.4154  2    0.02453 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
tmp.fixed_interactions <- update(tmp, . ~ . + as.factor(Rater)*Semester - Semester)
anova(tmp,tmp.fixed_interactions)
```

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

## Data: tall.nonmissing[tall.nonmissing$Rubric == "SelMeth", ]
## Models:
## tmp: as.numeric(Rating) ~ as.factor(Rater) + Semester + Sex + (1 |
## tmp:     Artifact) - 1
## tmp.fixed_interactions: as.numeric(Rating) ~ as.factor(Rater) + Sex + (1 | Artifact) +
## tmp.fixed_interactions:     as.factor(Rater):Semester - 1
##                         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp                        8 144.52 166.62 -64.260   128.52
## tmp.fixed_interactions    10 145.77 173.40 -62.887   125.77 2.7467  2     0.2533
```

The fixed-effect interactions are not needed based on the result of anova.

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

```
## Error: number of observations (=117) <= number of random effects (=182) for term (Semester | Artifac
```

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

```
## Error in update(mA, . ~ . - (1 | Artifact)): object 'mA' not found
```

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

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

```r
m0 <- tmp ## Null hypothesis
mA <- update(m0, . ~ . + (as.factor(Rater)|Artifact)) #
```

```
## Error: number of observations (=117) <= number of random effects (=273) for term (as.factor(Rater) |
```

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

```
## Error in update(mA, . ~ . - (1 | Artifact)): object 'mA' not found
```

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

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

```r
summary(tmp)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + Semester + Sex + (1 |
##     Artifact) - 1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "SelMeth", ]
##
## REML criterion at convergence: 144.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.09631 -0.34555 -0.06849  0.33489  2.66067
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.09013  0.3002
##  Residual             0.10714  0.3273
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1   3.2227     0.4531   7.113
## as.factor(Rater)2   3.1946     0.4530   7.051
## as.factor(Rater)3   3.0000     0.4441   6.755
## SemesterS19        -0.3195     0.1025  -3.119
## SexF               -1.0352     0.4536  -2.282
## SexM               -0.9136     0.4523  -2.020
##
```

```
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2 a.(R)3 SmsS19 SexF
## as.fctr(R)2  0.981
## as.fctr(R)3  0.980  0.980
## SemesterS19  0.000  0.002  0.000
## SexF        -0.980 -0.980 -0.979 -0.097
## SexM        -0.981 -0.982 -0.982 -0.035  0.978
```

It appeared that these two random effects are not needed either, therefore our final model here is still the original model we obtained in C.2

C.3.4 Visual Organization

```
fla <- formula(model.formula.alldata[["VisOrg"]])
tmp <- lmer(fla,data=tall.nonmissing[tall.nonmissing$Rubric=="VisOrg",])
```

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

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

## Data: tall.nonmissing[tall.nonmissing$Rubric == "VisOrg", ]
## Models:
## tmp.single_intercept: as.numeric(Rating) ~ (1 | Artifact)
## tmp: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##                      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    3 228.95 237.21 -111.47   222.95
## tmp                     5 222.97 236.74 -106.48   212.97 9.9784  2   0.006811
##
## tmp.single_intercept
## tmp                  **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA stated that we really need "Rater" as a factor.

```
m0 <- tmp ## Null hypothesis
mA <- update(m0, . ~ . + (as.factor(Rater)|Artifact))
```

```
## Error: number of observations (=116) <= number of random effects (=270) for term (as.factor(Rater) |
```

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

```
## Error in update(mA, . ~ . - (1 | Artifact)): object 'mA' not found
```

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

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

```
summary(tmp)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "VisOrg", ]
##
## 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
## as.factor(Rater)1  2.38148    0.09652   24.67
## as.factor(Rater)2  2.65269    0.09558   27.75
## as.factor(Rater)3  2.29935    0.09558   24.06
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2
## as.fctr(R)2 0.265
## as.fctr(R)3 0.265  0.264
```

The random effect here are not needed.

