Title: An evaluation of the Freshman Statistics “General Education” experimental results conducted by Dietrich College at Carnegie Mellon University

Author: Daniel Nason
Department of Statistics and Data Science, Carnegie Mellon University
dnason@andrew.cmu.edu

Abstract

In this paper, we address research questions related to the experiment conducted by Carnegie Mellon University’s Dietrich College to implement “General Education” programs for undergraduate students in Freshman Statistics to assist the Associate Dean in developing future curriculum. We analyze the data from this experiment, which includes the ratings assigned to 91 “artifacts” (project papers from the course) by raters from three different departments within the College along with the seven rubrics used to rate the paper and other categorical information (sex, semester, etc.). To evaluate these questions, we employed exploratory data analysis (EDA), calculated measures of agreement using intraclass correlation (ICC) and percent exact agreement (PAE), developed a regression model with fixed and random effects, and investigated other interesting relationships in the data. Our findings indicate that ratings vary across rubric categories and that raters utilize the differently when assessing each artifact, suggesting that the variation in ratings could be accounted for by a rater’s subjectivity rather than just the differences in the quality of the artifacts. These results should be of interest to the Associate Dean since they illustrate that the background of the raters may influence their ratings and that more training is necessary to ensure more consistency by the raters in case future experiments are conducted to evaluate the General Education program.

1 Introduction

To provide a more holistic educational experience for its students, Dietrich College at Carnegie Mellon University has begun implementing a new "General Education" (Gen Ed) program for undergraduate students. This program details educational requirements that their undergraduates must satisfy by taking various courses outside of their major. Exposure to such outside topics increases the breadth of their educational experience, allowing students to broaden their horizons by stepping out of their academic comfort zone and helping prepare them for success after graduation. One such course being evaluated is Freshman Statistics, as the College has experimented using raters from three different departments to rate a total of 91 project ("artifacts") that were randomly sampled from a Fall and Spring section of this course.

The associate dean in charge of this experiment is interested in the results and how it relates to the Gen Ed program for the College. Understanding these results can help to guide future experiments and tailor educational requirements so that they can provide a more targeted, holistic academic experience for their students. To examine these results, we address the following research questions:

1. *Distribution of Ratings by Rubric and Rater:* Is this distribution of ratings for each of the 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?
2. *Agreement Among the Raters:* 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?
3. *Relationships between Ratings and Factors in the Experiment:* 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?
4. *Other Interesting Findings in the Data:* Is there anything else interesting to say about this data?

2 Data

The data for this paper come from the Gen Ed experiment conducted by the Dietrich College at Carnegie Mellon University (Junker 2021). It includes 91 project papers (artifacts) that were randomly sampled from a Fall and Spring section of Freshman Statistics and blindly rated by three raters from different departments at the College. Of the 91 artifacts, 13 were rated by all three raters while the remaining 78 were rated only by a single rater, resulting a total of 117 observations in the data set. The data is also manipulated into a "tall" format where the rubrics and their associated ratings by rater are pivoted into two columns; as a result, the count of specific observations for a given category may exceed 117. Tables A, B, and C provide information on the definitions of the variables in the data set and what they measure:

Table A: Rubric for rating Freshman Statistics Projects

Short Name	Full Name	Description
rsrch_q	Research Question	Given a scenario, the student generates, critiques, or evaluates a relevant empirical research question.
crit_des	Critique Design	Given an empirical research question, the student critiques or evaluates to what extent a study design convincingly answer that question.

init_eda	Initial EDA	Given a data set, the student appropriately describes the data and provides initial Exploratory Data Analysis.
sel_meth	Select Method(s)	Given a data set and a research question, the student selects appropriate method(s) to analyze the data.
interp_res	Interpret Results	The student appropriately interprets the results of the selected method(s).
vis_org	Visual Organization	The student communicates in an organized, coherent, and effective fashion with visual elements (charts, graphs, tables, etc.).
txt_org	Text Organization	The student communicates in an organized, coherent, and effective fashion with text elements (words, sentences, paragraphs, section and subsection titles, etc.).

Table B: Rating scale used for all rubrics

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

Table C: Variable Names, Values, and Descriptions

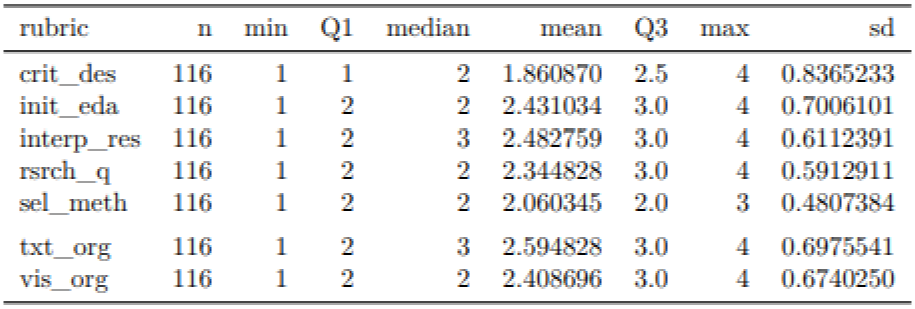
Variable Name	Values	Description
Rater	1, 2, or 3	Which of the three raters gave a rating
Semester	Fall or Spring	Which semester the artifact came from
Sex	M or F	Sex or gender of student who created the artifact
rsrch_q	1, 2, 3, or 4	Rating on Research Question
crit_des	1, 2, 3, or 4	Rating on Critique Design
init_eda	1, 2, 3, or 4	Rating on Initial EDA
sel_meth	1, 2, 3, or 4	Rating on Select Method(s)
interp_res	1, 2, 3, or 4	Rating on Interpret Results
vis_org	1, 2, 3, or 4	Rating on Visual Organization
txt_org	1, 2, 3, or 4	Rating on Text Organization
Artifact	(text labels)	Unique Identifier for each artifiact
Repeated	0 or 1	1 = this is one of the 13 artifacts seen by all 3 raters

Table D: Categorical Variable Summaries

Variable Name	Level 1: Count	Level 2: Count	Level 3: Count
Rater	Rater 1: 39 artifacts	Rater 2: 39 artifacts	Rater 3: 38 artifacts
Semester	Fall: 82 artifacts	Spring: 34 artifacts	N/A
Sex	Female: 64 artifacts	Male: 52 artifacts	N/A
Repeated	Not Repeated: 77 artifacts	Repeated: 39 artifacts	N/A

Table D summarizes the counts of the factor variables in the experiment. Of the 117 observations in the data set, 64 are from female students and 52 are from male students with one artifact having a missing value for sex; this value is dropped from the data set (Appendix, section 7.A.1.i). After dropping this observation, two of the raters reviewed 39 artifacts and one reviewed 38. 82 were randomly sampled from the Fall semester and 34 were sampled from the Spring semester. There is one missing observation in the categories Critique Design and Visual Organization.

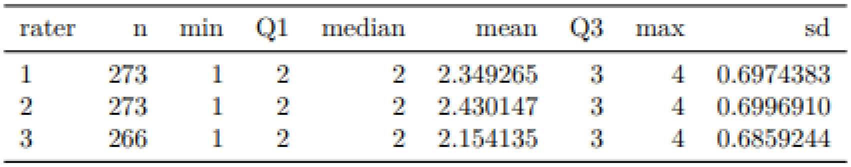
Table 1: Summary statistics of ratings by rubric



rubric	n	min	Q1	median	mean	Q3	max	sd
crit_des	116	1	1	2	1.860870	2.5	4	0.8365233
init_eda	116	1	2	2	2.431034	3.0	4	0.7006101
interp_res	116	1	2	3	2.482759	3.0	4	0.6112391
rsrch_q	116	1	2	2	2.344828	3.0	4	0.5912911
sel_meth	116	1	2	2	2.060345	2.0	3	0.4807384
txt_org	116	1	2	3	2.594828	3.0	4	0.6975541
vis_org	116	1	2	2	2.408696	3.0	4	0.6740250

Table 1 and Figure 1 (Appendix, section 7.A.1.i.a) provide summary statistics and visual summaries for ratings across the rubric categories for the full data set. Table 2 and Figure 2 also provide these metrics and visualizations for ratings by rater (Appendix, section 7.A.1.i.b).

Table 2: Summary statistics of ratings by rater



rater	n	min	Q1	median	mean	Q3	max	sd
1	273	1	2	2	2.349265	3	4	0.6974383
2	273	1	2	2	2.430147	3	4	0.6996910
3	266	1	2	2	2.154135	3	4	0.6859244

The subset of the data that includes only the artifacts reviewed by all three raters is also explored to see whether it differs from the original data set. This includes both the ratings by rubric and by rater. Table 3 and Figure 3 (Appendix, section 7.A.1.ii.a) provide summary statistics and visual summaries of the ratings by rubric, while Table 4 and Figure 4 provide this information for rater on the data subset (Appendix, section 7.A.1.ii.b).

3 Methods

To investigate the research questions of interest, we outline the approach of how each question will be addressed. We proceeded with our analysis after dropping the missing observation for sex but keeping the missing observations for the rubric categories Critique Design and Visual Organization. We used methods outlined in the Sheather (2009) textbook for exploratory data analysis and regression modeling with fixed and random effects.

Research Question 1 - Distribution of Ratings by Rubric and Rater

We built upon the tables and graphics in the Data section to investigate the distribution of ratings by rubric as well as ratings by rater. Further EDA was conducted by generating summary statistics and histograms of ratings by rubric after grouping by rater for both the original and the data where all raters reviewed the same artifacts. The results guided our approach to the other research questions and for modeling the data to investigate how these variables interact with one another and whether the effects are fixed, random, or both.

Research Question 2 - Agreement Among the Raters

To investigate the agreement between raters on their ratings of the artifacts, we examined the Intraclass Correlation (ICC) for the raters for each rubric category. ICC measures the common correlation among the raters' ratings for each artifact, so it provided a quantitative metric of agreement between the raters for both the original data and the subset where all raters reviewed the same artifacts. The Percent Exact Agreement (PAE) was also considered for each pair of raters within each rubric category. PAE provided more specific information than ICC since it allowed shows how much two raters agreed in a specific category and which rater contributes to the general disagreement among all raters. However, since PAE requires that both raters reviewed the same artifact, this metric was only calculated for the subset of the data where there was overlap by raters for a specific artifact.

Research Question 3 - Relationships between Ratings and Factors in the Experiment

We utilized multilevel modeling to examine how the various factors in the experiment are related to ratings and whether they interacted in any interesting ways. The process for determining this was as follows:

1. We explored whether it was appropriate to include fixed effects in the seven rubric-specific models by artifact for only the data where all three raters rated the same artifacts.
2. We applied this approach the seven rubric specific models for the original data to see if any interesting relationships arose based on AIC values and the Likelihood Ratio Test (LRT).
    - These metrics were used to evaluate adding new terms to the model instead of BIC since they were more likely to favor models with additional terms that could be potentially important.
3. Interactions between the fixed effects and random effects were also included in the models.
    - These were only considered in the models where AIC and LRT suggested including a fixed effect term was appropriate rather than simply the intercept-only model.
4. The results of these rubric specific models were then compared when looking at the combined data in the "tall" format using a similar approach to evaluate whether including additional terms was appropriate.
    - Using AIC and LRT as selection criteria fixed effects were first added to base model that only included a random effects interaction term between rubric category and artifact.
    - Interaction effects between the random effects were then included, along with random effects for the selected fixed effects terms.

This allowed us to determine whether there were any interesting relationships between the factor variables in the tall data set.

Research Question 4 - Other Interesting Findings in the Data

To answer this question, we conducted further EDA to investigate other interesting relationships in the data between the factor variables and ratings. These included graphics and summary statistics of the factor variables sex and semester to determine if there was an association between them and ratings.

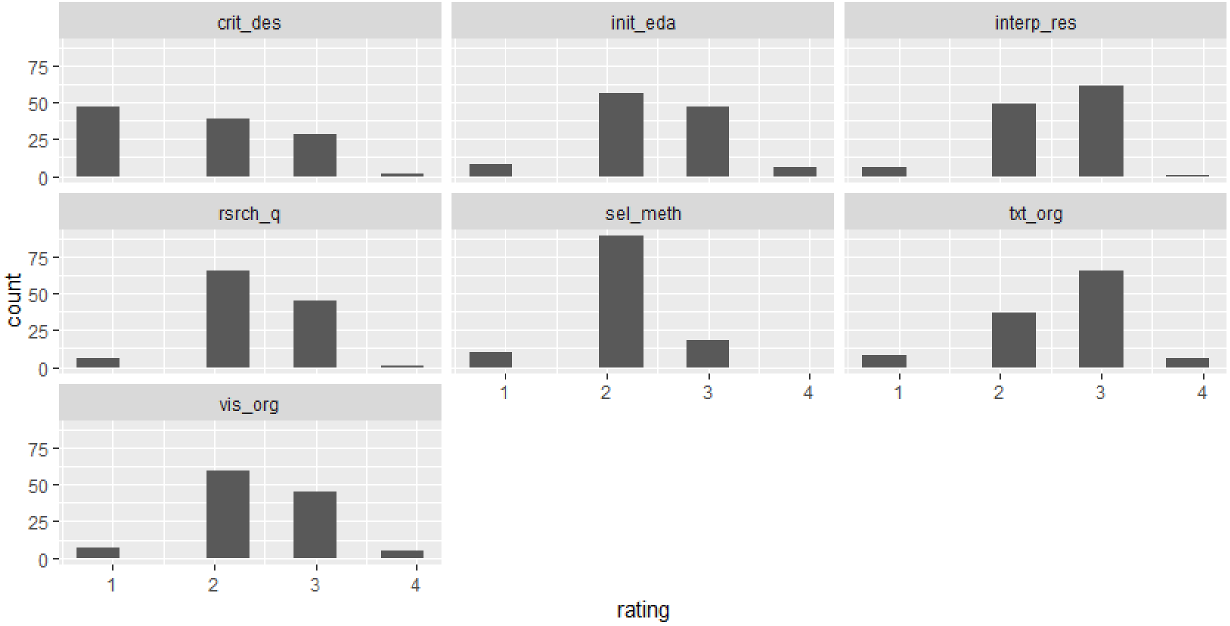
These were also investigated when grouping by rubric category to see if the results aligned with the model selected for Research Question 3.

4 Results

Because there was no noticeable difference in the distribution of ratings even after being disaggregated by the various factor variables, we drop the observation with a different value for sex. The missing observations for the rubric categories Critique Design and Organization are also kept in the data set since there was no noticeable difference in the distribution of ratings for these categories based on whether the missing values were present.

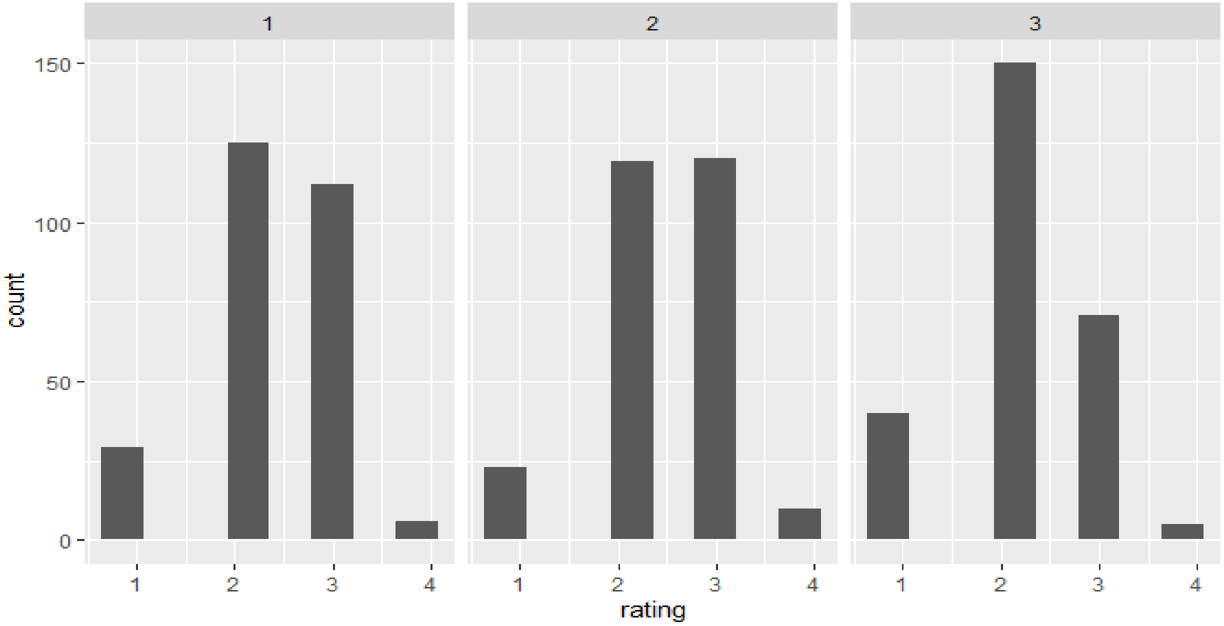
Research Question 1 - Distribution of Ratings by Rubric and Rater

Figure 1: Histograms of rating by rubric



Expanding upon the EDA in the Data section, we see from Figure 1 that the distribution by rubrics is not indistinguishable from other rubrics. Critique Design is the only rubric category whose distribution is not roughly normal, with the most artifacts receiving a rating of 1 and a decreasing number of students in the higher categories. As a result, it has the lowest overall mean. Also, Select Method is the only category where no artifact receives a rating of 4; the remaining categories have at least one artifact that received a rating of 4 and have means that are roughly centered between 2 and 3 as shown in Table 1.

