# Carnegie Mellon Freshman Statistics project papers: analyzing rating by different raters on different types of rubrics

Julia Keating
jhkeatin@andrew.cmu.edu

## 1 Abstract

Dietrich College at Carnegie Mellon University has been experimenting with rating work in Freshman Statistics, using three raters from three different departments and seven different rubrics. Data collected from the experiment consists of 91 project papers, referred to as artifacts, and includes the ratings assigned by each rubric and the raters who assigned them. We explore the effect of Rubric and Rater on Rating through the use of bar charts, and use backward elimination methods to fit models predicting Rating from factors such as Rater, Rubric, Semester and Sex, also testing for interactions and random effects. We find that the ratings of artifacts relates to the rubric and rater used to rate it, but that rater most strongly impacts rating. We suggest that Dietrich College carefully consider which raters and rubrics they want to use to assess student work in the future to ensure fair and accurate ratings.

## 2 Introduction

Dietrich College at Carnegie Mellon University has been experimenting with rating work in Freshman Statistics, using three raters from three different departments and seven different rubrics. In this analysis, we would like to answer the following questions:

1. How do the distributions of ratings for each rubric compare? How do the distributions of ratings given by each rater compare?
2. For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees with the others, or do they all disagree?
3. How are Rater, Semester, Sex, Repeated and Rubric related to the ratings? Do these factors interact in any interesting ways?
4. Is there anything else interesting to say about this data?

## 3 Data

Our data consists of 91 project papers, referred to as artifacts, randomly sampled from a Fall and a Spring section of Freshman Statistics at Carnegie Mellon University (Junker, 2021). Thirteen of the 91 artifacts were rated by all three raters, and each of the remaining 78 artifacts were rated by only rater.

The variables used in this analysis are defined in Table 1. Table 2 describes the seven different rubrics used to rate artifacts, and Table 3 provides the rating scale used for all rubrics in this experiment. Table 4 provides summary statistics of ratings for each of the seven rubrics. Observe that all rubrics have a minimum rating of 1 and all ratings except for SelMeth have a maximum rating of 4. Also, the CritDes and SelMeth rubrics has a much lower mean rating than the other rubrics. This suggests that the distribution of ratings may differ across the different rubrics, but we will explore this more thoroughly later in our analysis.

## 4 Methods

### 4.1 How do the distributions of ratings for each rubric compare? How do the distributions of ratings given by each rater compare?

To compare the distribution of ratings for each rubric, we will create bar plots of rating for each rubric, using both the fill data set and only the 13 artifacts seen by all three raters, and compare the plots.

| Variable number | Variable name | Definition |
|---|---|---|
| 1 | Rater | Which of the raters gave a rating (1, 2, or 3) |
| 2 | Semester | Which semester the artifact came from (Fall or Spring) |
| 3 | Sex | Sex or gender of student who created the artifact (M or F) |
| 4 | RsrchQ | Rating on Research Question |
| 5 | CritDes | Rating on Critique Design |
| 6 | InitEDA | Rating on Initial EDA |
| 7 | SelMeth | Rating on Select Method(s) |
| 8 | InterpRes | Rating on Interpret Results |
| 9 | VisOrg | Rating on Visual Organization |
| 10 | TxtOrg | Rating on Text Organization |
| 11 | Artifact | Unique identifier for each artifact |
| 12 | Repeated | Equals 1 if one of the 13 artifacts seen by all 3 raters, or 0 otherwise |

Table 1: Definitions of variables in data set

| Rubric | Description |
|---|---|
| Research Question | Given a scenario, the student generates, critiques or evaluates a relevant empirical research question. |
| Critique Design | Given an empirical research question, the student critiques or evaluates to what extent a study design convincingly answer that question. |
| Initial EDA | Given a data set, the student appropriately describes the data and provides initial Exploratory Data Analysis. |
| Select Method(s) | Given a data set and a research question, the student selects appropriate method(s) to analyze the data. |
| Interpret Results | The student appropriately interprets the results of the selected method(s). |
| Visual Organization | The student communicates in an organized, coherent and effective fashion with visual elements (charts, graphs, tables, etc.) |
| 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: Descriptions of different rubrics used in this experiment

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

Table 3: Rating scale used for all rubrics in this experiment

| Rubric | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| RsrchQ | 1.00 | 2.00 | 2.00 | 2.35 | 3.00 | 4.00 |
| CritDes | 1.00 | 1.00 | 2.00 | 1.87 | 3.00 | 4.00 |
| InitEDA | 1.00 | 2.00 | 2.00 | 2.44 | 3.00 | 4.00 |
| SelMeth | 1.00 | 2.00 | 2.00 | 2.07 | 2.00 | 3.00 |
| InterpRes | 1.00 | 2.00 | 3.00 | 2.49 | 3.00 | 4.00 |
| VisOrg | 1.00 | 2.00 | 2.00 | 2.41 | 3.00 | 4.00 |
| TxtOrg | 1.00 | 2.00 | 3.00 | 2.60 | 3.00 | 4.00 |

Table 4: Summary statistics of ratings for each rubric

## 4.2 For each rubric, do raters generally agree on their scores? If not, is there one rater who disagrees with the others, or do they all disagree?

To determine if raters agree on their scores, we will use only the 13 artifacts seen by all three raters to calculate intraclass correlation (ICC) values for each rubric. For rubrics where raters seem to disagree, i.e. rubrics with small ICC values, we will calculate the percent exact agreements between each pair of raters to determine if there is one rater who disagrees with the others or if they all disagree for each of those rubrics. We also repeat the calculation of ICC values using the full data set to see if the values agree between the full data set and the subset corresponding to the 13 artifacts seen by all three raters.

### 4.3 How are Rater, Semester, Sex, Repeated and Rubric related to the ratings? Do these factors interact in any interesting ways?

To determine how Rater, Semester, Sex and Repeated are related to Rating, we first used backwards elimination to fit seven models (one for each rubric) predicting Rater with Rater, Semester, Sex and Repeated as fixed effects using the full data set after removing any observations with missing data. For rubrics where the chosen model was not just the simple random-intercept model, we tested for interactions and new random effects. We repeated this process using only the 13 artifacts seen by all three raters without Repeated as a predictor. Lastly, we fit one combined model for all rubrics using the full data set after removing any observations with missing data. We used backwards elimination to choose this model, and tested for interactions and new random effects in the chosen model.

### 4.4 Is there anything else interesting to say about this data?

To further explore the data, we first created bar plots of the proportion of ratings for each semester to compare the distribution of ratings between each semester. Similarly, we created bar plots of the proportion of ratings for each sex to compare the distribution of ratings between each sex. Lastly, we created bar plots for each rubric of the proportion of ratings for each sex to compare the distribution of ratings between each sex for each rubric.

## 5 Results

### 5.1 How do the distributions of ratings for each rubric compare?

Figure 3 shows the distribution of ratings for each rubric using the full dataset. It appears that the distribution of ratings within the full data set is not the same for all of the rubrics. First, note that the number of ratings of 1 assigned is around 10 for all rubrics except for CritDes, which gives about 50 ratings of 1. Also, note that the number of ratings of 2 assigned is around 50 for all of the rubrics except for SelMeth, which gives over 75 ratings of 2. Note that the number of ratings of 4 assigned is greater than 0 for all rubrics except for SelMeth, which does not assign any ratings of 4. Lastly, we can see that there is some missing data for Rating in the full data set, which we will need to be careful with in fitting out model. See pages 1-3 of the Technical Appendix for more details.

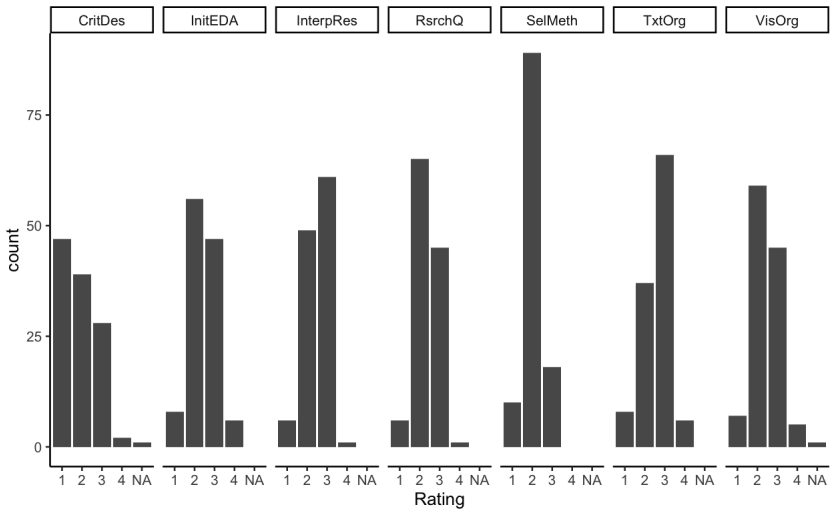



Figure 1: Distribution of ratings for each rubric using the full data set

Figure 4 shows the distribution of ratings for each rubric using only the 13 artifacts seen by all three raters. Again, it appears that the distribution of ratings is not the same for all of the rubrics. First, note that the number of ratings of 1 assigned is around 2 for all rubrics except for CritDes, which gives about 17 ratings of 1. Also, note that the number of ratings of 2 assigned is around 20 for all of the rubrics except for SelMeth and TxtOrg, which give close to 30 ratings of 2. Lastly, note that the number ratings of 4 assigned is 0 for all rubrics except for InterpRes and TxtOrg, which have at least one rating of 4. We can see in these plots that there is no missing data for rating in this subset of the data, so we will not need to worry about that in any models we fit using this subset. See pages 5-7 of the Technical Appendix for more details.

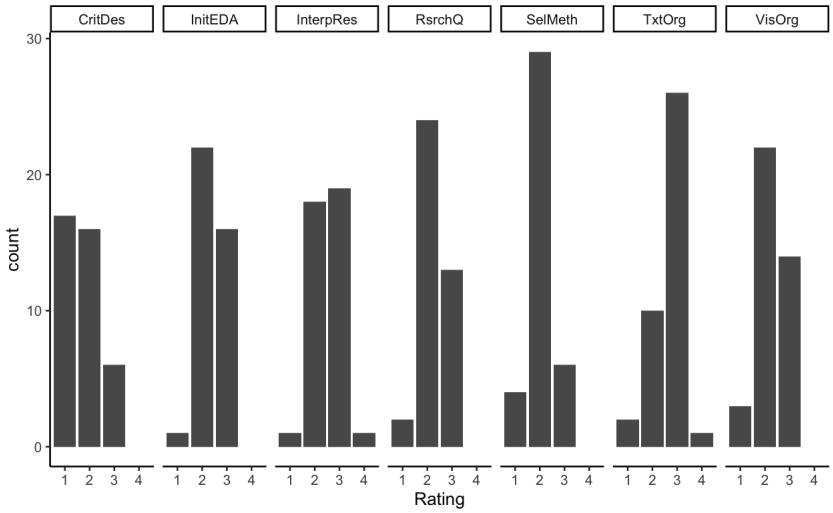



Figure 2: Distribution of ratings for each rubric using only the 13 artifacts seen by all three raters

## 5.2 How do the distributions of ratings given by each rater compare?

Figure 3 shows the distribution of ratings for each rater using the full dataset. It appears that the distribution of ratings given by each rater is pretty much indistinguishable from the other raters. Rater 3 gave slightly more 2's and slightly less 3's than the other two raters, but the distribution of ratings is pretty similar across all three raters. See pages 3-5 of the Technical Appendix for more details.

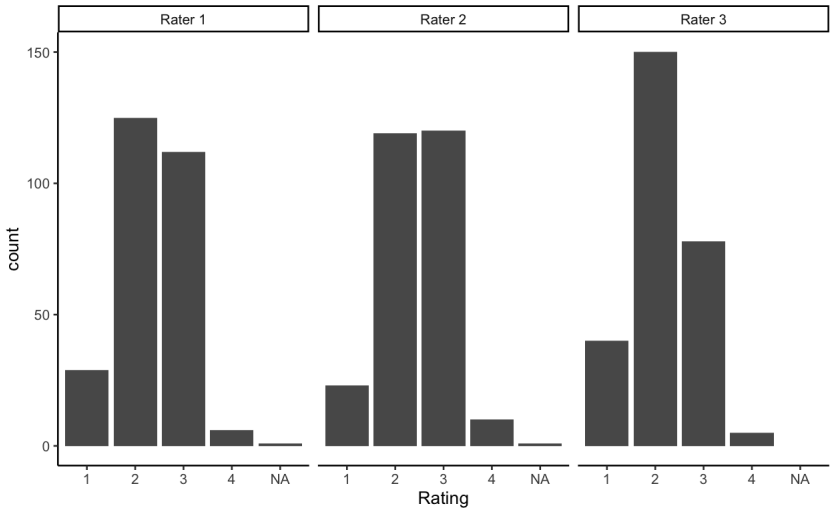



Figure 3: Distribution of ratings for each rater using the full data set

Figure 4 shows the distribution of ratings for each rater using only the 13 artifacts seen by all three raters. Again, it appears that the distribution of ratings given by each rater is pretty much indistinguishable from the other raters. Rater 3 gave slightly more 2's and slightly less 3's than the other two raters, but the distribution of ratings is pretty similar across all three raters. See pages 7-8 of the Technical Appendix for more details.

## 5.3 For each rubric, do raters generally agree on their scores?

Table 5 shows the ICC values calculated using either the full data set or only the 13 artifacts seen by all three raters. Note that ICC values are comparable between both the full data set and the subset of the data, which suggests that the 13 artifacts seen by all three raters are representative of the full data set.

We see that the rubrics CritDes, InitEDA, SelMeth and VisOrg have fairly high ICC values which suggests that the three raters generally agree on these rubrics. The rubrics RsrchQ, InterpRes and TxtOrg have fairly low ICC values which suggests that the three raters generally do not agree on these rubrics. See pages 8-10 and 13-16 of the Technical Appendix for more details. Next, we will look at these three rubrics more closely.

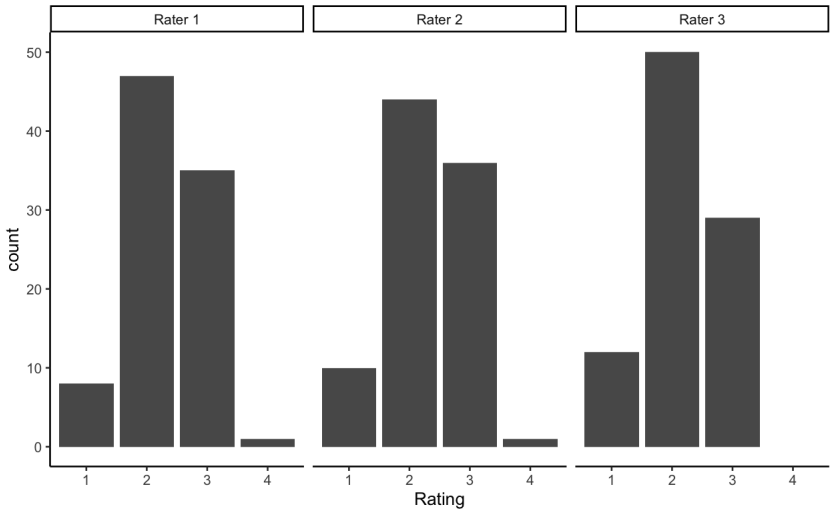



Figure 4: Distribution of ratings for each rater using only the 13 artifacts seen by all three raters

| Rubric | ICC value for full data set | ICC value for common artifacts |
|---|---|---|
| RsrchQ | 0.2096170 | 0.1891711 |
| CritDes | 0.6730667 | 0.5725587 |
| InitEDA | 0.6867283 | 0.4929391 |
| SelMeth | 0.4719928 | 0.5211740 |
| InterpRes | 0.2200243 | 0.2295545 |
| VisOrg | 0.6607148 | 0.5924793 |
| TxtOrg | 0.1879828 | 0.1428337 |

Table 5: ICC values for each rubric using either the full data set or only the 13 artifacts seen by all three raters

## 5.4 When raters do not agree on scores, is there one rater who disagrees with the others, or do they all disagree?

Table 6 shows the percent exact agreement between each pair of raters for the rubrics RsrchQ, InterpRes and TxtOrg. Based on these percentages, it appears that Rater 1 contributes most to the disagreement on the RsrchQ rubric, and the three raters disagree fairly equally on the InterpRes rubric and on the TxtOrg rubric. See pages 11-13 of the Technical Appendix for more details.

| Rubric | Raters 1 and 2 | Raters 2 and 3 | Raters 1 and 3 |
|---|---|---|---|
| RsrchQ | 38% | 54% | 77% |
| InterpRes | 62% | 62% | 54% |
| TxtOrg | 69% | 54% | 62% |

Table 6: Percent exact agreement between raters for rubrics where raters disagree on ratings

## 5.5 How are Rater, Semester, Sex, Repeated and Rubric related to the ratings? Do Rater, Semester, Sex, Repeated and Rubric interact in any interesting ways?