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

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ 1 + (0 + Rubric | Artifact)
##    Data: tall.nonmissing
##
## REML criterion at convergence: 1481.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0251 -0.4969 -0.0753  0.5165  3.7820
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.6484   0.8052
```

```
##           RubricInitEDA   0.3779   0.6147   0.27
##           RubricInterpRes 0.2524   0.5024   0.02 0.79
##           RubricRsrchQ    0.1734   0.4164   0.40 0.51 0.74
##           RubricSelMeth   0.1034   0.3216   0.58 0.39 0.42 0.29
##           RubricTxtOrg    0.3946   0.6281   0.04 0.69 0.80 0.64 0.25
##           RubricVisOrg    0.3152   0.5615   0.19 0.78 0.77 0.60 0.31 0.79
##  Residual                 0.1942   0.4407
## Number of obs: 817, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.24700    0.04048   55.51
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00236116 (tol = 0.002, component 1)
```

The random effects for VisOrg and TxtOrg seem highly correlated with each other and with everything except for the random effect for SelMeth.
The random effects for InterpRes and InitEDA are highly correlated.
The random effects for RsrchQ and InterpRes are highly correlated etc.

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) +
##     Semester + Sex + Repeated + Rubric
##    Data: tall.nonmissing
##
## REML criterion at convergence: 1436.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1156 -0.5058 -0.0211  0.5203  3.8014
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.54872  0.7408
##           RubricInitEDA   0.34949  0.5912   0.47
##           RubricInterpRes 0.17485  0.4181   0.23 0.75
##           RubricRsrchQ    0.16889  0.4110   0.59 0.45 0.71
##           RubricSelMeth   0.06806  0.2609   0.40 0.61 0.75 0.42
##           RubricTxtOrg    0.26217  0.5120   0.34 0.62 0.71 0.57 0.67
##           RubricVisOrg    0.25621  0.5062   0.35 0.74 0.68 0.52 0.42 0.76
##  Residual                 0.18836  0.4340
## Number of obs: 817, groups:  Artifact, 91
##
## Fixed effects:
##                    Estimate Std. Error t value
## (Intercept)        2.820352   0.388481   7.260
## as.factor(Rater)2  0.002014   0.054802   0.037
## as.factor(Rater)3 -0.174685   0.054959  -3.178
## SemesterS19       -0.175032   0.087853  -1.992
## SexF              -0.802713   0.383749  -2.092
```

```
## SexM              -0.792258   0.382756  -2.070
## Repeated          -0.074405   0.098553  -0.755
## RubricInitEDA      0.541262   0.094891   5.704
## RubricInterpRes    0.580887   0.100024   5.807
## RubricRsrchQ       0.455982   0.086769   5.255
## RubricSelMeth      0.162876   0.093285   1.746
## RubricTxtOrg       0.685736   0.098736   6.945
## RubricVisOrg       0.524294   0.098256   5.336
```

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

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) +
##     Semester + Rubric
##    Data: tall.nonmissing
##
## REML criterion at convergence: 1435.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1189 -0.5102 -0.0149  0.5188  3.7768
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.56192  0.7496
##           RubricInitEDA   0.35041  0.5920   0.48
##           RubricInterpRes 0.17215  0.4149   0.25 0.75
##           RubricRsrchQ    0.17153  0.4142   0.60 0.45 0.72
##           RubricSelMeth   0.06984  0.2643   0.44 0.62 0.75 0.44
##           RubricTxtOrg    0.25478  0.5048   0.35 0.62 0.70 0.56 0.67
##           RubricVisOrg    0.26008  0.5100   0.37 0.74 0.69 0.53 0.44 0.76
##  Residual                 0.18910  0.4349
## Number of obs: 817, groups:  Artifact, 91
##
## Fixed effects:
##                    Estimate Std. Error t value
## (Intercept)        2.022481   0.098863  20.457
## as.factor(Rater)2 -0.001366   0.054995  -0.025
## as.factor(Rater)3 -0.168823   0.054994  -3.070
## SemesterS19       -0.184473   0.084144  -2.192
## RubricInitEDA      0.541715   0.094889   5.709
## RubricInterpRes    0.580226   0.100003   5.802
## RubricRsrchQ       0.453253   0.086692   5.228
## RubricSelMeth      0.156549   0.092719   1.688
## RubricTxtOrg       0.685878   0.098769   6.944
## RubricVisOrg       0.522884   0.098208   5.324
##
## Correlation of Fixed Effects:
##             (Intr) a.(R)2 a.(R)3 SmsS19 RbIEDA RbrcIR RbrcRQ RbrcSM RbrcTO
## as.fctr(R)2 -0.282
## as.fctr(R)3 -0.282  0.500
```