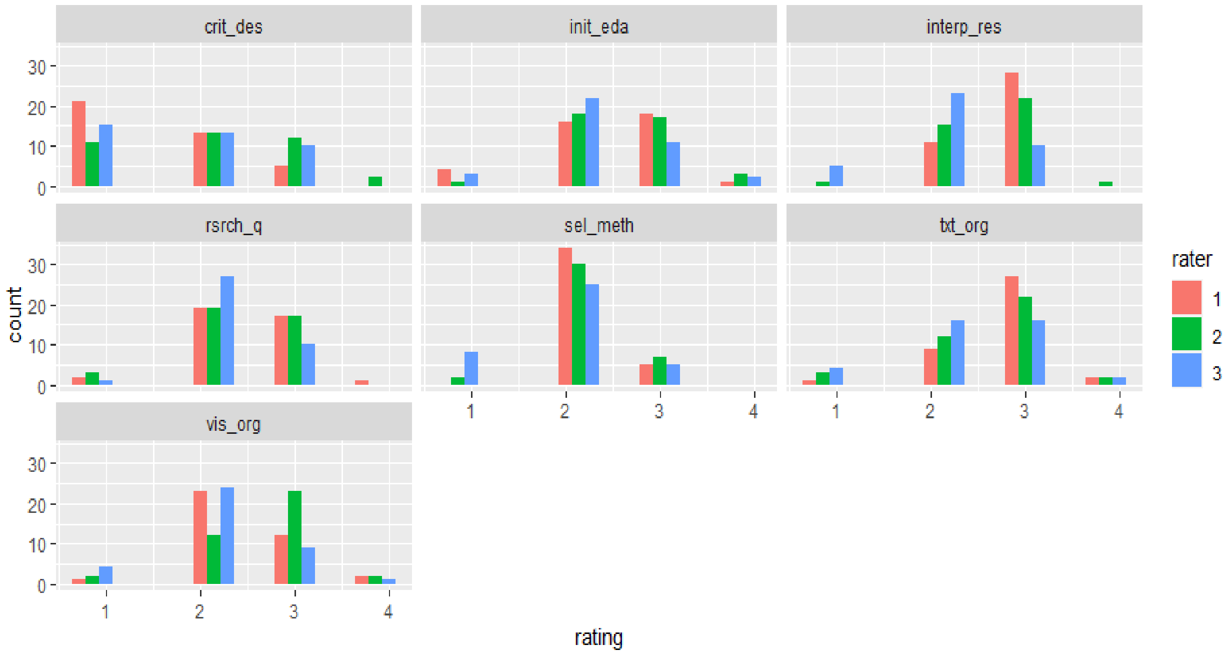
Figure 2: Histograms of rating by rater



The ratings of each rater are also examined. Table 2 and Figure 2 (Appendix, section 7.A.1.i.b) provide summary statistics and visual summaries for each of the raters. Both raters 1 and 2 have relatively similar mean rating scores and distributions for ratings, while Rater 3 has a lower average rating and assigns more artifacts with a rating of 2 than either of the other raters. The distributions of raters 1 and 2 also more closely resemble a normal distribution than rater 3 based on the histograms shown in Figure 2.

The subset of the data that includes only the artifacts reviewed by all three raters is also examined to see whether it differs from the original data set. Comparing this to the full data using Table 3 and Figure 3 (Appendix, section 7.A.1.ii.a), we see that the distributions are roughly identical to the full data; however, one notable difference is that five of the seven rubric categories have a maximum rating of 3, as opposed to only one category for the full data. The ratings by rater for this subset are also considered in Table 4 and Figure 4, and the results parallel the findings for Table 3 and Figure 3 in that they generally align with the overall data.

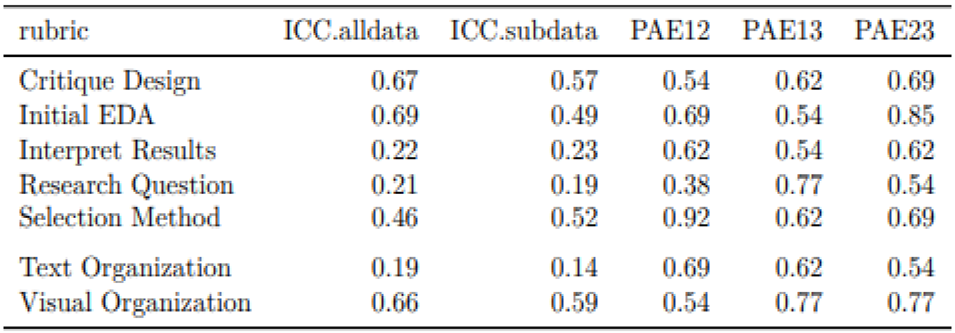
Figure 5: Histogram of rating by rubric across rater



We also assess the distribution of rating by rubric when grouped by rater to explore their agreement in the ratings by rubric. Table 5 and Figure 5 (Appendix, section 7.B.1.i) provide the summary statistics and visualizations by rater across each of the rubric categories. Figure 5 shows that except for Critique Design, rater 3 assigns the fewest ratings with a score of 3 or higher when looking at each of the rubrics. While generally rater 1 and rater 2 assign similar ratings for the artifacts they reviewed, this is not necessarily true for each of the rubric categories. Critique Design and Visual Organization have the most visible differences in the ratings between these two raters. For example, in Visual Organization rater 2 assigns more ratings of 3 and less ratings of 2 compared to rater 1 to the artifacts. Figure 6 and Table 6 (Appendix, section 7.B.1.ii) also demonstrate that these relationships are generally consistent in the subset of the data where each of the raters reviewed the same artifacts. This EDA demonstrates that there is a non-uniform relationship between ratings by both rubric and by rater in both the full data and the subset where each rater evaluated the same artifact, and these relationships are further explored in the Questions 2 and 3.

Research Question 2 - Agreement Among the Raters

Table 7: Intraclass Correlation (ICC) and Percent Exact Agreement (PAE) between raters by rubric



rubric	ICC.alldata	ICC.subdata	PAE12	PAE13	PAE23
Critique Design	0.67	0.57	0.54	0.62	0.69
Initial EDA	0.69	0.49	0.69	0.54	0.85
Interpret Results	0.22	0.23	0.62	0.54	0.62
Research Question	0.21	0.19	0.38	0.77	0.54
Selection Method	0.46	0.52	0.92	0.62	0.69
Text Organization	0.19	0.14	0.69	0.62	0.54
Visual Organization	0.66	0.59	0.54	0.77	0.77

Table 7 (Appendix, section 7.B.2) summarizes the ICC by rubric category for both the full data set and the subset of data where each rater reviewed the same artifact. Comparing the calculations across these data sets, the ICC is larger for every category except for Select Method and Interpret Results. The largest difference between the two ICCs is in the Initial EDA rubric category where the ICC total data set is 0.2 higher, while the smallest difference is in the Interpret Results category where the ICC for the total data set is 0.01 lower.

The findings in Table 7 also portray PAE between raters by rubric category for the subset of the data. PAE measures how frequently a pair of raters reached the same conclusion for a given rubric across the artifacts they both rated to identify which rubric two raters disagree on and how it contributes to the overall disagreement between raters. The highest PAE between any two raters is for raters 1 and 2 for the Select Methods category (0.92), while lowest agreement is for raters 1 and 2 for the Research Question rubric category (0.38). Of the 21 pairwise comparisons conducted across the 7 rubrics for the 3 raters, there is only one instance when two raters who reviewed the same artifact agreed on less than half of the ratings (PAE of 0.38 for raters 1 and 2 with Research Question). On the other hand, there are 5 instances when the two raters who reviewed the same artifact agreed on at least 75% of the ratings. The majority of the PAE measures (15/21) for the rubrics lie between 50% and 75%, and none of the ICCs exceed .7 for the total data and .6 for the subset of data.

Table 7 confirms the EDA conducted for Question 1 in that it identifies some of the disagreement among raters for the rubric categories, and these results are considered when investigating the relationships between ratings and the factor variables in the experiment for Question 3.

Research Question 3 - Relationships between Ratings and Factors in the Experiment

Looking only at the subset of the data where all 3 raters rated the same artifacts (Appendix, section 7.B.3.i), we find that for each rubric category, the intercept-only model is adequate. None of the likelihood ratio tests for nested fixed effects yield p-values that are statistically significant at the 5% level, so no interaction terms and random effects are not considered for this subset of the data.

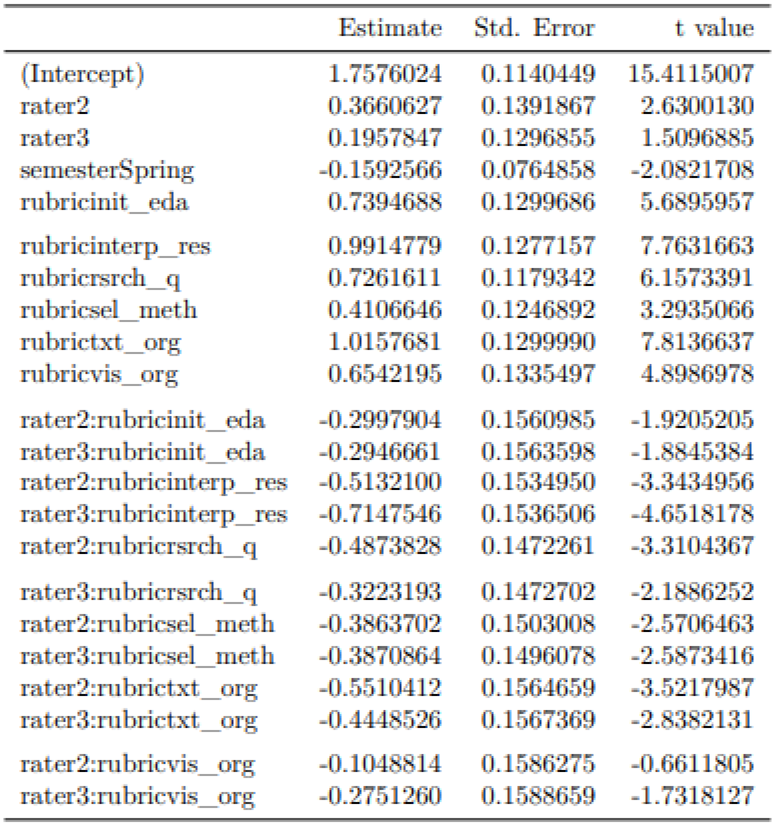
Table 15: Coefficient and Variance Estimates for Rubric-Specific Models							
Variable	crit_des	init_eda	interp_res	rsrch_q	sel_meth	txt_org	vis_org
Intercept	N/A	2.4425	N/A	2.3517	N/A	2.5875	N/A
Rater 1	1.6863	N/A	2.7042	N/A	2.1875	N/A	2.3779
Rater 2	2.1129	N/A	2.5857	N/A	2.1594	N/A	2.6489
Rater 3	1.8908	N/A	2.1392	N/A	1.9648	N/A	2.2836
Fall	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Spring	N/A	N/A	N/A	N/A	-0.3196	N/A	N/A
Female	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Male	N/A	N/A	N/A	N/A	0.1216	N/A	N/A
Repeat	N/A	N/A	N/A	N/A	N/A	N/A	N/A
No Repeat	N/A	N/A	N/A	N/A	N/A	N/A	N/A
σ^2	0.2473	0.1655	0.2525	0.2783	0.1071	0.3957	0.1467
τ^2	0.4349	0.3651	0.0622	0.0728	0.0901	0.0937	0.2907

We expand our investigation to the seven rubric-specific models for the full data. Table 15 illustrates the coefficient estimates for the rubric-specific models. It also includes estimates the variance for individual ratings by rubric (σ^2) and the variance of the predicted rating of an individual artifact compared to the mean rating for each rubric (τ^2). We see that the random effects intercept-only model is selected for the categories Initial EDA, Research Question, and Text Organization (Appendix, section 7.B.3.ii). Three other rubrics (Critique Design, Interpret Results, and Visual Organization) prefer also including rubric as a fixed effect, while the remaining category (Select Method) includes sex, semester, and their interaction (Appendix, section 7.B.3.iii). The ICCs for these models are approximately the same as the ICC calculated in Table 7, and random effects for the fixed effects included in these rubric-specific models were not included due to insufficient sample size (Appendix, sections 7.B.3.iii.a, 7.B.3.iii.b, 7.B.3.iii.c, and 7.B.3.iii.d). For the rubric categories, we see that the τ^2 values are around or greater than 0.29 for the rubric categories Critique Design, Initial EDA, and Visual Organization, suggesting there is more variability in the artifacts for these categories compared to each rubric's average. Alternatively, the τ^2 for the remaining categories is less than 0.1, implying that there is less variability in the ratings for each artifact in these categories compared to the average score for each rubric.

However, the rubric-specific models assume independence between the ratings of the rubrics. This is an unrealistic assumption since a well-written (or poorly written) artifact is likely to have ratings that are correlated across multiple categories. Therefore, to explore how interactions between rubric and the various factors in the experiment are related to ratings and relax the independence assumption between rubric categories, we include both fixed and random effects in the model that associates rating with rubric utilizing the "tall" data set. The final model includes fixed effects for rater, semester, rubric, and the interaction between rater and rubric, as well as random effects interactions between artifact and rater and artifact and rubric. Table 8 portrays the coefficient estimates (Appendix, section 7.B.3.iv) for the equation of the final model (represented in terms of R code):

$$Final\ Model: Rating \sim Rater + Rubric + Semester + Rater{:}Rubric + (0 + Rater \mid Artifact) \\ + (0 + Rubric \mid Artifact)$$

Table 8: Final Model Coefficient and Standard Error Estimates



	Estimate	Std. Error	t value
(Intercept)	1.7576024	0.1140449	15.4115007
rater2	0.3660627	0.1391867	2.6300130
rater3	0.1957847	0.1296855	1.5096885
semesterSpring	-0.1592566	0.0764858	-2.0821708
rubricinit_eda	0.7394688	0.1299686	5.6895957
rubricinterp_res	0.9914779	0.1277157	7.7631663
rubricrsrch_q	0.7261611	0.1179342	6.1573391
rubricsel_meth	0.4106646	0.1246892	3.2935066
rubrictxt_org	1.0157681	0.1299990	7.8136637
rubricvis_org	0.6542195	0.1335497	4.8986978
rater2:rubricinit_eda	-0.2997904	0.1560985	-1.9205205
rater3:rubricinit_eda	-0.2946661	0.1563598	-1.8845384
rater2:rubricinterp_res	-0.5132100	0.1534950	-3.3434956
rater3:rubricinterp_res	-0.7147546	0.1536506	-4.6518178
rater2:rubricrsrch_q	-0.4873828	0.1472261	-3.3104367
rater3:rubricrsrch_q	-0.3223193	0.1472702	-2.1886252
rater2:rubricsel_meth	-0.3863702	0.1503008	-2.5706463
rater3:rubricsel_meth	-0.3870864	0.1496078	-2.5873416
rater2:rubrictxt_org	-0.5510412	0.1564659	-3.5217987
rater3:rubrictxt_org	-0.4448526	0.1567369	-2.8382131
rater2:rubricvis_org	-0.1048814	0.1586275	-0.6611805
rater3:rubricvis_org	-0.2751260	0.1588659	-1.7318127

From Table 8 we see that the fixed effect terms included from variable selection are all included as fixed effects in at least one of the seven rubric-specific models for the full data set. For the non-interaction fixed effects terms, we see that every coefficient except for rater 3 is statistically significant at the 5% level. Additionally, the interaction terms between both rater 2 and rater 3 for the rubric categories Interpret Results, Research Question, Select Method, and Text Organization are statistically significant at the 5% level.