Table 7 shows the models chosen by backwards elimination for each rubric using the full data set after removing observations with missing data. See pages 16-23 of the Technical Appendix for more details. Note that the models chosen for the rubrics InitEDA, RsrchQ and TxtOrg are just the simple random-intercept model. None of the interactions or new random effects tested for the rubrics CritDes, InterpRes, SelMeth and VisOrg significantly improved the respective models. See page PAGE NUMBER of the Technical Appendix for more details. Thus, the final models we choose to predict Rating for each rubric using the full data set are the ones listed in Table 7. Since all models which are not just the simple random intercept model include Rater as a fixed effect, it appears that Rater is most strongly related to Rating.

Table 8 provides the estimated coefficients for the fixed effects in the chosen model for each rubric. Based on these coefficients, it appears that ratings with the CritDes rubric are lowest for Rater 1 and highest for Rater 2. It also appears that the ratings with the SelMeth and InterpRes rubrics are lowest for Rater 3 and highest for

Rater 1. Additionally, the ratings with the SelMeth rubric are higher in the Fall semester than in the Spring semester. Lastly, it appears that ratings with the VisOrg rubric are lowest for Rater 3 and highest for Rater 2.

| Rubric | Model | Model Number |
|---|---|---|
| RsrchQ | $\text{Rating} \sim (1\|\text{Artifact})$ | (1) |
| CritDes | $\text{Rating} \sim \text{Rater} + (1\|\text{Artifact}) - 1$ | (2) |
| InitEDA | $\text{Rating} \sim (1\|\text{Artifact})$ | (3) |
| SelMeth | $\text{Rating} \sim \text{Rater} + \text{Semester} + (1\|\text{Artifact}) - 1$ | (4) |
| InterpRes | $\text{Rating} \sim \text{Rater} + (1\|\text{Artifact}) - 1$ | (5) |
| VisOrg | $\text{Rating} \sim \text{Rater} + (1\|\text{Artifact}) - 1$ | (6) |
| TxtOrg | $\text{Rating} \sim (1\|\text{Artifact})$ | (7) |

Table 7: Models chosen for each rubric using the full dataset by backward elimination

| Rubric | Intercept | Rater 1 | Rater 2 | Rater 3 | Spring Semester |
|---|---|---|---|---|---|
| RsrchQ | 2.352 | - | - | - | - |
| CritDes | - | 1.686 | 2.113 | 1.891 | - |
| InitEDA | 2.442 | - | - | - | - |
| SelMeth | - | 2.250 | 2.226 | 2.033 | -0.359 |
| InterpRes | - | 2.704 | 2.586 | 2.139 | - |
| VisOrg | - | 2.378 | 2.649 | 2.284 | - |
| TxtOrg | 2.587 | - | - | - | - |

Table 8: Estimated fixed effects for the final model for each rubric

We repeated the model selection process using only the 13 artifacts seen by all three raters, but the model chosen for each rubric was just the simple random-effect model so we did not test for interactions or new random effects for this subset of the data. See pages 23-24 of the Technical Appendix for more details.

Equation (8) is the combined model predicting Rating for all rubrics chosen by backwards elimination with the full data set after removing incomplete observations and testing for interactions and new random effects. See pages 24-29 of the Technical Appendix for more details. Note that only the interaction between Rater and Rubric is kept by backwards elimination, which we found significantly improves the model. We also found significant evidence that including random effects for Rubric, but none of the other predictors, improves the model.

$$\text{Rating} \sim (0 + \text{Rubric}|\text{Artifact}) + \text{Semester} + \text{Rater} * \text{Rubric} \tag{8}$$

Table 9 provides the estimated coefficients for the fixed effects in the chosen combined model, along with the associated standard errors and t-values. Based on these coefficients we predict:

- With Rater and Rubric fixed, ratings are higher in the Fall semester than in the Spring semester
- With Semester fixed, ratings for rubric RsrchQ are lowest for Rater 3 and highest for Rater 1
- With Semester fixed, ratings for rubric CritDes are lowest for Rater 1 and highest for Rater 2
- With Semester fixed, ratings for rubric InitEDA are lowest for Rater 3 and highest for Rater 2
- With Semester fixed, ratings for rubric SelMeth are lowest for Rater 3 and highest for Rater 1
- With Semester fixed, ratings for the rubric InterpRes are lowest for Rater 3 and highest for Rater 1
- With Semester fixed, ratings for the rubric VisOrg are lowest for Rater 3 and highest for Rater 2
- With Semester fixed, ratings for the rubric TxtOrg are lowest for Rater 3 and highest for Rater 1
- With Semester fixed, ratings for Rater 1 increase by rubric in the following order: CritDes, SelMeth, VisOrg, RsrchQ, InitEDA, InterpRes, TxtOrg
- With Semester fixed, ratings for Rater 2 increase by rubric in the following order: CritDes, SelMeth, RsrchQ, InitEDA, TxtOrg, InterpRes, VisOrg

- With Semester fixed, ratings for Rater 3 increase by rubric in the following order: CritDes, SelMeth, InterpRes, VisOrg, RsrchQ, InitEDA, TxtOrg

Note that for almost all rubrics, our model predicts that Rater 3 will give the lowest rating. Additionally, our model predicts that rubrics CritDes and SelMeth will have the lowest ratings for all three raters.

| Variable | Estimate | Standard Error | t-value |
|---|---|---|---|
| (Intercept) | 2.33250840 | 0.10170831 | 22.9333122 |
| Spring Semester | -0.15918237 | 0.07647761 | -2.0814245 |
| Rater 1 * RsrchQ | 0.15122961 | 0.12505496 | 1.2093052 |
| Rater 2 * RsrchQ | 0.02986976 | 0.12773693 | 0.2338381 |
| Rater 3 * RsrchQ | 0.02477928 | 0.11412693 | 0.2171203 |
| Rater 1 * CritDes | -0.57494785 | 0.14367924 | -4.0016070 |
| Rater 2 * CritDes | -0.20890482 | 0.14628476 | -1.4280696 |
| Rater 3 * CritDes | -0.37903926 | 0.13418031 | -2.8248501 |
| Rater 1 * InitEDA | 0.16454200 | 0.12894375 | 1.2760758 |
| Rater 2 * InitEDA | 0.23078272 | 0.13112179 | 1.7600639 |
| Rater 3 * InitEDA | 0.06572843 | 0.11288966 | 0.5822361 |
| Rater 1 * SelMeth | -0.16427065 | 0.11896097 | -1.3808786 |
| Rater 2 * SelMeth | -0.18459440 | 0.12231371 | -1.5091882 |
| Rater 3 * SelMeth | -0.35551658 | 0.11391256 | -3.1209602 |
| Rater 1 * InterpRes | 0.41657075 | 0.11861779 | 3.5118742 |
| Rater 2 * InterpRes | 0.26938064 | 0.12135056 | 2.2198549 |
| Rater 3 * InterpRes | -0.10237522 | 0.10563258 | -0.9691633 |
| Rater 1 * VisOrg | 0.07930650 | 0.11732284 | 0.6759681 |
| Rater 2 * VisOrg | 0.34044562 | 0.11964179 | 2.8455411 |
| Rater 1 * TxtOrg | 0.44083158 | 0.12545589 | 3.5138372 |
| Rater 2 * TxtOrg | 0.25583093 | 0.12779231 | 2.0019274 |
| Rater 3 * TxtOrg | 0.19185163 | 0.10885046 | 1.7625247 |

Table 9: Estimated fixed effects with associated standard errors and t-values for the final combined model

### 5.6 Is there anything else interesting to say about this data?

Figure 5 shows the proportion of ratings assigned in each semester using the full data set. Note that each semester has about the same proportion of artifacts with rating 4, but the Spring has a lower proportion of artifacts with rating 2 and 3 than the Fall semester and a higher proportion of artifacts with rating 1 than the Fall semester. See pages 29-31 of the Technical Appendix for more details.

Figure 6 shows the proportion of ratings assigned for each sex using the full data set. Note that the distribution of ratings for males and females is about the same. See pages 31-32 of the Technical Appendix for more details.

Lastly, Figure 7 shows the proportion of ratings assigned for each sex for each rubric using the full data set. See pages 32-38 of the Technical Appendix for more details. We see that females seem to do worse on the CritDes rubric than males since females have a much higher proportion of 1's than males with this rubric. Also, females seem to do more mediocre on the SelMeth rubric than males since they have a much higher proportion of 2's and lower proportion of 1's and 3's than males on this rubric. It seems that females seem to do better than males on the InterpRes rubric since they score a much higher proportion of 3's and much lower proportion of 2's than males on this rubric.

## 6 Discussion

We found that the distribution of ratings is not the same for each rubric, but the distribution of ratings is somewhat similar for each rater. We found that the three raters seem to generally agree on scores for the rubrics CritDes, InitEDA, SelMeth and VisOrg. Additionally, we would that Rater 1 contributes most to the disagreement for the rubric RsrchQ whereas the three raters disagree fairly equally on the rubrics InterpRes and TxtOrg. We found that rubric is most strongly related to ratings, and Rater and Rubric do interact significantly. Rater 3 appears to give the lowest rating of all three raters for almost all rubrics. Additionally, rubrics CritDes and SelMeth appear to assign the lowest rating for all three raters. Lastly, we found that the distribution of ratings is not the same for each semester, but is fairly similar for each sex when considering all rubrics. There

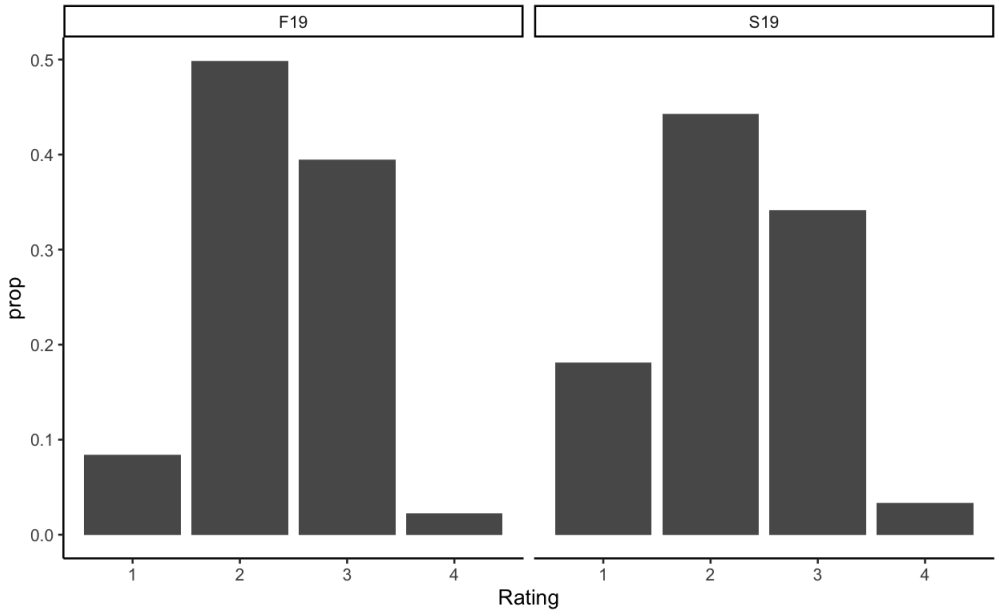



Figure 5: Proportion of ratings assigned in each semester for the full dataset

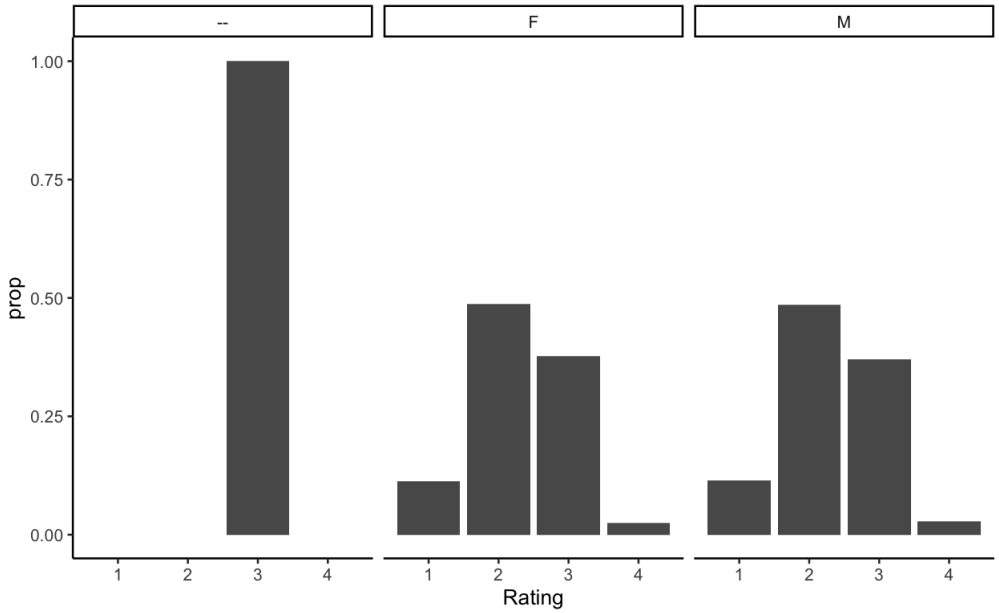



Figure 6: Proportion of ratings assigned for each sex for the full data set

are slight differences in the distribution of ratings by sex when considering each rubric individually.

Ultimately, we have found that the ratings of artifacts does relate to the rubric and rater used to rate it, so we suggest that Dietrich College carefully consider which raters and rubrics they want to use to assess student work in the future to ensure fair and accurate ratings. Note that we did not assess the validity of our models in this analysis, which we might want to do in future research in order to further support the results of our analysis. It also might be interesting, based on our results to research question 4, to fit separate models for each Semester or for each Sex and see if the chosen models differ.