```
## SemesterS19 -0.267  0.016  0.016
## RubrcIntEDA -0.610 -0.001  0.000 -0.001
## RbrcIntrpRs -0.732 -0.001  0.000  0.000  0.734
## RubrcRsrchQ -0.698 -0.001  0.000  0.002  0.588  0.756
## RubricSlMth -0.777  0.000  0.000  0.006  0.663  0.779  0.689
## RubrcTxtOrg -0.679 -0.001  0.000 -0.001  0.676  0.751  0.684  0.728
## RubricVsOrg -0.672 -0.001 -0.001  0.000  0.716  0.745  0.668  0.682  0.751
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00201387 (tol = 0.002, component 1)
```

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

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00404493 (tol = 0.002, component 1)
```

This didn't quite converge, so we will try switching optimizers and increasing the number of iterations allowed.

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

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

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) +
##     Semester + Rubric + as.factor(Rater):Semester + as.factor(Rater):Rubric +
##     Semester:Rubric + as.factor(Rater):Semester:Rubric
##    Data: tall.nonmissing
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))
##
## REML criterion at convergence: 1435.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9270 -0.5054 -0.0614  0.4975  3.6575
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.48972  0.6998
##           RubricInitEDA   0.35173  0.5931   0.43
##           RubricInterpRes 0.15058  0.3880   0.34 0.80
##           RubricRsrchQ    0.16853  0.4105   0.67 0.44 0.74
##           RubricSelMeth   0.06777  0.2603   0.47 0.65 0.80 0.53
##           RubricTxtOrg    0.25248  0.5025   0.45 0.65 0.67 0.61 0.63
##           RubricVisOrg    0.25614  0.5061   0.36 0.73 0.69 0.58 0.38 0.76
##  Residual                 0.18788  0.4334
## Number of obs: 817, groups:  Artifact, 91
##
## Fixed effects:
```

```
##                                              Estimate Std. Error t value
## (Intercept)                                  1.751456   0.136454  12.835
## as.factor(Rater)2                            0.302533   0.154869   1.953
## as.factor(Rater)3                            0.250656   0.154869   1.619
## SemesterS19                                 -0.142029   0.250413  -0.567
## RubricInitEDA                                0.762359   0.164514   4.634
## RubricInterpRes                              0.975227   0.161435   6.041
## RubricRsrchQ                                 0.707102   0.146783   4.817
## RubricSelMeth                                0.456381   0.154440   2.955
## RubricTxtOrg                                 1.007110   0.160332   6.281
## RubricVisOrg                                 0.644127   0.165802   3.885
## as.factor(Rater)2:SemesterS19                0.267582   0.303451   0.882
## as.factor(Rater)3:SemesterS19               -0.082426   0.300220  -0.275
## as.factor(Rater)2:RubricInitEDA             -0.324263   0.203593  -1.593
## as.factor(Rater)3:RubricInitEDA             -0.383677   0.203593  -1.885
## as.factor(Rater)2:RubricInterpRes           -0.469370   0.200517  -2.341
## as.factor(Rater)3:RubricInterpRes           -0.712152   0.200517  -3.552
## as.factor(Rater)2:RubricRsrchQ              -0.447408   0.188810  -2.370
## as.factor(Rater)3:RubricRsrchQ              -0.475448   0.188810  -2.518
## as.factor(Rater)2:RubricSelMeth             -0.301920   0.193075  -1.564
## as.factor(Rater)3:RubricSelMeth             -0.354612   0.193075  -1.837
## as.factor(Rater)2:RubricTxtOrg              -0.448517   0.200471  -2.237
## as.factor(Rater)3:RubricTxtOrg              -0.422441   0.200471  -2.107
## as.factor(Rater)2:RubricVisOrg               0.008513   0.204448   0.042
## as.factor(Rater)3:RubricVisOrg              -0.293212   0.204448  -1.434
## SemesterS19:RubricInitEDA                   -0.046340   0.300300  -0.154
## SemesterS19:RubricInterpRes                  0.133408   0.294525   0.453
## SemesterS19:RubricRsrchQ                     0.138291   0.266735   0.518
## SemesterS19:RubricSelMeth                   -0.081130   0.281445  -0.288
## SemesterS19:RubricTxtOrg                     0.171896   0.292275   0.588
## SemesterS19:RubricVisOrg                     0.152067   0.301182   0.505
## as.factor(Rater)2:SemesterS19:RubricInitEDA  0.020173   0.391289   0.052
## as.factor(Rater)3:SemesterS19:RubricInitEDA  0.258867   0.388223   0.667
## as.factor(Rater)2:SemesterS19:RubricInterpRes -0.268664   0.384277  -0.699
## as.factor(Rater)3:SemesterS19:RubricInterpRes -0.152973   0.381562  -0.401
## as.factor(Rater)2:SemesterS19:RubricRsrchQ   -0.218376   0.359363  -0.608
## as.factor(Rater)3:SemesterS19:RubricRsrchQ    0.354425   0.355606   0.997
## as.factor(Rater)2:SemesterS19:RubricSelMeth  -0.404633   0.368931  -1.097
## as.factor(Rater)3:SemesterS19:RubricSelMeth  -0.203130   0.365910  -0.555
## as.factor(Rater)2:SemesterS19:RubricTxtOrg   -0.542944   0.384068  -1.414
## as.factor(Rater)3:SemesterS19:RubricTxtOrg   -0.305748   0.381005  -0.802
## as.factor(Rater)2:SemesterS19:RubricVisOrg   -0.604387   0.391633  -1.543
## as.factor(Rater)3:SemesterS19:RubricVisOrg   -0.183661   0.388822  -0.472

