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

Caleb Pena, cpea@andrew.cmu.edu

December 10, 2021

Abstract

This paper investigates the effectiveness and equity of Carnegie Mellon's grading system. We analyze data from a recent grading experiment conducted by CMU's Dietrich College of Humanities and Social Sciences. Using histograms and hypothesis testing, we examined how grades vary by rater, subject, and semester. We also fit a Mixed Effects model to determine whether sex or semester impact grades. Our study found no evidence of discrimination and very little evidence of grade inflation. Since these results are confined to this controlled experiment, we recommend similar data be analyzed on real students' grades.

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.

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.).

In addition to the ratings themselves, the dataset tracks the sex of the students and the semester the paper was from. 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.

Methods

We broke our analysis into four topics related to the questions posed to us. Each topic will be addressed in a separate subsection. 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 tests of independence on the counts of grades across rubrics and raters. Specifically, we conducted a chi-squared test and a Fisher's exact test.

Next, we looked more closely at the artifacts that were evaluated by multiple raters. Using this subset of the data, we 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.

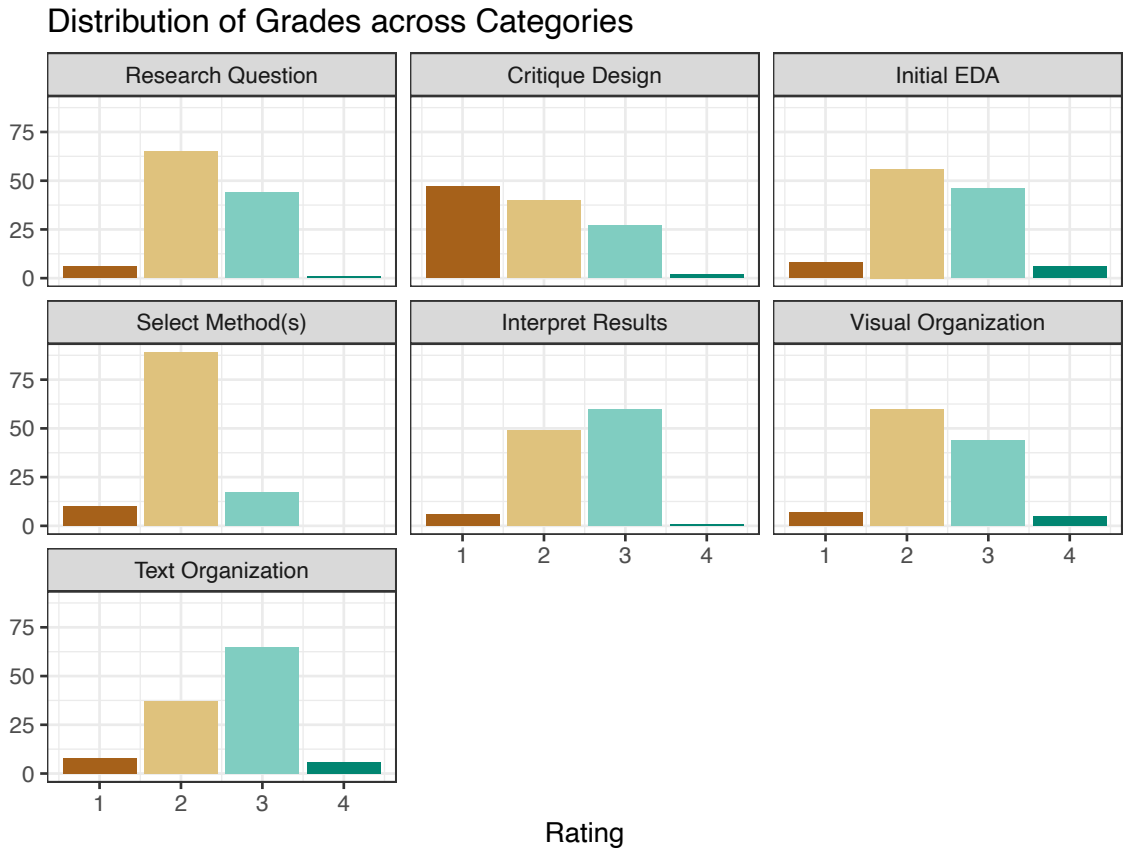
To identify how the other factors (e.g. semester, sex, etc.) were related to the ratings, we built a mixed effects model grouped by artifact. Using forwards stepwise selection, we identified the most useful fixed effects and we interpret their coefficients in the Results section. We also explored a number of candidate models using different combinations of random effects. For full details of how model selection was performed, 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

Topic 1: Distributional differences amongst rubrics and raters

The histogram in Figure 1 illustrates how the three raters evaluated the different rubric items. Most of these rubrics show a similar pattern: most students have a roughly equal chance of receiving either a 2 or a 3 with a few outstanding or unsatisfactory artifacts in the tails. This pattern appears to break down in a few categories. Raters grade the `Critique Design` section much harsher with 1s accounting for just over 40% of the total grades. They are also much more stringent in `Select Method(s)`. Although almost no papers are marked unsatistfactory, over 76% were given 2s denoting significant flaws.



(Figure 1)

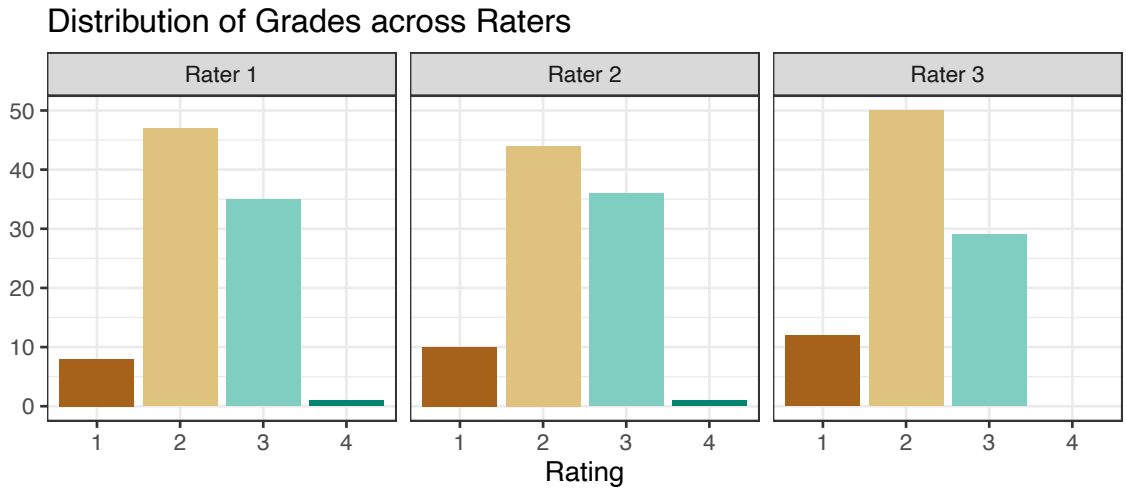
However, since the overall sample size is still relatively small ($n = 117$), an off-the-cuff look at a visualization might be misleading. To ensure these differences are not just due to random variation, we conducted two tests for independence of the counts of ratings: a χ^2-test and a Fisher exact test. The results summarized in Table 2 show that both methods support our belief that raters are treating some rubrics differently from the others.