# 7 References

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

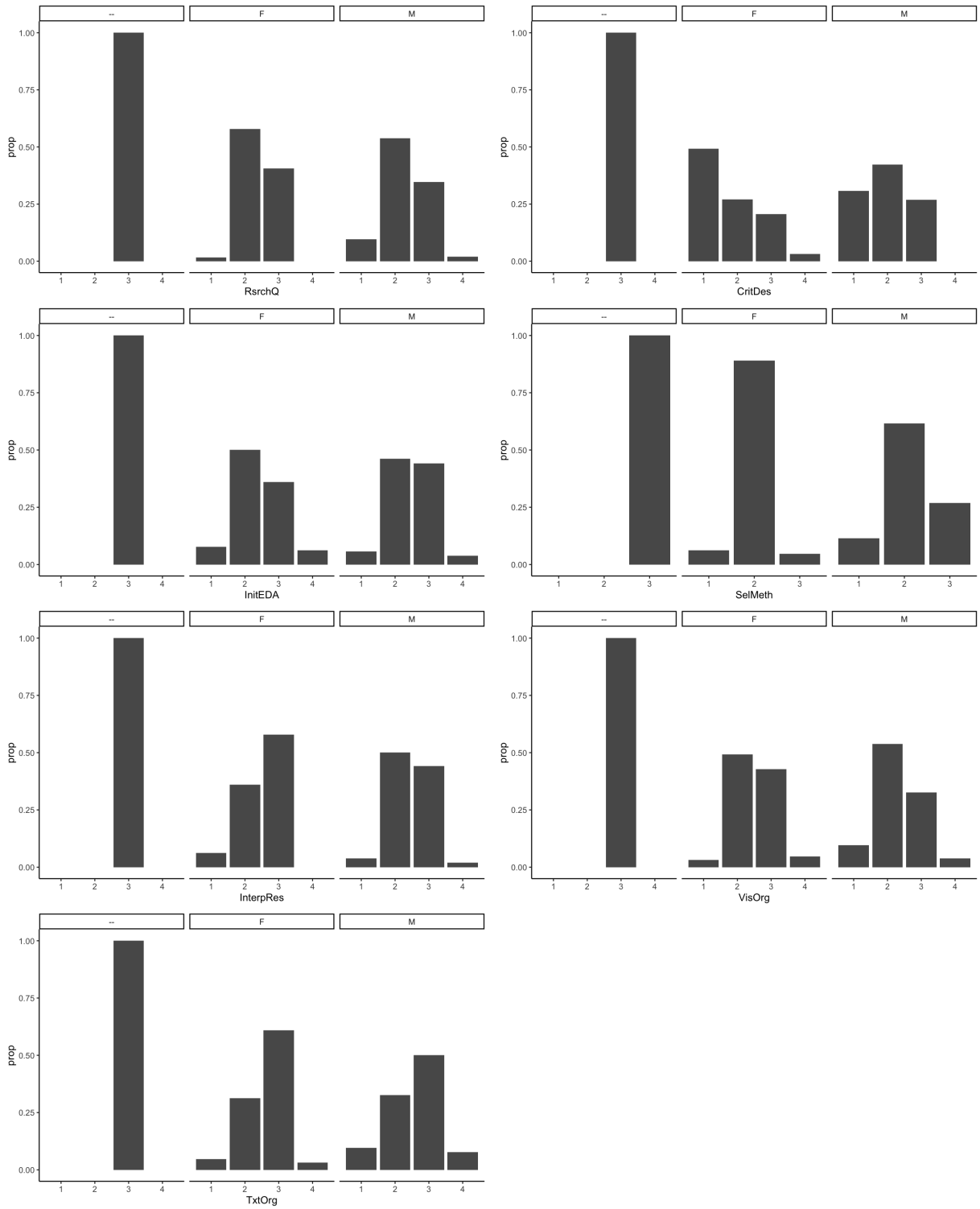



Figure 7: Proportion of ratings assigned for each sex for each rubric for the full data set

# Technical Appendix

## Contents



```
ratings.dat <- read.csv("ratings.csv")
tall.dat <- read.csv("tall.csv")
tall.dat$Rating <- factor(tall.dat$Rating, levels=1:4)

# make missing Sex values consistent between ratings.dat and tall.dat
tall.dat$Sex[nchar(tall.dat$Sex)==0] <- "--"
```

## Is the distribution of ratings for each rubrics pretty much indistinguishable from the other rubrics, or are there rubrics that tend to get especially high or low ratings? Is the distribution of ratings given by each rater pretty much indistinguishable from the other raters, or are there raters that tend to give especially high or low ratings?

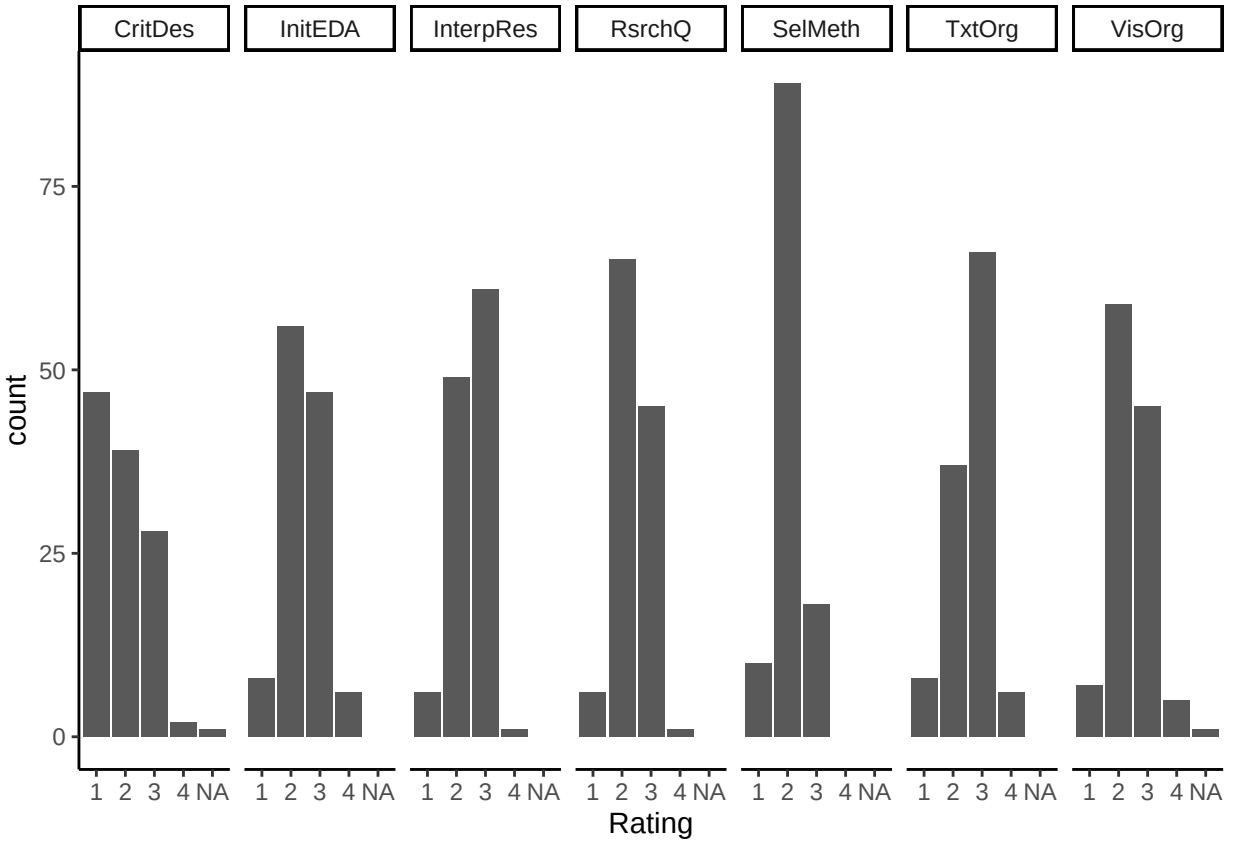
```
summary(ratings.dat[7:13])
```

```
##      RsrchQ        CritDes         InitEDA         SelMeth        InterpRes
##  Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.00   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.00   Median :2.000   Median :2.000   Median :2.000   Median :3.000
```

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

The minimum score for all rubrics is 1. The maximum score is 4 for all rubrics except for SelMeth, which has maximum score 3. The mean score is between 2 and 2.5 for all of the rubrics except for CritDes, which has mean 1.871, and TxtOrg, which has mean 2.598.

```
ggplot(tall.dat, aes(x=Rating, group=Rubric)) +
  geom_bar(na.rm=TRUE) +
  facet_grid(.~Rubric) +
  theme_classic()
```




```
tmp <- tall.dat %>%
  dplyr::group_by(Rubric, Rating) %>%
  dplyr::summarise(count = n())
```

```
## `summarise()` has grouped output by 'Rubric'. You can override using the `.groups` argument.
```

```
tmp
```

```
## # A tibble: 29 x 3
## # Groups:   Rubric [7]
##    Rubric    Rating count
##    <chr>     <fct>  <int>
##  1 CritDes   1         47
##  2 CritDes   2         39
##  3 CritDes   3         28
##  4 CritDes   4          2
##  5 CritDes   <NA>       1
##  6 InitEDA   1          8
##  7 InitEDA   2         56
##  8 InitEDA   3         47
##  9 InitEDA   4          6
## 10 InterpRes 1          6
## # ... with 19 more rows
```

It appears that the distribution of ratings is not the same for all of the rubrics. First, note that the number of ratings of 1 assigned is around 10 for all rubrics except for CritDes, which gives about 50 ratings of 1. Also, note that the number of ratings of 2 assigned is around 50 for all of the rubrics except for SelMeth, which gives over 75 ratings of 2. Lastly, note that the number of ratings of 4 assigned is greater than 0 for all rubrics except for SelMeth, which does not assign any ratings of 4.

Also note that there is some missing data for Rating in the full dataset, which we will need to be careful with in fitting out model.

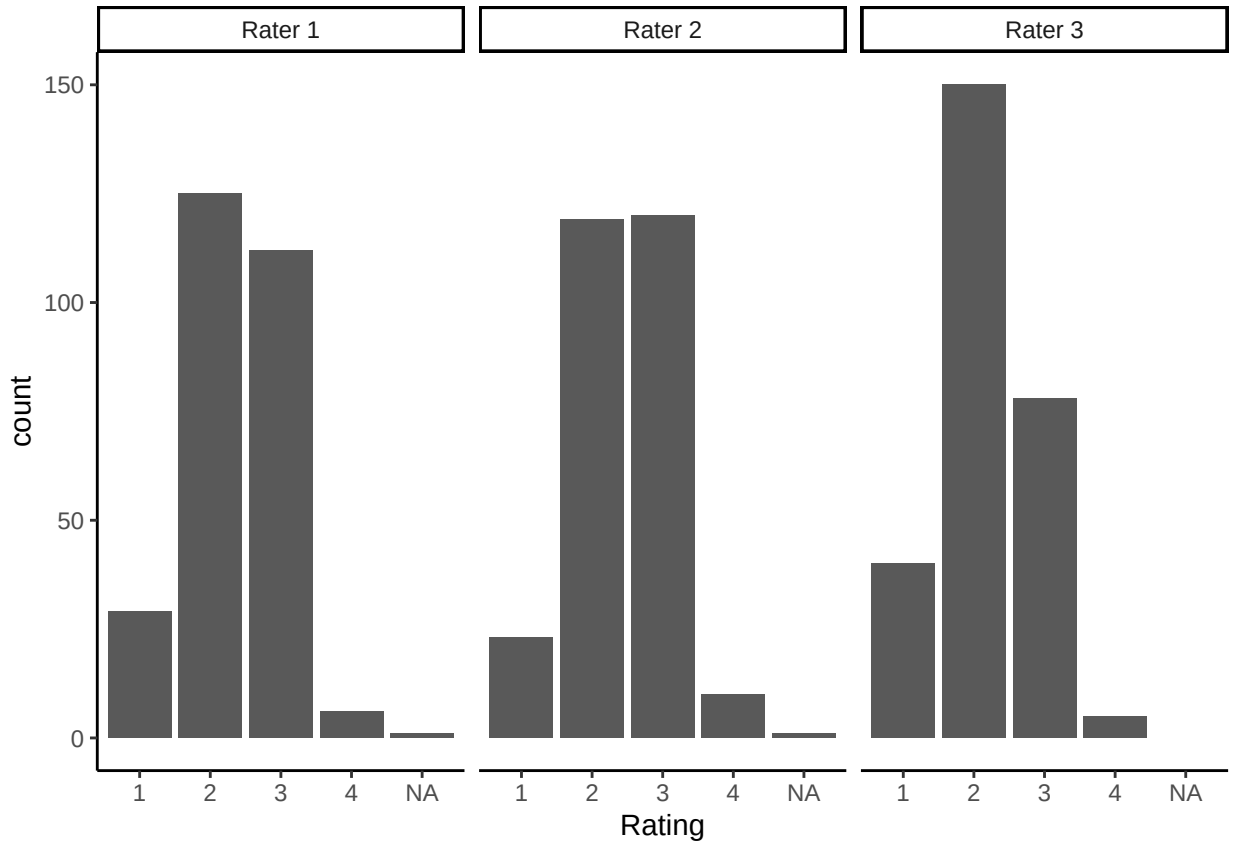
```
table(ratings.dat$Rater)
```

```
##
##  1  2  3
## 39 39 39
```

Each rater rated 39 artifacts (not necessarily unique to that rater).

```
rater.name <- function(x) { paste("Rater",x) }

ggplot(tall.dat, aes(x=Rating, group=Rater)) +
  geom_bar() +
  facet_grid(.~Rater, labeller=labeller(Rater=rater.name)) +
  theme_classic()
```




```
tmp <- tall.dat %>%
  dplyr::group_by(rater.name(Rater), Rating) %>%
  dplyr::summarise(count = n())
```

```
## `summarise()` has grouped output by 'rater.name(Rater)'. You can override using the `.groups` argume
```

```
tmp
```

```
## # A tibble: 14 x 3
## # Groups:   rater.name(Rater) [3]
##    `rater.name(Rater)` Rating count
##    <chr>               <fct>  <int>
##  1 Rater 1             1         29
##  2 Rater 1             2        125
##  3 Rater 1             3        112
##  4 Rater 1             4          6
##  5 Rater 1             <NA>       1
##  6 Rater 2             1         23
##  7 Rater 2             2        119
##  8 Rater 2             3        120
##  9 Rater 2             4         10
## 10 Rater 2             <NA>       1
## 11 Rater 3             1         40
## 12 Rater 3             2        150
## 13 Rater 3             3         78
## 14 Rater 3             4          5
```

It appears that the distribution of ratings given by each rater is pretty much indistinguishable from the other raters. Rater 3 gave slightly more 2's and slightly less 3's than the other two raters, but the distribution of

ratings is pretty similar across all three raters.

Now we will consider only the 13 artifacts seen by all three raters.

