Grade A Performance: Examining the Equity and Effectiveness of CMU's Grading System

Caleb Pena, cpea@andrew.cmu.edu

November 29, 2021

Abstract

To be written.

Introduction

The struggle to earn and maintain good grades is an essential part of the college experience. Grades matter to students because they can influence the decisions of graduate school admissions boards and of potential employers. But grades also matter to the universities themselves. Maintaining databases of grades allows schools to understand where and how students are struggling, and whether conscious or unconscious biases are influencing professors' evaluations.

As Carnegie Mellon redesigns its general education program, Dietrich College has a unique opportunity to reassess its grading practices to determine whether students are being adequately and fairly assisted. This paper analyzes rated papers from an undergraduate statistics course to identify trends and find potential areas for improvement. In particular, the dean has asked us to focus our research on answering the following questions:

- Is the distribution of ratings for each rubric more or less indistinguishable from the other rubrics, or are there rubrics that tend to have especially high or low ratings? Is the distribution of ratings given by each rater more or less indistinguishable from the other raters, or are there raters that tend to give especially high or low ratings?
- 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?
- 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?
- Is there anything else interesting to say about this data?

Data

Our data comes from a recent experiment conducted through the Dietrich College of Humanities and Social Sciences in the Spring and Fall semesters of 2019. The college asked three raters, each from a different department, to review papers submitted for a freshman statistics class. They were asked to rate the students' performance across seven areas on a scale of one to four. A full description of these seven rubrics is provided in Table 1.

The dataset tracks the sex of the students and the semester the paper was from in addition to the ratings themselves. In total, 91 papers (known in the experiment as "artifacts") were reviewed. 13 of these were reviewed by all three raters for a total of 117 unique evaluations.

A deeper breakdown of the data including detailed descriptions of the grading distributions may be found in the results section.

Table 1: Description of Rubrics

Full Name	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.).

Methods

We broke our analysis into four subsections related to each of the questions posed to us. First, we were asked to identify whether or not the distribution of ratings depended on either the rater or the rubric. To examine this relationship, we built histograms to visually inspect the conditional rating distributions. We also performed two sets of hypothesis tests. We conducted a chi-squared test on the rubric question and conducted a chi-squared test and a Fisher's exact test on the rater question.

Next, we looked more closely at the artifacts that were evaluated by multiple raters. As in the first subsection, we built conditional histograms to see how each rater graded these papers. We also computed a metric known as intra-class correlation. This measure tells us the pairwise correlation between the ratings of different raters scoring the same artifact. In addition, we reported the percentage of the ratings that pairs of raters scored identically. Taken together, these methods give us a good idea of which raters if any behaved differently from the rest.

To identify how the other factors (e.g. semester, sex, etc.) were related to the ratings, we built a mixed effects model grouped by individual artifacts. Using backwards elimination, we identified the most useful fixed effects and we interpret their coefficients in the Results section. We also explored a number of candidate models using different combinations of random effects. For these alternate models, please consult Part C of the technical appendix.

Finally, the client asked whether we uncovered any other worthwhile information in our analysis. Using an approach similar to what we did for the first question, we analyzed the difference in ratings across the two semesters. We inspected the distribution visually using histograms and conducted a series of chi-squared tests (one for each rubric) to evaluate if there were any distributional differences across the semesters.

Results

Aaaanndd... unfortunately, that's all I've got at the moment. I would appreciate if you looked over the technical appendix and gave feedback on the work there. Thank you to the reviewers! You guys are the best.

Technical Appendix

Part A

Question: *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??*

The table below and the collection of bar graphs show the spread of ratings for each rubric. Let's highlight a few important takeaways:

- Raters give out 4s sparingly. Aside from cases of truly exceptional work, raters will typically give out grades no higher than 3.
- Raters show a similar reluctance to hand out grades of 1 everywhere except in Critique Design. In that rubric, 1s are actually the most common rating given.
- Very few students selected their methods appropriately. More than 3/4 of SelMeth ratings were 2s.

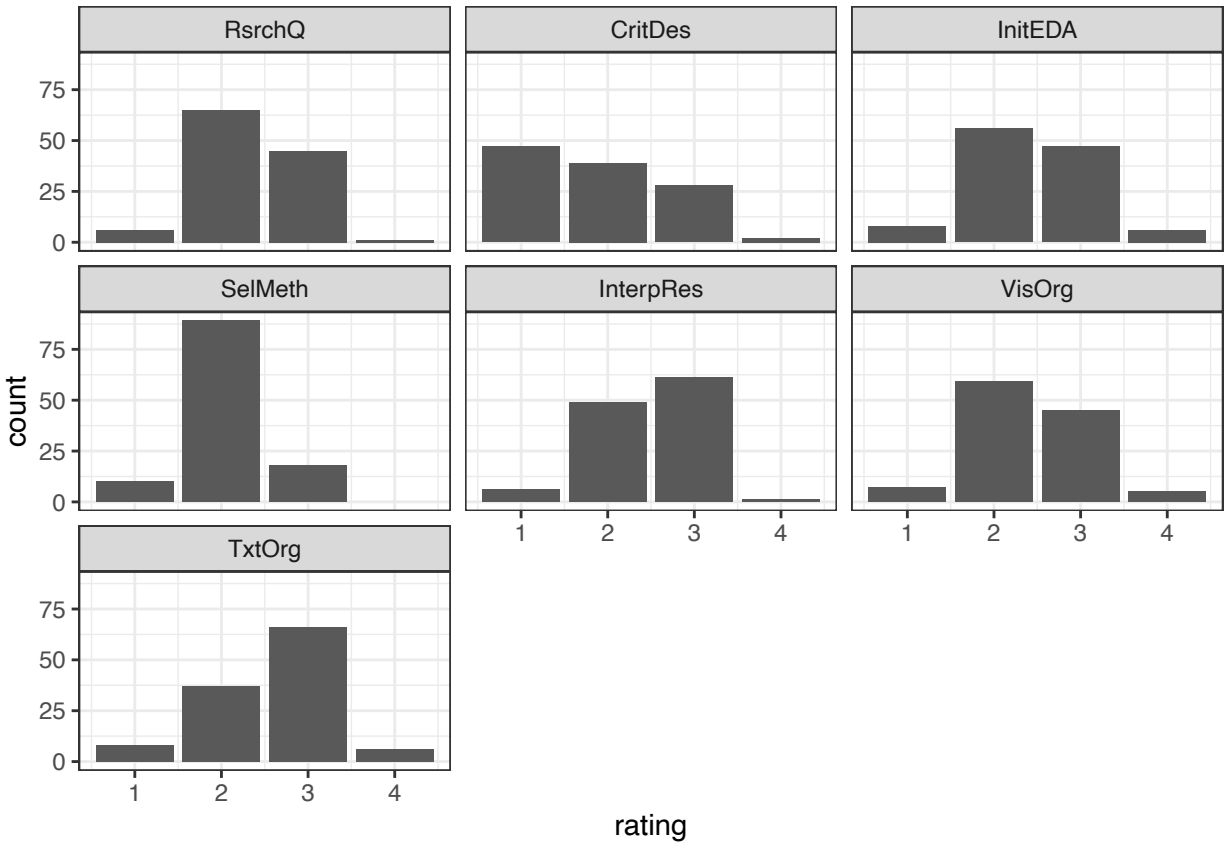
```
ratings_tall %>%
  group_by(rubric, rating) %>%
  summarise(count = n(),
            percent = count/117)
```