Table 2: Hypothesis Test Results - Rubrics

Test	P-value	Result
Chi Squared Test	2.168 e-30	Reject the Null Hypothesis
Fisher Exact Test	4.998 e-4	Reject the Null Hypothesis

Figure 2 attempts to capture differences in grading at the level of the rater. As in Figure 1, we see a clear pattern. 2s are the most common grade across the board followed by 3s, 1s, and finally 4s. If any single rater is different, rater 3 is the most likely candidate. They appear more likely to have given low grades than the other raters.

Once again, we turn to hypothesis testing to evaluate whether these differences are random or reflect a true pattern. This time, both methods are much more skeptical of whether the count of each rating is dependent on the rater. As table 3 shows, neither test produced a p-value anywhere near $\alpha = 0.05$.



(Figure 2)

Table 3: Hypothesis Test Results - Raters

Test	P-value	Result
Chi Squared Test	0.8034	Fail to Reject the Null Hypothesis
Fisher Exact Test	0.8171	Fail to Reject the Null Hypothesis

Topic 2: Agreement across the three raters

In Topic 1, we looked at the overall distributions of grades by rater. But a pair of raters can have a similar distribution of grades while still disagreeing on how they assess individual artifacts. Intra-class correlation can help address this shortcoming. Table 4 breaks down these metrics by rubric. Here we see weak correlations for `Research Question`, `Interpret Results`, and `Text Organization`. On the other hand, `Critique Design`, `Initial EDA`, `Select Method(s)`, and `Visual Organization` all have high correlations. This indicates a low/high rating by one rater probably means the other raters will follow suit. Most of the variation is on the individual level, not the grader level.

Table 4: Intra-class correlations

Rubric	ICC
Research Question	0.1891892
Critique Design	0.5725594
Initial EDA	0.4929577
Select Method(s)	0.5212766
Interpret Results	0.2295720
Visual Organization	0.5924529
Text Organization	0.1428571

A more precise way of identifying agreement is to simply find the percentage of rubric items a pair of raters graded the exact same way. For example, Table 5 shows that despite the low correlation in `Research Question` raters 1 and 3 agreed nearly three quarters of the time. Disagreements between raters 1 and 2 seems to be driving the correlation downwards. The complete table may be found in part B of the technical appendix.

Table 5: Percent Exact Agreement

Rubric	Pair	Agreement
Research Question	Raters 1 and 2	38.5%
Research Question	Raters 1 and 3	76.9%
Research Question	Raters 2 and 3	53.8%

Topic 3: Relationship of other factors to the ratings

Despite considering a large number of possible combinations of fixed and random effects (including interaction effects), the only variables that significantly improved the performance of our mixed effects model were `Rubric` and `Rater`. The sex of the student, the semester under consideration, and whether or not multiple graders were evaluating the artifact do not appear to be useful predictors.

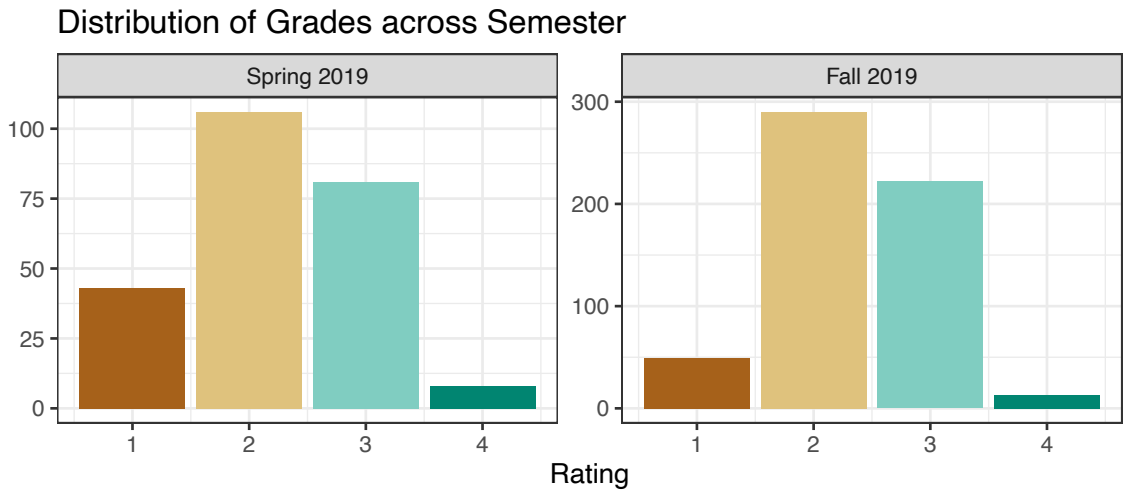
Our final model includes fixed effects for each level of `Rubric`, a random intercept, and a random slope for the `Rater` variable. Table 6 summarizes the fixed effects. As we expected from the first and second topics, the `Critique Design` and `Select Method(s)` categories have the lowest coefficients, with an expected grade of 1.9 and 2.1 respectively. By contrast, the model anticipates most students will perform well on `Text Organization`.

Table 6: Fixed Effects

Term	Estimate	Standard Error	T-value
I(Research Question)	2.386172	0.0581828	41.01163
I(Critique Design)	1.935161	0.0877906	22.04291
I(Initial EDA)	2.508855	0.0718694	34.90851
I(Select Method(s))	2.122391	0.0464798	45.66267
I(Interpret Results)	2.548309	0.0563429	45.22854
I(Visual Organization)	2.493839	0.0660854	37.73660
I(Text Organization)	2.652282	0.0648843	40.87710

Topic 4: Grading patterns across semesters

Although we found evidence in Topic 3 that `semester` does not impact rating, the histograms in Figure 3 do show a decrease in the overall number of 1s issued by the graders. Table 7 shows the results of our multiple testing. The only rubric that shows a dependent relationship between `rating` and `semester` is `Select Method(s)`. Further inspection showed why this might be the case. In the Spring, no student earned more than a 2 in this category. But in the Fall, 17 students earned 3s.



(Figure 3)

Table 7: Multiple Testing Results - Semester

Rubric	P-value	Bonferonni P	Significant
Research Question	0.1449836	1.0148849	FALSE
Critique Design	0.4707299	3.2951095	FALSE
Initial EDA	0.5369035	3.7583242	FALSE
Select Method(s)	0.0027488	0.0192417	TRUE
Interpret Results	0.1255156	0.8786095	FALSE
Visual Organization	0.0075332	0.0527325	FALSE
Text Organization	0.4636152	3.2453063	FALSE

Discussion

Now that we have discussed the results of our analysis, we are prepared to answer the questions set out at the beginning. First, we learned that ratings are heavily dependent on which rubric is being looked at. Although Carnegie Mellon undergraduates tend to organize and interpret their results well, they appear to struggle with designing their analysis and choosing appropriate methods. Furthermore, this conclusion is not driven by any one particular grader. High ICCs in these two categories suggest this problem is not due to a single rater being especially harsh. This could be an indication that instructors should spend more time giving real world applications and explaining what tools are most useful in those situations.