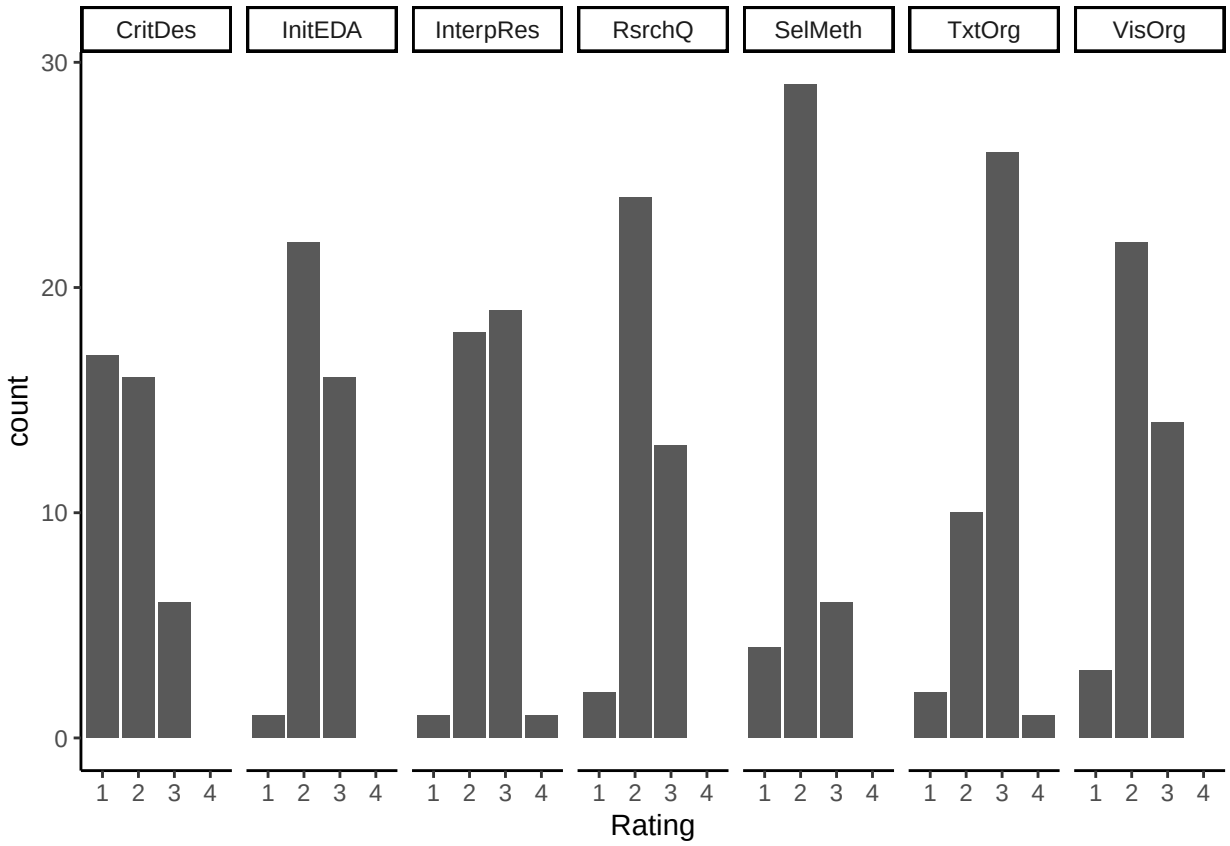
```
sub.ratings <- subset(ratings.dat, ratings.dat$Overlap!='NA')
sub.tall <- subset(tall.dat, tall.dat$Repeated==1)
```

```
summary(sub.ratings[7:13])
```

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

Again, the minimum score for all rubrics is 1. However, the maximum score is 3 for all rubrics except for InterpRes and TxtOrg, which have maximum score 4. The mean score is between 2 and 2.5 for all of the rubrics except for CritDes, which has mean 1.718, and InterpRes and TxtOrg, which have means 2.513 and 2.667, respectively.

```
ggplot(sub.tall, aes(x=Rating, group=Rubric)) +
  geom_bar() +
  facet_grid(.~Rubric) +
  theme_classic()
```




```
tmp <- sub.tall %>%
  dplyr::group_by(Rubric, Rating) %>%
  dplyr::summarise(count = n())
```

```
## `summarise()` has grouped output by 'Rubric'. You can override using the `.groups` argument.
```

```
tmp
```

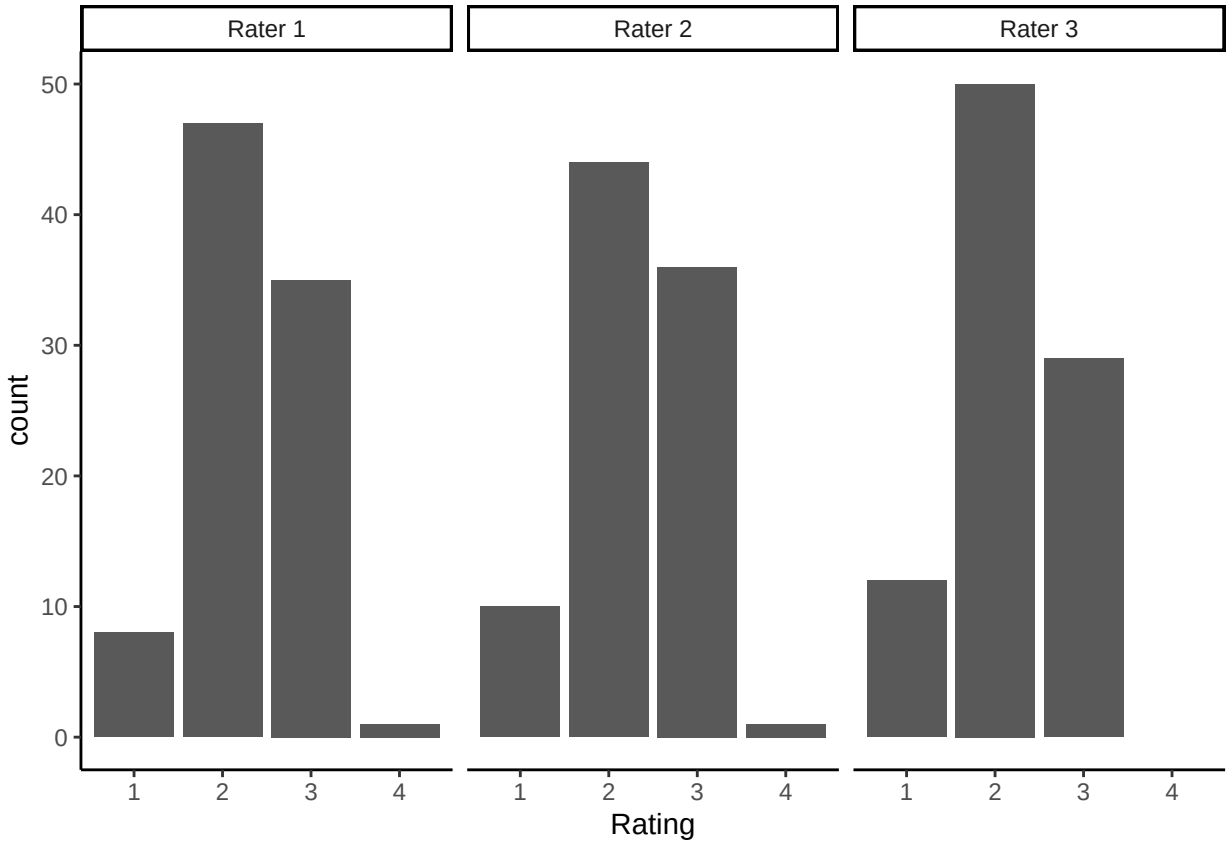
```
## # A tibble: 23 x 3
## # Groups:   Rubric [7]
##    Rubric    Rating count
##    <chr>     <fct>  <int>
##  1 CritDes   1         17
##  2 CritDes   2         16
##  3 CritDes   3          6
##  4 InitEDA   1          1
##  5 InitEDA   2         22
##  6 InitEDA   3         16
##  7 InterpRes 1          1
##  8 InterpRes 2         18
##  9 InterpRes 3         19
## 10 InterpRes 4          1
## # ... with 13 more rows
```

Again, it appears that the distribution of ratings is not the same for all of the rubrics. First, note that the number of ratings of 1 assigned is around 2 for all rubrics except for CritDes, which gives about 17 ratings of 1. Also, note that the number of ratings of 2 assigned is around 20 for all of the rubrics except for SelMeth and TxtOrg, which give close to 30 ratings of 2. Lastly, note that the number ratings of 4 assigned is 0 for all rubrics except for InterpRes and TxtOrg, which have at least one rating of 4.

Note that there is no missing data for rating in this subset of the data, so we will not need to worry about that in any models we fit using this subset.

```
rater.name <- function(x) { paste("Rater",x) }

ggplot(sub.tall, aes(x=Rating, group=Rater)) +
  geom_bar() +
  facet_grid(.~Rater, labeller=labeller(Rater=rater.name)) +
  theme_classic()
```




```
tmp <- sub.tall %>%
  dplyr::group_by(rater.name(Rater), Rating) %>%
  dplyr::summarise(count = n())
```

```
## `summarise()` has grouped output by 'rater.name(Rater)'. You can override using the `.groups` argumer
```

```
tmp
```

```
## # A tibble: 11 x 3
## # Groups:   rater.name(Rater) [3]
##    `rater.name(Rater)` Rating count
##    <chr>               <fct>  <int>
##  1 Rater 1             1          8
##  2 Rater 1             2         47
##  3 Rater 1             3         35
##  4 Rater 1             4          1
##  5 Rater 2             1         10
##  6 Rater 2             2         44
##  7 Rater 2             3         36
```

```
##  8 Rater 2             4          1
##  9 Rater 3             1         12
## 10 Rater 3             2         50
## 11 Rater 3             3         29
```

Again, it appears that the distribution of ratings given by each rater is pretty much indistinguishable from the other raters. Rater 3 gave slightly more 2's and slightly less 3's than the other two raters, but the distribution of ratings is pretty similar across all three raters.

Based on these observations, it does appear that the 13 artifacts seen by all three raters are a pretty good representation of the whole set of 91 artifacts. We see almost identical distributions of ratings by rubric and by rater when looking at only the 13 artifacts seen by all three raters as when looking at all 91 artifacts.

We do have one observation with missing data for Sex, but I do not see a reason to remove it from the dataset so we will leave it as a third sex category.

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

Looking at only the 13 artifacts which were seen by all three graders, we will calculate ICC values.

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: RsrchQ.ratings
## REML criterion at convergence: 66.1533
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.2446
##  Residual             0.5064
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       2.282
```

```
(0.2446^2)/((0.2446^2)+(0.5064^2))
```

```
## [1] 0.1891711
```

```
CritDes.ratings <- sub.tall[sub.tall$Rubric=="CritDes",]
lmer(Rating ~ 1 + (1|Artifact), data=CritDes.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: CritDes.ratings
## REML criterion at convergence: 75.1397
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.5560
##  Residual             0.4804
## Number of obs: 39, groups:  Artifact, 13
```

```
## Fixed Effects:
## (Intercept)
##       1.718
```

```
(0.5560^2)/((0.5560^2)+(0.4804^2))
```

```
## [1] 0.5725587
```

```
InitEDA.ratings <- sub.tall[sub.tall$Rubric=="InitEDA",]
lmer(Rating ~ 1 + (1|Artifact), data=InitEDA.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InitEDA.ratings
## REML criterion at convergence: 56.7573
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.3867
##  Residual             0.3922
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       2.385
```

```
(0.3867^2)/((0.3867^2)+(0.3922^2))
```

```
## [1] 0.4929391
```

```
SelMeth.ratings <- sub.tall[sub.tall$Rubric=="SelMeth",]
lmer(Rating ~ 1 + (1|Artifact), data=SelMeth.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: SelMeth.ratings
## REML criterion at convergence: 50.8562
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.3736
##  Residual             0.3581
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       2.051
```

```
(0.3736^2)/((0.3736^2)+(0.3581^2))
```

```
## [1] 0.521174
```

```
InterpRes.ratings <- sub.tall[sub.tall$Rubric=="InterpRes",]
lmer(Rating ~ 1 + (1|Artifact), data=InterpRes.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InterpRes.ratings
## REML criterion at convergence: 71.0715
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.2899
##  Residual             0.5311
```

```
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       2.513
```

```
(0.2899^2)/((0.2899^2)+(0.5311^2))
```

```
## [1] 0.2295545
```

```
VisOrg.ratings <- sub.tall[sub.tall$Rubric=="VisOrg",]
lmer(Rating ~ 1 + (1|Artifact), data=VisOrg.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: VisOrg.ratings
## REML criterion at convergence: 60.5245
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.4729
##  Residual             0.3922
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       2.282
```

```
(0.4729^2)/((0.4729^2)+(0.3922^2))
```

```
## [1] 0.5924793
```

```
TxtOrg.ratings <- sub.tall[sub.tall$Rubric=="TxtOrg",]
lmer(Rating ~ 1 + (1|Artifact), data=TxtOrg.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: TxtOrg.ratings
## REML criterion at convergence: 74.6212
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.2357
##  Residual             0.5774
## Number of obs: 39, groups:  Artifact, 13
## Fixed Effects:
## (Intercept)
##       2.667
```

```
(0.2357^2)/((0.2357^2)+(0.5774^2))
```

```
## [1] 0.1428337
```

The ICC values are as follows:

RsrchQ: 0.1891711 CritDes: 0.5725587
InitEDA: 0.4929391 SelMeth: 0.521174 InterpRes: 0.2295545 VisOrg: 0.5924793 TxtOrg: 0.1428337

Based on these values, it appears that the graders agree for the most part on the rubrics CritDes, InitEDA, SelMeth and VisOrg and do not agree on the rubrics RsrchQ, InterpRes and TxtOrg. We will now look more closely at the three rubrics that the graders do not appear to agree upon.

First, we will look at the rubric RsrchQ.

```
raters_1_and_2_on_RsrchQ <- data.frame(r1=sub.ratings$RsrchQ[sub.ratings$Rater==1],
                                       r2=sub.ratings$RsrchQ[sub.ratings$Rater==2],
                                       a1=sub.ratings$Artifact[sub.ratings$Rater==1],
                                       a2=sub.ratings$Artifact[sub.ratings$Rater==2])
r1 <- factor(raters_1_and_2_on_RsrchQ$r1,levels=1:4)
r2 <- factor(raters_1_and_2_on_RsrchQ$r2,levels=1:4)
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
```

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

```
raters_1_and_3_on_RsrchQ <- data.frame(r1=sub.ratings$RsrchQ[sub.ratings$Rater==1],
                                       r3=sub.ratings$RsrchQ[sub.ratings$Rater==3],
                                       a1=sub.ratings$Artifact[sub.ratings$Rater==1],
                                       a3=sub.ratings$Artifact[sub.ratings$Rater==3])
r1 <- factor(raters_1_and_3_on_RsrchQ$r1,levels=1:4)
r3 <- factor(raters_2_and_3_on_RsrchQ$r3,levels=1:4)
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
```

For the RsrchQ rubric: The percent exact agreement between raters 1 and 2 is 5/13, which is about 38%. The percent exact agreement between raters 2 and 3 is 7/13, which is about 54%. The percent exact agreement between raters 1 and 3 is 10/13, which is about 77%. So, it appears that rater 1 contributes most to the disagreement on the RsrchQ rubric.

Now, we will look at the rubric InterpRes.

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

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

```
raters_1_and_3_on_InterpRes <- data.frame(r1=sub.ratings$InterpRes[sub.ratings$Rater==1],
                                          r3=sub.ratings$InterpRes[sub.ratings$Rater==3],
                                          a1=sub.ratings$Artifact[sub.ratings$Rater==1],
                                          a3=sub.ratings$Artifact[sub.ratings$Rater==3])
r1 <- factor(raters_1_and_3_on_InterpRes$r1,levels=1:4)
r3 <- factor(raters_2_and_3_on_InterpRes$r3,levels=1:4)
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
```

For the InterpRes rubric: The percent exact agreement between raters 1 and 2 is 8/13, which is about 62%. The percent exact agreement between raters 2 and 3 is 8/13, which is about 62%. The percent exact agreement between raters 1 and 2 is 7/13, which is about 54%. So, it appears the three raters disagree fairly equally on the InterpRes rubric.

Lastly, we will look at the rubric TxtOrg.