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

```
## # A tibble: 29 x 4
## # Groups:   rubric [7]
##    rubric    rating count percent
##    <chr>      <dbl> <int>   <dbl>
##  1 CritDes        1    47 0.402
##  2 CritDes        2    39 0.333
##  3 CritDes        3    28 0.239
##  4 CritDes        4     2 0.0171
##  5 CritDes       NA     1 0.00855
##  6 InitEDA        1     8 0.0684
##  7 InitEDA        2    56 0.479
##  8 InitEDA        3    47 0.402
##  9 InitEDA        4     6 0.0513
## 10 InterpRes      1     6 0.0513
## # ... with 19 more rows
```

```
ratings_tall %>%
  mutate(rubric = factor(rubric, levels = unique(rubric))) %>%
  ggplot(aes(x = rating)) +
  geom_bar() +
  theme_bw() +
  facet_wrap(vars(rubric))
```

```
## Warning: Removed 2 rows containing non-finite values (stat_count).
```



The above graph provides strong evidence that different rubrics come with different rating expectations. To add a little more statistical rigor to this conclusion, we can consider the results of a chi-square test to evaluate the spread of the counts. The test does provide a small p-value but it comes with a major caveat. Since the same artifacts have several ratings spread across the rubrics, the data is not truly independent. Further work needs to be done to evaluate this assumption.

```
chisq.test(table(ratings_tall$rubric, ratings_tall$rating))
```

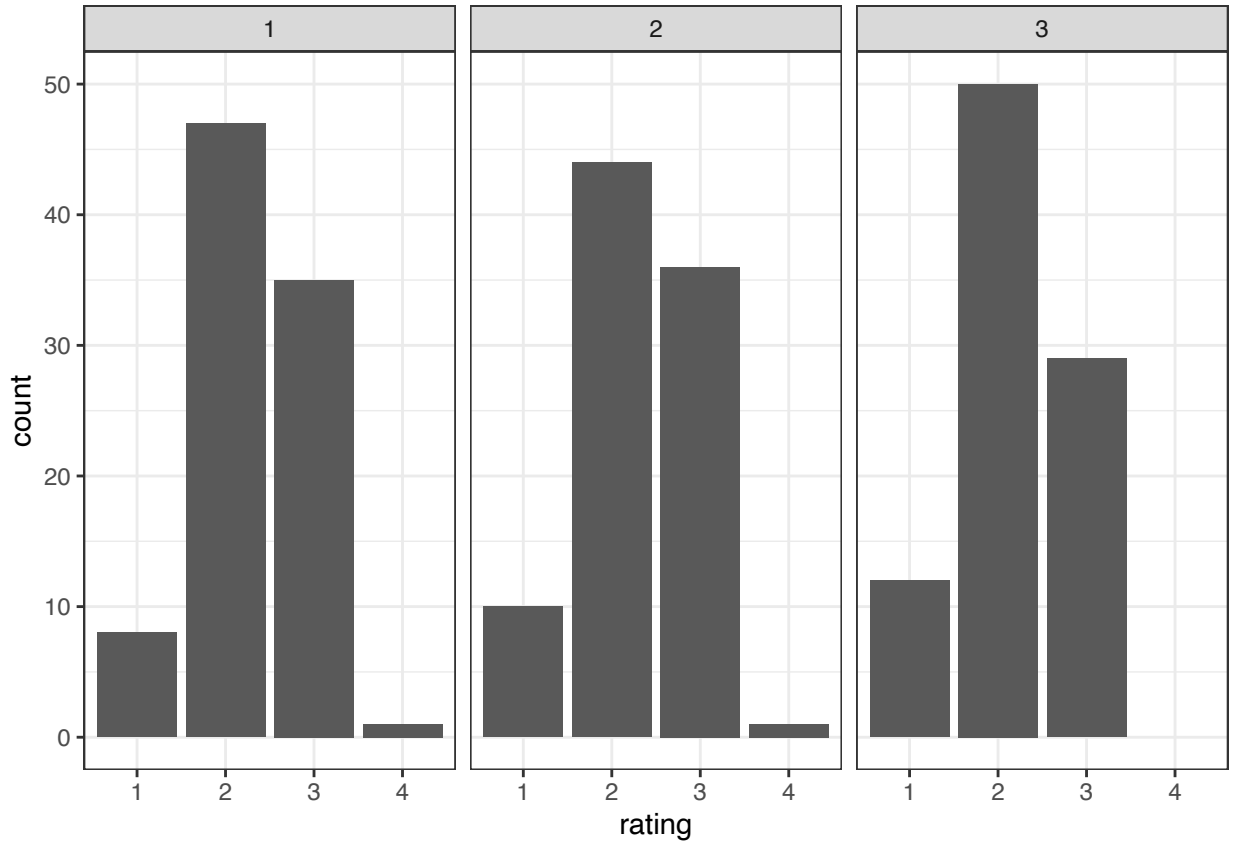
```
## Warning in stats::chisq.test(x, y, ...): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(ratings_tall$rubric, ratings_tall$rating)
## X-squared = 188.47, df = 18, p-value < 2.2e-16
```