On a more positive note, the training given to raters to ensure an even grading scale seems to be effective. Despite the fact that the three raters came from different departments (Junker 2021), they all had very similar grade distributions. This should give CMU students confidence that they are being fairly assessed.

The second question asked whether the raters tended to agree with one another on the rubric level. For most rubrics they did, but three showed much more variation: `Research Question`, `Text Organization`, and `Interpret Results`. The first two are not surprising. These tend to be far more subjective questions

without definitive right or wrong answers. That the raters tend to disagree on interpreting results is more concerning and warrants further investigation beyond the scope of this paper.

In the third question, we were asked to identify whether any other variables impacted rating. Three variables in particular are of concern to us. `Sex` is important because Dietrich College works hard to avoid discrimination. To this end, assignments are often assessed blind. That is, the rater does not see the name of the student. Nevertheless, there might still be subtle indicators of the student's gender present in the paper (e.g., differences in word choice or tone). Thus unconscious discrimination might still be possible even if the grader doesn't sneak a peak at the name. Second, `Semester` is important because it is in the college's best interests to avoid grade inflation. If grades trend endlessly upward, they become less meaningful. However, since this experiment only covers two semesters, we must be careful to avoid overstating any conclusions drawn from this variable. Finally, the `Repeated` variable tracks whether one or multiple graders reviewed the artifact of interest. This shouldn't have any effect on the ratings. If it does, we need to reevaluate the mechanism used to assign raters as this is evidence that we have biased our results. Thankfully, despite a thorough search of the feature space, none of these variables were informative after controlling for the rubric and the rater.

Finally, using a multiple testing approach, we looked and (for most rubrics) were unable to find evidence of substantial changes in grading between the two semesters covered. The `Select Method(s)` rubric is somewhat alarming but the changes might not be a bad thing. Overly strict grading can be just as damaging as overly lenient grading.

Although these results are largely encouraging, two limitations should be noted. First, our data only looks at data from a single undergraduate course. Mathematics classes in general have very different grade distributions from, for example, a literature class. We should be careful before we generalize our results to other disciplines. Secondly, grades are not typically assigned in the context of an experiment. Typically, professors and TAs are not aware of any additional scrutiny and are facing grading deadlines and other pressures. Thus, our results might rely on data that does not reflect the "chaos" of the semester.

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 Dec 10, 2021 from https://canvas.cmu.edu/courses/25337/files/folder/Project02

Contents

Technical Appendix

Part A - Distributional differences amongst rubrics and raters

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.

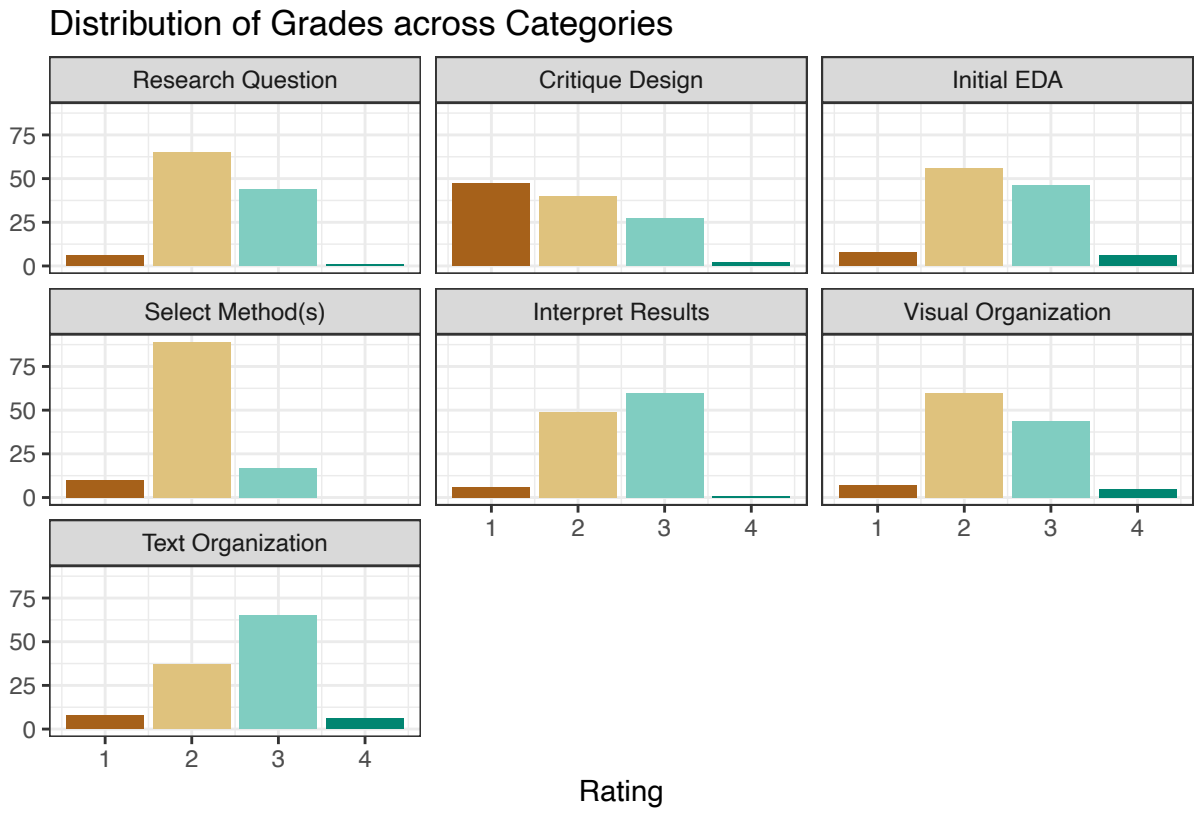
```
clean_ratings %>%
  group_by(rubric, rating) %>%
  summarise(count = n(),
            percent = count/117)
```

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

```
## # A tibble: 27 x 4
## # Groups:   rubric [7]
##    rubric            rating count percent
##    <fct>              <dbl> <int>   <dbl>
##  1 Research Question      1     6 0.0513
##  2 Research Question      2    65 0.556
##  3 Research Question      3    44 0.376
##  4 Research Question      4     1 0.00855
##  5 Critique Design        1    47 0.402
##  6 Critique Design        2    40 0.342
##  7 Critique Design        3    27 0.231
##  8 Critique Design        4     2 0.0171
##  9 Initial EDA            1     8 0.0684
## 10 Initial EDA            2    56 0.479
## # ... with 17 more rows
```

```
clean_ratings %>%
  ggplot(aes(x = rating, fill = factor(rating))) +
  geom_bar(show.legend = F) +
  theme_bw() +
  labs(x = "Rating", y = "",
       title = "Distribution of Grades across Categories", caption = "(Figure 1)") +
```

```
  facet_wrap(vars(rubric)) +
  scale_fill_brewer(palette = "BrBG")
```



(Figure 1)