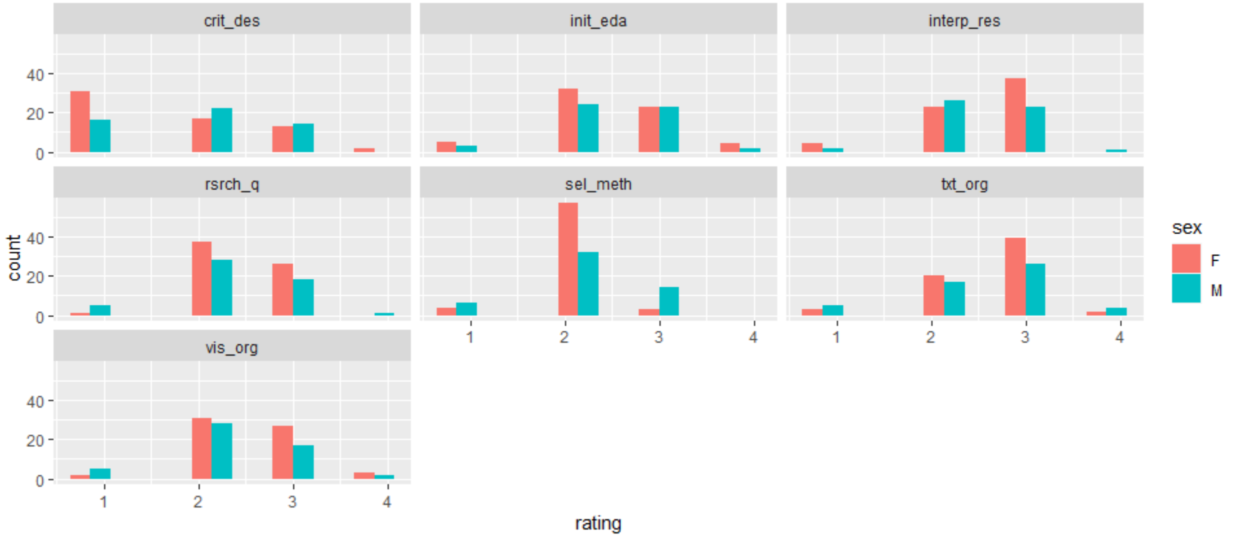
Interpreting some of the statistically significant fixed effects of the model portrayed in Table 8, the artifacts from the Spring semester have an approximately 0.16 lower score on average compared to those from the fall semester. However, the presence of interaction terms between rater and rubric implies that there is not a meaningful interpretation of the coefficients for an individual rater or rubric category. Instead, these terms are combined with the interaction terms for rater and rubric to provide predictions for average rating.

Examining these interactions, we see all the terms between rater and rubric have negative coefficient estimates. These terms account for the average difference between ratings for an artifact in a given rubric category rated by either rater 2 or 3 compared to the average rating given by rater 1. The average ratings for artifacts can be calculated by adding up the coefficients for the appropriate values of the categorical variables. For example, the rating for an artifact rated by rater 2 using the Select Methods rubric is calculated by summing the terms 1.76 + 0.37 + 0.41 – 0.39 = 2.15. Similarly, the rating for an artifact rated by rater 3 using the Interpret Results rubric is calculated by adding up the coefficient estimates 1.76 + 0.20 + 0.73 – 0.71 = 1.98. The fixed effects coefficients for average ratings for the other rubric categories can be computed in a similar way.

We also examine how the random effects terms influence the interpretation of the model. The random effects interactions are specific to each artifact across each of the rubric and rater categories, so we examine some of the more noteworthy coefficients (Appendix, section 7.B.3.iv). For example, the random effect estimate for artifact 62 in the Critique Design rubric category is 1.38; this means that the estimated rating score for artifact 62 when reviewed by rater 1 in the Fall semester is 1.76 + 1.38 = 3.14. Likewise, the random effects estimates for artifact O2 are 0.33 and -1.02 for the category Initial EDA and rater 2, respectively; utilizing the model, this means that the estimated average rating for artifact O2 in the Fall semester reviewed by rater 2 in the Initial EDA category is 1.76 + 0.37 + 0.74 – 0.30 + 0.33 – 1.03 = 1.87. The random effects coefficients of the model can be utilized similarly to generate an estimate for the rating of a specific artifact.

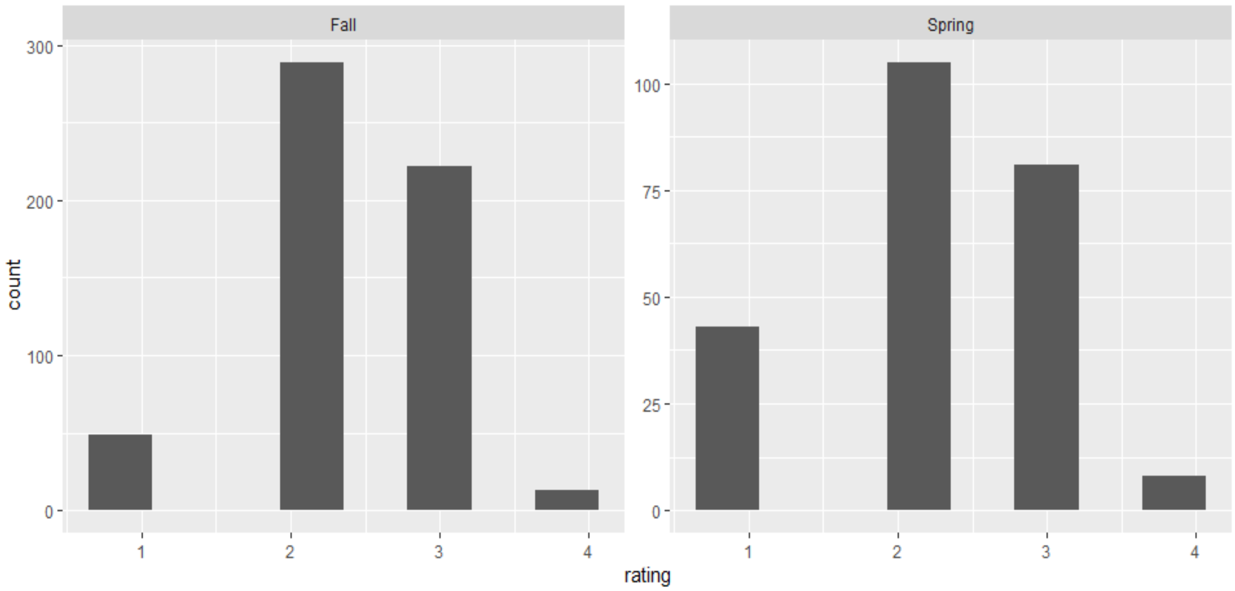
Research Question 4 - Other Interesting Findings in the Data

Figure 8: Histogram of ratings by rubric and sex



Figures 7 and 8 display how sex relates to ratings and whether these ratings differ by rubric category (Appendix, section 7.B.4). While the distribution of ratings is roughly identical by sex, there are some differences in ratings by sex when disaggregating the results across each rubric category. The most noteworthy difference is in the Select Method rubric category, where many more female students received ratings of 2 than males, and males receive more ratings of 3 than females despite having fewer observations in the data set. This is also confirmed in Table 10 (Appendix, section 7.B.4), as the average score for males in this category is 2.15 compared to 1.98 for females. Similar relationships are present in the subset of the data where all raters reviewed the same artifacts, as illustrated in Figures 9 and 10 (Appendix, section 7.B.4). The similarity in the distribution of ratings when disaggregated by sex or by sex and rubric may explain why it was not selected in the final model as either a fixed or random effect.

Figure 11: Histogram of ratings by semester



Similar EDA is conducted when looking at ratings by semester and by rubric and semester in Figures 11 and 12 (Appendix, section 7.B.4). Figure 11 compares the distribution of ratings by semester, and we see that a larger proportion of artifacts received a rating of 1 in the Spring Semester relative to the Fall semester. This is also verified by Table 13 (Appendix, section 7.B.4), as the average for the Spring is approximately 0.12 lower than the average for the Fall. However, when disaggregating by rubric and comparing across semesters, we see that the distributions across rubric categories are roughly identical apart from Select Method, where no artifact received a rating of 3 in the Spring. Note that this analysis could not also be applied to the subset of the data since all the artifacts that were reviewed by all three raters were only from the Fall semester. The difference in the distribution of ratings by semester helps to explain why semester was included as a fixed effect only in the final model.

5 Discussion

The results of this experiment yield some interesting insights for the Associate Dean about assessing the performance of students Freshman Statistics within the Gen Ed curriculum. We find evidence that the variation in ratings across artifacts is related to both the rubric category and the rater evaluating the artifact, as raters did not reach similar results when evaluating the same artifact. Our analysis suggests that this disagreement was influenced by how raters utilized each rubric category, and we think these results are driven by the difference in academic specialties of the raters within Dietrich College. The presence of these trends in the data limit the generalizability of the results since it is difficult to ascertain whether the rating on an artifact due to its quality or the subjectivity of how the rater used a rubric to evaluate that artifact. The relationships between the demographic factor variables such as sex or semester were relatively consistent with our expectations.

Our findings suggest that more steps must be taken to train raters before assigning them to assess the artifacts. More uniformity in the ratings would increase our confidence in the results of the experiment and allow us to make valid inferences about the performance of the Gen Ed program. We can also consider employing other modeling techniques to see if there are any other relationships that exist in the experiment that should be kept in mind for future experiments to evaluate these programs. Having a valid assessment of the performance of the Gen Ed programs would help the Associate Dean determine whether such a program is necessary and adequately improves the educational experience of undergraduates in the College.

Research Question 1 - Distribution of Ratings by Rubric and Rater

Critique Design and Select Methods appear to be the rubric categories that are consistently associated with the lowest rating scores for the artifacts. This is true for both the full data set and the subset where the raters reviewed the same artifacts. While the distribution of Select Method resembles the other categories in that it is approximately normally distributed, Critique Design is not normally distributed and has both the lowest average rating and the greatest number of artifacts that receive a rating of 1. These results could be driven by a couple of factors related to the course. Based on the rubric descriptions, both Critique Design and Select Methods are categories that rely more on critical thinking than on simply applying an analytical method that is typically taught in a Freshman Statistics course. First-year students, especially first semester freshman, likely have not developed these abstract reasoning and communication skills to the point where they can competently defend their arguments to an expert rater. Alternatively, the curriculum could be failing to sufficiently cover or emphasize the importance of these topics for statistical writing. Instead, the course could be focused on the rote aspects of introductory statistics such as the arithmetic and algebra for hypothesis testing instead of more abstract topics like experimental design or understanding how different variable types impact the approach and tools needed to conduct valid statistical analyses and generalize the results.

The EDA conducted also illustrates that the distribution of ratings by rater is not identical for both the full data set and its subset. In both data sets, rater 3 appears to assign lower scores to artifacts on average than either of the other 2 raters. Disaggregating the data across rubric by rater for both data sets also reveals that there are disagreements between each rater by the type of rubric as well. While little information is known about the raters, we know that they all come from different departments within Dietrich College; as a result, their backgrounds could explain the variability in the ratings across rubric categories. The College contains a mix of academic disciplines, from more analytical and technical fields like Statistics and Data Science to more subjective fields such as English and History. Academics in these varied fields may not necessarily focus on statistical approaches with the sets of assumptions. The lack of consistent training may result in some raters focusing on more mechanical aspects of the artifacts such as the Interpret Results or Initial EDA rubrics; alternatively, other raters may give less attention to these categories and focus more on the composition of the artifact and the strength of the reasoning and argument provided. Such differences are inherent in a subjective rating process and may help us understand the discrepancies in the results of the experiment.

Research Question 2 - Agreement Among the Raters

Table 7 shows that based on the ICCs by rubric category, the raters agree less on the artifacts that they all viewed relative to the full data set. This is evident by the larger ICC value associated with a given rubric category for five of the seven rubric categories; the only two that have lower ICCs are Interpret Results and Select Method. In general, raters do tend agree with each other more often than not when examining the PAE for the artifacts that each of them reviewed. However, since most of the PAEs are between 50-75%, this suggests that the agreement among raters is not exceptionally strong for the rubrics that they all reviewed.

There are specific rubrics that indicate some disagreement between raters. The rubrics for Research Question and Text Organization appear to have the least amount of agreement among the raters. This is illustrated by either the relatively lower measures of agreement between ICC and PAE (Text Organization) or a combination of low ICC scores and variability in the PAE scores depending on the pair of raters (Research Question). These differences suggests that overall agreement between the raters does not imply that every rater agrees with one another for each rubric category.

One potential reason for the differences in ratings for the full data set compared to the subset could be discrepancy in sample size. While each rater reviewed 39 artifacts in total, only a third of them were reviewed by each of the three raters. If there had been more overlap in the number of artifacts reviewed by each rater, it is possible that the ICCs would converge more closely with one another for the full data and the subset. The differences in the backgrounds of the raters could also be driving the relatively weak levels of agreement between the raters. The lack of consistent training by each of the raters results in each of them emphasizing different rubrics and therefore causing discrepancies in the categories they assign for a different rating.

It is important to note the limitations of both ICC and PAE. ICC only measures the correlation between raters for a given rubric category. For example, if one rater gave ratings of 2 only and the other rater gave ratings of 3 only for a certain rubric type, the ICC for that rubric would be large even though they did not agree with one another. On the other hand, the ICC could be small for a given rubric category even if there is at least some overlap in the ratings between the raters. PAE provides an alternative to ICC because it measures the proportion times that two raters agreed on the same rating for a given rubric. However, PAE is limited as a measure of agreement since it requires both raters to rate the rubric categories of the same artifact. There also does not appear to be a consistent relationship between PAE and ICC; that is, a large PAE across the three rater pairs does not necessarily correspond to a high ICC value for the subset of data or vice versa.

Research Question 3 - Relationships between Ratings and Factors in the Experiment

$$Final\ Model: Rating \sim Rater + Rubric + Semester + Rater{:}Rubric + (0 + Rater \mid Artifact) \\ + (0 + Rubric \mid Artifact)$$

While some of the terms in the final model effects are expected due to the nature of the experiment, others are noteworthy and should be considered in the context of the results of the experiment. The statistically significant negative coefficient of the fixed effect portrayed in Table 8 for semester is not surprising, since Freshman level statistics courses typically require at least some remedial high school mathematics. Therefore, if a student was not prepared to enroll upon entering the College, they could have been taking the prerequisite course for this class in the Fall, suggesting that they were not adequately for the curriculum from their secondary education.

The EDA and results for Question 1 suggest that there would be a fixed effect for rubric as well as a random effect interaction between rubric and artifact. Since the rubrics are assessing different skills, the ratings were likely to vary by rubric category. In some instances, such as Initial EDA or Interpret Results, these skills are largely focused on in an introductory statistics course to understand the basic aspects of statistical analysis. Undergraduates prepared to take Freshman Statistics are likely adequately prepared by their high school education to learn these concepts with minimal difficulty. Alternatively, skills assessed in rubrics such as Select Method require more critical thinking skills about the principles of statistics, and it is less likely that first-year freshman have adequately developed these skills to argue their case to an expert rater. This would explain the larger fixed effects coefficient estimates for Initial EDA and Interpret Results relative to Select Method, even though all of them are statistically significant at the 5% level and positive. Similarly, we would expect a random effect interaction between rubric and artifact; since each artifact is of different quality, the ratings for each rubric would vary across artifact.

The more interesting result identified from the final model is the fixed effect for rater as well as the interaction between the fixed effects for rater and rubric. Also noteworthy is the random effects interaction between rater and artifact. Table 8 displays the statistically significant positive coefficient for rater 2 which suggests that rater 2 gives a higher rating score on average compared to rater 1. Additionally, the coefficients that capture the interaction between the rubrics Interpret Results, Research Question, Select Method, and Text Organization and raters 2 and 3 are all statistically significant and negative. Coupling this with the inclusion of a random effects interaction, these results suggest that the raters are not rating the artifacts in the same way when assessing the artifacts. This is problematic for the results of the experiment because the variation in ratings across the artifacts is associated with both who is rating the artifact and the quality of the artifact. Such findings dampen the utility of the results obtained by the experiment.