```
raters_1_and_2_on_TxtOrg <- data.frame(r1=sub.ratings$TxtOrg[sub.ratings$Rater==1],
                                       r2=sub.ratings$TxtOrg[sub.ratings$Rater==2],
                                       a1=sub.ratings$Artifact[sub.ratings$Rater==1],
                                       a2=sub.ratings$Artifact[sub.ratings$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)
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
```

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

```
raters_1_and_3_on_TxtOrg <- data.frame(r1=sub.ratings$TxtOrg[sub.ratings$Rater==1],
                                       r3=sub.ratings$TxtOrg[sub.ratings$Rater==3],
                                       a1=sub.ratings$Artifact[sub.ratings$Rater==1],
                                       a3=sub.ratings$Artifact[sub.ratings$Rater==3])
r1 <- factor(raters_1_and_3_on_TxtOrg$r1,levels=1:4)
r3 <- factor(raters_2_and_3_on_TxtOrg$r3,levels=1:4)
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
```

For the TxtOrg rubric: The percent exact agreement between raters 1 and 2 is 9/13, which is about 69%. The percent exact agreement between raters 2 and 3 is 7/13, which is about 54%. The percent exact agreement between raters 1 and 2 is 8/13, which is about 62%. So, it appears the three raters disagree fairly equally on the TxtOrg rubric.

Now, we will calculate the ICC values again but using the full dataset.

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: RsrchQ.ratings
## REML criterion at convergence: 211.0659
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.2715
##  Residual             0.5272
## Number of obs: 117, groups:  Artifact, 91
## Fixed Effects:
## (Intercept)
```

```
##       2.358
```

```
(0.2715^2)/((0.2715^2)+(0.5272^2))
```

```
## [1] 0.209617
```

```
CritDes.ratings <- tall.dat[tall.dat$Rubric=="CritDes",]
lmer(Rating ~ 1 + (1|Artifact), data=CritDes.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: CritDes.ratings
## REML criterion at convergence: 277.8691
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.7045
##  Residual             0.4910
## Number of obs: 116, groups:  Artifact, 90
## Fixed Effects:
## (Intercept)
##       1.907
```

```
(0.7045^2)/((0.7045^2)+(0.4910^2))
```

```
## [1] 0.6730667
```

```
InitEDA.ratings <- tall.dat[tall.dat$Rubric=="InitEDA",]
lmer(Rating ~ 1 + (1|Artifact), data=InitEDA.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InitEDA.ratings
## REML criterion at convergence: 240.7763
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.6023
##  Residual             0.4068
## Number of obs: 117, groups:  Artifact, 91
## Fixed Effects:
## (Intercept)
##       2.448
```

```
(0.6023^2)/((0.6023^2)+(0.4068^2))
```

```
## [1] 0.6867283
```

```
SelMeth.ratings <- tall.dat[tall.dat$Rubric=="SelMeth",]
lmer(Rating ~ 1 + (1|Artifact), data=SelMeth.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: SelMeth.ratings
## REML criterion at convergence: 157.7375
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.3329
##  Residual             0.3521
## Number of obs: 117, groups:  Artifact, 91
## Fixed Effects:
```

```
## (Intercept)
##       2.072
```

```
(0.3329^2)/((0.3329^2)+(0.3521^2))
```

```
## [1] 0.4719928
```

```
InterpRes.ratings <- tall.dat[tall.dat$Rubric=="InterpRes",]
lmer(Rating ~ 1 + (1|Artifact), data=InterpRes.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InterpRes.ratings
## REML criterion at convergence: 217.9031
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.2867
##  Residual             0.5398
## Number of obs: 117, groups:  Artifact, 91
## Fixed Effects:
## (Intercept)
##       2.484
```

```
(0.2867^2)/((0.2867^2)+(0.5398^2))
```

```
## [1] 0.2200243
```

```
VisOrg.ratings <- tall.dat[tall.dat$Rubric=="VisOrg",]
lmer(Rating ~ 1 + (1|Artifact), data=VisOrg.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: VisOrg.ratings
## REML criterion at convergence: 226.4172
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.5561
##  Residual             0.3985
## Number of obs: 116, groups:  Artifact, 90
## Fixed Effects:
## (Intercept)
##       2.445
```

```
(0.5561^2)/((0.5561^2)+(0.3985^2))
```

```
## [1] 0.6607148
```

```
TxtOrg.ratings <- tall.dat[tall.dat$Rubric=="TxtOrg",]
lmer(Rating ~ 1 + (1|Artifact), data=TxtOrg.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: TxtOrg.ratings
## REML criterion at convergence: 249.0106
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.3024
##  Residual             0.6285
## Number of obs: 117, groups:  Artifact, 91
```

```
## Fixed Effects:
## (Intercept)
##       2.591
```

```
(0.3024^2)/((0.3024^2)+(0.6285^2))
```

```
## [1] 0.1879828
```

The ICC values (using full dataset vs using 13 common artifacts) are as follows:

RsrchQ: 0.209617 vs 0.1891711 CritDes: 0.6730667 vs 0.5725587
InitEDA: 0.6867283 vs 0.4929391 SelMeth: 0.4719928 vs 0.521174 InterpRes: 0.2200243 vs 0.2295545 VisOrg: 0.6607148 vs 0.5924793 TxtOrg: 0.1879828 vs 0.1428337

The ICC values for each rubric calculated using the full dataset are fairly close (within about 0.10) to the ICC values calculated using the 13 artifacts seen by all three raters with the exception of rubric InitEDA, which has a difference of about 0.20.

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

First using the full dataset, we will first try to incorporate Rater, Semester, Sex and Repeated as fixed effects using model selection methods.

```
#delete missing data for model fitting
tall.dat.tmp <- tall.dat[-c(161,684),]
tall.dat.tmp <- tall.dat.tmp[tall.dat.tmp$Sex!="--",]
```

```
Rubric.names <- sort(unique(tall.dat.tmp$Rubric))
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.dat.tmp[tall.dat.tmp$Rubric==i,]
    tmp <- lmer(Rating ~ -1 + as.factor(Rater) + Semester + Sex + Repeated + (1|Artifact), data=rubric.
  ## do backwards elimination
    tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)
  ## check to see if the raters are significantly different from one another
    tmp.single_intercept <- update(tmp.back_elim, . ~ . + 1 - as.factor(Rater))
    pval <- anova(tmp.single_intercept,tmp.back_elim)$"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
```

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

Now we will look at each of these models more closely.

## RsrchQ

```
RsrchQ.ratings <- tall.dat.tmp[tall.dat.tmp$Rubric=="RsrchQ",]
lmer(Rating ~ (1 | Artifact), data=RsrchQ.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (1 | Artifact)
##    Data: RsrchQ.ratings
## REML criterion at convergence: 209.0797
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.2697
##  Residual             0.5275
## Number of obs: 116, groups:  Artifact, 90
## Fixed Effects:
## (Intercept)
##       2.352
```

```
(0.2697^2)/((0.2697^2)+(0.5275^2))
```

```
## [1] 0.2072344
```

```
summary(lmer(Rating ~ (1 | Artifact), data=RsrchQ.ratings))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (1 | Artifact)
##    Data: RsrchQ.ratings
##
## REML criterion at convergence: 209.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2694 -0.5285 -0.3736  0.9743  2.4770
```

```
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.07276  0.2697
##  Residual             0.27825  0.5275
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.35169    0.05794   40.59
```

```
fixef(lmer(Rating ~ (1 | Artifact), data=RsrchQ.ratings))
```

```
## (Intercept)
##    2.351689
```

### CritDes

```
CritDes.ratings <- tall.dat.tmp[tall.dat.tmp$Rubric=="CritDes",]

# calculate ICC value
CritDes.fit <- lmer(Rating ~ as.factor(Rater) + (1 | Artifact) - 1, data=CritDes.ratings)
CritDes.fit
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ as.factor(Rater) + (1 | Artifact) - 1
##    Data: CritDes.ratings
## REML criterion at convergence: 270.9688
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.6595
##  Residual             0.4972
## Number of obs: 115, groups:  Artifact, 89
## Fixed Effects:
## as.factor(Rater)1  as.factor(Rater)2  as.factor(Rater)3
##             1.686              2.113              1.891
```

```
(0.6595^2)/((0.6595^2)+(0.4972^2))
```

```
## [1] 0.6376039
```

```
# check if fixed effects are significant
summary(CritDes.fit)$coef
```

```
##                   Estimate Std. Error  t value
## as.factor(Rater)1 1.686325  0.1206556 13.97635
## as.factor(Rater)2 2.112884  0.1218849 17.33508
## as.factor(Rater)3 1.890793  0.1218849 15.51294
```

Based on the t values, all of the fixed effects make sense in this model.

```
# check if including Rater significantly improves the model
CritDes.fit2 <- update(CritDes.fit, .~. + 1 - as.factor(Rater))
anova(CritDes.fit2, CritDes.fit)
```

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

```
## Data: CritDes.ratings
## Models:
## CritDes.fit2: Rating ~ (1 | Artifact)
## CritDes.fit: Rating ~ as.factor(Rater) + (1 | Artifact) - 1
##              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## CritDes.fit2    3 277.68 285.91 -135.84   271.68
## CritDes.fit     5 273.62 287.35 -131.81   263.62 8.0535  2    0.01783 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including Rater as a predictor does significantly improve the model.

The model mA cannot be fit, so we do not need to test for random effects for Rater. Thus, our final model for the CritDes rubric is the one produced from backwards elimination earlier.

## InitEDA

```
InitEDA.ratings <- tall.dat.tmp[tall.dat.tmp$Rubric=="InitEDA",]
lmer(Rating ~ (1 | Artifact), data=InitEDA.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (1 | Artifact)
##    Data: InitEDA.ratings
## REML criterion at convergence: 238.9824
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.6042
##  Residual             0.4068
## Number of obs: 116, groups:  Artifact, 90
## Fixed Effects:
## (Intercept)
##       2.442
```

```
(0.6042^2)/((0.6042^2)+(0.4068^2))
```

```
## [1] 0.6880819
```

## SelMeth

```
SelMeth.ratings <- tall.dat.tmp[tall.dat.tmp$Rubric=="SelMeth",]

# calculate ICC value
SelMeth.fit <- lmer(Rating ~ as.factor(Rater) + Semester + (1 | Artifact) - 1, data=SelMeth.ratings)
SelMeth.fit
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ as.factor(Rater) + Semester + (1 | Artifact) - 1
##    Data: SelMeth.ratings
## REML criterion at convergence: 143.5577
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.2996
##  Residual             0.3293
## Number of obs: 116, groups:  Artifact, 90
```

```
## Fixed Effects:
## as.factor(Rater)1  as.factor(Rater)2  as.factor(Rater)3        SemesterS19
##            2.2504             2.2265             2.0332            -0.3586
```

```
(0.2996^2)/((0.2996^2)+(0.3293^2))
```

```
## [1] 0.4528798
```

```
# check if fixed effects are significant
summary(SelMeth.fit)$coef
```

```
##                      Estimate Std. Error   t value
## as.factor(Rater)1  2.2503734 0.07503131 29.992456
## as.factor(Rater)2  2.2265337 0.07423995 29.991047
## as.factor(Rater)3  2.0331606 0.07521048 27.032944
## SemesterS19       -0.3586022 0.09796206 -3.660623
```

Based on the t values, all of the fixed effects make sense in this model.

```
# check if including Rater significantly improves the model
SelMeth.fit2 <- update(SelMeth.fit, .~. + 1 - as.factor(Rater))
anova(SelMeth.fit2, SelMeth.fit)
```

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

```
## Data: SelMeth.ratings
## Models:
## SelMeth.fit2: Rating ~ Semester + (1 | Artifact)
## SelMeth.fit: Rating ~ as.factor(Rater) + Semester + (1 | Artifact) - 1
##              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## SelMeth.fit2    4 145.07 156.08 -68.534   137.07
## SelMeth.fit     6 142.05 158.58 -65.027   130.05 7.0146  2    0.02998 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including Rater as a predictor does significantly improve the model.

```
# check if including Semester significantly improves the model
SelMeth.fit2 <- update(SelMeth.fit, .~. + 1 - Semester)
anova(SelMeth.fit2, SelMeth.fit)
```

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

```
## Data: SelMeth.ratings
## Models:
## SelMeth.fit2: Rating ~ as.factor(Rater) + (1 | Artifact)
## SelMeth.fit: Rating ~ as.factor(Rater) + Semester + (1 | Artifact) - 1
##              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## SelMeth.fit2    5 153.09 166.85 -71.543   143.09
## SelMeth.fit     6 142.05 158.58 -65.027   130.05 13.031  1  0.0003063 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including Semester as a predictor does significantly improve the model.

```
# check for interactions between Rater and Semester
SelMeth.fit2 <- update(SelMeth.fit, .~. + as.factor(Rater)*Semester)
anova(SelMeth.fit2, SelMeth.fit)
```

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

```
## Data: SelMeth.ratings
## Models:
## SelMeth.fit: Rating ~ as.factor(Rater) + Semester + (1 | Artifact) - 1
## SelMeth.fit2: Rating ~ as.factor(Rater) + Semester + (1 | Artifact) + as.factor(Rater):Semester - 1
##              npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## SelMeth.fit     6 142.05 158.58 -65.027   130.05
## SelMeth.fit2    8 143.46 165.49 -63.731   127.46 2.592  2     0.2736
```

Including the interaction between Rater and Semester does not significantly improve the model, so we will leave it out.

The model mA cannot be fit, so we do not need to test for random effects for Semester.

Again, the model mA cannot be fit, so we do not need to test for random effects for Rater. Thus, our final model for the SelMeth rubric is the one produced from backwards elimination earlier.

### InterpRes

```
InterpRes.ratings <- tall.dat.tmp[tall.dat.tmp$Rubric=="InterpRes",]

# calculate ICC value
InterpRes.fit <- lmer(Rating ~ as.factor(Rater) + (1 | Artifact) - 1, data=InterpRes.ratings)
InterpRes.fit
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ as.factor(Rater) + (1 | Artifact) - 1
##    Data: InterpRes.ratings
## REML criterion at convergence: 199.6794
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.2495
##  Residual             0.5025
## Number of obs: 116, groups:  Artifact, 90
## Fixed Effects:
## as.factor(Rater)1  as.factor(Rater)2  as.factor(Rater)3
##             2.704              2.586              2.139
```

```
(0.2495^2)/((0.2495^2)+(0.5025^2))
```

```
## [1] 0.1977727
```

```
# check if fixed effects are significant
summary(InterpRes.fit)$coef
```

```
##                   Estimate Std. Error  t value
## as.factor(Rater)1 2.704214 0.08912484 30.34186
## as.factor(Rater)2 2.585742 0.08912484 29.01259
## as.factor(Rater)3 2.139182 0.09026675 23.69845
```

Based on the t values, all of the fixed effects make sense in this model.

```
# check if including Rater significantly improves the model
InterpRes.fit2 <- update(InterpRes.fit, .~. + 1 - as.factor(Rater))
anova(InterpRes.fit2, InterpRes.fit)
```

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

```
## Data: InterpRes.ratings
## Models:
## InterpRes.fit2: Rating ~ (1 | Artifact)
## InterpRes.fit: Rating ~ as.factor(Rater) + (1 | Artifact) - 1
##                npar    AIC    BIC   logLik deviance  Chisq Df Pr(>Chisq)
## InterpRes.fit2    3 218.53 226.79 -106.263   212.53
## InterpRes.fit     5 200.66 214.43  -95.331   190.66 21.864  2  1.787e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including Rater as a predictor does significantly improve the model.

The model mA cannot be fit, so we do not need to test for random effects for Rater. Thus, our final model for the InterpRes rubric is the one produced from backwards elimination earlier.