Next we look at the distribution of ratings across raters. The overall pattern appears to be the same across raters. Rater 3 appears to be a slightly harsher grader but not significantly so. That these differences are relatively minor is confirmed by the results of the chi-squared and fisher tests run below. Note the same caveat as before.

```
ratings_repeated <- ratings_tall %>% filter(repeated == 1) %>% mutate(rating = factor(rating))
ratings_repeated %>%
  ggplot(aes(x = rating)) +
```

```
  geom_bar() +
  theme_bw() +
  facet_wrap(vars(rater))
```




```
chisq.test(table(ratings_repeated$rater, ratings_repeated$rating))
```

```
## Warning in stats::chisq.test(x, y, ...): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(ratings_repeated$rater, ratings_repeated$rating)
## X-squared = 3.043, df = 6, p-value = 0.8034
```

```
fisher.test(table(ratings_repeated$rater, ratings_repeated$rating))
```

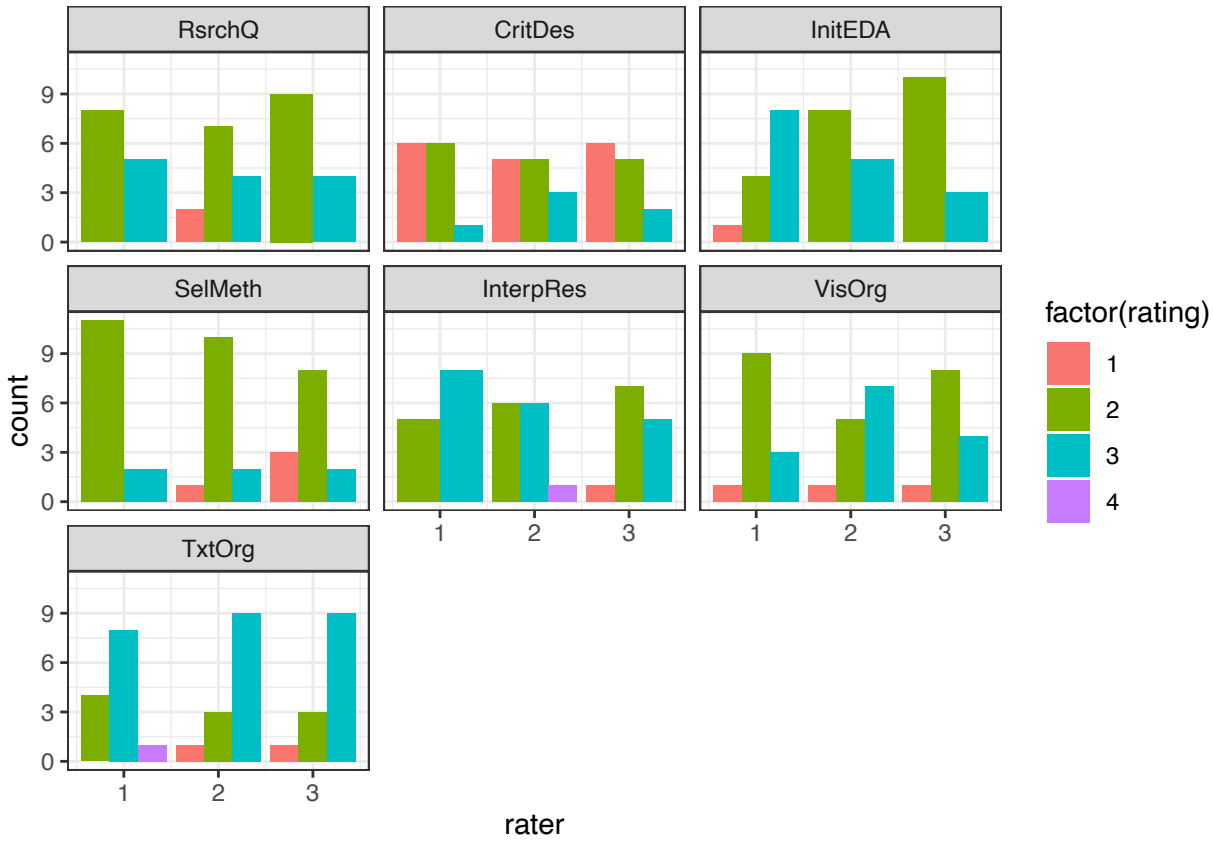
```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(ratings_repeated$rater, ratings_repeated$rating)
## p-value = 0.8069
## alternative hypothesis: two.sided
```