##
## 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
```

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

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) +
##     Semester + Rubric + as.factor(Rater):Rubric
##    Data: tall.nonmissing
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))
##
## REML criterion at convergence: 1430.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9356 -0.5143 -0.0409  0.4895  3.5799
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.50832  0.7130
##           RubricInitEDA   0.35400  0.5950   0.46
##           RubricInterpRes 0.15704  0.3963   0.38 0.82
##           RubricRsrchQ    0.18267  0.4274   0.64 0.45 0.73
##           RubricSelMeth   0.07263  0.2695   0.45 0.62 0.76 0.41
##           RubricTxtOrg    0.25909  0.5090   0.43 0.64 0.68 0.56 0.65
##           RubricVisOrg    0.25612  0.5061   0.36 0.72 0.69 0.53 0.41 0.78
##  Residual                 0.18484  0.4299
## Number of obs: 817, groups:  Artifact, 91
##
## Fixed effects:
##                                  Estimate Std. Error t value
## (Intercept)                       1.77051    0.11793  15.013
## as.factor(Rater)2                 0.36517    0.13281   2.750
## as.factor(Rater)3                 0.22498    0.13235   1.700
## SemesterS19                      -0.18904    0.08384  -2.255
## RubricInitEDA                     0.74459    0.13625   5.465
## RubricInterpRes                   1.01239    0.13425   7.541
## RubricRsrchQ                      0.74689    0.12376   6.035
## RubricSelMeth                     0.42339    0.12972   3.264
## RubricTxtOrg                      1.04783    0.13506   7.758
## RubricVisOrg                      0.68170    0.13885   4.910
## as.factor(Rater)2:RubricInitEDA  -0.30781    0.17212  -1.788
## as.factor(Rater)3:RubricInitEDA  -0.30463    0.17167  -1.775
## as.factor(Rater)2:RubricInterpRes -0.53771    0.16967  -3.169
## as.factor(Rater)3:RubricInterpRes -0.75345    0.16929  -4.451
## as.factor(Rater)2:RubricRsrchQ   -0.50228    0.16112  -3.117
## as.factor(Rater)3:RubricRsrchQ   -0.37510    0.16057  -2.336
## as.factor(Rater)2:RubricSelMeth  -0.39806    0.16417  -2.425
## as.factor(Rater)3:RubricSelMeth  -0.40351    0.16374  -2.464
## as.factor(Rater)2:RubricTxtOrg   -0.58389    0.17104  -3.414
## as.factor(Rater)3:RubricTxtOrg   -0.49781    0.17061  -2.918
## as.factor(Rater)2:RubricVisOrg   -0.14568    0.17393  -0.838
## as.factor(Rater)3:RubricVisOrg   -0.33879    0.17354  -1.952
```

```
##
## 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
```

```
formula(comb.inter.u)
```

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

```
formula(comb.inter_elim)
```

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

```
formula(comb.back_elim)
```