## VisOrg

```
VisOrg.ratings <- tall.dat.tmp[tall.dat.tmp$Rubric=="VisOrg",]

# calculate ICC value
VisOrg.fit <- lmer(Rating ~ as.factor(Rater) + (1 | Artifact) - 1, data=VisOrg.ratings)
VisOrg.fit
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ as.factor(Rater) + (1 | Artifact) - 1
##    Data: VisOrg.ratings
## REML criterion at convergence: 219.5832
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.5392
##  Residual             0.3830
## Number of obs: 115, groups:  Artifact, 89
## Fixed Effects:
## as.factor(Rater)1  as.factor(Rater)2  as.factor(Rater)3
##             2.378              2.649              2.284
```

```
(0.5392^2)/((0.5392^2)+(0.3830^2))
```

```
## [1] 0.6646539
```

```
# check if fixed effects are significant
summary(VisOrg.fit)$coef
```

```
##                   Estimate Std. Error  t value
## as.factor(Rater)1 2.377941 0.09658396 24.62045
## as.factor(Rater)2 2.648913 0.09563943 27.69687
## as.factor(Rater)3 2.283545 0.09658396 23.64311
```

Based on the t values, all of the fixed effects make sense in this model.

```
# check if including Rater significantly improves the model
VisOrg.fit2 <- update(VisOrg.fit, .~. + 1 - as.factor(Rater))
anova(VisOrg.fit2, VisOrg.fit)
```

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

```
## Data: VisOrg.ratings
```

```
## Models:
## VisOrg.fit2: Rating ~ (1 | Artifact)
## VisOrg.fit: Rating ~ as.factor(Rater) + (1 | Artifact) - 1
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## VisOrg.fit2    3 227.21 235.44 -110.60   221.21
## VisOrg.fit     5 220.82 234.54 -105.41   210.82 10.392  2   0.005539 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including Rater as a predictor does significantly improve the model.

The model mA cannot be fit, so we do not need to test for random effects for Rater. Thus, our final model for the VisOrg rubric is the one produced from backwards elimination earlier.

### TxtOrg

```
TxtOrg.ratings <- tall.dat.tmp[tall.dat.tmp$Rubric=="TxtOrg",]
lmer(Rating ~ (1 | Artifact), data=TxtOrg.ratings)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (1 | Artifact)
##    Data: TxtOrg.ratings
## REML criterion at convergence: 247.5354
## Random effects:
##  Groups   Name        Std.Dev.
##  Artifact (Intercept) 0.3061
##  Residual             0.6291
## Number of obs: 116, groups:  Artifact, 90
## Fixed Effects:
## (Intercept)
##       2.587
```

```
(0.3061^2)/((0.3061^2)+(0.6291^2))
```

```
## [1] 0.1914282
```

```
fixef(lmer(Rating ~ (1 | Artifact), data=TxtOrg.ratings))
```

```
## (Intercept)
##    2.587453
```

The ICC values using full dataset for the final models (which are the same as the ones produced by backwards elimination) are as follows:

RsrchQ: 0.2072344 CritDes: 0.6376039
InitEDA: 0.6880819 SelMeth: 0.4528798 InterpRes: 0.1977727 VisOrg: 0.6646539 TxtOrg: 0.1914282

The ICC's from the final models agree for the most part with the earlier ICC's.

Based on these analyses, it appears that Rater is most strongly related to Rating since it appears in all models which are not just the simple random-intercept models.

Now, we will consider only the 13 artifacts seen by all three raters (we remove the Repeated variable from consideration now because it is the same for all observations in the subset of the dataset).

```
Rubric.names <- sort(unique(sub.tall$Rubric))
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 <- sub.tall[sub.tall$Rubric==i,]
    tmp <- lmer(Rating ~ -1 + as.factor(Rater) + Semester + Sex + (1|Artifact), data=rubric.data,REML=FA
  ## do backwards elimination
    tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)
  ## check to see if the raters are significantly different from one another
    tmp.single_intercept <- update(tmp.back_elim, . ~ . + 1 - as.factor(Rater))
    pval <- anova(tmp.single_intercept,tmp.back_elim)$"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
## Rating ~ (1 | Artifact)
##
## $InitEDA
## Rating ~ (1 | Artifact)
##
## $InterpRes
## Rating ~ (1 | Artifact)
##
## $RsrchQ
## Rating ~ (1 | Artifact)
##
## $SelMeth
## Rating ~ (1 | Artifact)
##
## $TxtOrg
## Rating ~ (1 | Artifact)
##
## $VisOrg
## Rating ~ (1 | Artifact)
```

Backwards elimination chose the simple random-intercept model for each rubric. Because of this, we do not need to check for interactions or random effects. (The ICC values would be identical to the ones calculated earlier)

Lastly, we will consider fitting one combined model for all rubrics using the full dataset

```
combined.fit <- lmer(Rating ~ 1 + (0+Rubric|Artifact), data=tall.dat.tmp)
```

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

```
summary(combined.fit)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (0 + Rubric | Artifact)
```

```
##    Data: tall.dat.tmp
##
## 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
```

```
combined.fit2 <- update(combined.fit, .~. + as.factor(Rater) + Semester + Sex + Repeated + Rubric)
summary(combined.fit2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester +
##     Sex + Repeated + Rubric
##    Data: tall.dat.tmp
##
## 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
```

```
## as.factor(Rater)2  0.001977   0.054887   0.036
## as.factor(Rater)3 -0.174867   0.055045  -3.177
## SemesterS19       -0.175017   0.087850  -1.992
## SexM               0.010506   0.081271   0.129
## Repeated          -0.073586   0.098522  -0.747
## RubricInitEDA      0.547054   0.095710   5.716
## RubricInterpRes    0.587091   0.100893   5.819
## RubricRsrchQ       0.460875   0.087516   5.266
## RubricSelMeth      0.164863   0.094265   1.749
## RubricTxtOrg       0.692880   0.099523   6.962
## RubricVisOrg       0.530182   0.099136   5.348
##
## Correlation of Fixed Effects:
##             (Intr) a.(R)2 a.(R)3 SmsS19 SexM   Repetd RbIEDA RbrcIR RbrcRQ
## as.fctr(R)2 -0.245
## as.fctr(R)3 -0.237  0.499
## SemesterS19 -0.361  0.008  0.000
## SexM        -0.398 -0.026 -0.035  0.302
## Repeated    -0.154  0.001 -0.003  0.079  0.009
## RubrcIntEDA -0.552 -0.001  0.000 -0.001  0.000  0.007
## RbrcIntrpRs -0.660 -0.001  0.000 -0.001  0.000 -0.009  0.734
## RubrcRsrchQ -0.626 -0.001  0.000 -0.001  0.000 -0.039  0.585  0.756
## RubricSlMth -0.689 -0.001  0.000 -0.001  0.000 -0.088  0.659  0.777  0.689
## RubrcTxtOrg -0.611 -0.001  0.000 -0.001  0.000  0.005  0.674  0.751  0.682
## RubricVsOrg -0.607 -0.001 -0.001 -0.002 -0.001 -0.021  0.715  0.745  0.668
##             RbrcSM RbrcTO
## as.fctr(R)2
## as.fctr(R)3
## SemesterS19
## SexM
## Repeated
## RubrcIntEDA
## RbrcIntrpRs
## RubrcRsrchQ
## RubricSlMth
## RubrcTxtOrg  0.725
## RubricVsOrg  0.680  0.750
```

```
combined.back_elim <- fitLMER.fnc(combined.fit2, log.file.name = FALSE)
```

```
summary(combined.back_elim)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester +
##     Rubric
##    Data: tall.dat.tmp
##
## 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
## as.factor(Rater)2  0.0003231  0.0547446   0.006
## as.factor(Rater)3 -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) a.(R)2 a.(R)3 SmsS19 RbIEDA RbrcIR RbrcRQ RbrcSM RbrcTO
## as.fctr(R)2 -0.281
## as.fctr(R)3 -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
```

```
# test for interactions
comb.inter <- update(combined.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| = 0.00431172 (tol = 0.002, component 1)
```

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

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

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

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

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

Only the interaction between Rater and Rubric is kept by backwards elimination

```
anova(combined.back_elim,comb.inter_elim)
```

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

```
## Data: tall.dat.tmp
## Models:
## combined.back_elim: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric
## comb.inter_elim: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric + as.factor
##                     npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## combined.back_elim   39 1464.0 1647.2 -693.02   1386.0
## comb.inter_elim      51 1454.5 1694.1 -676.26   1352.5 33.526 12   0.000801 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significant evidence that including the interaction between Rater and Rubric improves the model.

```
# check for random effects
m0 <- comb.inter_elim
mA <- lmer(Rating ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) | Artifact) + as.factor(Rater) + Se
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00347545 (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.dat.tmp
## Models:
## m0: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric + as.factor(Rater):Rubri
## mA: Rating ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) | Artifact) + as.factor(Rater) + Semest
##    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m0   51 1454.5 1694.1 -676.26   1352.5
## mA   57 1415.9 1683.6 -650.94   1301.9 50.647  6  3.487e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significant evidence that including random effects for Rater improves the model.

```
mA <- lmer(Rating ~ (0 + Rubric | Artifact) + (0 + Semester | Artifact) + as.factor(Rater) + Semester +
```

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

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

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

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

```
## Data: tall.dat.tmp
## Models:
## m0: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric + as.factor(Rater):Rubri
## mA: Rating ~ (0 + Rubric | Artifact) + (0 + Semester | Artifact) + as.factor(Rater) + Semester + Rubi
##    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m0   51 1454.5 1694.1 -676.26   1352.5
```

```
## mA   54 1458.4 1712.0 -675.18   1350.4 2.1534  3     0.5412
```

No significant evidence that including random effects for Semester improves the model.

The model mA cannot be fit, so we do not need to worry about testing for this random effect.

Thus, the final model includes random effects for Rater and the interaction between Rater and Rubric.

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

```
## fixed-effect model matrix is rank deficient so dropping 1 column / coefficient
```

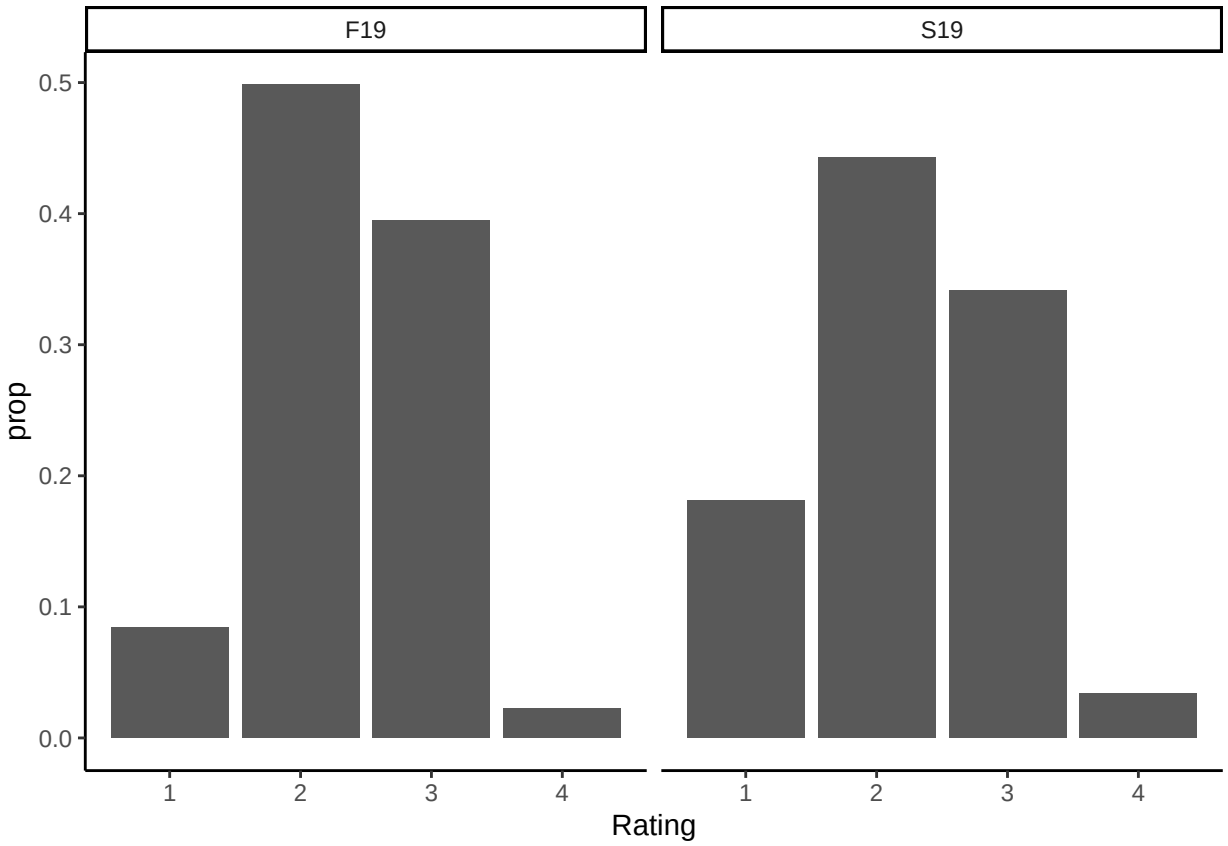
```
## boundary (singular) fit: see ?isSingular
```

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

```
##                                    Estimate Std. Error    t value
## (Intercept)                      2.33250840 0.10170831 22.9333122
## SemesterS19                     -0.15918237 0.07647761 -2.0814245
## as.factor(Rater)1:RubricCritDes -0.57494785 0.14367924 -4.0016070
## as.factor(Rater)2:RubricCritDes -0.20890482 0.14628476 -1.4280696
## as.factor(Rater)3:RubricCritDes -0.37903926 0.13418031 -2.8248501
## as.factor(Rater)1:RubricInitEDA  0.16454200 0.12894375  1.2760758
## as.factor(Rater)2:RubricInitEDA  0.23078272 0.13112179  1.7600639
## as.factor(Rater)3:RubricInitEDA  0.06572843 0.11288966  0.5822361
## as.factor(Rater)1:RubricInterpRes  0.41657075 0.11861779  3.5118742
## as.factor(Rater)2:RubricInterpRes  0.26938064 0.12135056  2.2198549
## as.factor(Rater)3:RubricInterpRes -0.10237522 0.10563258 -0.9691633
## as.factor(Rater)1:RubricRsrchQ   0.15122961 0.12505496  1.2093052
## as.factor(Rater)2:RubricRsrchQ   0.02986976 0.12773693  0.2338381
## as.factor(Rater)3:RubricRsrchQ   0.02477928 0.11412693  0.2171203
## as.factor(Rater)1:RubricSelMeth -0.16427065 0.11896097 -1.3808786
## as.factor(Rater)2:RubricSelMeth -0.18459440 0.12231371 -1.5091882
## as.factor(Rater)3:RubricSelMeth -0.35551658 0.11391256 -3.1209602
## as.factor(Rater)1:RubricTxtOrg   0.44083158 0.12545589  3.5138372
## as.factor(Rater)2:RubricTxtOrg   0.25583093 0.12779231  2.0019274
## as.factor(Rater)3:RubricTxtOrg   0.19185163 0.10885046  1.7625247
## as.factor(Rater)1:RubricVisOrg   0.07930650 0.11732284  0.6759681
## as.factor(Rater)2:RubricVisOrg   0.34044562 0.11964179  2.8455411
```

## Is there anything else interesting to say about this data?