Part B

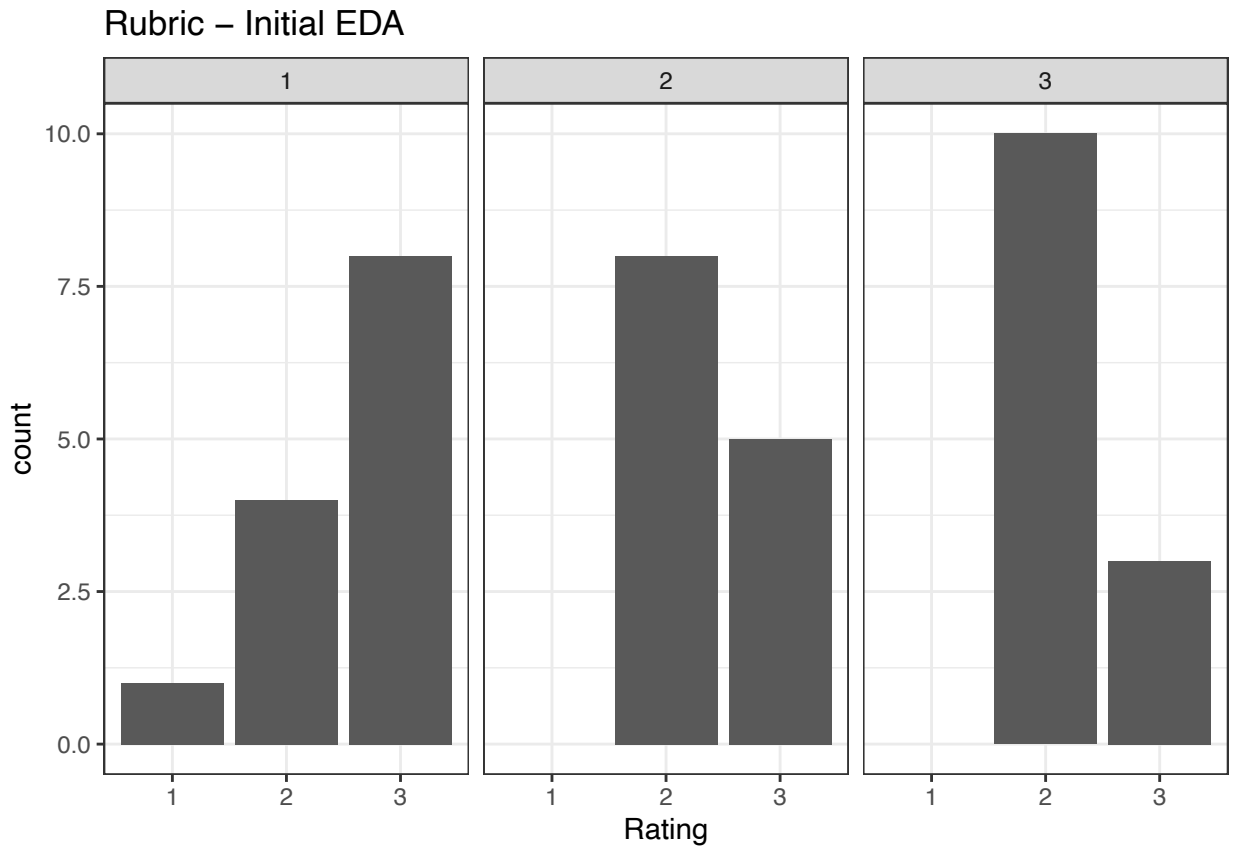
Question: *For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees with the others? Or do they all disagree?*

We are interested in answering the same question as before only subsetted by rubric. The graph below gives some idea of the differences in spread. The strongest differences emerge in the `InitEDA` and `VisOrg` categories. However, this is not a foolproof method to evaluate whether the raters tended to agree or disagree. Distributions might look similar even though raters are giving artifacts very different scores.

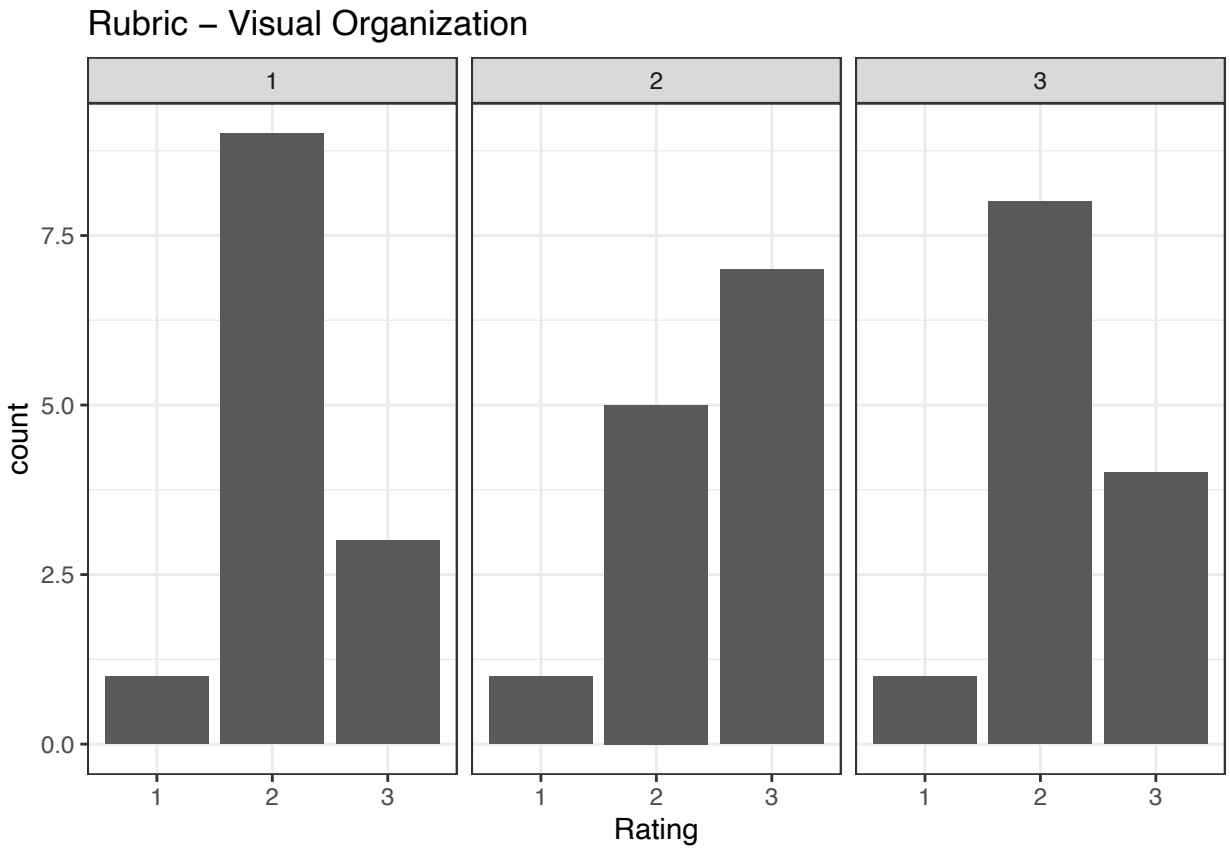
```
ratings_repeated %>%
  mutate(rubric = factor(rubric, levels = unique(ratings_repeated$rubric))) %>%
  ggplot() +
  geom_bar(aes(x = rater, fill = factor(rating)),
           position = "dodge") +
  theme_bw() +
  facet_wrap(vars(rubric))
```




```
ratings_repeated %>%
  filter(rubric == "InitEDA") %>%
  ggplot() +
  geom_bar(aes(x = rating)) +
  theme_bw() +
  labs(x = "Rating",y = "count", title = "Rubric - Initial EDA") +
  facet_wrap(vars(rater))
```




```
ratings_repeated %>%
  filter(rubric == "VisOrg") %>%
  ggplot() +
  geom_bar(aes(x = rating)) +
  theme_bw() +
  labs(x = "Rating",y = "count", title = "Rubric - Visual Organization") +
  facet_wrap(vars(rater))
```