These relationships should be of interest to the Associate Dean for future experiments. While it is evident that certain rubric categories require more emphasis in the core curriculum of Freshman Statistics, it is also important to highlight the differences in the background of the raters, as the subjectivity of the review could reveal that the differences in training across raters affect how they assess evidence. More training for the raters may be necessary to ensure that the raters are interpreting the artifacts similarly and not being implicitly biased toward assigning ratings on rubric categories that more closely align with their academic backgrounds.

Research Question 4 - Other Interesting Findings in the Data

In addition to investigating the relationship between ratings, rubric, and rater, we also explore how other factors such as semester and sex relate to ratings for a given artifact. From section 7.B.4 of the Appendix we find no noticeable differences between ratings by sex even after disaggregating the data by rubric and examining the relationship in the subset of the full data. This is also evident from the final model discussed in Question 3 since sex was not included as either a fixed or random effect term. These results are noteworthy for multiple reasons. First, the lack of a notable difference for sex suggests that males and females perform relatively similarly according to the raters and there is no gender gap in performance, even across the rubric categories. It should also be noted that a larger number of females sampled for the data than males. Since we don’t have information on how the data were sampled, if we assume that the sampling was proportionate to the count of males and females in the Freshman Statistics classes, the results suggest that more females are taking courses Statistics. This could potentially indicate that more females are taking STEM courses and considering additional coursework in a STEM-related field. However, it could also be explained by less females placing out of Freshman Statistics, or that more females are taking the course to satisfy a Gen Ed requirement.

We explore the relationship between rating and semester through similar summary statistics and graphics. As shown in Section 7.B.4 of the Appendix, average rating for the Fall semester is slightly higher than the Spring, and this true for almost every rubric category as well. This aligns with our expectation since students taking Freshman level statistics courses in the Spring typically require at least one semester of remedial mathematics in their first semester of college to satisfy the course prerequisites. As a result, they are likely to be less skilled in mathematics and would therefore receive lower ratings than those who were prepared to take Freshman Statistics in the fall. The difference is also accounted for in the final model from Question 3 as a fixed effect. These results could be of note to the dean if more information is available about the mathematics background and coursework of students who took Freshman Statistics in the Spring versus the Fall. If many of these students took a remedial mathematics course at Carnegie Mellon University in preparation, it may suggest that the course is not adequately preparing these students for the introductory statistics. Outcomes for these students could also be improved by emphasizing resources available on campus such as peer tutoring or office hours to ensure that the students get the support necessary to succeed in appropriately learning the curriculum.

While these results help contextualize some of the relationships found between rater and the factor variables in the data set, more work can be to investigate this data to find interesting results. Future steps could include more thorough EDA between the factor variables themselves, or additional groupings between factor variables and ratings. Other models could be fit to the data such as additional multilevel models such as the multilevel logit model or ordinary and generalized linear regression models and compared to what association they capture between variables in the data set. Any insights obtained from these next steps can help inform the Associate Dean on how to develop the appropriate curriculum for Freshman Statistics to ensure that students who pass the course are adequately prepared for future studies in statistics within and outside of the Statistics Department.

6 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 8, 2021 from https://canvas.cmu.edu/courses/25337/files/folder/Project02

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

Project 2 Technical Appendix

Daniel Nason

Contents

7 Technical Appendix

```
library(tidyverse)
library(arm)
library(janitor)
library(LMERConvenienceFunctions, warn.conflicts=F, quietly=T)
library(lme4, warn.conflicts=F, quietly=T)
library(latex2exp)
library(kableExtra)
setwd("C:/Users/Owner/CMU/Fall/36-617/Project 2")
```

7.A Data

7.A.1 Research Question 1 - Distribution of Ratings by Rubric and Rater

7.A.1.i Full Data

```
ratings <- read_csv("ratings.csv") %>%
  janitor::clean_names() %>%
  dplyr::select(-c(x1, sample, overlap)) %>%
  mutate(
    sex = as.factor(sex),
    repeated = as.factor(repeated),
    semester = as.factor(semester),
    rater = as.factor(rater)
    )

# drop the row with the error for sex
ratings <- ratings %>%
  filter(!(sex == "--"))

colSums(is.na(ratings)) # NAs: 1 in crit_des and 1 in vis_org
```

7.A.1.i.a Rubrics

```
##       rater    semester         sex     rsrch_q    crit_des    init_eda    sel_meth
##           0           0           0           0           1           0           0
## interp_res     vis_org     txt_org    artifact    repeated
##           0           1           0           0           0
```

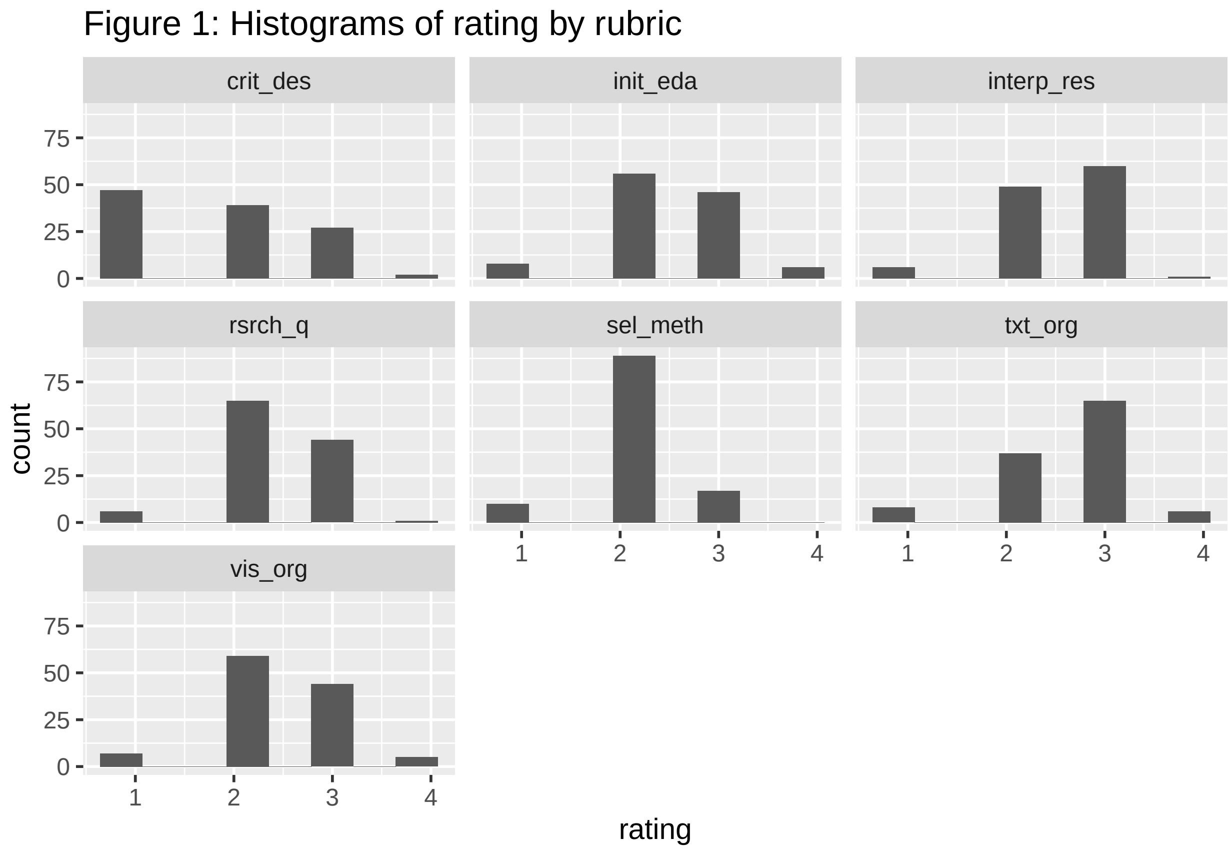
```
addmargins(table(ratings$rater))
```

```
##
##   1   2   3 Sum
##  39  39  38 116
```

```
# names for the tables, to be used in the paper
row_name_vec <- c("Research Question", "Critique Design", "Initial EDA", "Select Method", "Interpret Res

# distribution of the rubrics
```

```r
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  ggplot(aes(x = rating)) +
  geom_histogram(bins = 8, position = "dodge") +
  facet_wrap(~ rubric) +
  labs(title = "Figure 1: Histograms of rating by rubric")
```



```r
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  group_by(rubric) %>%
  dplyr::summarise(
    n = length(rating),
    min = min(rating, na.rm = T),
    Q1 = quantile(rating, 0.25, na.rm = T),
    median = median(rating, na.rm = T),
    mean = mean(rating, na.rm = T),
    Q3 = quantile(rating, 0.75, na.rm = T),
    max = max(rating, na.rm = T),
    sd = sd(rating, na.rm = T)
  ) %>%
  kbl(booktabs=T, caption = "Summary statistics of ratings by rubric") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 1: Summary statistics of ratings by rubric

rubric	n	min	Q1	median	mean	Q3	max	sd
crit_des	116	1	1	2	1.860870	2.5	4	0.8365233
init_eda	116	1	2	2	2.431034	3.0	4	0.7006101
interp_res	116	1	2	3	2.482759	3.0	4	0.6112391
rsrch_q	116	1	2	2	2.344828	3.0	4	0.5912911
sel_meth	116	1	2	2	2.060345	2.0	3	0.4807384
txt_org	116	1	2	3	2.594828	3.0	4	0.6975541
vis_org	116	1	2	2	2.408696	3.0	4	0.6740250

We see that each of the raters evaluated 39 rubrics; however, since there was an error in the “sex” data, this observation is dropped from the data set. The rubrics for Critique Design and Select Method have lower means than the others, and for Critique Design no student received a rating of 4. The summary statistics show that the distribution of the ratings for each of the rubrics is not identical. The Critique Design and Selection Method rubrics have a lower rating on average compared to the other rubrics. The standard deviation is also higher for Critique Design, but lower for Select Method compared to the other rubric categories.

The histograms of rating by rubric category also confirm that the distributions are not identical across the rubric categories. The ratings for Initial EDA, Research Question, and Visual Organization are roughly similar in that the most frequent value given by the rater’s is 2 followed closely by 3, and less than 10 students each receive either a 1 or a 4. The distributions for Interpret Results and Text Organization have similar distributions except that the most frequent value is 3 then 2, with the scores 1 and 4 each occurring less than 10 times. All of these distributions somewhat resemble the normal distribution in that the majority of the data lies near the center (2 or 3) with little data for the more extreme values (1 or 4) resembling tails. Selection Method also resembles a normal distribution, except that no student received a 4 in this category from any of the raters.

For Critique Design, we see that the most frequent rating given is 1, followed by 2 and then 3, with hardly any students receiving a 4. In general, while few students earn a score of 4 for any of the categories, Critique Design has the lowest score on average and the most amount of ratings of 1, suggesting that students struggled the most with this category.

```
table(ratings$sex)
```

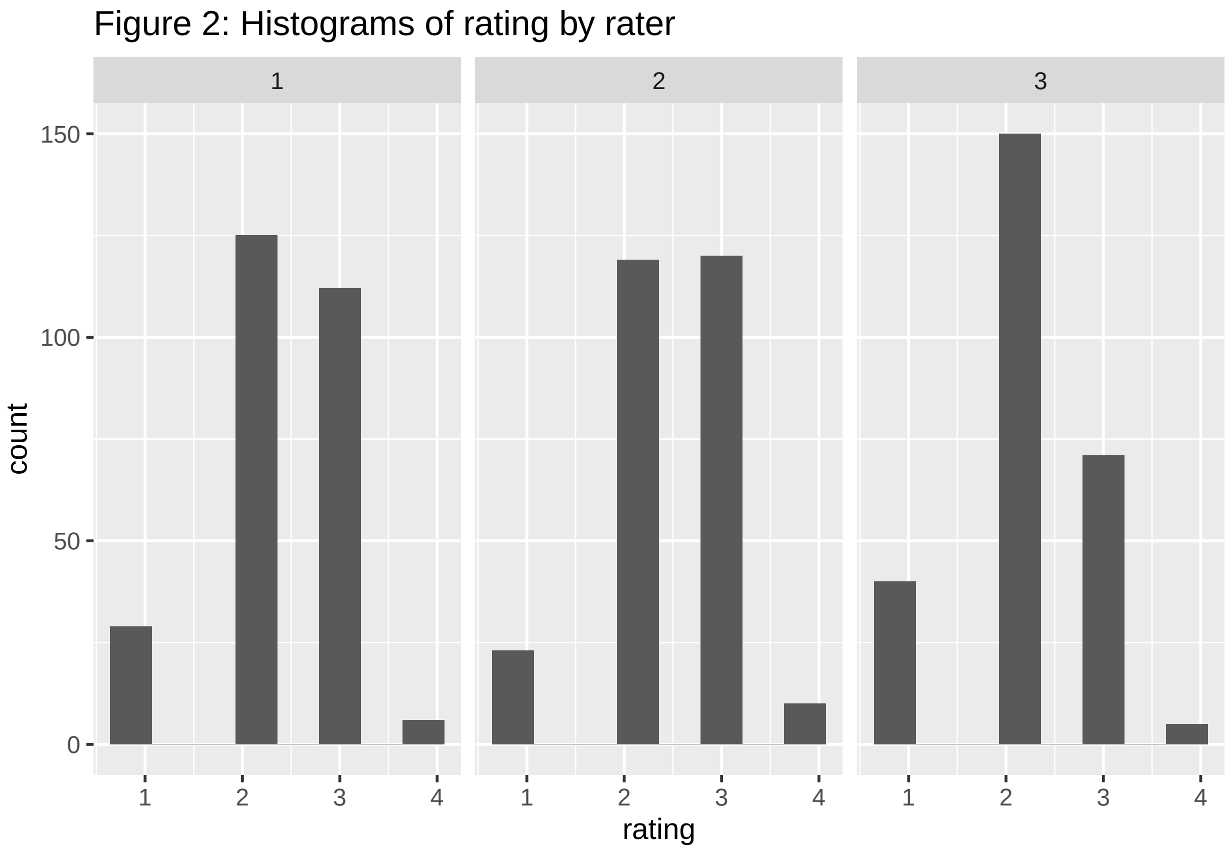
7.A.1.i.b Raters

```
##
## --  F  M
##  0 64 52
```

```
table(ratings$semester)
```

```
##
##   Fall Spring
##     82     34
```

```
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  ggplot(aes(x = rating)) +
  geom_histogram(bins = 8, position = "dodge") +
  facet_wrap(~ as.factor(rater)) +
  labs(title = "Figure 2: Histograms of rating by rater")
```



```
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  group_by(rater) %>%
  dplyr::summarise(
    n = length(rating),
    min = min(rating, na.rm = T),
    Q1 = quantile(rating, 0.25, na.rm = T),
    median = median(rating, na.rm = T),
    mean = mean(rating, na.rm = T),
    Q3 = quantile(rating, 0.75, na.rm = T),
    max = max(rating, na.rm = T),
    sd = sd(rating, na.rm = T)
  ) %>%
  kbl(booktabs=T, caption = "Summary statistics of ratings by rater") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 2: Summary statistics of ratings by rater

rater	n	min	Q1	median	mean	Q3	max	sd
1	273	1	2	2	2.349265	3	4	0.6974383
2	273	1	2	2	2.430147	3	4	0.6996910
3	266	1	2	2	2.154135	3	4	0.6859244

Examining the categorical variables, we see that more of the ratings were evaluated from students who took the class in the Fall compared to the Spring. There are slightly more females than males in the class, and about one in three of the students who took the class had to repeat it. Since there is a blank entry for sex, this observation is dropped from the data set.

The summary statistics show that the distribution of the ratings for each of the raters is relatively similar, but there are some differences. Rater 3 gives lower ratings on average compared to the other raters, but the standard deviation is relatively similar across raters, and the five number summaries are identical.

The histogram by rater also illustrate that the distribution is not identical across raters. The rating distribution for raters 1 and 2 are similar and approximately resemble a normal distribution, but the distribution for rater 3 is somewhat different in that rater 3 is the least like to give a 3 or 4 rating.