```
ggplot(tall.dat, aes(x=Rating, group=Semester)) +
  geom_bar(na.rm=TRUE, mapping = aes(x = Rating, y = ..prop.., group = Semester), stat = "count") +
  facet_grid(.~Semester) +
  theme_classic()
```




```
tmp <- tall.dat %>%
  dplyr::group_by(Semester, Rating) %>%
  dplyr::summarise(count = n())
```

```
## `summarise()` has grouped output by 'Semester'. You can override using the `.groups` argument.
```

```
tmp
```

```
## # A tibble: 10 x 3
## # Groups:   Semester [2]
##    Semester Rating count
##    <chr>     <dbl> <int>
##  1 F19           1    49
##  2 F19           2   289
##  3 F19           3   229
##  4 F19           4    13
##  5 F19          NA     1
##  6 S19           1    43
##  7 S19           2   105
##  8 S19           3    81
##  9 S19           4     8
## 10 S19          NA     1
```

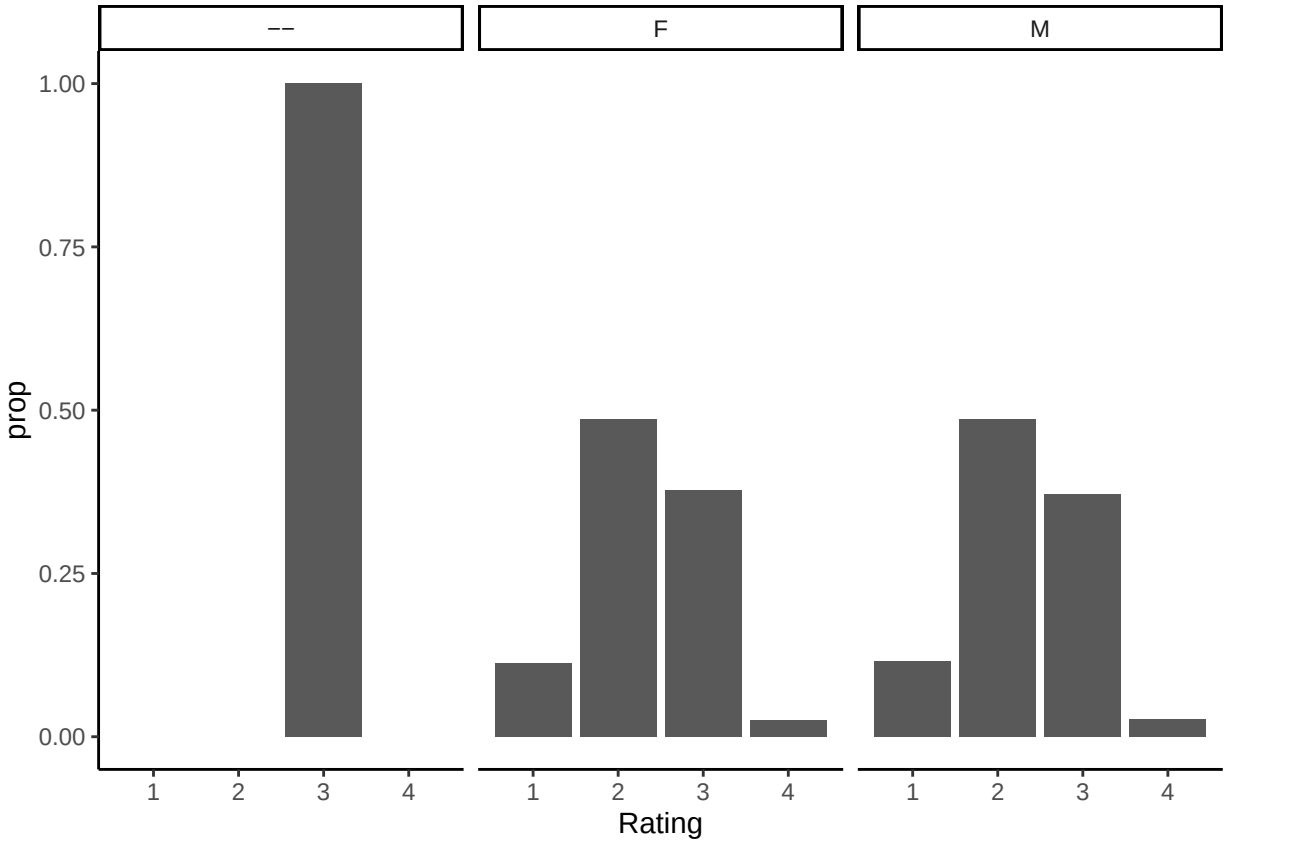
```
tmp <- tall.dat %>%
  dplyr::group_by(Semester) %>%
  dplyr::summarise(count = n())
tmp
```

```
## # A tibble: 2 x 2
```

```
##   Semester count
##   <chr>    <int>
## 1 F19        581
## 2 S19        238
```

We have a lot less papers for the Spring semester than for the Fall semester. Each semester has about the same number of artifacts with rating 1 and 4, but the Spring has a lot less artifacts with rating 2 and 3 than the Fall semester.

```
ggplot(tall.dat, aes(x=Rating, group=Sex)) +
  geom_bar(na.rm=TRUE,mapping = aes(x = Rating, y = ..prop.., group = Sex), stat = "count") +
  facet_grid(.~Sex) +
  theme_classic()
```




```
tmp <- tall.dat %>%
  dplyr::group_by(Sex, Rating) %>%
  dplyr::summarise(count = n())
```

```
## `summarise()` has grouped output by 'Sex'. You can override using the `.groups` argument.
```

```
tmp
```

```
## # A tibble: 10 x 3
## # Groups:   Sex [3]
##    Sex   Rating count
##    <chr>  <dbl> <int>
##  1 --         3     7
##  2 F          1    50
##  3 F          2   217
```

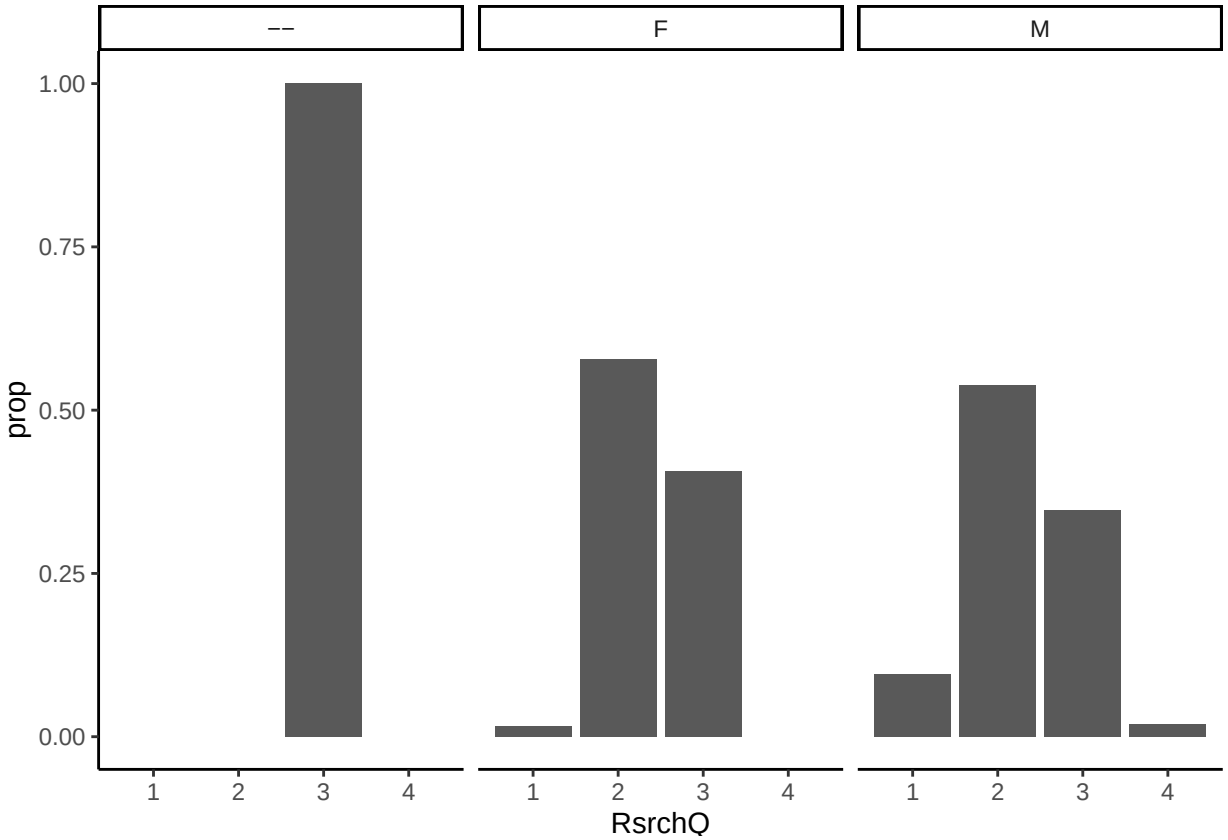
```
##  4 F          3   168
##  5 F          4    11
##  6 F         NA     2
##  7 M          1    42
##  8 M          2   177
##  9 M          3   135
## 10 M          4    10
```

```
tmp <- tall.dat %>%
  dplyr::group_by(Sex) %>%
  dplyr::summarise(count = n())
tmp
```

```
## # A tibble: 3 x 2
##   Sex   count
##   <chr> <int>
## 1 --        7
## 2 F       448
## 3 M       364
```

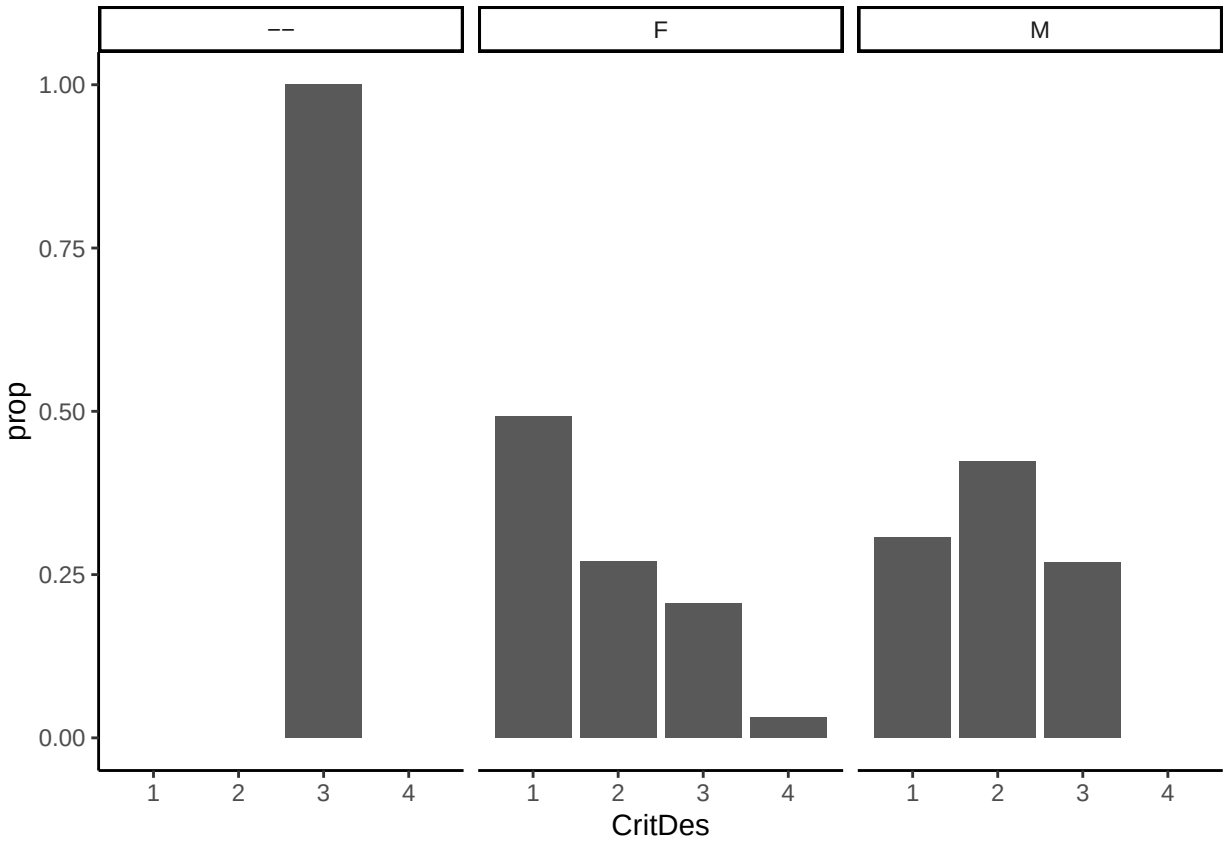
About the same distribution of ratings for males and females.

```
ggplot(ratings.dat, aes(x=RsrchQ, group=Sex)) +
  geom_bar(na.rm=TRUE,mapping = aes(x = RsrchQ, y = ..prop.., group = Sex), stat = "count")+
  facet_grid(.~Sex) +
  theme_classic()
```

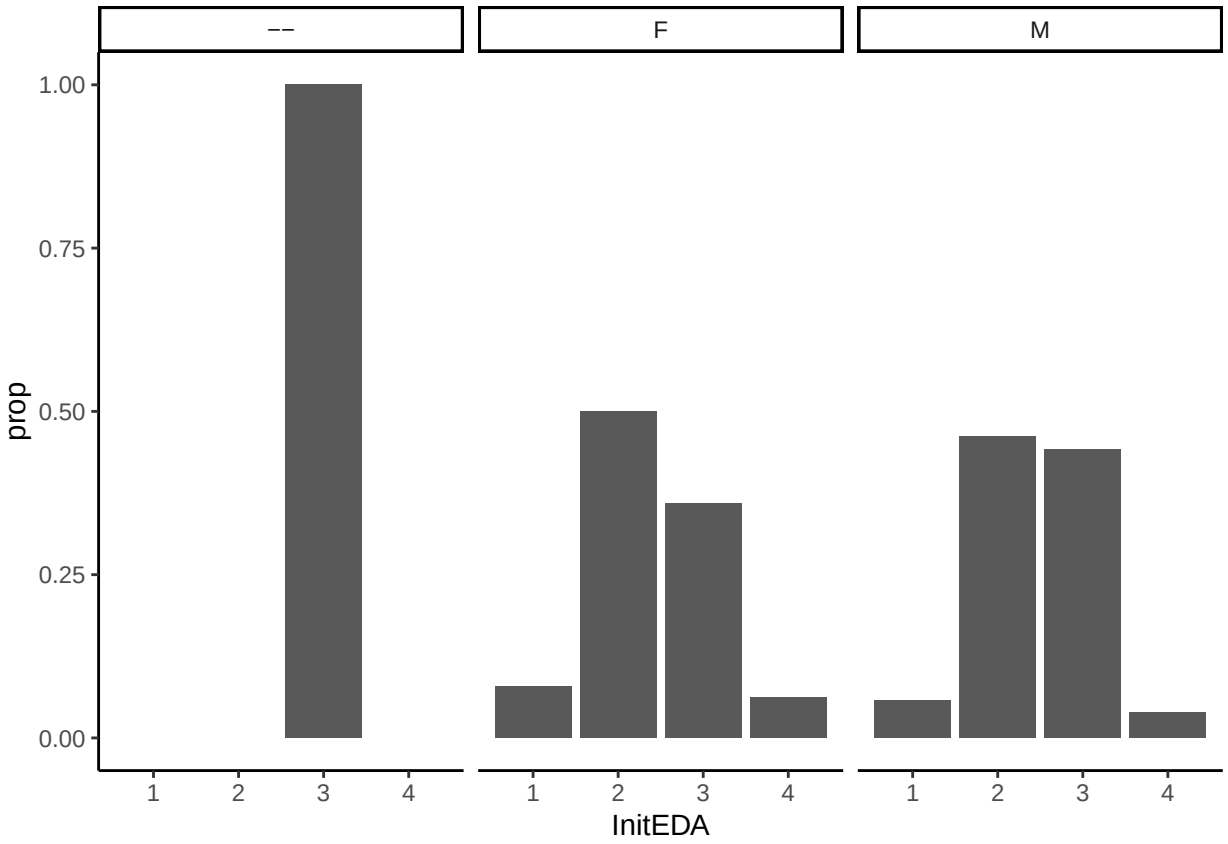