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 and a fisher exact test to evaluate the spread of the counts. The tests do provide small p-values but this comes with a 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(clean_ratings$rubric, clean_ratings$rating))
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  table(clean_ratings$rubric, clean_ratings$rating)
## X-squared = 188.29, df = 18, p-value < 2.2e-16
```

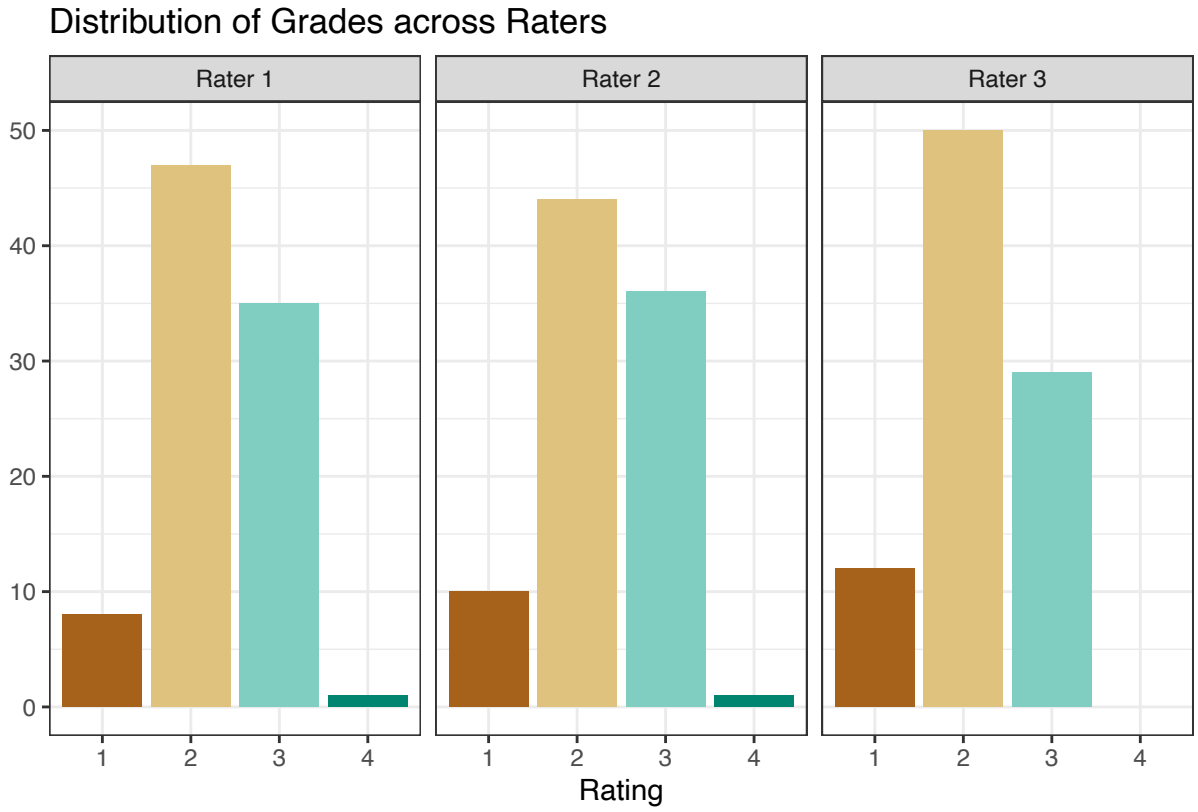
```
fisher.test(table(clean_ratings$rubric, clean_ratings$rating),
                            simulate.p.value = T)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
```

```
## data:  table(clean_ratings$rubric, clean_ratings$rating)
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

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 <- clean_ratings %>% filter(repeated == 1) %>% mutate(rating = factor(rating))
ratings_repeated %>%
  mutate(rater = paste("Rater", rater)) %>%
  ggplot(aes(x = rating, fill = rating)) +
  geom_bar(show.legend = F) +
  labs(x = "Rating", y = "",
       title = "Distribution of Grades across Raters", caption = "(Figure 2)") +
  theme_bw() +
  facet_wrap(vars(rater)) +
  scale_fill_brewer(palette = "BrBG")
```



(Figure 2)