7.A.1.ii Subsetted Data - Artifacts reviewed by all raters

```
sub_ratings <- read_csv("ratings.csv") %>%
  janitor::clean_names() %>%
  dplyr::select(-c(x1, sample)) %>%
  filter(!is.na(overlap)) %>%
  mutate(
    sex = as.factor(sex),
    repeated = as.factor(repeated),
    semester = as.factor(repeated),
    rater = as.factor(rater)
    )
```

7.A.1.ii.a Rubrics

```
## New names:
## * ‘‘ -> ...1

## Rows: 117 Columns: 15

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): Semester, Sex, Artifact
## dbl (12): ...1, Rater, Sample, Overlap, RsrchQ, CritDes, InitEDA, SelMeth, I...

##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

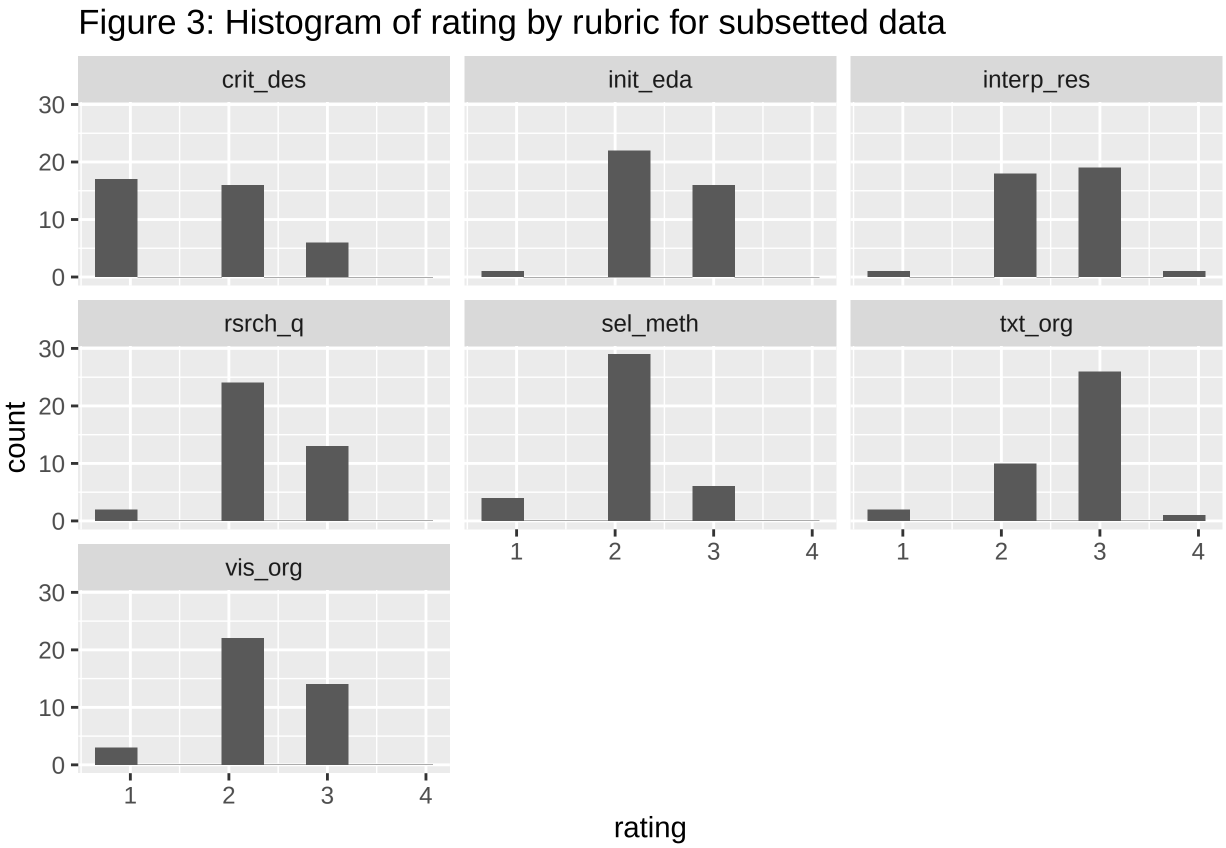
```
colSums(is.na(sub_ratings)) # none based on filtering
```

```
##       rater     overlap    semester         sex     rsrch_q    crit_des    init_eda
##           0           0           0           0           0           0           0
##    sel_meth interp_res     vis_org     txt_org    artifact    repeated
##           0           0           0           0           0           0
```

```
addmargins(table(sub_ratings$rater))
```

```
##
##    1   2   3 Sum
##   13  13  13  39
```

```
sub_ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  ggplot(aes(x = rating)) +
  geom_histogram(bins = 8, position = "dodge") +
  facet_wrap(~ rubric) +
  labs(title = "Figure 3: Histogram of rating by rubric for subsetted data")
```



```
sub_ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  group_by(rubric) %>%
  dplyr::summarise(
    n = length(rating),
    min = min(rating, na.rm = T),
    Q1 = quantile(rating, 0.25, na.rm = T),
    median = median(rating, na.rm = T),
    mean = mean(rating, na.rm = T),
    Q3 = quantile(rating, 0.75, na.rm = T),
    max = max(rating, na.rm = T),
    sd = sd(rating, na.rm = T)
  ) %>%
  kbl(booktabs=T, caption = "Summary statistics of ratings by rubric (common artifacts only)") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 3: Summary statistics of ratings by rubric (common artifacts only)

rubric	n	min	Q1	median	mean	Q3	max	sd
crit_des	39	1	1	2	1.717949	2	3	0.7236137
init_eda	39	1	2	2	2.384615	3	3	0.5436419
interp_res	39	1	2	3	2.512821	3	4	0.6013929
rsrch_q	39	1	2	2	2.282051	3	3	0.5595448
sel_meth	39	1	2	2	2.051282	2	3	0.5103517
txt_org	39	1	2	3	2.666667	3	4	0.6212607
vis_org	39	1	2	2	2.282051	3	3	0.6047495

We limit the data to examine just the 13 artifacts seen by all 13 raters, which reduces the number of observations to 39, and compare the results to the original data. The summary statistics show that the distribution of the ratings for each of the rubrics is not identical, similar to the full data set. The average of the ratings by rubric item is relatively consistent with the full data set; however, we see that fewer rubric categories receive a score of 4 in this data. While the original data set has no students receiving a rating of 4 for Select Method, in the data set where all 13 raters reviewed each paper, this rubric category as well as Critique Design, Initial EDA, Research Question, and Visual Organization have no instances where an artifact receives a rating of 4 for their paper. The standard deviations are also smaller in this data for almost every rubric category, suggesting less disagreement between the raters where they all reviewed the same papers.

The histograms by rubric also confirm that the distributions are not identical across the rubric categories, similar to original data set. The distributions across each of the rubrics remains similar when looking at the original data compared to the subsetted data, with the main distinction being the lack of ratings of 4 for the categories Critique Design, Initial EDA, Research Question, and Visual Organization.

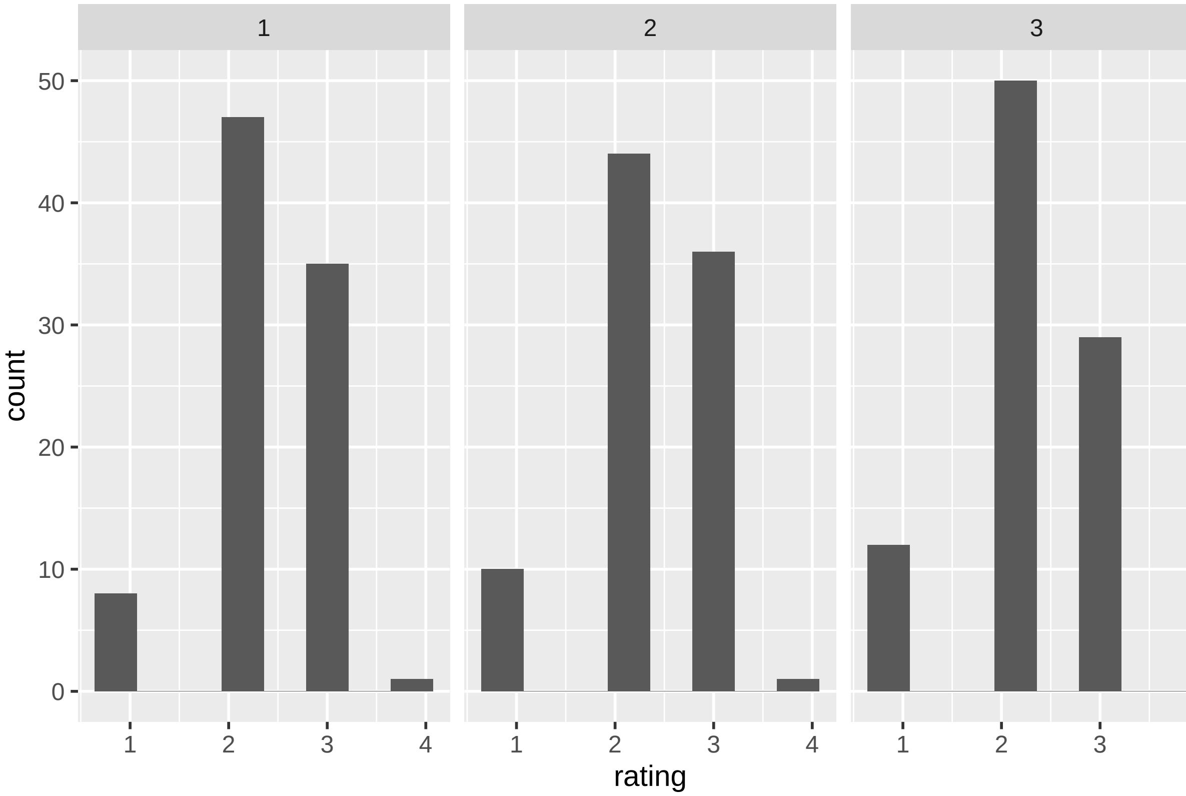
```
sub_ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  group_by(rater) %>%
  dplyr::summarise(
    n = length(rating),
    min = min(rating, na.rm = T),
    Q1 = quantile(rating, 0.25, na.rm = T),
    median = median(rating, na.rm = T),
    mean = mean(rating, na.rm = T),
    Q3 = quantile(rating, 0.75, na.rm = T),
    max = max(rating, na.rm = T),
    sd = sd(rating, na.rm = T)
  ) %>%
  kbl(booktabs=T, caption = "Summary statistics of ratings by rater (common artifacts only)") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 4: Summary statistics of ratings by rater (common artifacts only)

rater	n	min	Q1	median	mean	Q3	max	sd
1	91	1	2	2	2.318681	3	4	0.6477160
2	91	1	2	2	2.307692	3	4	0.6781070
3	91	1	2	2	2.186813	3	3	0.6482813

```
sub_ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  ggplot(aes(x = rating)) +
  geom_histogram(bins = 8, position = "dodge") +
  facet_wrap(~ rater) +
  labs(title = "Figure 4: Histogram of rating by rater for subsetted data")
```

Figure 4: Histogram of rating by rater for subsetted data



7.A.1.ii.b Raters

The summary statistics and histograms show that the distribution of the ratings for each of the raters is close to the overall data set in that are relatively similar. We see that rater 3 gives lower ratings on average; however, the distribution has slightly changed for rater 2 and as a result their average for the subset of the data is slightly lower. The standard deviation is relatively similar across raters and slightly smaller relative to the overall data set. The five number summaries are also similar to the overall data set, except that rater 3 gives no ratings of 4 for the data where all raters reviewed the same papers.

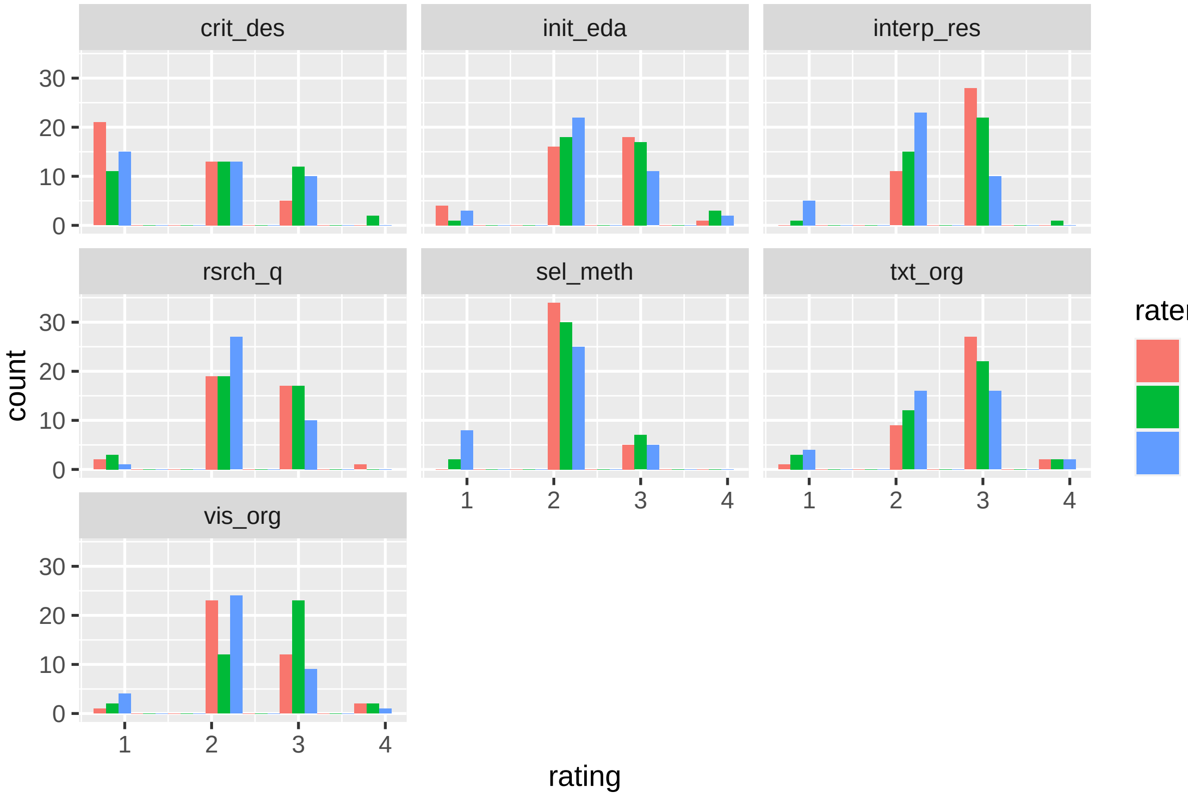
These results suggest that rater 3 disagrees with the other raters for overall ratings, but the distribution of ratings more closely aligns with the other raters in the subset compared to the overall data set.

7.B Results

7.B.1 Research Question 1 - Distribution of Ratings by Rubric and Rater

```r
# EDA for agreement across rubrics by rater
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  ggplot(aes(x = rating, fill = rater)) +
  geom_histogram(bins = 8, position = "dodge") +
  facet_wrap(~ rubric) +
  labs(title = "Figure 5: Histogram of rating by rubric for each rater")
```

Figure 5: Histogram of rating by rubric for each rater



7.B.1.i Full Data

```
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  group_by(rater, rubric) %>%
  dplyr::summarise(
    n = length(rating),
    min = min(rating, na.rm = T),
    Q1 = quantile(rating, 0.25, na.rm = T),
    median = median(rating, na.rm = T),
    mean = mean(rating, na.rm = T),
    Q3 = quantile(rating, 0.75, na.rm = T),
    max = max(rating, na.rm = T),
    sd = sd(rating, na.rm = T)
  ) %>%
  kbl(booktabs=T, caption = "Summary statistics of ratings by rater and rubric") %>%
  kable_classic(latex_options = "HOLD_position")
```