We calculate the intra-class correlations below. These represent the correlation between the different raters' grades of each artifact. Contrary to our expectations from the above graphs, here we see weak correlations for `RsrchQ`, `InterpRes`, and `TxtOrg`. Meanwhile the two rubrics we were concerned about, `InitEDA` and `VisOrg`, have high correlations indicating the raters agreed more than the overall distribution of ratings might indicate.

```
get_ICCs <- function(the_rubric){
  data <- ratings_repeated %>%
    filter(rubric == the_rubric) %>%
    mutate(rating = as.numeric(rating))
  model <- lmer(rating ~ 1 + (1|artifact), data=data)
  tau_2 <- as.data.frame(VarCorr(model))$vcov[1]
  sigma_2 <- as.data.frame(VarCorr(model))$vcov[2]
  return(tau_2/(tau_2 + sigma_2))
}

tibble(Rubric = unique(ratings_repeated$rubric),
       ICC = map_dbl(Rubric, get_ICCs)) %>%
  knitr::kable(caption = "Intra-class correlations")
```

Table 1: Intra-class correlations

Rubric	ICC
RsrchQ	0.1891892
CritDes	0.5725594

Rubric	ICC
InitEDA	0.4929577
SelMeth	0.5212766
InterpRes	0.2295720
VisOrg	0.5924529
TxtOrg	0.1428571

The source of these agreements/disagreements can be pinned down better in the figure below. For example, we can see that the disagreements of how to rate the research questions largely came down to difference between raters 1 and 2.

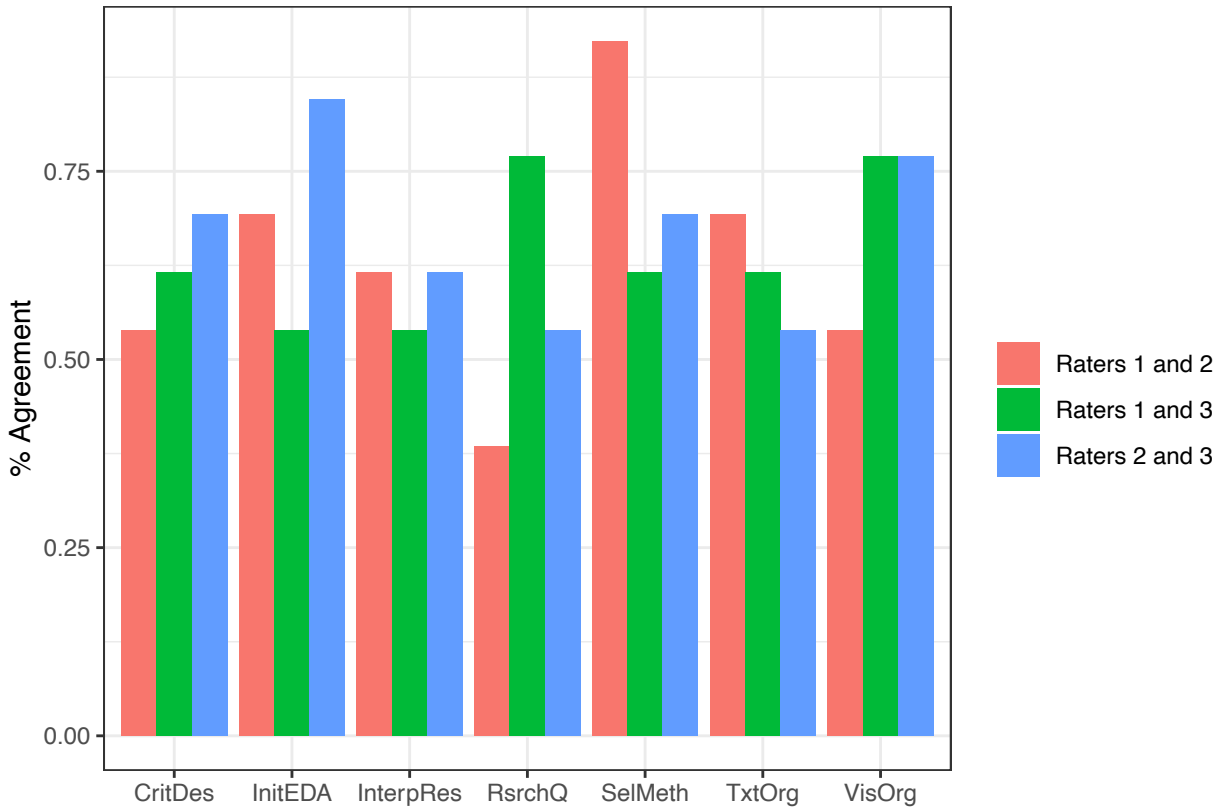
```
get_pct_agreement <- function(rater_1, rater_2, the_rubric){
  data <- (ratings_repeated %>% filter(rubric == the_rubric))
  mean(data[data$rater == rater_1,"rating"] == data[data$rater == rater_2,"rating"])
}

get_pct_agreement(1, 2, "InitEDA")
```

```
## [1] 0.6923077
```

```
get_pairs_agreement <- function(rubric){
  c(get_pct_agreement(1, 2, rubric),
    get_pct_agreement(1, 3, rubric),
    get_pct_agreement(2, 3, rubric))
}
get_summary <- function(rubric){
  tibble(rubric = rep(rubric, 3),
         pair = c("Raters 1 and 2", "Raters 1 and 3", "Raters 2 and 3"),
         pct_agreement = get_pairs_agreement(rubric))
}

map_df(unique(ratings_repeated$rubric), get_summary) %>%
  ggplot() +
  geom_col(aes(x = rubric, y = pct_agreement, fill = pair), position = "dodge") +
  labs(x = "", y = "% Agreement", fill = "") +
  theme_bw()
```