```
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))
```

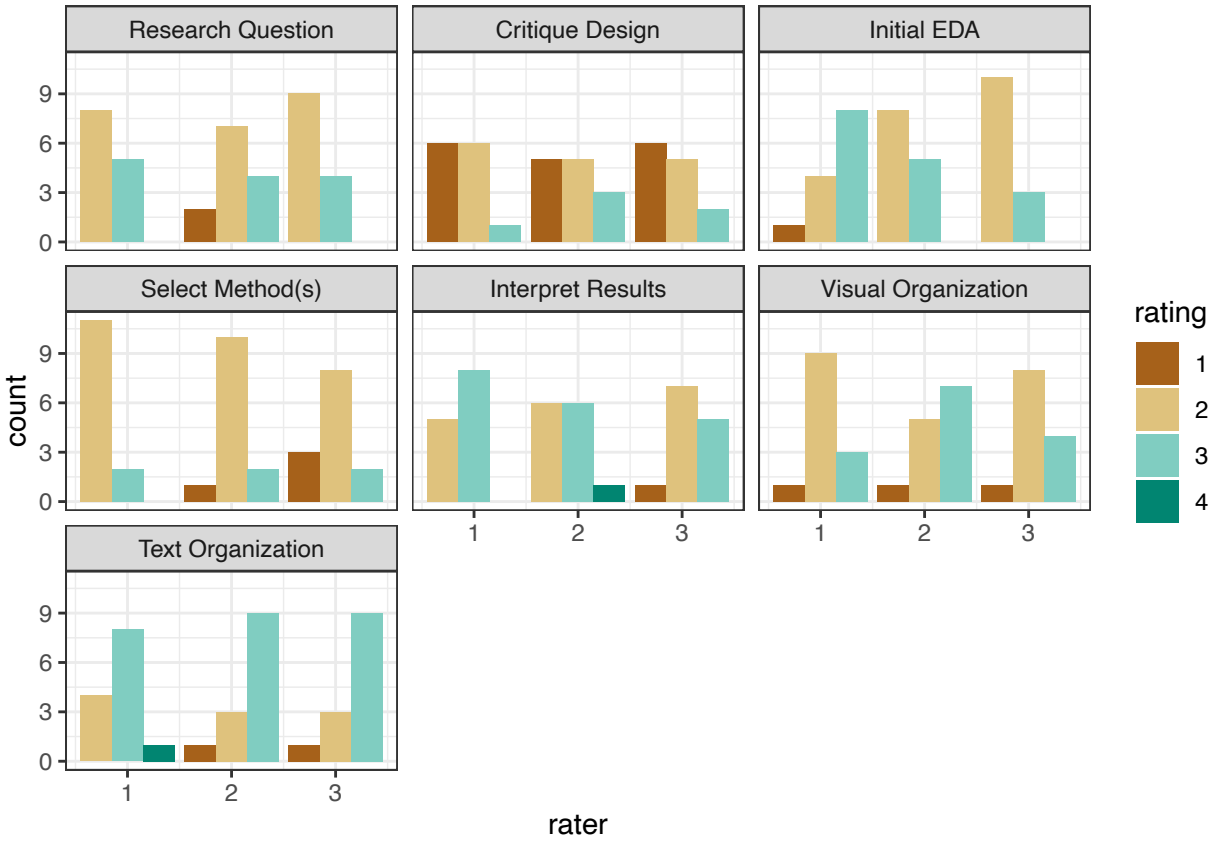
```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(ratings_repeated$rater, ratings_repeated$rating)
## p-value = 0.8069
## alternative hypothesis: two.sided
```

Part B - Agreement across the three raters

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?*

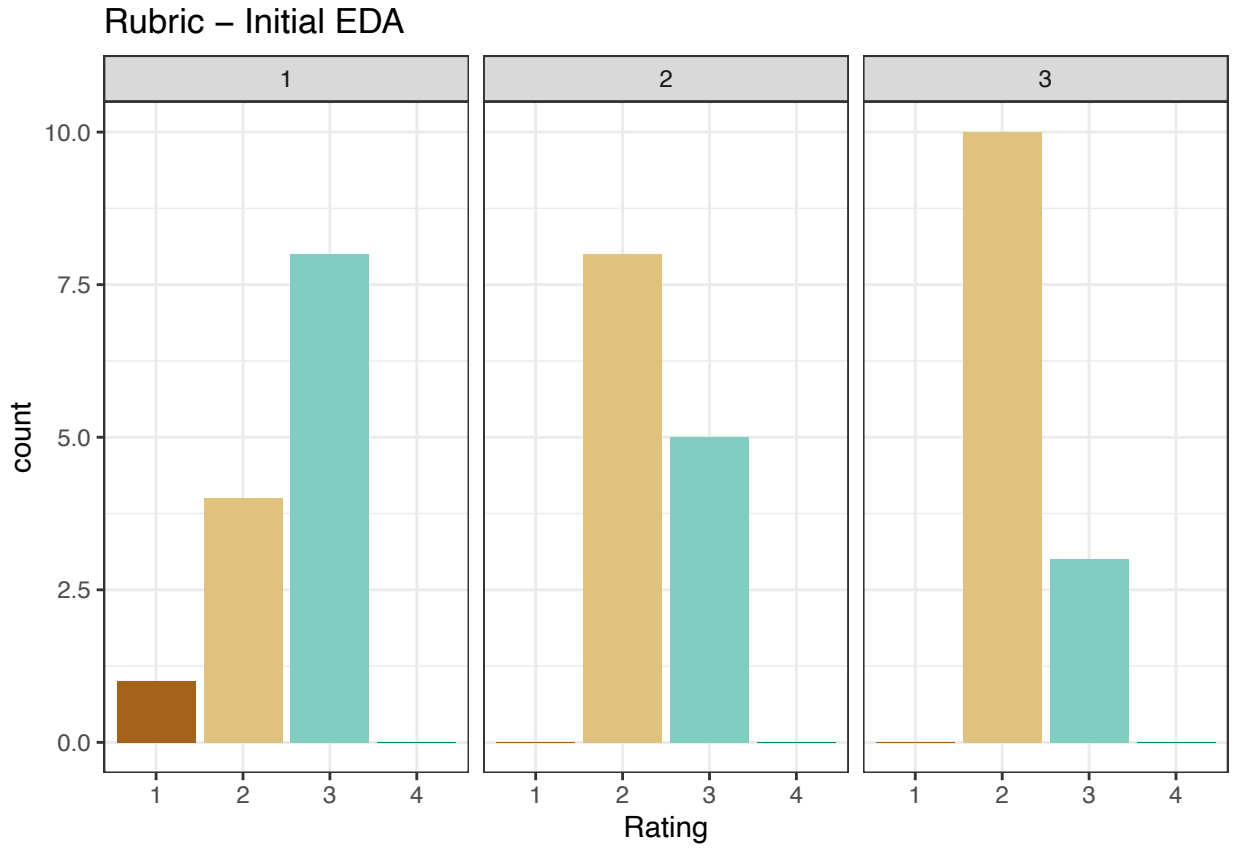
We are interested in answering the same question as before only subsetted by rubric. The graph below gives some idea of the differences in spread. The strongest differences emerge in the `InitEDA` and `VisOrg` categories. However, this is not a foolproof method to evaluate whether the raters tended to 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 = rating),
           position = position_dodge(preserve = "single")) +
  theme_bw() +
  facet_wrap(vars(rubric)) +
  scale_fill_brewer(palette = "BrBG")
```

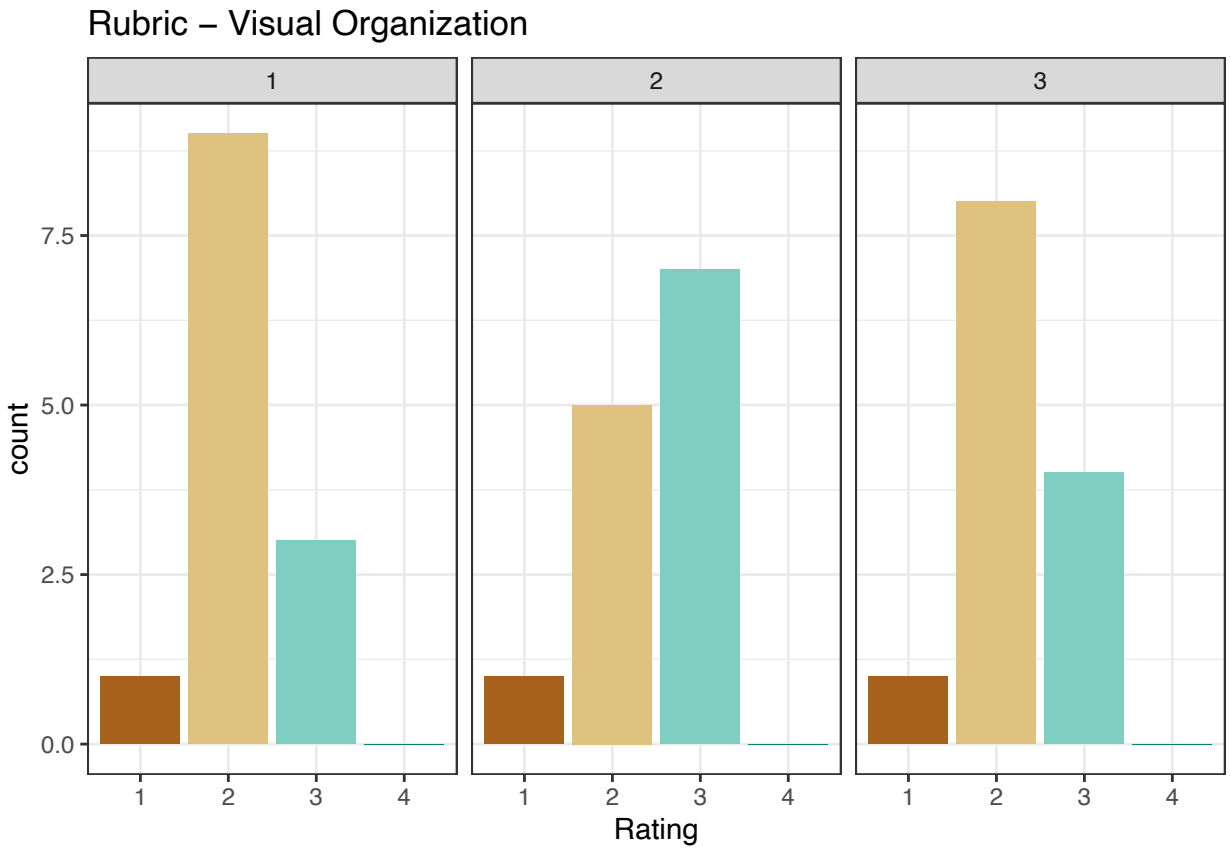



```
ratings_repeated %>%
  filter(rubric == "Initial EDA") %>%
  count(rater, rating) %>%
  complete(rater, rating, fill = list(n = 0)) %>%
  ggplot() +
  geom_bar(aes(x = rating, y = n, fill = rating),
           stat = "identity", show.legend = F) +
  theme_bw() +
```

```
  labs(x = "Rating",y = "count", title = "Rubric - Initial EDA") +
  facet_wrap(vars(rater)) +
  scale_fill_brewer(palette = "BrBG")
```




```
ratings_repeated %>%
  filter(rubric == "Visual Organization") %>%
  count(rater, rating) %>%
  complete(rater, rating, fill = list(n = 0)) %>%
  ggplot() +
  geom_bar(aes(x = rating, y = n, fill = rating),
           stat = "identity", show.legend = F) +
  theme_bw() +
  labs(x = "Rating",y = "count", title = "Rubric - Visual Organization") +
  facet_wrap(vars(rater)) +
  scale_fill_brewer(palette = "BrBG")
```