```
## ‘summarise()‘ has grouped output by ’rater’. You can override using the ‘.groups‘ argument.
```

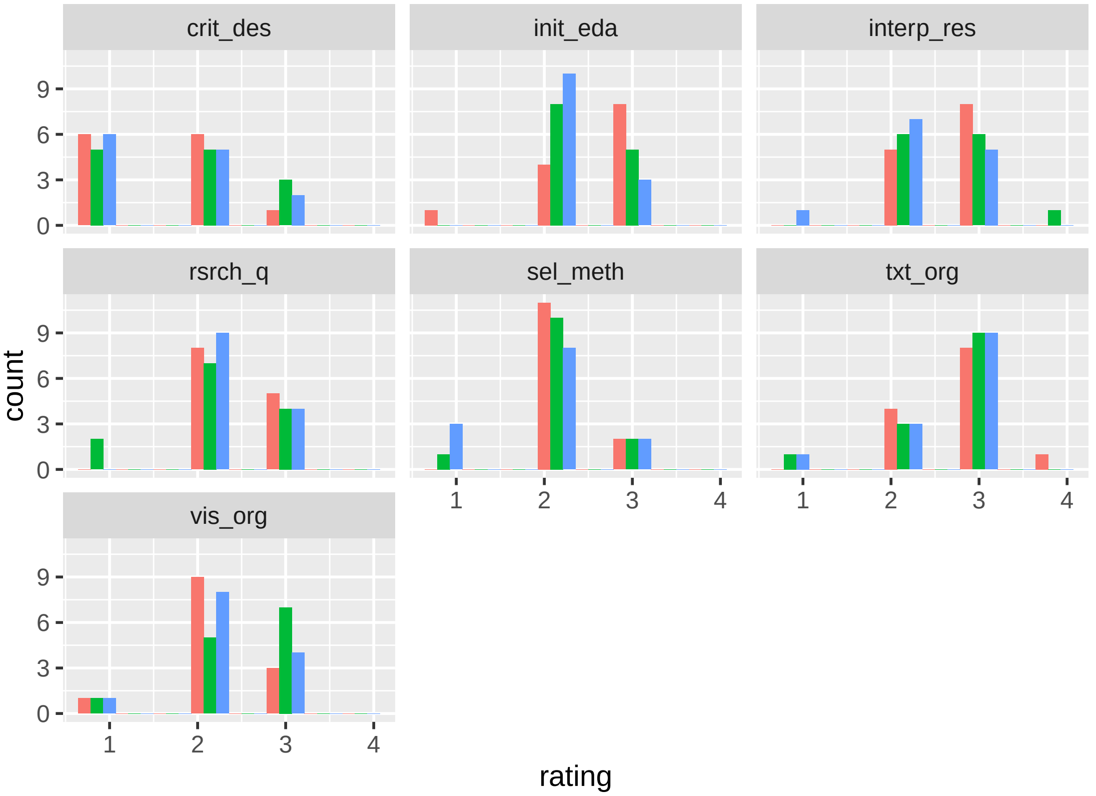
Table 5: Summary statistics of ratings by rater and rubric

rater	rubric	n	min	Q1	median	mean	Q3	max	sd
1	crit_des	39	1	1.0	1	1.589744	2.00	3	0.7151720
1	init_eda	39	1	2.0	2	2.410256	3.00	4	0.7151720
1	interp_res	39	2	2.0	3	2.717949	3.00	3	0.4558808
1	rsrch_q	39	1	2.0	2	2.435897	3.00	4	0.6405126
1	sel_meth	39	2	2.0	2	2.128205	2.00	3	0.3386884
1	txt_org	39	1	2.5	3	2.769231	3.00	4	0.5831646
1	vis_org	39	1	2.0	2	2.394737	3.00	4	0.6383879
2	crit_des	39	1	1.0	2	2.131579	3.00	4	0.9055699
2	init_eda	39	1	2.0	3	2.564103	3.00	4	0.6803587
2	interp_res	39	1	2.0	3	2.589744	3.00	4	0.5946228
2	rsrch_q	39	1	2.0	2	2.358974	3.00	3	0.6277436
2	sel_meth	39	1	2.0	2	2.128205	2.00	3	0.4690128
2	txt_org	39	1	2.0	3	2.589744	3.00	4	0.7151720
2	vis_org	39	1	2.0	3	2.641026	3.00	4	0.6683514
3	crit_des	38	1	1.0	2	1.868421	2.75	3	0.8111071
3	init_eda	38	1	2.0	2	2.315790	3.00	4	0.7015528
3	interp_res	38	1	2.0	2	2.131579	2.75	3	0.6225949
3	rsrch_q	38	1	2.0	2	2.236842	2.75	3	0.4895784
3	sel_meth	38	1	2.0	2	1.921053	2.00	3	0.5873246
3	txt_org	38	1	2.0	2	2.421053	3.00	4	0.7580765
3	vis_org	38	1	2.0	2	2.184210	2.75	4	0.6516201

We see the differences across between rater when looking across each of the rubric categories. For each of the rubrics we see that rater 3 gives the fewest ratings with a score of 3 or higher, except Critique Design where rater 1 gives the lowest average ratings and the least amount of ratings with a score of 3 or higher. However, the plots also show that raters 1 and 2 do not necessarily agree by rubric category in the ratings that they give. For example, rater 1 is more likely than rater 2 to give a rating of 3 for some categories (Interpret Results, Text Organization), but not for others (Visual Organization). These results suggest that there is disagreement between all the raters when looking at ratings across each of the rubric categories.

```r
sub_ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  ggplot(aes(x = rating, fill = rater)) +
  geom_histogram(bins = 8, position = "dodge") +
  facet_wrap(~ rubric) +
  labs(title = "Figure 6: Histogram of rating by rubric for each rubric for subsetted data")
```

Figure 6: Histogram of rating by rubric for each rubric for subsetted



7.B.1.ii Subsetted Data

```
sub_ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  group_by(rater, rubric) %>%
  dplyr::summarise(
    n = length(rating),
    min = min(rating, na.rm = T),
    Q1 = quantile(rating, 0.25, na.rm = T),
    median = median(rating, na.rm = T),
    mean = mean(rating, na.rm = T),
    Q3 = quantile(rating, 0.75, na.rm = T),
    max = max(rating, na.rm = T),
    sd = sd(rating, na.rm = T)
  ) %>%
  kbl(booktabs=T, caption = "Summary statistics of ratings by rater and rubric (common artifacts only)"
  kable_classic(latex_options = "HOLD_position")
```

```
## ‘summarise()‘ has grouped output by ’rater’. You can override using the ‘.groups‘ argument.
```

Table 6: Summary statistics of ratings by rater and rubric (common artifacts only)

rater	rubric	n	min	Q1	median	mean	Q3	max	sd
1	crit_des	13	1	1	2	1.615385	2	3	0.6504436
1	init_eda	13	1	2	3	2.538461	3	3	0.6602253
1	interp_res	13	2	2	3	2.615385	3	3	0.5063697
1	rsrch_q	13	2	2	2	2.384615	3	3	0.5063697
1	sel_meth	13	2	2	2	2.153846	2	3	0.3755338
1	txt_org	13	2	2	3	2.769231	3	4	0.5991447
1	vis_org	13	1	2	2	2.153846	2	3	0.5547002
2	crit_des	13	1	1	2	1.846154	2	3	0.8006408
2	init_eda	13	2	2	2	2.384615	3	3	0.5063697
2	interp_res	13	2	2	3	2.615385	3	4	0.6504436
2	rsrch_q	13	1	2	2	2.153846	3	3	0.6887372
2	sel_meth	13	1	2	2	2.076923	2	3	0.4935481
2	txt_org	13	1	2	3	2.615385	3	3	0.6504436
2	vis_org	13	1	2	3	2.461539	3	3	0.6602253
3	crit_des	13	1	1	2	1.692308	2	3	0.7510676
3	init_eda	13	2	2	2	2.230769	2	3	0.4385290
3	interp_res	13	1	2	2	2.307692	3	3	0.6304252
3	rsrch_q	13	2	2	2	2.307692	3	3	0.4803845
3	sel_meth	13	1	2	2	1.923077	2	3	0.6405126
3	txt_org	13	1	2	3	2.615385	3	3	0.6504436
3	vis_org	13	1	2	2	2.230769	3	3	0.5991447

When looking at the ratings by rater across each of the rubric categories, we see that the subsetted data has a relatively similar distribution to the overall data. For each of the rubrics we see that rater 3 gives the fewest ratings with a score of 3 or higher, except Critique Design and Visual Organization. The histograms also show that raters 1 and 2 do not necessarily agree by rubric category in the ratings that they give in that rater 1 is more likely than rater 2 to give a rating of 3 for some categories (Research Question, Initial EDA, Interpret Results), but not for others (Text Organization, Critique Design). These results suggest that there is disagreement between all the raters when looking at ratings across rubric categories, which is similar to the overall data set.

7.B.2 Research Question 2 - Agreement Among the Raters

```r
# subset data
ICC_vec_sub <- NULL
for (i in 1:7){
  x <- apply(sub_ratings[,i+4], 2, as.numeric)
  tmp <- lmer(x ~ 1 + (1 | artifact), data = sub_ratings)
  s2 <- summary(tmp)$sigma^2
  t2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
  ICC <- t2 / (t2 + s2)
  ICC_vec_sub <- c(ICC_vec_sub, ICC)
}

# percent exact agreement, subset of data
sub_rubrics <- as.data.frame(sub_ratings) %>%
  dplyr::select(rater, rsrch_q, crit_des, init_eda, sel_meth, interp_res, vis_org, txt_org)

r1 <- c()
r2 <- c()
rubric <- c()
tab_vec <- c()
combo <- combn(3,2)
rub_nam <- colnames(ratings)[4:10]
for (i in 1:dim(combo)[2]) {
  idx1 <- combo[1, i]
  idx2 <- combo[2, i]

  for (j in 1:length(colnames(sub_rubrics)[-1])) {
    x <- table(
          sub_rubrics[sub_rubrics$rater == idx1, rub_nam[j]],
          sub_rubrics[sub_rubrics$rater == idx2, rub_nam[j]]
          )
    y <- x
    if (min(as.numeric(rownames(x))) > min(as.numeric(colnames(x)))) {
      y <- y[, -1]
    }
    else if (min(as.numeric(rownames(x))) < min(as.numeric(colnames(x)))) {
      y <- y[-1, ]
    }
    tab_vec <- c(tab_vec, sum(diag(y)) / sum(x))
    r1 <- c(r1, idx1)
    r2 <- c(r2, idx2)
    rubric <- c(rubric, rub_nam[j])
  }
}
agree_12 <- tab_vec[1:7]
agree_13 <- tab_vec[8:14]
agree_23 <- tab_vec[15:21]

# full data
ICC_vec_full <- NULL
for (i in 1:7){
  x <- apply(ratings[,i+3], 2, as.numeric)
  tmp <- lmer(x ~ 1 + (1 | artifact), data = ratings)
  s2 <- summary(tmp)$sigma^2
  t2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
  ICC <- t2 / (t2 + s2)
  ICC_vec_full <- c(ICC_vec_full, ICC)
}

tab_df <- data.frame(Rubric = row_name_vec, ICC.alldata = round(ICC_vec_full, 2), ICC.subdata = round(I
  arrange(Rubric)
tab_df %>%
  kbl(booktabs=T, caption = "Intraclass Correlation (ICC) and Percent Exact Agreement (PAE) between rat
  kable_classic(latex_options = "HOLD_position")
```

Table 7: Intraclass Correlation (ICC) and Percent Exact Agreement (PAE) between raters by rubric

Rubric	ICC.alldata	ICC.subdata	PAE12	PAE13	PAE23
Critique Design	0.67	0.57	0.54	0.62	0.69
Initial EDA	0.69	0.49	0.69	0.54	0.85
Interpret Results	0.22	0.23	0.62	0.54	0.62
Research Question	0.21	0.19	0.38	0.77	0.54
Select Method	0.46	0.52	0.92	0.62	0.69
Text Organization	0.19	0.14	0.69	0.62	0.54
Visual Organization	0.66	0.59	0.54	0.77	0.77

Comparing the intraclass correlation (ICC) calculations for the full data set and the subsetted data, we see that the ICC is larger for every rubric category except for Select Method and Interpret Results. Since this is a measure of agreement among the raters, the results suggest that the raters generally agree less on the artifacts that they all viewed relative to the full data set.

The percent exact agreement (PAE) also measures agreement among the raters since it specifically examines what percentage of the time a pair of raters reached the same conclusion for a given rubric on an artifact they both rated. These calculations also help us to identify which specific rubric any two raters disagree on and how it contributes to the overall disagreement between raters. From the table, we see that the highest percent exact agreement is between raters 1 and 2 for the Select Methods category, while lowest agreement is between raters 1 and 2 for the Research Question rubric category.

However, we cannot perform the percent exact agreement calculations for the entire data set because the calculation requires that both raters reviewed the same artifact. Only the smaller data set allowed us to perform this calculation since the observations included were the ones in which all 3 raters reviewed the same artifact.

7.B.3 Research Question 3 - Relationships between Ratings and Factors in the Experiment

```r
sub_tall <- sub_ratings %>%
  pivot_longer(cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating")

rubric.names <- sort(unique(sub_tall$rubric))
model.formula.13 <- as.list(rep(NA,7))
names(model.formula.13) <- rubric.names
# drop semester b/c only one level for subset of data
for (i in rubric.names) {
  ## fit each base model
  rubric.data <- sub_tall[sub_tall$rubric==i,]
  tmp <- lmer(rating ~ -1 + rater + sex + (1|artifact), data = rubric.data, REML = FALSE)
  ## do backwards elimination
  tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)
  ## check to see if the raters are significantly different from one another
  tmp.single_intercept <- update(tmp.back_elim, . ~ . + 1 - rater)
  pval <- anova(tmp.single_intercept, tmp.back_elim)$"Pr(>Chisq)"[2]
  ## choose the best model
  if (pval <= 0.05) {
    tmp_final <- tmp.back_elim
  } else {
    tmp_final <- tmp.single_intercept
  }
    ## and add to list...
  model.formula.13[[i]] <- formula(tmp_final)
}
```

7.B.3.i Fixed Effects - 7 Rubric-Specific Models (subsetted data)

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

```
model.formula.13
```

```
## $crit_des
## rating ~ (1 | artifact)
##
## $init_eda
## rating ~ (1 | artifact)
##
## $interp_res
## rating ~ (1 | artifact)
##
## $rsrch_q
## rating ~ (1 | artifact)
##
## $sel_meth
## rating ~ (1 | artifact)
##
## $txt_org
## rating ~ (1 | artifact)
##
## $vis_org
## rating ~ (1 | artifact)
```

For the subsetted data set that only looks at the instances where all 3 raters looked at the same artifacts entries we find that for each rubric category, the intercept only model is adequate. None of the likelihood ratio tests for nested fixed effects yield p-values that are statistically significant at the 5% level. Since this is the case, interaction terms are not considered for this subsetted data.

```
full_tall <- ratings %>%
  pivot_longer(cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating")

rubric.names <- sort(unique(full_tall$rubric))
model.formula.full <- as.list(rep(NA,7))
names(model.formula.full) <- rubric.names
```

```
for (i in rubric.names) {
  rubric.data <- full_tall[full_tall$rubric==i,]
  tmp <- lmer(rating ~ -1 + rater + sex + semester + (1|artifact), data = rubric.data, REML = FALSE)
  tmp.back_elim <- fitLMER.fnc(tmp, set.REML.FALSE = TRUE,log.file.name = FALSE)
  tmp.single_intercept <- update(tmp.back_elim, . ~ . + 1 - rater)
  pval <- anova(tmp.single_intercept, tmp.back_elim)$"Pr(>Chisq)"[2]
  if (pval <= 0.05) {
    tmp_final <- tmp.back_elim
  } else {
    tmp_final <- tmp.single_intercept
  }
  model.formula.full[[i]] <- formula(tmp_final)
}
```

7.B.3.ii Fixed Effects - 7 Rubric-Specific Models (full data)

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

```
model.formula.full
```

```
## $crit_des
## rating ~ rater + (1 | artifact) - 1
##
## $init_eda
## rating ~ (1 | artifact)
##
## $interp_res
## rating ~ rater + (1 | artifact) - 1
##
## $rsrch_q
## rating ~ (1 | artifact)
##
## $sel_meth
## rating ~ rater + sex + semester + (1 | artifact) - 1
##
## $txt_org
## rating ~ (1 | artifact)
##
## $vis_org
## rating ~ rater + (1 | artifact) - 1
```

From the output, we see that for three of the rubrics (Initial EDA, Research Question, and Text Organization), the intercept only model is selected. None of the likelihood ratio tests for nested fixed effects yield p-values that are statistically significant at the 5% level for these rubrics. Of the other four rubric categories, three of them (Critique Design, Interpret Results, and Visual Organization) prefer including only rater as a fixed effect. The other rubric category (Select Method) included fixed effects for sex and semester and rater. The categories that have these additional terms included are investigated in further detail.

7.B.3.iii Interactions and Random Effects - 7 Rubric-Specific Models (full data)

```
# testing for critique design
fla <- formula(model.formula.full[["crit_des"]])
tmp <- lmer(fla, data = full_tall %>% filter(rubric == "crit_des"))
summary(tmp)$coef
```

7.B.3.iii.a - Critique Design

```
##        Estimate Std. Error  t value
## rater1 1.686325  0.1206556 13.97635
## rater2 2.112884  0.1218849 17.33508
## rater3 1.890793  0.1218849 15.51294
```

```
# statistically significant
tmp.single_intercept <- update(tmp, . ~ . + 1 - rater)
anova(tmp.single_intercept,tmp)
```

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

```
## Data: full_tall %>% filter(rubric == "crit_des")
## Models:
## tmp.single_intercept: rating ~ (1 | artifact)
## tmp: rating ~ rater + (1 | artifact) - 1
##                      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    3 277.68 285.91 -135.84   271.68
## tmp                     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
```

```
# adding random effects for rater only
m0 <- tmp.single_intercept
# cannot be tested since too many terms in the model
mA <- update(m0, . ~ . + 1 + rater + (rater|artifact))
```

```
## Error: number of observations (=115) <= number of random effects (=267) for term (rater | artifact);
```

```
summary(tmp)
```

```
## Linear mixed model fit by REML [’lmerMod’]
## Formula: rating ~ rater + (1 | artifact) - 1
##    Data: full_tall %>% filter(rubric == "crit_des")
##
## REML criterion at convergence: 271
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.55495 -0.50027 -0.08228  0.64663  1.60935
```

```
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  artifact (Intercept) 0.4349   0.6595
##  Residual             0.2473   0.4972
## Number of obs: 115, groups:  artifact, 89
##
## Fixed effects:
##        Estimate Std. Error t value
## rater1   1.6863     0.1207   13.98
## rater2   2.1129     0.1219   17.34
## rater3   1.8908     0.1219   15.51
##
## Correlation of Fixed Effects:
##        rater1 rater2
## rater2 0.244
## rater3 0.244  0.246
```

```
s2 <- summary(tmp)$sigma^2
t2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
ICC <- t2 / (t2 + s2)
ICC
```

```
## (Intercept)
##   0.6375475
```

Since there are insufficient terms to add random effects to this model, the final model for Critique Design includes only a fixed effects term for rater. The ICC for this model is approximately the same as the ICC calculated in Table 7.