Part C

Question: *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?*

Typical functions designed to perform variable selection on mixed effects models do not perform well on this dataset. We cannot exhaustively search the predictor space. Instead, we will hardcode backwards stewpise regression to get an idea of the appropriate form of the model. We begin with the random intercepts model containing all possible fixed effects. As the output below shows, this leads us to quickly throw out the rubric, repeated, and semester variables. In the last comparison, removing rater produces a slightly lower BIC but not enough to convince us it's absence is meaningfully better so we leave it in.

```
ratings_tall <- ratings_tall %>% mutate(rater = factor(rater))
full_re_model <- lmer(rating ~ sex + semester + rater + repeated + rubric + (1 | rubric),
                      data = ratings_tall)

AIC(full_re_model,
    lmer(rating ~ sex + semester + rater + rubric + (1 | rubric),
                      data = ratings_tall),
    lmer(rating ~ sex + semester + repeated + rubric + (1 | rubric),
                      data = ratings_tall),
    lmer(rating ~ sex + rater + repeated + rubric + (1 | rubric),
                      data = ratings_tall),
    lmer(rating ~ semester + rater + repeated + rubric + (1 | rubric),
                      data = ratings_tall),
    lmer(rating ~ sex + semester + rater + repeated + (1 | rubric),
                      data = ratings_tall),
```

```
    k = log(2*nrow(ratings_tall)))
```

```
##                                                                                 df
## full_re_model                                                                   14
## lmer(rating ~ sex + semester + rater + rubric + (1 | rubric), data = ratings_tall)      13
## lmer(rating ~ sex + semester + repeated + rubric + (1 | rubric), data = ratings_tall)   12
## lmer(rating ~ sex + rater + repeated + rubric + (1 | rubric), data = ratings_tall)      13
## lmer(rating ~ semester + rater + repeated + rubric + (1 | rubric), data = ratings_tall) 13
## lmer(rating ~ sex + semester + rater + repeated + (1 | rubric), data = ratings_tall)     8
##                                                                                    AIC
## full_re_model                                                                  1745.040
## lmer(rating ~ sex + semester + rater + rubric + (1 | rubric), data = ratings_tall)      1735.625
## lmer(rating ~ sex + semester + repeated + rubric + (1 | rubric), data = ratings_tall)   1746.615
## lmer(rating ~ sex + rater + repeated + rubric + (1 | rubric), data = ratings_tall)      1740.391
## lmer(rating ~ semester + rater + repeated + rubric + (1 | rubric), data = ratings_tall) 1751.499
## lmer(rating ~ sex + semester + rater + repeated + (1 | rubric), data = ratings_tall)    1703.363
```

```
AIC(lmer(rating ~ sex + semester + rater + repeated + (1 | rubric),
                    data = ratings_tall),
    lmer(rating ~ sex + semester + rater + (1 | rubric),
                    data = ratings_tall),
    lmer(rating ~ sex + semester + repeated + (1 | rubric),
                    data = ratings_tall),
    lmer(rating ~ sex + rater + repeated + (1 | rubric),
                    data = ratings_tall),
    lmer(rating ~ semester + rater + repeated + (1 | rubric),
                    data = ratings_tall),
    k = log(2*nrow(ratings_tall)))
```

```
##                                                                                df
## lmer(rating ~ sex + semester + rater + repeated + (1 | rubric), data = ratings_tall)  8
## lmer(rating ~ sex + semester + rater + (1 | rubric), data = ratings_tall)             7
## lmer(rating ~ sex + semester + repeated + (1 | rubric), data = ratings_tall)          6
## lmer(rating ~ sex + rater + repeated + (1 | rubric), data = ratings_tall)             7
## lmer(rating ~ semester + rater + repeated + (1 | rubric), data = ratings_tall)        7
##                                                                                   AIC
## lmer(rating ~ sex + semester + rater + repeated + (1 | rubric), data = ratings_tall) 1703.363
## lmer(rating ~ sex + semester + rater + (1 | rubric), data = ratings_tall)            1693.948
## lmer(rating ~ sex + semester + repeated + (1 | rubric), data = ratings_tall)         1704.950
## lmer(rating ~ sex + rater + repeated + (1 | rubric), data = ratings_tall)            1698.702
## lmer(rating ~ semester + rater + repeated + (1 | rubric), data = ratings_tall)       1709.719
```

```
AIC(lmer(rating ~ sex + semester + rater + (1 | rubric),
                    data = ratings_tall),
    lmer(rating ~ sex + semester + (1 | rubric),
                    data = ratings_tall),
    lmer(rating ~ sex + rater + (1 | rubric),
                    data = ratings_tall),
    lmer(rating ~ semester + rater + (1 | rubric),
                    data = ratings_tall),
    k = log(2*nrow(ratings_tall)))
```

```
##                                                                       df
## lmer(rating ~ sex + semester + rater + (1 | rubric), data = ratings_tall)  7
## lmer(rating ~ sex + semester + (1 | rubric), data = ratings_tall)          5
```

```
## lmer(rating ~ sex + rater + (1 | rubric), data = ratings_tall)             6
## lmer(rating ~ semester + rater + (1 | rubric), data = ratings_tall)        6
##                                                                            AIC
## lmer(rating ~ sex + semester + rater + (1 | rubric), data = ratings_tall) 1693.948
## lmer(rating ~ sex + semester + (1 | rubric), data = ratings_tall)         1695.615
## lmer(rating ~ sex + rater + (1 | rubric), data = ratings_tall)            1688.611
## lmer(rating ~ semester + rater + (1 | rubric), data = ratings_tall)       1701.003
```

```
AIC(lmer(rating ~ sex + rater + (1 | rubric),
                    data = ratings_tall),
    lmer(rating ~ sex + (1 | rubric),
                    data = ratings_tall),
    lmer(rating ~ rater + (1 | rubric),
                    data = ratings_tall),
    k = log(2*nrow(ratings_tall)))
```

```
##                                                              df      AIC
## lmer(rating ~ sex + rater + (1 | rubric), data = ratings_tall)  6 1688.611
## lmer(rating ~ sex + (1 | rubric), data = ratings_tall)          4 1690.352
## lmer(rating ~ rater + (1 | rubric), data = ratings_tall)        5 1696.111
```