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

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

```
##  Groups   Name            Std.Dev. Corr
##  Artifact RubricCritDes   0.69980
##           RubricInitEDA   0.59307  0.427
##           RubricInterpRes 0.38804  0.345 0.802
##           RubricRsrchQ    0.41052  0.667 0.444 0.736
##           RubricSelMeth   0.26032  0.474 0.651 0.802 0.527
##           RubricTxtOrg    0.50247  0.445 0.655 0.672 0.611 0.630
##           RubricVisOrg    0.50610  0.364 0.733 0.685 0.580 0.378 0.762
##  Residual                 0.43345
```

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

```
##  Groups   Name            Std.Dev. Corr
##  Artifact RubricCritDes   0.71296
##           RubricInitEDA   0.59498  0.455
##           RubricInterpRes 0.39628  0.376 0.817
##           RubricRsrchQ    0.42740  0.642 0.454 0.727
##           RubricSelMeth   0.26949  0.455 0.616 0.758 0.407
##           RubricTxtOrg    0.50901  0.427 0.643 0.681 0.558 0.645
##           RubricVisOrg    0.50608  0.355 0.722 0.688 0.527 0.409 0.778
##  Residual                 0.42993
```

```r
summary(comb.back_elim)$varcor
```

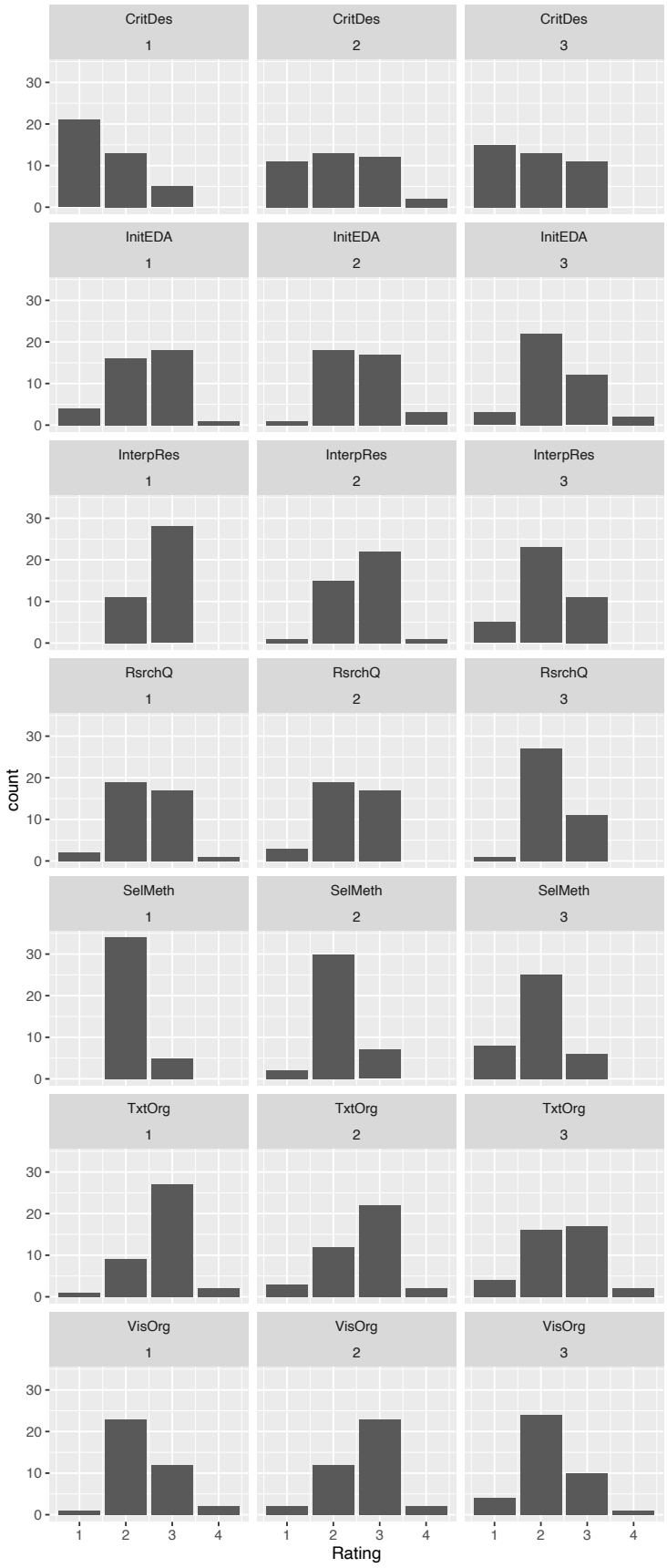
```
##  Groups   Name            Std.Dev. Corr
##  Artifact RubricCritDes   0.74961
##           RubricInitEDA   0.59195  0.477
##           RubricInterpRes 0.41491  0.252 0.755
##           RubricRsrchQ    0.41416  0.602 0.452 0.715
##           RubricSelMeth   0.26428  0.436 0.620 0.751 0.440
##           RubricTxtOrg    0.50475  0.346 0.622 0.699 0.564 0.665
##           RubricVisOrg    0.50998  0.365 0.739 0.689 0.531 0.442 0.760
##  Residual                 0.43485
```