```


# testing for critique design
fla <- formula(model.formula.full[["interp_res"]])
tmp <- lmer(fla, data = full_tall %>% filter(rubric == "interp_res"))
summary(tmp)$coef
```

7.B.3.iii.b - Interpret Results

```
##        Estimate Std. Error  t value
## rater1 2.704214 0.08912484 30.34186
## rater2 2.585742 0.08912484 29.01259
## rater3 2.139182 0.09026675 23.69845
```

```
# statistically significant
tmp.single_intercept <- update(tmp, . ~ . + 1 - rater)
anova(tmp.single_intercept,tmp)
```

```
## refitting model(s) with ML (instead of REML)
## Data: full_tall %>% filter(rubric == "interp_res")
## Models:
## tmp.single_intercept: rating ~ (1 | artifact)
## tmp: rating ~ rater + (1 | artifact) - 1
##                      npar    AIC    BIC   logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    3 218.53 226.79 -106.263   212.53
## tmp                     5 200.66 214.43  -95.331   190.66 21.864  2  1.787e-05
##
## tmp.single_intercept
## tmp                  ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# adding random effects for rater only
m0 <- tmp.single_intercept
# cannot be tested since too many terms in the model
mA <- update(m0, . ~ . + 1 + rater + (rater|artifact))
```

```
## Error: number of observations (=116) <= number of random effects (=270) for term (rater | artifact);
```

```
summary(tmp)
```

```
## Linear mixed model fit by REML [’lmerMod’]
## Formula: rating ~ rater + (1 | artifact) - 1
##    Data: full_tall %>% filter(rubric == "interp_res")
##
## REML criterion at convergence: 199.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5317 -0.7627  0.2635  0.6614  2.6535
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  artifact (Intercept) 0.06224  0.2495
##  Residual             0.25250  0.5025
## Number of obs: 116, groups:  artifact, 90
##
## Fixed effects:
##        Estimate Std. Error t value
## rater1  2.70421    0.08912   30.34
## rater2  2.58574    0.08912   29.01
## rater3  2.13918    0.09027   23.70
##
## Correlation of Fixed Effects:
##        rater1 rater2
## rater2 0.061
## rater3 0.062  0.062
```

```
s2 <- summary(tmp)$sigma^2
t2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
ICC <- t2 / (t2 + s2)
ICC
```

```
## (Intercept)
##   0.1977433
```

Since there are insufficient terms to add random effects to this model, the final model for Interpret Results includes only a fixed effects term for rater. The ICC for this model is approximately the same as the ICC calculated in Table 7.

```
fla <- formula(model.formula.full[["vis_org"]])
tmp <- lmer(fla, data = full_tall %>% filter(rubric == "vis_org"))
summary(tmp)$coef
```

7.B.3.iii.c - Visual Organization

```
##        Estimate Std. Error  t value
## rater1 2.377941 0.09658396 24.62045
## rater2 2.648913 0.09563943 27.69687
## rater3 2.283545 0.09658396 23.64311
```

```
tmp.single_intercept <- update(tmp, . ~ . + 1 - rater)
anova(tmp.single_intercept,tmp)
```

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

```
## Data: full_tall %>% filter(rubric == "vis_org")
## Models:
## tmp.single_intercept: rating ~ (1 | artifact)
## tmp: rating ~ rater + (1 | artifact) - 1
##                      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    3 227.21 235.44 -110.60   221.21
## tmp                     5 220.82 234.54 -105.41   210.82 10.392  2   0.005539
##
## tmp.single_intercept
## tmp                  **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# adding random effects for rater only
m0 <- tmp.single_intercept
# cannot be tested since too many terms in the model
mA <- update(m0, . ~ . + 1 + rater + (rater|artifact))
```

```
## Error: number of observations (=115) <= number of random effects (=267) for term (rater | artifact);
```

```
summary(tmp)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rating ~ rater + (1 | artifact) - 1
##    Data: full_tall %>% filter(rubric == "vis_org")
```

```
##
## REML criterion at convergence: 219.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5004 -0.3365 -0.2483  0.3841  1.8552
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  artifact (Intercept) 0.2907   0.5392
##  Residual             0.1467   0.3830
## Number of obs: 115, groups:  artifact, 89
##
## Fixed effects:
##        Estimate Std. Error t value
## rater1  2.37794    0.09658   24.62
## rater2  2.64891    0.09564   27.70
## rater3  2.28355    0.09658   23.64
##
## Correlation of Fixed Effects:
##        rater1 rater2
## rater2 0.263
## rater3 0.265  0.263
```

```
s2 <- summary(tmp)$sigma^2
t2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
ICC <- t2 / (t2 + s2)
ICC
```

```
## (Intercept)
##   0.6645938
```

Since there are insufficient terms to add random effects to this model, the final model for Visual Organization includes only a fixed effects term for rater. The ICC for this model is approximately the same as the ICC calculated in Table 7.

```
# testing for critique design
fla <- formula(model.formula.full[["sel_meth"]])
tmp <- lmer(fla, data = full_tall %>% filter(rubric == "sel_meth"))
summary(tmp)$coef
```

7.B.3.iii.d - Select Method

```
##                  Estimate Std. Error   t value
## rater1          2.1875329 0.08956093 24.425081
## rater2          2.1594176 0.09071042 23.805619
## rater3          1.9648158 0.09211804 21.329328
## sexM            0.1215888 0.09502187  1.279587
## semesterSpring -0.3195481 0.10246192 -3.118701
```

```r
# statistically significant
tmp.single_intercept <- update(tmp, . ~ . + 1 - sex - semester - rater)
anova(tmp.single_intercept,tmp)
```

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

## Data: full_tall %>% filter(rubric == "sel_meth")
## Models:
## tmp.single_intercept: rating ~ (1 | artifact)
## tmp: rating ~ rater + sex + semester + (1 | artifact) - 1
##                      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    3 155.37 163.63 -74.687   149.37
## tmp                     7 142.35 161.63 -64.178   128.35 21.018  4   0.000314
##
## tmp.single_intercept
## tmp                  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# including interaction for sex and semester, only instance where result is significant of all interact
tmp.fixed_interaction <- update(tmp, . ~ . + sex:semester)
anova(tmp, tmp.fixed_interaction)
```

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

## Data: full_tall %>% filter(rubric == "sel_meth")
## Models:
## tmp: rating ~ rater + sex + semester + (1 | artifact) - 1
## tmp.fixed_interaction: rating ~ rater + sex + semester + (1 | artifact) + sex:semester - 1
##                       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp                      7 142.35 161.63 -64.178   128.35
## tmp.fixed_interaction    8 139.75 161.78 -61.877   123.75 4.6022  1    0.03193
##
## tmp
## tmp.fixed_interaction *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# adding random effects for rater only
m0 <- tmp.single_intercept
# cannot be tested since too many terms in the model
mA <- update(m0, . ~ . + 1 + sex * semester + (sex|artifact))
```

```
## Error: number of observations (=116) <= number of random effects (=180) for term (sex | artifact); t
```

```r
mA <- update(m0, . ~ . + 1 + sex * semester + (semester|artifact))
```

```
## Error: number of observations (=116) <= number of random effects (=180) for term (semester | artifac
summary(tmp.fixed_interaction)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rating ~ rater + sex + semester + (1 | artifact) + sex:semester -
##     1
##    Data: full_tall %>% filter(rubric == "sel_meth")
##
## REML criterion at convergence: 141.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1759 -0.2823 -0.1189  0.3820  2.7330
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  artifact (Intercept) 0.0850   0.2915
##  Residual             0.1057   0.3251
## Number of obs: 116, groups:  artifact, 90
##
## Fixed effects:
##                      Estimate Std. Error t value
## rater1                2.12357    0.09314  22.799
## rater2                2.10377    0.09310  22.598
## rater3                1.90761    0.09459  20.167
## sexM                  0.22948    0.10647   2.155
## semesterSpring       -0.18123    0.12004  -1.510
## sexM:semesterSpring -0.46524    0.22049  -2.110
##
## Correlation of Fixed Effects:
##             rater1 rater2 rater3 sexM   smstrS
## rater2       0.554
## rater3       0.556  0.558
## sexM        -0.611 -0.622 -0.624
## semstrSprng -0.568 -0.538 -0.540  0.482
## sxM:smstrSp  0.324  0.285  0.286 -0.482 -0.545
```

```
s2 <- summary(tmp.fixed_interaction)$sigma^2
t2 <- attr(summary(tmp.fixed_interaction)$varcor[[1]],"stddev")^2
ICC <- t2 / (t2 + s2)
ICC
```

```
## (Intercept)
##   0.4457239
```

Since there are insufficient terms to add random effects to this model, the final model for Select Method includes only fixed effects for sex, semester, and their interaction. The ICC for this model is approximately the same as the ICC calculated in Table 7.

From these results, we see that rater, semester, and sex (and the interaction between semester and sex) are possibly relevant terms to consider as fixed effects; however, these results are only relevant for some of the rubric categories. For the categories with additional fixed effects included, there is no notable change in the ICCs calculated in these models compared to the ICCs calculated in Table 7. Additionally, random effects interactions for these terms could not be considered since the number of coefficients for the model exceeded the number of observations. Since these variables improve the model as fixed effects, their interactions for both fixed and random effects are considered when looking at the data to model rating by artifact.

```
full_tall <- ratings %>%
  pivot_longer(cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating")

# base model
lmer.tall.base <- lmer(rating ~ 1 + (0 + rubric| artifact), data = full_tall)
```

7.B.3.iv Fixed Effects, Interactions, and Random Effects - All Data

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

```
summary(lmer.tall.base)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rating ~ 1 + (0 + rubric | artifact)
##    Data: full_tall
##
## REML criterion at convergence: 1471.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0218 -0.4939 -0.0752  0.5271  3.7760
##
## Random effects:
##  Groups   Name              Variance Std.Dev. Corr
##  artifact rubriccrit_des    0.6407   0.8004
##           rubricinit_eda    0.3829   0.6188   0.26
##           rubricinterp_res 0.2566   0.5065   0.00 0.79
##           rubricrsrch_q    0.1740   0.4171   0.38 0.50 0.74
##           rubricsel_meth   0.0962   0.3102   0.56 0.37 0.41 0.26
##           rubrictxt_org    0.4042   0.6358   0.03 0.69 0.80 0.64 0.24
##           rubricvis_org    0.3188   0.5646   0.17 0.78 0.76 0.60 0.29 0.79
##  Residual                  0.1948   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
```

```
display(lmer.tall.base)
```

```
## lmer(formula = rating ~ 1 + (0 + rubric | artifact), data = full_tall)
## coef.est  coef.se
##     2.23     0.04
##
```

```
## Error terms:
##  Groups   Name              Std.Dev. Corr
##  artifact rubriccrit_des    0.80
##           rubricinit_eda    0.62     0.26
##           rubricinterp_res  0.51     0.00 0.79
##           rubricrsrch_q     0.42     0.38 0.50 0.74
##           rubricsel_meth    0.31     0.56 0.37 0.41 0.26
##           rubrictxt_org     0.64     0.03 0.69 0.80 0.64 0.24
##           rubricvis_org     0.56     0.17 0.78 0.76 0.60 0.29 0.79
##  Residual                   0.44
## ---
## number of obs: 810, groups: artifact, 90
## AIC = 1531.7, DIC = 1462.5
## deviance = 1467.1
```

```
# all fixed effects
lmer.tall.FE <- lmer(rating ~ rater + sex + semester + repeated + rubric + (0 + rubric| artifact), data
```

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

```
summary(lmer.tall.FE)
```

```
## Linear mixed model fit by REML [’lmerMod’]
## Formula: rating ~ rater + sex + semester + repeated + rubric + (0 + rubric |
##     artifact)
##    Data: full_tall
##
## REML criterion at convergence: 1429.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1133 -0.5129 -0.0218  0.5320  3.7567
##
## Random effects:
##  Groups   Name              Variance Std.Dev. Corr
##  artifact rubriccrit_des    0.55199  0.7430
##           rubricinit_eda    0.35049  0.5920   0.47
##           rubricinterp_res  0.17313  0.4161   0.24 0.76
##           rubricrsrch_q     0.16845  0.4104   0.59 0.44 0.72
##           rubricsel_meth    0.06723  0.2593   0.39 0.61 0.75 0.42
##           rubrictxt_org     0.25973  0.5096   0.34 0.62 0.74 0.55 0.67
##           rubricvis_org     0.25592  0.5059   0.35 0.74 0.68 0.53 0.42 0.77
##  Residual                   0.19093  0.4370
## Number of obs: 810, groups:  artifact, 90
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)       2.013503   0.109172  18.443
## rater2            0.001915   0.055082   0.035
## rater3           -0.175284   0.055241  -3.173
## sexM              0.010212   0.081319   0.126
## semesterSpring   -0.174931   0.087901  -1.990
## repeated1        -0.072055   0.098560  -0.731
## rubricinit_eda    0.547891   0.095695   5.725
## rubricinterp_res  0.586268   0.100821   5.815
## rubricrsrch_q     0.461887   0.087567   5.275
## rubricsel_meth    0.165503   0.094304   1.755
## rubrictxt_org     0.693415   0.099206   6.990
## rubricvis_org     0.529839   0.099161   5.343
##
## Correlation of Fixed Effects:
##             (Intr) rater2 rater3 sexM   smstrS reptd1 rbrcnt_d rbrcntr_ rbrcr_
## rater2      -0.245
## rater3      -0.238  0.499
## sexM        -0.398 -0.026 -0.035
## semstrSprng -0.361  0.008  0.000  0.302
## repeated1   -0.154  0.001 -0.003  0.009  0.079
## rubricint_d -0.552 -0.001  0.000  0.000 -0.001  0.008
## rbrcntrp_rs -0.660 -0.001  0.000  0.000 -0.001 -0.010  0.736
## rbrcrsrch_q -0.627 -0.001  0.000  0.000 -0.001 -0.038  0.585    0.757
## rubrcsl_mth -0.689 -0.001  0.000  0.000 -0.001 -0.088  0.659    0.776    0.690
## rubrctxt_rg -0.611 -0.001  0.000  0.000 -0.001  0.004  0.672    0.759    0.677
## rubricvs_rg -0.607 -0.001 -0.001 -0.001 -0.002 -0.021  0.718    0.743    0.671
##             rbrcs_ rbrct_
## rater2
## rater3
## sexM
## semstrSprng
## repeated1
## rubricint_d
## rbrcntrp_rs
## rbrcrsrch_q
## rubrcsl_mth
## rubrctxt_rg  0.723
## rubricvs_rg  0.681  0.754
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
display(lmer.tall.FE) # 1354.2 - DIC
```