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
Research Question	0.1891892
Critique Design	0.5725594

Rubric	ICC
Initial EDA	0.4929577
Select Method(s)	0.5212766
Interpret Results	0.2295720
Visual Organization	0.5924529
Text Organization	0.1428571

The source of these agreements/disagreements can be pinned down better in the table 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_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) %>%
  mutate(pct_agreement = paste0(round(pct_agreement*100, digits = 1), "%")) %>%
  rename(Rubric = rubric, Pair = pair, `Agreement` = pct_agreement) %>%
  knitr::kable(caption = "Percent Exact Agreement")
```

Table 2: Percent Exact Agreement

Rubric	Pair	Agreement
Research Question	Raters 1 and 2	38.5%
Research Question	Raters 1 and 3	76.9%
Research Question	Raters 2 and 3	53.8%
Critique Design	Raters 1 and 2	53.8%
Critique Design	Raters 1 and 3	61.5%
Critique Design	Raters 2 and 3	69.2%
Initial EDA	Raters 1 and 2	69.2%
Initial EDA	Raters 1 and 3	53.8%
Initial EDA	Raters 2 and 3	84.6%
Select Method(s)	Raters 1 and 2	92.3%
Select Method(s)	Raters 1 and 3	61.5%
Select Method(s)	Raters 2 and 3	69.2%
Interpret Results	Raters 1 and 2	61.5%
Interpret Results	Raters 1 and 3	53.8%
Interpret Results	Raters 2 and 3	61.5%
Visual Organization	Raters 1 and 2	53.8%

Rubric	Pair	Agreement
Visual Organization	Raters 1 and 3	76.9%
Visual Organization	Raters 2 and 3	76.9%
Text Organization	Raters 1 and 2	69.2%
Text Organization	Raters 1 and 3	61.5%
Text Organization	Raters 2 and 3	53.8%

Part C - Relationship of other factors to the ratings

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?*

To identify how the other factors (e.g. semester, sex, etc.) were related to the ratings, we decided to build a mixed effects model. Two possible grouping variables were considered: `rubric` and `artifact`. To decide between these two grouping variables, we considered the variance at each level of the models. τ^2 represents the variance at the group level and σ^2 represents the remaining unexplained variance. By calculating $\frac{\tau^2}{\tau^2+\sigma^2}$, we can estimate the proportion of variance occurring at the group level. From the output below, we see that grouping by artifact allows us to explain more of the variance in ratings.

```
init_re_rubric <- lmer(rating ~ (1 | rubric), data = clean_ratings)
init_re_artifact <- lmer(rating ~ (1 | artifact), data = clean_ratings)

vars_rubric <- as.data.frame(VarCorr(init_re_rubric))$vcov
vars_artifact <- as.data.frame(VarCorr(init_re_artifact))$vcov

vars_rubric[1]/(vars_rubric[1] + vars_rubric[2])
```

```
## [1] 0.1247899
```

```
vars_artifact[1]/(vars_artifact[1] + vars_artifact[2])
```

```
## [1] 0.2563061
```

From here, we used forward stepwise selection using BIC to identify whether to add any fixed effects. Since we are comparing models with different fixed effects, we first refit the model using maximum likelihood instead of REML. Then we tried adding in `rubric`, `semester`, `sex`, `repeated`, and `rater`. Of these, `rubric` performed the best with BIC declining from 1652.8 to 1562.1 On the next step of the algorithm, none of the new variables were able to improve performance.

```
init_re_artifact <- lmer(rating ~ (1 | artifact), data = clean_ratings, REML = F)
AIC(init_re_artifact,
    lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F),
    lmer(rating ~ semester + (1 | artifact), data = clean_ratings, REML = F),
    lmer(rating ~ sex + (1 | artifact), data = clean_ratings, REML = F),
    lmer(rating ~ rater + (1 | artifact), data = clean_ratings, REML = F),
    lmer(rating ~ repeated + (1 | artifact), data = clean_ratings, REML = F),
    k = log(2*nrow(clean_ratings))) %>% rename(BIC = AIC)
```

```
##                                                                          df
## init_re_artifact                                                          3
## lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F)    9
## lmer(rating ~ semester + (1 | artifact), data = clean_ratings, REML = F)  4
```

```
## lmer(rating ~ sex + (1 | artifact), data = clean_ratings, REML = F)       4
## lmer(rating ~ rater + (1 | artifact), data = clean_ratings, REML = F)     4
## lmer(rating ~ repeated + (1 | artifact), data = clean_ratings, REML = F)  4
##                                                                              BIC
## init_re_artifact                                                        1652.799
## lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F)   1562.145
## lmer(rating ~ semester + (1 | artifact), data = clean_ratings, REML = F) 1658.845
## lmer(rating ~ sex + (1 | artifact), data = clean_ratings, REML = F)      1660.138
## lmer(rating ~ rater + (1 | artifact), data = clean_ratings, REML = F)    1654.902
## lmer(rating ~ repeated + (1 | artifact), data = clean_ratings, REML = F) 1659.915
```

```
step_2 <- lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F)
AIC(step_2,
    lmer(rating ~ rubric + semester + (1 | artifact), data = clean_ratings,
         REML = F),
    lmer(rating ~ rubric + rater + (1 | artifact), data = clean_ratings,
         REML = F),
    lmer(rating ~ rubric + repeated + (1 | artifact), data = clean_ratings,
         REML = F),
    lmer(rating ~ rubric + sex + (1 | artifact), data = clean_ratings,
         REML = F),
    k = log(2*nrow(clean_ratings))) %>% rename(BIC = AIC)
```

```
##                                                                              df
## step_2                                                                        9
## lmer(rating ~ rubric + semester + (1 | artifact), data = clean_ratings, REML = F) 10
## lmer(rating ~ rubric + rater + (1 | artifact), data = clean_ratings, REML = F)    10
## lmer(rating ~ rubric + repeated + (1 | artifact), data = clean_ratings, REML = F) 10
## lmer(rating ~ rubric + sex + (1 | artifact), data = clean_ratings, REML = F)      10
##                                                                                  BIC
## step_2                                                                      1562.145
## lmer(rating ~ rubric + semester + (1 | artifact), data = clean_ratings, REML = F) 1568.206
## lmer(rating ~ rubric + rater + (1 | artifact), data = clean_ratings, REML = F)    1563.835
## lmer(rating ~ rubric + repeated + (1 | artifact), data = clean_ratings, REML = F) 1569.271
## lmer(rating ~ rubric + sex + (1 | artifact), data = clean_ratings, REML = F)      1569.481
```

Next we considered adding different interactions effects. We tried interacting `rubric` with `semester`, `sex`, `repeated`, and `rater`. None of these improved the BIC.