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

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

```
## Data: tall.nonmissing
## Models:
## comb.back_elim: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) +
## comb.back_elim:     Semester + Rubric
## comb.inter_elim: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) +
## comb.inter_elim:     Semester + Rubric + as.factor(Rater):Rubric
## comb.inter.u: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) +
## comb.inter.u:     Semester + Rubric + as.factor(Rater):Semester + as.factor(Rater):Rubric +
## comb.inter.u:     Semester:Rubric + as.factor(Rater):Semester:Rubric
##                 npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## comb.back_elim    39 1475.2 1658.7 -698.58   1397.2
## comb.inter_elim   51 1465.5 1705.5 -681.76   1363.5 33.653 12   0.000765 ***
## comb.inter.u      71 1481.8 1815.9 -669.91   1339.8 23.694 20   0.256027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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


CritDes
1
CritDes
2
CritDes
3
InitEDA
1
InitEDA
2
InitEDA
3
InterpRes
1
InterpRes
2
InterpRes
3
RsrchQ
1
RsrchQ
2
RsrchQ
3
SelMeth
1
SelMeth
2
SelMeth
3
TxtOrg
1
TxtOrg
2
TxtOrg
3
VisOrg
1
VisOrg
2
VisOrg
3
count
0
10
20
30
1
2
3
4
Rating

```
m0 <- comb.inter_elim
mA <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) | Artifact) + as.factor
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00278442 (tol = 0.002, component 1)
```

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

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

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

```
## Data: tall.nonmissing
## Models:
## m0: as.numeric(Rating) ~ (0 + Rubric | Artifact) + as.factor(Rater) +
## m0:     Semester + Rubric + as.factor(Rater):Rubric
## mA: as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) |
## mA:     Artifact) + as.factor(Rater) + Semester + Rubric + as.factor(Rater):Rubric
##    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m0   51 1465.5 1705.5 -681.76   1363.5
## mA   57 1425.9 1694.1 -655.94   1311.9 51.624  6  2.219e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m0 <- comb.inter_elim
mA <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) | Artifact) + (0 + as.fa
```

```
## Error: number of observations (=817) <= number of random effects (=1911) for term (0 + as.factor(Rat
```

```
comb.final <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) | Artifact) + as
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00278442 (tol = 0.002, component 1)
```