Next we considered adding different interaction terms. None of them improved the model so we stick with just `sex` and `rater` as our fixed effects.

```
AIC(lmer(rating ~ sex + rater + (1 | rubric),
                    data = ratings_tall, REML = F),
    lmer(rating ~ sex*rater + (1| rubric),
                    data = ratings_tall, REML = F),
    lmer(rating ~ sex*semester + rater + (1| rubric),
                    data = ratings_tall, REML = F),
    lmer(rating ~ sex*repeated + rater + (1| rubric),
                    data = ratings_tall, REML = F),
    k = log(2*nrow(ratings_tall)))
```

```
##                                                                                  df
## lmer(rating ~ sex + rater + (1 | rubric), data = ratings_tall, REML = F)           6
## lmer(rating ~ sex * rater + (1 | rubric), data = ratings_tall, REML = F)           8
## lmer(rating ~ sex * semester + rater + (1 | rubric), data = ratings_tall, REML = F)  8
## lmer(rating ~ sex * repeated + rater + (1 | rubric), data = ratings_tall, REML = F)  8
##                                                                                      AIC
## lmer(rating ~ sex + rater + (1 | rubric), data = ratings_tall, REML = F)           1673.287
## lmer(rating ~ sex * rater + (1 | rubric), data = ratings_tall, REML = F)           1687.440
## lmer(rating ~ sex * semester + rater + (1 | rubric), data = ratings_tall, REML = F) 1676.054
## lmer(rating ~ sex * repeated + rater + (1 | rubric), data = ratings_tall, REML = F) 1684.443
```

Using REML, we can perform a series of likelihood ratio tests to identify whether adding any additional random effects can improve the model. The output below shows that adding random slopes for `rater` provides significant improvement to the model.

```
anova(lmer(rating ~ sex + rater + (1 | rubric),
                      data = ratings_tall, REML = T),
      lmer(rating ~ sex + rater + (1 + sex | rubric),
                      data = ratings_tall, REML = T)) %>% broom::tidy()
```

```
## # A tibble: 2 x 9
##   term                  npar   AIC   BIC logLik deviance statistic    df p.value
##   <chr>                <dbl> <dbl> <dbl>  <dbl>    <dbl>     <dbl> <dbl>   <dbl>
## 1 lmer(rating ~ sex +~     6 1641. 1669.  -814.    1629.     NA       NA NA
## 2 lmer(rating ~ sex +~     8 1640. 1678.  -812.    1624.      4.71     2  0.0949
```

```
anova(lmer(rating ~ sex + rater + (1 | rubric),
                      data = ratings_tall, REML = T),
      lmer(rating ~ sex + rater + (1 + repeated | rubric),
                      data = ratings_tall, REML = T)) %>% broom::tidy()
```

```
## # A tibble: 2 x 9
##   term                  npar   AIC   BIC logLik deviance statistic    df p.value
##   <chr>                <dbl> <dbl> <dbl>  <dbl>    <dbl>     <dbl> <dbl>   <dbl>
## 1 lmer(rating ~ sex +~     6 1641. 1669.  -814.    1629.     NA       NA  NA
## 2 lmer(rating ~ sex +~     8 1643. 1680.  -813.    1627.      2.12     2   0.347
```

```
anova(lmer(rating ~ sex + rater + (1 | rubric),
                      data = ratings_tall, REML = T),
      lmer(rating ~ sex + rater + (1 + rater | rubric),
                      data = ratings_tall, REML = T)) %>% broom::tidy()
```

```
## # A tibble: 2 x 9
##   term                  npar   AIC   BIC logLik deviance statistic    df p.value
##   <chr>                <dbl> <dbl> <dbl>  <dbl>    <dbl>     <dbl> <dbl>   <dbl>
## 1 lmer(rating ~ sex +~     6 1641. 1669.  -814.    1629.      NA      NA NA
## 2 lmer(rating ~ sex +~    11 1637. 1689.  -808.    1615.      13.7     5  0.0178
```

```
anova(lmer(rating ~ sex + rater + (1 | rubric),
                      data = ratings_tall, REML = T),
      lmer(rating ~ sex + rater + (1 + semester| rubric),
                      data = ratings_tall, REML = T)) %>% broom::tidy()
```

```
## # A tibble: 2 x 9
##   term                  npar   AIC   BIC logLik deviance statistic    df p.value
##   <chr>                <dbl> <dbl> <dbl>  <dbl>    <dbl>     <dbl> <dbl>   <dbl>
## 1 lmer(rating ~ sex +~     6 1641. 1669.  -814.    1629.     NA       NA  NA
## 2 lmer(rating ~ sex +~     8 1643. 1680.  -813.    1627.      2.12     2   0.347
```

This leads us to the final model:

$$Rating_i = \alpha_{0j[i]} + \alpha_{1j[i]} Rater_i + \beta_2 I(Male) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \tag{1}$$

$$\alpha_{0j} = \beta_0 + \eta_{0j}, \eta_{0j} \overset{iid}{\sim} N(0, \tau_0^2) \tag{2}$$

$$\alpha_{1j} = \beta_1 + \eta_{1j}, \eta_{1j} \overset{iid}{\sim} N(0, \tau_1^2) \tag{3}$$