```
AIC(lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F),
    lmer(rating ~ rubric*semester + (1 | artifact), data = clean_ratings,
         REML = F),
    k = log(2*nrow(clean_ratings))) %>% rename(BIC = AIC)
```

```
##                                                                              df
## lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F)        9
## lmer(rating ~ rubric * semester + (1 | artifact), data = clean_ratings, REML = F) 16
##                                                                                  BIC
## lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F)      1562.145
## lmer(rating ~ rubric * semester + (1 | artifact), data = clean_ratings, REML = F) 1603.692
```

```
AIC(lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F),
    lmer(rating ~ rubric*sex + (1 | artifact), data = clean_ratings, REML = F),
    k = log(2*nrow(clean_ratings))) %>% rename(BIC = AIC)
```

```
##                                                                        df
## lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F)  9
## lmer(rating ~ rubric * sex + (1 | artifact), data = clean_ratings, REML = F) 16
##                                                                       BIC
## lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F) 1562.145
## lmer(rating ~ rubric * sex + (1 | artifact), data = clean_ratings, REML = F) 1603.092
```

```
AIC(lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F),
    lmer(rating ~ rubric*repeated + (1 | artifact), data = clean_ratings,
         REML = F),
    k = log(2*nrow(clean_ratings))) %>% rename(BIC = AIC)
```

```
##                                                                        df
## lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F)  9
## lmer(rating ~ rubric * repeated + (1 | artifact), data = clean_ratings, REML = F) 16
##                                                                       BIC
## lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F) 1562.145
## lmer(rating ~ rubric * repeated + (1 | artifact), data = clean_ratings, REML = F) 1606.676
```

```
AIC(lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F),
    lmer(rating ~ rubric*rater + (1 | artifact), data = clean_ratings, REML = F),
    k = log(2*nrow(clean_ratings))) %>% rename(BIC = AIC)
```

```
##                                                                      df
## lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F)  9
## lmer(rating ~ rubric * rater + (1 | artifact), data = clean_ratings, REML = F) 16
##                                                                     BIC
## lmer(rating ~ rubric + (1 | artifact), data = clean_ratings, REML = F) 1562.145
## lmer(rating ~ rubric * rater + (1 | artifact), data = clean_ratings, REML = F) 1582.074
```

```
step_2 <- lmer(rating ~ rubric + (1 | artifact), data = clean_ratings)
```

Next, we switched back to REML to consider adding in new random effects. Of the variables tried, `rater` and `rubric` improved the model the most.

```
test <- lmer(rating ~ rubric + (1 + sex | artifact), data = clean_ratings)
anova(step_2,
      test,
      refit = F)
```

```
## Data: clean_ratings
## Models:
## step_2: rating ~ rubric + (1 | artifact)
## test: rating ~ rubric + (1 + sex | artifact)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## step_2    9 1540.5 1582.8 -761.26   1522.5
## test     11 1544.4 1596.1 -761.20   1522.4 0.1188  2     0.9423
```

```
test <- lmer(rating ~ rubric + (1 + repeated | artifact), data = clean_ratings)
anova(step_2,
      test,
      refit = F)
```

```
## Data: clean_ratings
## Models:
## step_2: rating ~ rubric + (1 | artifact)
```

```
## test: rating ~ rubric + (1 + repeated | artifact)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## step_2    9 1540.5 1582.8 -761.26   1522.5
## test     11 1542.5 1594.2 -760.24   1520.5 2.0376  2      0.361
```

```
test <- lmer(rating ~ rubric + (1 + rater | artifact), data = clean_ratings)
anova(step_2,
      test,
      refit = F)
```

```
## Data: clean_ratings
## Models:
## step_2: rating ~ rubric + (1 | artifact)
## test: rating ~ rubric + (1 + rater | artifact)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## step_2    9 1540.5 1582.8 -761.26   1522.5
## test     11 1532.1 1583.8 -755.04   1510.1 12.446  2   0.001984 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
test <- lmer(rating ~ rubric + (1 + semester | artifact), data = clean_ratings)
anova(step_2,
      test,
      refit = F)
```

```
## Data: clean_ratings
## Models:
## step_2: rating ~ rubric + (1 | artifact)
## test: rating ~ rubric + (1 + semester | artifact)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## step_2    9 1540.5 1582.8 -761.26   1522.5
## test     11 1538.6 1590.3 -758.30   1516.6 5.9208  2     0.0518 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
with_rubric <- lmer(rating ~ rubric + (1 + rubric | artifact), data = clean_ratings,
           control = lmerControl(optimizer = "bobyqa",
                                                 optCtrl = list(maxfun = 2e6)))
anova(step_2,
      with_rubric,
      refit = F)
```

```
## Data: clean_ratings
## Models:
## step_2: rating ~ rubric + (1 | artifact)
## with_rubric: rating ~ rubric + (1 + rubric | artifact)
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## step_2         9 1540.5 1582.8 -761.26   1522.5
## with_rubric   36 1503.8 1672.9 -715.88   1431.8 90.768 27  8.005e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
test <- lmer(rating ~ rubric + (1 + sex + rubric| artifact), data = clean_ratings)
anova(with_rubric,
      test,
      refit = F)
```

```
## Data: clean_ratings
## Models:
## with_rubric: rating ~ rubric + (1 + rubric | artifact)
## test: rating ~ rubric + (1 + sex + rubric | artifact)
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## with_rubric   36 1503.8 1672.9 -715.88   1431.8
## test          44 1511.7 1718.5 -711.85   1423.7 8.0533  8     0.4283
```

```
test <- lmer(rating ~ rubric + (1 + repeated + rubric| artifact), data = clean_ratings)
anova(with_rubric,
      test,
      refit = F)
```

```
## Data: clean_ratings
## Models:
## with_rubric: rating ~ rubric + (1 + rubric | artifact)
## test: rating ~ rubric + (1 + repeated + rubric | artifact)
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## with_rubric   36 1503.8 1672.9 -715.88   1431.8
## test          44 1513.3 1720.1 -712.65   1425.3 6.4604  8     0.5958
```

```
test <- lmer(rating ~ rubric + (1 + rater + rubric| artifact), data = clean_ratings)
anova(with_rubric,
      test,
      refit = F)
```