```
formula(comb.final)
```

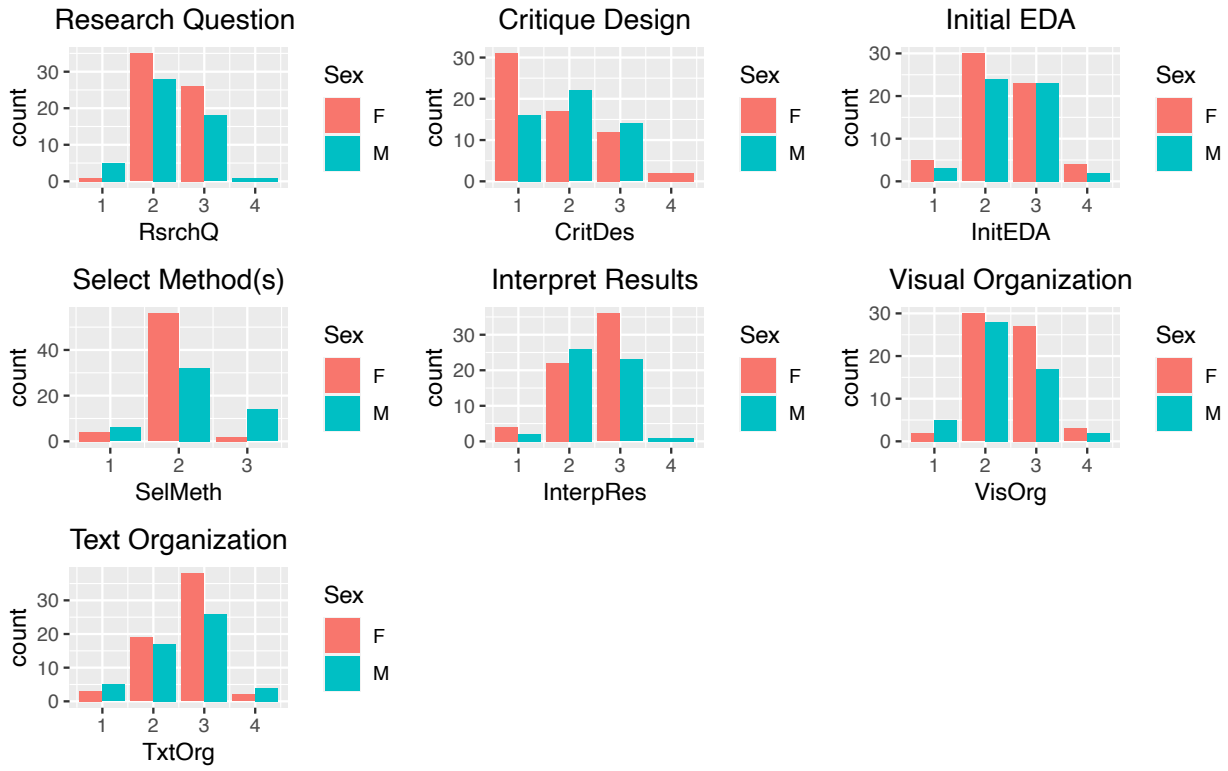
```
## as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) |
##     Artifact) + as.factor(Rater) + Semester + Rubric + as.factor(Rater):Rubric
```

D