```
ggplot(ratings.dat, aes(x=CritDes, group=Sex)) +
  geom_bar(na.rm=TRUE,mapping = aes(x = CritDes, y = ..prop.., group = Sex), stat = "count")+
```

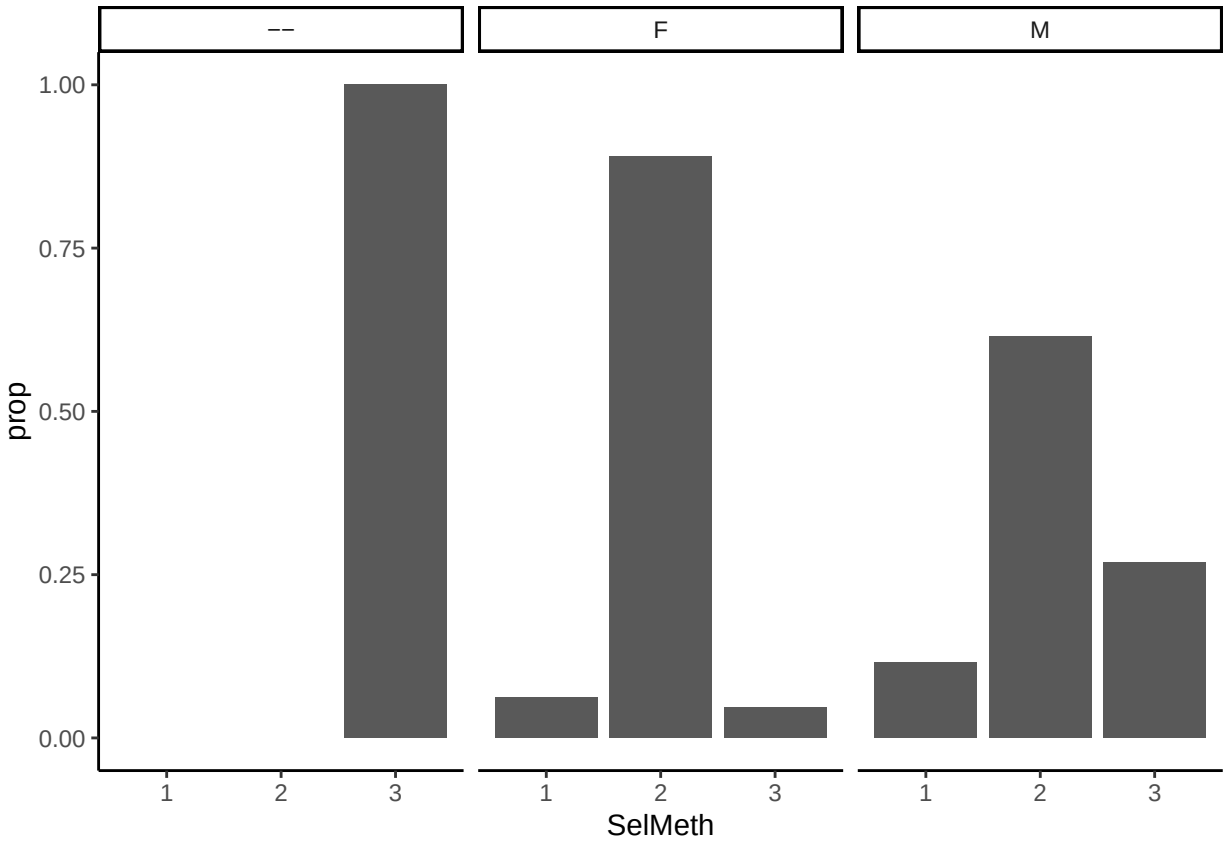
```
  facet_grid(.~Sex) +
  theme_classic()
```



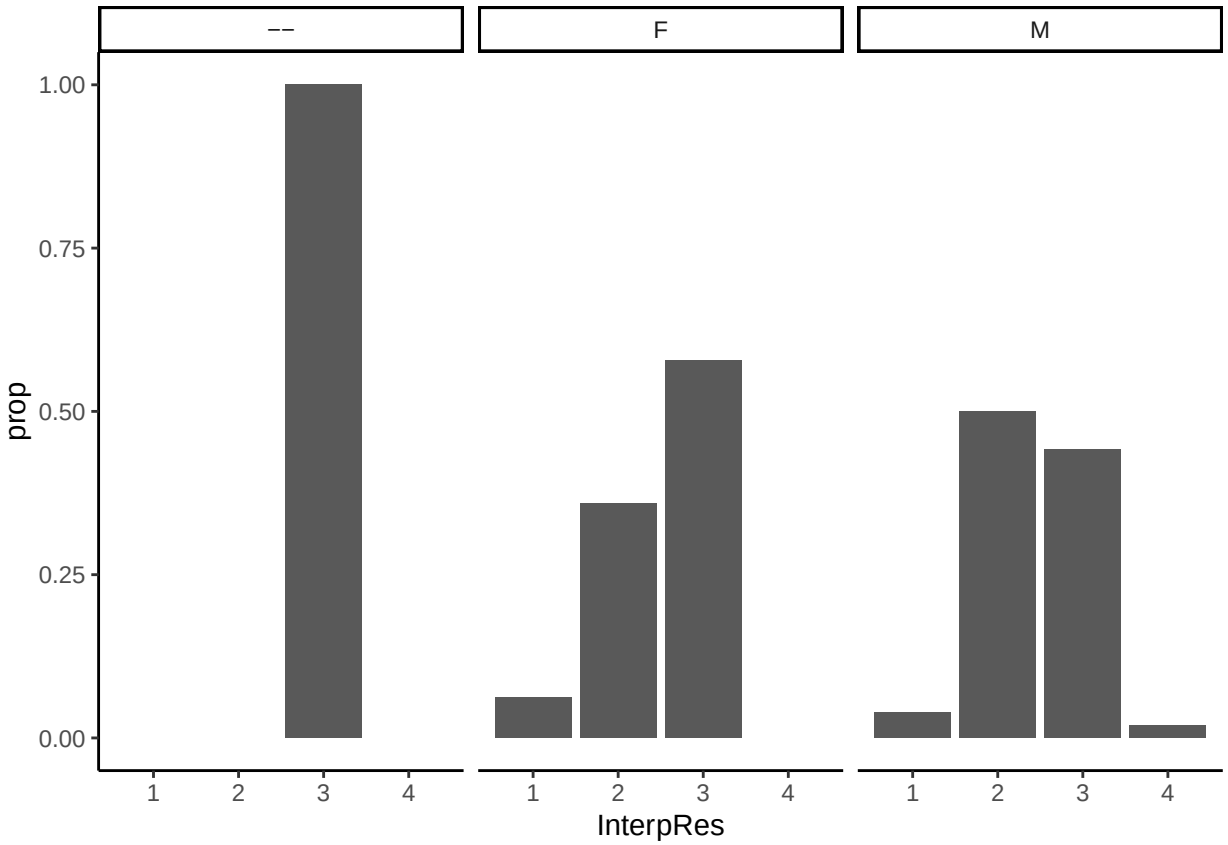

```
ggplot(ratings.dat, aes(x=InitEDA, group=Sex)) +
  geom_bar(na.rm=TRUE,mapping = aes(x = InitEDA, y = ..prop.., group = Sex), stat = "count")+
  facet_grid(.~Sex) +
  theme_classic()
```



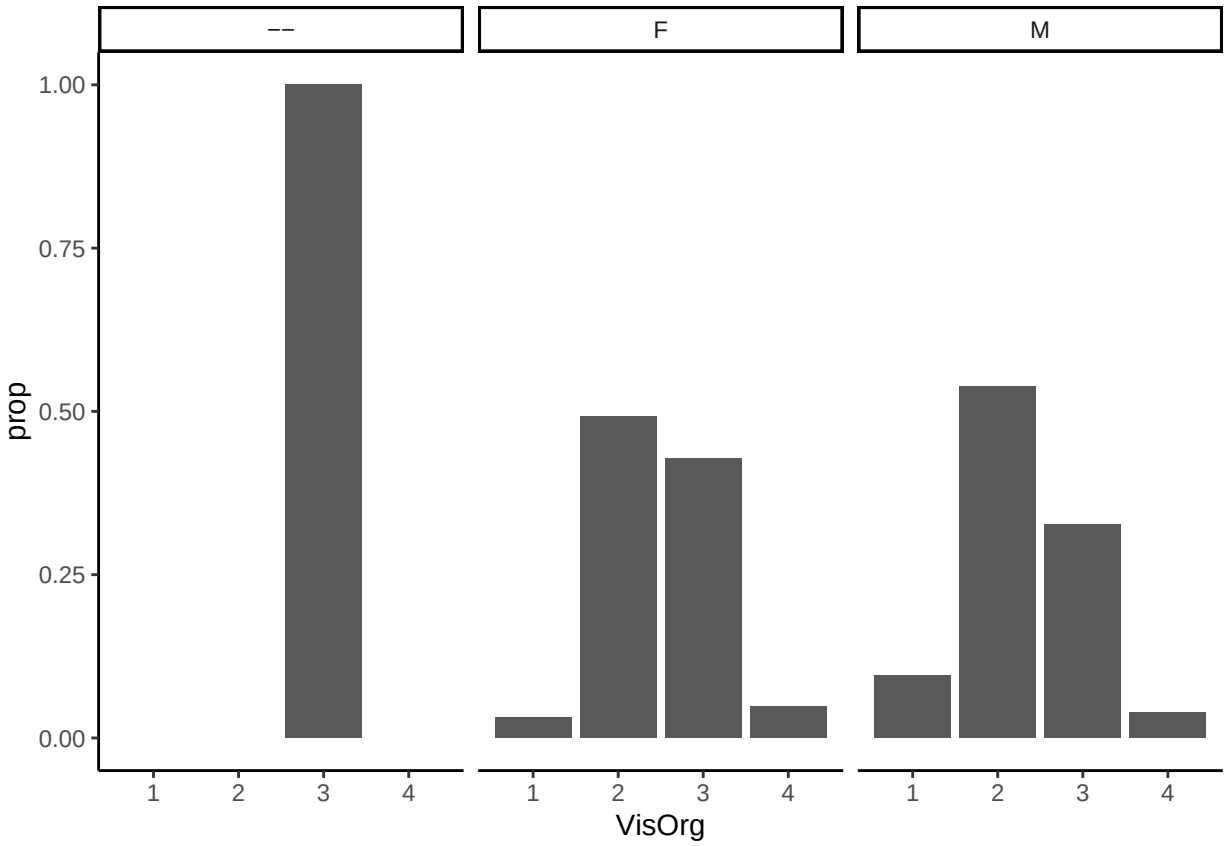

```
ggplot(ratings.dat, aes(x=SelMeth, group=Sex)) +
  geom_bar(na.rm=TRUE,mapping = aes(x = SelMeth, y = ..prop.., group = Sex), stat = "count")+
  facet_grid(.~Sex) +
  theme_classic()
```



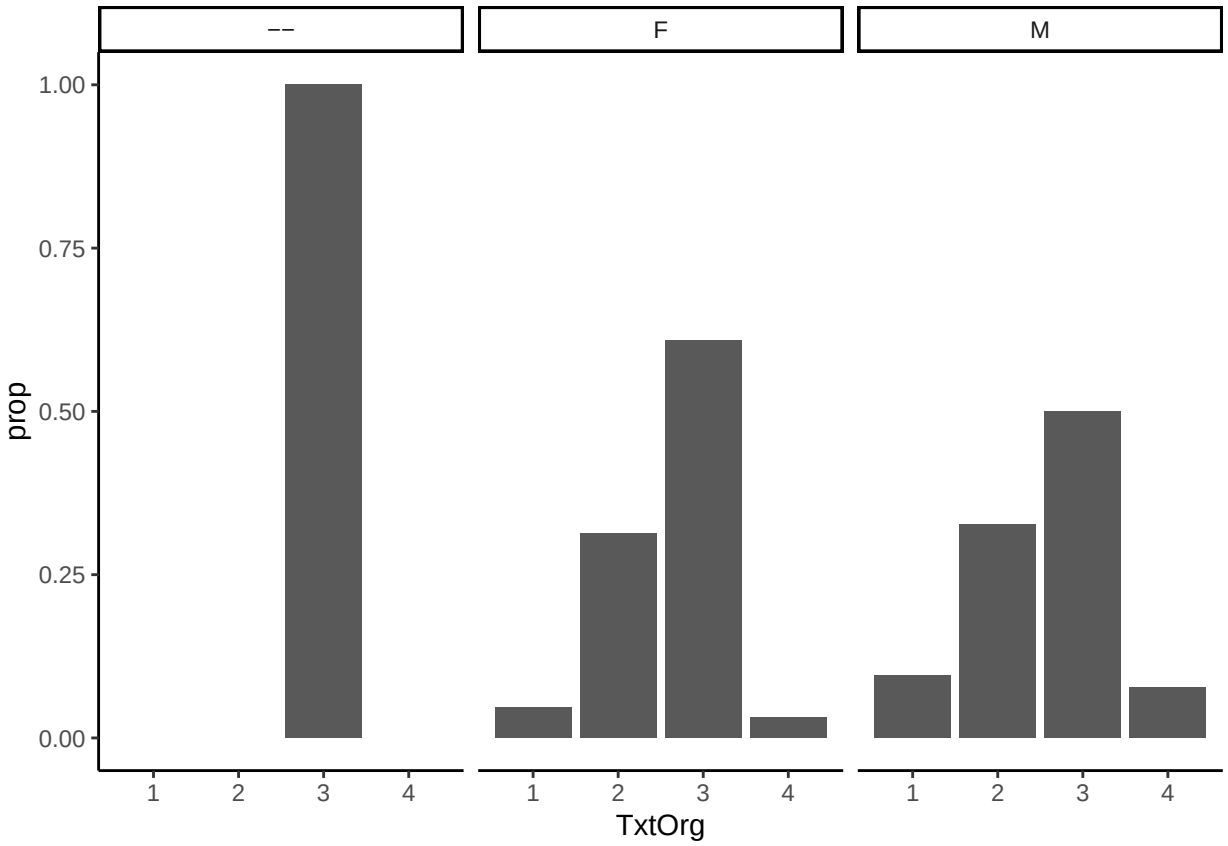

```
ggplot(ratings.dat, aes(x=InterpRes, group=Sex)) +
  geom_bar(na.rm=TRUE,mapping = aes(x = InterpRes, y = ..prop.., group = Sex), stat = "count")+
  facet_grid(.~Sex) +
  theme_classic()
```




```
ggplot(ratings.dat, aes(x=VisOrg, group=Sex)) +
  geom_bar(na.rm=TRUE,mapping = aes(x = VisOrg, y = ..prop.., group = Sex), stat = "count")+
  facet_grid(.~Sex) +
  theme_classic()
```




```
ggplot(ratings.dat, aes(x=TxtOrg, group=Sex)) +
  geom_bar(na.rm=TRUE,mapping = aes(x = TxtOrg, y = ..prop.., group = Sex), stat = "count")+
  facet_grid(.~Sex) +
  theme_classic()
```




Looking at proportions to account for the different number of males and females in the dataset, females seem to do worse on the CritDes rubric than males since females have a much higher proportion of 1's than males with this rubric. Females seem to do more mediocre on the SelMeth rubric than males since they have a much higher proportion of 2's and lower proportion of 1's and 3's than males on this rubric. It seems that females seem to do better than males on the InterpRes rubric since they score a much higher proportion of 3's and much lower proportion of 2's than males on this rubric. Overall, makes and females perform fairly equally on all rubrics, so it will be difficult to determine the Sex of the artifact with missing Sex data from just ratings.