```
## Data: clean_ratings
## Models:
## with_rubric: rating ~ rubric + (1 + rubric | artifact)
## test: rating ~ rubric + (1 + rater + rubric | artifact)
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## with_rubric   36 1503.8 1672.9 -715.88   1431.8
## test          44 1492.7 1699.4 -702.33   1404.7 27.104  8  0.0006782 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
test <- lmer(rating ~ rubric + (1 + semester + rubric| artifact), data = clean_ratings)
anova(with_rubric,
      test,
      refit = F)
```

```
## Data: clean_ratings
## Models:
## with_rubric: rating ~ rubric + (1 + rubric | artifact)
## test: rating ~ rubric + (1 + semester + rubric | artifact)
##             npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## with_rubric   36 1503.8 1672.9 -715.88   1431.8
## test          44 1505.9 1712.7 -708.95   1417.9 13.85  8    0.08577 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This leads us to the final model:

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

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

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

$$\alpha_{2j} = \eta_{2j}, \eta_{2j} \overset{iid}{\sim} N(0, \tau_2^2) \quad (4)$$

```
final_me_model <- lmer(rating ~ rubric + (1 + rater + rubric | artifact), data = clean_ratings)


summary(final_me_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rating ~ rubric + (1 + rater + rubric | artifact)
##    Data: clean_ratings
##
## REML criterion at convergence: 1404.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0507 -0.4772 -0.0505  0.5229  3.3503
##
## Random effects:
##  Groups   Name                       Variance Std.Dev. Corr
##  artifact (Intercept)                0.2842   0.5331
##           rater                      0.0371   0.1926   -0.68
##           rubricCritique Design      0.3794   0.6159   -0.02  0.03
##           rubricInitial EDA          0.3407   0.5837   -0.56  0.45  0.28
##           rubricSelect Method(s)     0.1879   0.4335   -0.96  0.57  0.13  0.55
##           rubricInterpret Results    0.1213   0.3482   -0.80  0.73 -0.26  0.72
##           rubricVisual Organization 0.2436   0.4936   -0.63  0.56  0.07  0.68
##           rubricText Organization    0.2302   0.4798   -0.65  0.56  0.04  0.54
##  Residual                            0.1726   0.4155
##
##
##
##
##
##
##
##   0.73
##   0.44  0.59
##   0.57  0.54  0.66
##
## Number of obs: 812, groups:  artifact, 90
##
## Fixed effects:
##                            Estimate Std. Error t value
## (Intercept)                 2.38554    0.05818  41.000
## rubricCritique Design      -0.44962    0.08628  -5.211
## rubricInitial EDA           0.12352    0.08250   1.497
## rubricSelect Method(s)     -0.26285    0.07029  -3.740
```

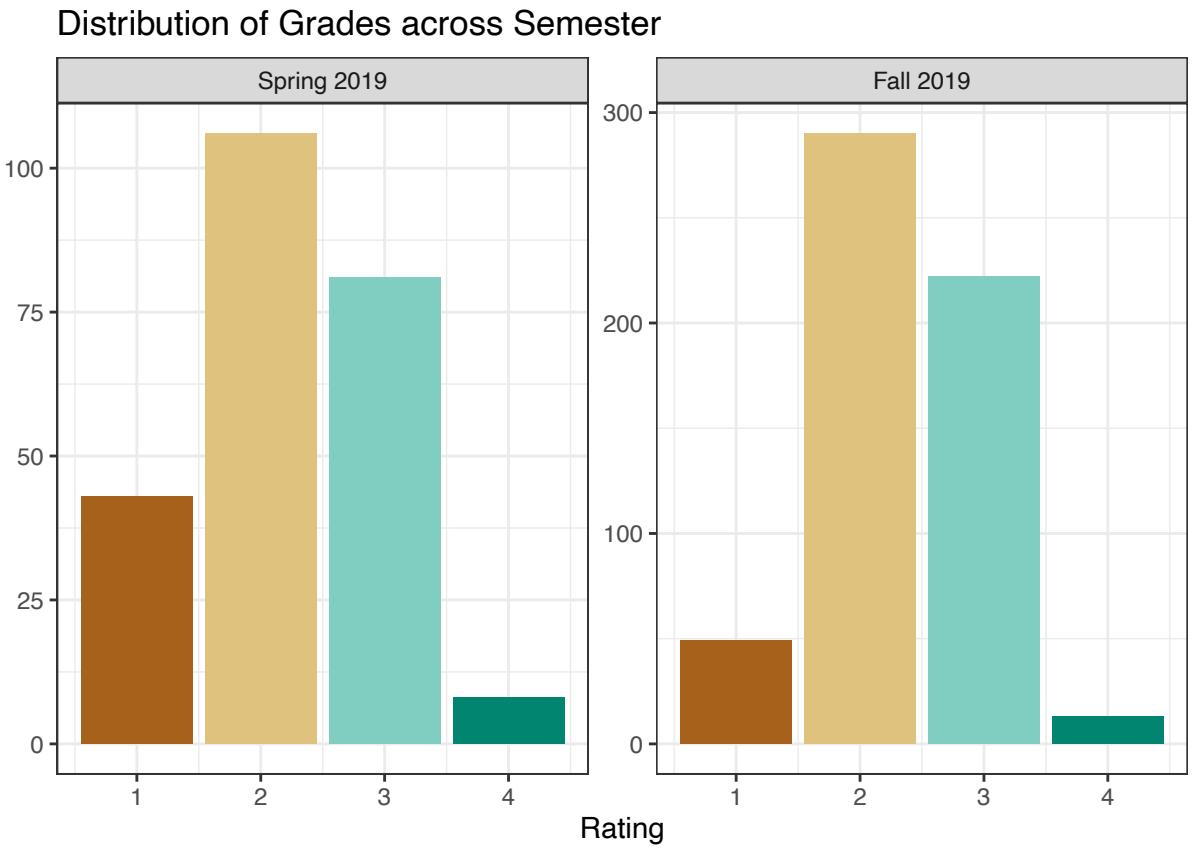
```
## rubricInterpret Results    0.16431    0.06414   2.562
## rubricVisual Organization  0.10901    0.07474   1.459
## rubricText Organization    0.26931    0.07393   3.643
##
## Correlation of Fixed Effects:
##             (Intr) rbrcCD rbIEDA rbSM() rbrcIR rbrcVO
## rbrcCrtqDsg -0.312
## rbrcIntlEDA -0.524  0.379
## rbrcSMthd() -0.753  0.322  0.496
## rbrcIntrprR -0.577  0.150  0.555  0.557
## rbrcVslOrgn -0.529  0.272  0.568  0.439  0.492
## rbrcTxtOrgn -0.541  0.259  0.493  0.505  0.473  0.549
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00605746 (tol = 0.002, component 1)
```

Part D - Grading patterns across semesters

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.

```
clean_ratings %>%
  mutate(semester = case_when(semester == "F19" ~ "Fall 2019",
                              semester == "S19" ~ "Spring 2019"),
         semester = fct_rev(factor(semester))) %>%
  ggplot(aes(x = rating, fill = factor(rating))) +
  geom_bar(show.legend = F) +
  facet_wrap(vars(semester), scales = "free") +
  labs(x = "Rating", y= "",
       title = "Distribution of Grades across Semester") +
  theme_bw() +
  scale_fill_brewer(palette = "BrBG")
```




```
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
```