```
Rsr_g = ggplot(ratings, aes(x = RsrchQ, fill = Sex)) +
  geom_bar(position = position_dodge()) + labs(title = 'Research Question ') + theme(plot.title = eleme
Cri_g = ggplot(ratings, aes(x = CritDes, fill = Sex)) +
  geom_bar(position = position_dodge()) + labs(title = 'Critique Design') + theme(plot.title = element_
Ini_g = ggplot(ratings, aes(x = InitEDA, fill = Sex)) +
```

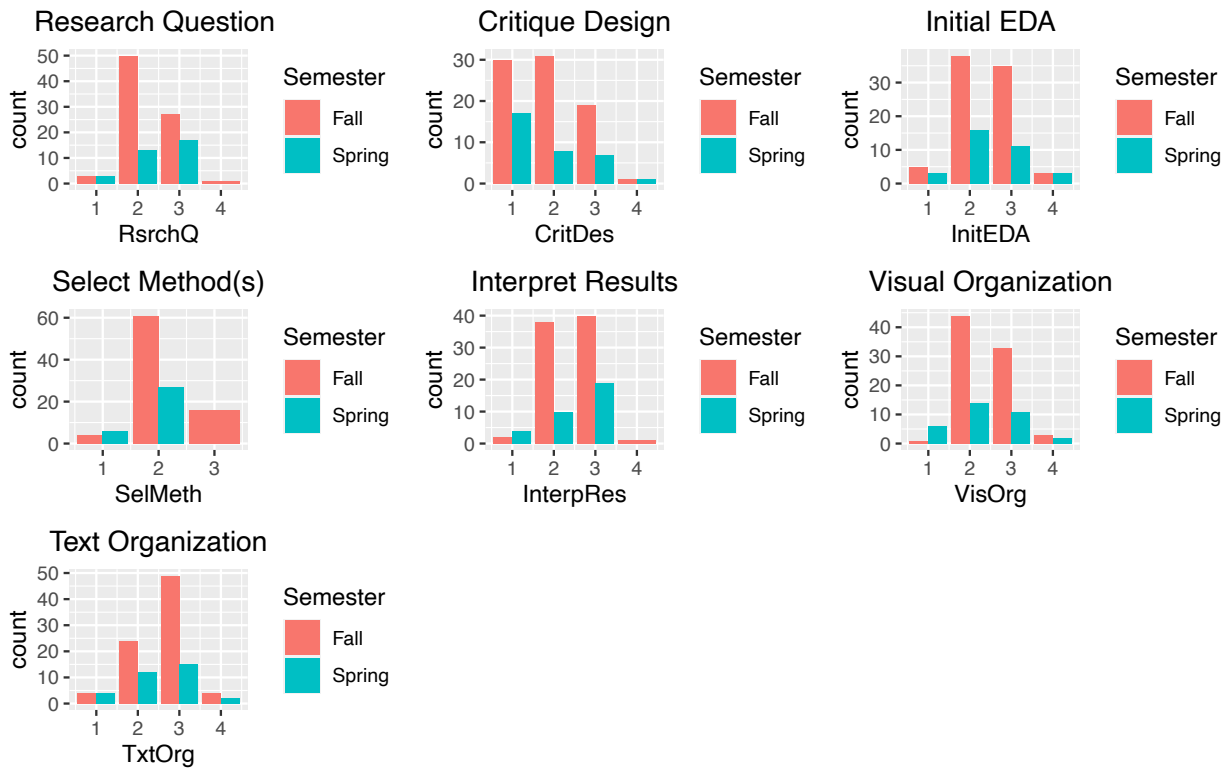
```
  geom_bar(position = position_dodge()) + labs(title = 'Initial EDA') + theme(plot.title = element_text
Sel_g = ggplot(ratings, aes(x = SelMeth, fill = Sex)) +
  geom_bar(position = position_dodge()) + labs(title = 'Select Method(s)') + theme(plot.title = element_
Int_g = ggplot(ratings, aes(x = InterpRes, fill = Sex)) +
  geom_bar(position = position_dodge()) + labs(title = 'Interpret Results') + theme(plot.title = elemen
Vis_g = ggplot(ratings, aes(x = VisOrg, fill = Sex)) +
  geom_bar(position = position_dodge()) + labs(title = 'Visual Organization') + theme(plot.title = eleme
Txt_g = ggplot(ratings, aes(x = TxtOrg, fill = Sex)) +
  geom_bar(position = position_dodge()) + labs(title = 'Text Organization') + theme(plot.title = elemen


fig7 = grid.arrange(Rsr_g, Cri_g, Ini_g, Sel_g, Int_g, Vis_g, Txt_g, ncol = 3)
```




```
ratings$Semester <- factor(ratings$Semester)
Rsr_g = ggplot(ratings, aes(x = RsrchQ, fill = Semester)) +
  geom_bar(position = position_dodge()) + labs(title = 'Research Question ') + theme(plot.title = eleme
Cri_g = ggplot(ratings, aes(x = CritDes, fill = Semester)) +
  geom_bar(position = position_dodge()) + labs(title = 'Critique Design') + theme(plot.title = element_
Ini_g = ggplot(ratings, aes(x = InitEDA, fill = Semester)) +
  geom_bar(position = position_dodge()) + labs(title = 'Initial EDA') + theme(plot.title = element_text
Sel_g = ggplot(ratings, aes(x = SelMeth, fill = Semester)) +
  geom_bar(position = position_dodge()) + labs(title = 'Select Method(s)') + theme(plot.title = element_
Int_g = ggplot(ratings, aes(x = InterpRes, fill = Semester)) +
  geom_bar(position = position_dodge()) + labs(title = 'Interpret Results') + theme(plot.title = elemen
Vis_g = ggplot(ratings, aes(x = VisOrg, fill = Semester)) +
  geom_bar(position = position_dodge()) + labs(title = 'Visual Organization') + theme(plot.title = eleme
Txt_g = ggplot(ratings, aes(x = TxtOrg, fill = Semester)) +
  geom_bar(position = position_dodge()) + labs(title = 'Text Organization') + theme(plot.title = elemen
```

```
fig6 = grid.arrange(Rsr_g, Cri_g, Ini_g, Sel_g, Int_g, Vis_g, Txt_g, ncol = 3)
```



The barplots of 7 rubrics show that there are significant differences between different genders and different semesters. However, all models on our final list from C.2 didn't suggest either of them to be included in our model, which we may need some further investigation on to see what exactly precluded these factors from being selected by the model or what could be wrong in our modeling process.