From the output below, we can see male students perform slightly better than their female classmates, but this difference might be due to random chance ($t = 0.041$). The more notable difference is that rater 3 appears to give out lower grades than the other raters after controlling for sex.

```
final_me_model <- lmer(rating ~ sex + rater + (1 + rater | rubric),
                       data = ratings_tall, REML = T)
```

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

```
summary(final_me_model)
```

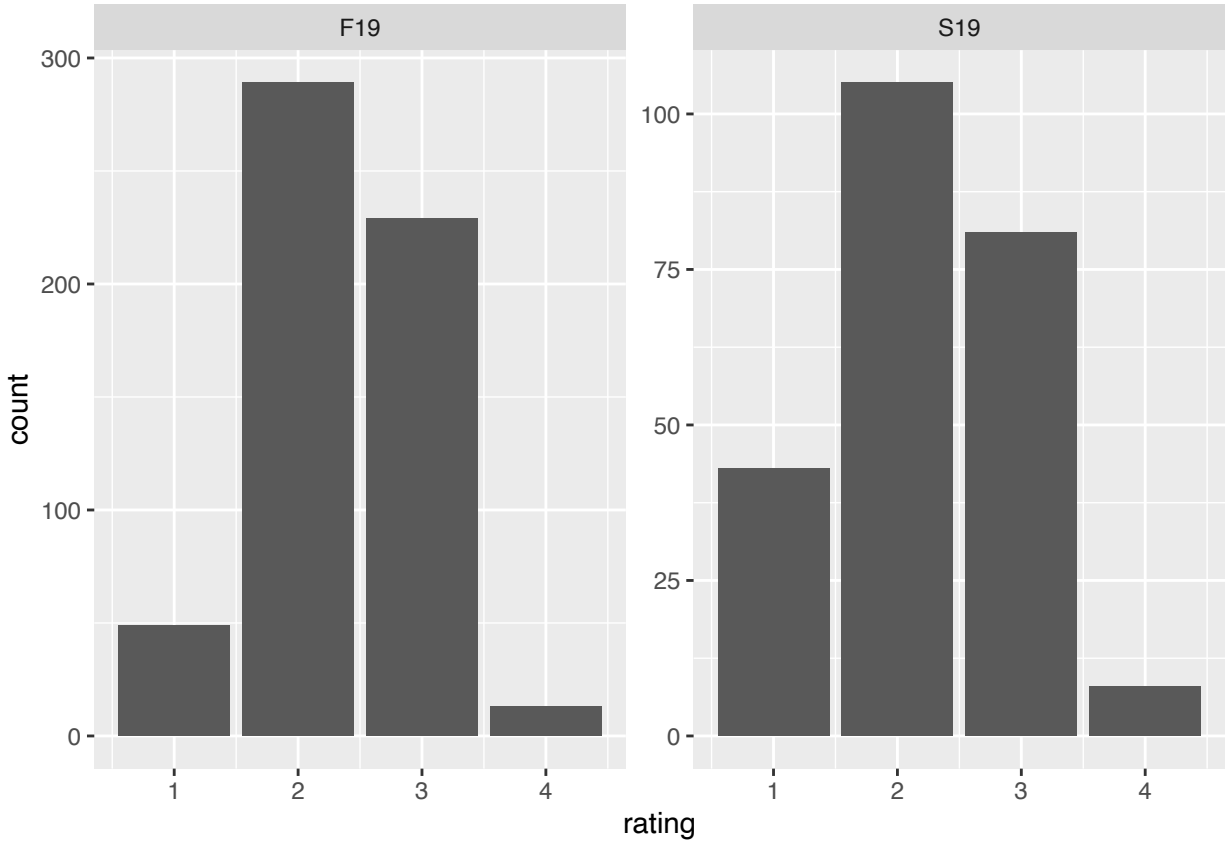
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rating ~ sex + rater + (1 + rater | rubric)
##    Data: ratings_tall
##
## REML criterion at convergence: 1630.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7103 -0.6841 -0.1120  0.7472  2.9258
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  rubric   (Intercept) 0.14785  0.3845
##           rater2      0.04366  0.2089   -0.94
##           rater3      0.04883  0.2210   -0.98  0.99
##  Residual             0.41974  0.6479
## Number of obs: 810, groups:  rubric, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.34880    0.15173  15.481
## sexM         0.00189    0.04583   0.041
## rater2       0.07928    0.09658   0.821
## rater3      -0.19556    0.10053  -1.945
##
## Correlation of Fixed Effects:
##        (Intr) sexM   rater2
## sexM   -0.124
## rater2 -0.841 -0.024
## rater3 -0.874 -0.028  0.834
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

Part D

Question: *Is there anything else interesting to say about this data?*

Overall, as we saw in part C, the semester does not seem to impact the ratings. However, there does appear to be some evidence that the semesterly ratings are different within the `SelMeth` and `VisOrg` rubrics.

```
ratings_tall %>% ggplot() +geom_bar(aes(x = rating)) + facet_wrap(vars(semester), scales = "free")
```




```
table(ratings_tall$semester, ratings_tall$rating)
```

```
##
##         1   2   3   4
##   F19  49 289 229  13
##   S19  43 105  81   8
```

```
tibble(rubric = unique(ratings_tall$rubric),
       chi_sq_p_value = map_dbl(rubric, function(x) chisq.test(
         table(ratings_tall[ratings_tall$rubric == x,]$semester,
               ratings_tall[ratings_tall$rubric == x,]$rating))$p.value),
       sig = chi_sq_p_value < 0.05)
```

```
## # A tibble: 7 x 3
##   rubric    chi_sq_p_value sig
##   <chr>              <dbl> <lgl>
## 1 RsrchQ           0.163   FALSE
## 2 CritDes          0.361   FALSE
## 3 InitEDA          0.507   FALSE
## 4 SelMeth          0.00217 TRUE
```

```
## 5 InterpRes        0.128   FALSE
## 6 VisOrg           0.00749 TRUE
## 7 TxtOrg           0.440   FALSE
```