```
## lmer(formula = rating ~ rater + sex + semester + repeated + rubric +
##     (0 + rubric | artifact), data = full_tall)
##                  coef.est coef.se
## (Intercept)       2.01     0.11
## rater2            0.00     0.06
## rater3           -0.18     0.06
## sexM              0.01     0.08
## semesterSpring   -0.17     0.09
## repeated1        -0.07     0.10
## rubricinit_eda    0.55     0.10
## rubricinterp_res  0.59     0.10
## rubricrsrch_q     0.46     0.09
## rubricsel_meth    0.17     0.09
## rubrictxt_org     0.69     0.10
## rubricvis_org     0.53     0.10
##
## Error terms:
```

```
##  Groups   Name             Std.Dev. Corr
##  artifact rubriccrit_des   0.74
##           rubricinit_eda   0.59     0.47
##           rubricinterp_res 0.42     0.24 0.76
##           rubricrsrch_q    0.41     0.59 0.44 0.72
##           rubricsel_meth   0.26     0.39 0.61 0.75 0.42
##           rubrictxt_org    0.51     0.34 0.62 0.74 0.55 0.67
##           rubricvis_org    0.51     0.35 0.74 0.68 0.53 0.42 0.77
##  Residual                  0.44
## ---
## number of obs: 810, groups: artifact, 90
## AIC = 1511.9, DIC = 1341.7
## deviance = 1385.8
```

```
# finding select fixed effects
lmer.tall.FE.back_select <- fitLMER.fnc(lmer.tall.FE, log.file.name = F)
```

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

```
formula(lmer.tall.FE.back_select) # rater, semester, rubric selected in model
```

```
## rating ~ rater + semester + rubric + (0 + rubric | artifact)
```

```
summary(lmer.tall.FE.back_select)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rating ~ rater + semester + rubric + (0 + rubric | artifact)
##    Data: full_tall
##
## REML criterion at convergence: 1424.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1200 -0.5126 -0.0173  0.5301  3.7752
##
## Random effects:
##  Groups   Name             Variance Std.Dev. Corr
##  artifact rubriccrit_des   0.55495  0.7450
##           rubricinit_eda   0.35066  0.5922   0.47
##           rubricinterp_res 0.16892  0.4110   0.23 0.75
##           rubricrsrch_q    0.16778  0.4096   0.59 0.44 0.70
##           rubricsel_meth   0.06498  0.2549   0.40 0.60 0.74 0.40
##           rubrictxt_org    0.25615  0.5061   0.33 0.61 0.69 0.55 0.66
##           rubricvis_org    0.25897  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)       2.0084164  0.0987614  20.336
## rater2            0.0003198  0.0547448   0.006
## rater3           -0.1771064  0.0548894  -3.227
## semesterSpring   -0.1730415  0.0826942  -2.093
## rubricinit_eda    0.5474729  0.0957142   5.720
## rubricinterp_res  0.5864552  0.1008605   5.815
## rubricrsrch_q     0.4584065  0.0874176   5.244
## rubricsel_meth    0.1590769  0.0937768   1.696
## rubrictxt_org     0.6930041  0.0995457   6.962
## rubricvis_org     0.5289038  0.0990947   5.337
##
## Correlation of Fixed Effects:
##             (Intr) rater2 rater3 smstrS rbrcnt_d rbrcntr_ rbrcr_ rbrcs_ rbrct_
## rater2      -0.281
## rater3      -0.277  0.499
## semstrSprng -0.264  0.017  0.011
## rubricint_d -0.610 -0.001  0.000 -0.002
## rbrcntrp_rs -0.735 -0.001  0.000  0.000  0.734
## rbrcrsrch_q -0.701 -0.001  0.000  0.002  0.586    0.756
## rubrcsl_mth -0.782  0.000  0.000  0.006  0.662    0.779    0.688
## rubrctxt_rg -0.679 -0.001  0.000 -0.001  0.674    0.751    0.682  0.728
## rubricvs_rg -0.675 -0.001 -0.001  0.000  0.715    0.745    0.667  0.681  0.750
```

```
display(lmer.tall.FE.back_select)
```

```
## lmer(formula = rating ~ rater + semester + rubric + (0 + rubric |
##     artifact), data = full_tall, REML = TRUE)
##                  coef.est coef.se
## (Intercept)       2.01     0.10
## rater2            0.00     0.05
## rater3           -0.18     0.05
## semesterSpring   -0.17     0.08
## rubricinit_eda    0.55     0.10
## rubricinterp_res  0.59     0.10
## rubricrsrch_q     0.46     0.09
## rubricsel_meth    0.16     0.09
## rubrictxt_org     0.69     0.10
## rubricvis_org     0.53     0.10
##
## Error terms:
##  Groups   Name             Std.Dev. Corr
##  artifact rubriccrit_des   0.74
##           rubricinit_eda   0.59     0.47
##           rubricinterp_res 0.41     0.23 0.75
##           rubricrsrch_q    0.41     0.59 0.44 0.70
##           rubricsel_meth   0.25     0.40 0.60 0.74 0.40
##           rubrictxt_org    0.51     0.33 0.61 0.69 0.55 0.66
##           rubricvis_org    0.51     0.35 0.73 0.68 0.52 0.41 0.75
##  Residual                  0.44
## ---
## number of obs: 810, groups: artifact, 90
## AIC = 1502.1, DIC = 1348
## deviance = 1386.0
```

```
# testing terms with backward selection
```

```r
lmer.tall.FE.interact <- lmer(rating ~ rater * semester * rubric + (0 + rubric| artifact), data = full_
lmer.tall.FE.interact.back_select <- fitLMER.fnc(lmer.tall.FE.interact, log.file.name = F)
```

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

```r
formula(lmer.tall.FE.interact.back_select) # rater rubric interaction only
```

```
## rating ~ rater + semester + rubric + (0 + rubric | artifact) +
##     rater:rubric
```

```r
summary(lmer.tall.FE.interact.back_select)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rating ~ rater + semester + rubric + (0 + rubric | artifact) +
##     rater:rubric
##    Data: full_tall
##
## REML criterion at convergence: 1419.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9280 -0.5123 -0.0446  0.4827  3.5856
##
## Random effects:
##  Groups   Name              Variance Std.Dev. Corr
##  artifact rubriccrit_des    0.50354  0.7096
##           rubricinit_eda    0.35486  0.5957   0.44
##           rubricinterp_res  0.15193  0.3898   0.35 0.82
##           rubricrsrch_q     0.17956  0.4237   0.63 0.44 0.72
##           rubricsel_meth    0.06728  0.2594   0.42 0.60 0.74 0.36
##           rubrictxt_org     0.26072  0.5106   0.42 0.64 0.67 0.55 0.64
##           rubricvis_org     0.25495  0.5049   0.34 0.71 0.68 0.51 0.38 0.77
##  Residual                   0.18517  0.4303
## Number of obs: 810, groups:  artifact, 90
##
## Fixed effects:
##                           Estimate Std. Error t value
## (Intercept)                1.75947    0.11785  14.929
## rater2                     0.36535    0.13295   2.748
## rater3                     0.21418    0.13297   1.611
## semesterSpring            -0.17780    0.08228  -2.161
## rubricinit_eda             0.74624    0.13676   5.456
## rubricinterp_res           1.01451    0.13479   7.527
## rubricrsrch_q              0.74924    0.12419   6.033
## rubricsel_meth             0.42669    0.13040   3.272
## rubrictxt_org              1.04964    0.13552   7.746
## rubricvis_org              0.68350    0.13948   4.900
## rater2:rubricinit_eda     -0.30842    0.17249  -1.788
## rater3:rubricinit_eda     -0.29521    0.17282  -1.708
## rater2:rubricinterp_res -0.53670    0.17008  -3.156
## rater3:rubricinterp_res -0.75243    0.17049  -4.413
## rater2:rubricrsrch_q    -0.50154    0.16150  -3.105
## rater3:rubricrsrch_q    -0.37064    0.16179  -2.291
## rater2:rubricsel_meth   -0.39599    0.16467  -2.405
## rater3:rubricsel_meth   -0.41321    0.16503  -2.504
## rater2:rubrictxt_org    -0.58377    0.17141  -3.406
## rater3:rubrictxt_org    -0.48644    0.17177  -2.832
## rater2:rubricvis_org    -0.14440    0.17442  -0.828
## rater3:rubricvis_org    -0.33374    0.17481  -1.909
```

```
##
## Correlation matrix not shown by default, as p = 22 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```
display(lmer.tall.FE.interact.back_select)
```

```
## lmer(formula = rating ~ rater + semester + rubric + (0 + rubric |
##     artifact) + rater:rubric, data = full_tall, REML = TRUE)
##                         coef.est coef.se
## (Intercept)              1.76     0.12
## rater2                   0.37     0.13
## rater3                   0.21     0.13
## semesterSpring          -0.18     0.08
## rubricinit_eda           0.75     0.14
## rubricinterp_res         1.01     0.13
## rubricrsrch_q            0.75     0.12
## rubricsel_meth           0.43     0.13
## rubrictxt_org            1.05     0.14
## rubricvis_org            0.68     0.14
## rater2:rubricinit_eda   -0.31     0.17
## rater3:rubricinit_eda   -0.30     0.17
## rater2:rubricinterp_res -0.54     0.17
## rater3:rubricinterp_res -0.75     0.17
## rater2:rubricrsrch_q    -0.50     0.16
## rater3:rubricrsrch_q    -0.37     0.16
## rater2:rubricsel_meth   -0.40     0.16
## rater3:rubricsel_meth   -0.41     0.17
## rater2:rubrictxt_org    -0.58     0.17
## rater3:rubrictxt_org    -0.49     0.17
## rater2:rubricvis_org    -0.14     0.17
## rater3:rubricvis_org    -0.33     0.17
##
## Error terms:
##  Groups   Name             Std.Dev. Corr
##  artifact rubriccrit_des   0.71
##           rubricinit_eda   0.60     0.44
##           rubricinterp_res 0.39     0.35 0.82
##           rubricrsrch_q    0.42     0.63 0.44 0.72
##           rubricsel_meth   0.26     0.42 0.60 0.74 0.36
##           rubrictxt_org    0.51     0.42 0.64 0.67 0.55 0.64
##           rubricvis_org    0.50     0.34 0.71 0.68 0.51 0.38 0.77
##  Residual                  0.43
## ---
## number of obs: 810, groups: artifact, 90
## AIC = 1521.6, DIC = 1285.4
## deviance = 1352.5
```

```
# nested model, use AIX and LRT to compare fixed effects
anova(lmer.tall.FE.back_select, lmer.tall.FE.interact.back_select, lmer.tall.FE.interact)
```

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

```
## Data: full_tall
## Models:
## lmer.tall.FE.back_select: rating ~ rater + semester + rubric + (0 + rubric | artifact)
## lmer.tall.FE.interact.back_select: rating ~ rater + semester + rubric + (0 + rubric | artifact) + ra
## lmer.tall.FE.interact: rating ~ rater * semester * rubric + (0 + rubric | artifact)
##                                   npar    AIC    BIC  logLik deviance  Chisq Df
## lmer.tall.FE.back_select            39 1464.0 1647.2 -693.02   1386.0
## lmer.tall.FE.interact.back_select   51 1454.5 1694.1 -676.26   1352.5 33.526 12
## lmer.tall.FE.interact               71 1471.4 1804.8 -664.68   1329.4 23.161 20
##                                   Pr(>Chisq)
## lmer.tall.FE.back_select
## lmer.tall.FE.interact.back_select   0.000801 ***
## lmer.tall.FE.interact               0.280962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# AIC and LRT prefer the model with rater, semester, and rubric and the interaction between rater and r
```

```
# random effects - rater, semester, rubric, and rater:rubric
lmer.tall.FRE.base <- lmer(rating ~ 1 + rater + semester + rubric + rater:rubric + (0 + rubric | artifa

lmer.tall.FRE.RE_select <- fitLMER.fnc(lmer.tall.FRE.base, ran.effects = c("(0 + rater|artifact)", "(0 +
```

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

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

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

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

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

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

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

```
## refitting model(s) with ML (instead of REML)
## refitting model(s) with ML (instead of REML)
## boundary (singular) fit: see ?isSingular
```

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

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

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

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

```
formula(lmer.tall.FRE.RE_select)
```

```
anova(lmer.tall.FRE.base, lmer.tall.FRE.RE_select)
```

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

```
## Data: full_tall
## Models:
## lmer.tall.FRE.base: rating ~ 1 + rater + semester + rubric + rater:rubric + (0 + rubric | artifact)
## lmer.tall.FRE.RE_select: rating ~ rater + semester + rubric + (0 + rubric | artifact) + (0 + rater |
##                          npar    AIC    BIC  logLik deviance  Chisq Df
## lmer.tall.FRE.base        51 1454.5 1694.1 -676.26   1352.5
## lmer.tall.FRE.RE_select   57 1415.9 1683.6 -650.94   1301.9 50.647  6
##                          Pr(>Chisq)
## lmer.tall.FRE.base
## lmer.tall.FRE.RE_select  3.487e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
final_model <- lmer(rating ~ rater + semester + rubric + (0 + rubric | artifact) + (0 + rater | artifac
```

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

```
summary(final_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rating ~ rater + semester + rubric + (0 + rubric | artifact) +
##     (0 + rater | artifact) + rater:rubric
##    Data: full_tall
##
## REML criterion at convergence: 1370.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.06406 -0.46877 -0.02989  0.45338  2.74004
##
## Random effects:
##  Groups     Name              Variance Std.Dev. Corr
##  artifact   rubriccrit_des   0.49649  0.7046
##             rubricinit_eda   0.31803  0.5639    0.32
##             rubricinterp_res 0.10227  0.3198    0.14  0.67
##             rubricrsrch_q    0.17904  0.4231    0.50  0.19  0.54
##             rubricsel_meth   0.03827  0.1956    0.15  0.23  0.38 -0.24
##             rubrictxt_org    0.25046  0.5005    0.27  0.44  0.37  0.31  0.22
##             rubricvis_org    0.23258  0.4823    0.18  0.50  0.45  0.28 -0.16
##  artifact.1 rater1           0.01273  0.1128
##             rater2           0.11165  0.3341   -0.49
##             rater3           0.09386  0.3064    0.33  0.66
##  Residual                    0.13470  0.3670
##
##
##
##
##
##
##
##   0.54
##
##
##
##
## Number of obs: 810, groups:  artifact, 90
##
## Fixed effects:
##                         Estimate Std. Error t value
## (Intercept)              1.75760    0.11404  15.412
## rater2                   0.36606    0.13919   2.630
## rater3                   0.19578    0.12969   1.510
## semesterSpring          -0.15926    0.07649  -2.082
## rubricinit_eda           0.73947    0.12997   5.690
## rubricinterp_res         0.99148    0.12772   7.763
## rubricrsrch_q            0.72616    0.11793   6.157
## rubricsel_meth           0.41066    0.12469   3.294
## rubrictxt_org            1.01577    0.13000   7.814
## rubricvis_org            0.65422    0.13355   4.899
## rater2:rubricinit_eda   -0.29979    0.15610  -1.921
## rater3:rubricinit_eda   -0.29467    0.15636  -1.885
## rater2:rubricinterp_res -0.51321    0.15350  -3.343
## rater3:rubricinterp_res -0.71475    0.15365  -4.652
## rater2:rubricrsrch_q    -0.48738    0.14723  -3.310
## rater3:rubricrsrch_q    -0.32232    0.14727  -2.189
## rater2:rubricsel_meth   -0.38637    0.15030  -2.571
## rater3:rubricsel_meth   -0.38709    0.14961  -2.587
## rater2:rubrictxt_org    -0.55104    0.15647  -3.522
## rater3:rubrictxt_org    -0.44485    0.15674  -2.838
## rater2:rubricvis_org    -0.10488    0.15863  -0.661
## rater3:rubricvis_org    -0.27513    0.15887  -1.732

##
## Correlation matrix not shown by default, as p = 22 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
ranef(final_model)
```

```
## $artifact
##     rubriccrit_des rubricinit_eda rubricinterp_res rubricrsrch_q rubricsel_meth
## 100    0.799464383   -0.260810599     -0.121684265  -0.164864784    0.232126321
## 101   -0.496691626    0.434086698     -0.166401273  -0.741288676    0.032441783
## 102   -0.771237818   -0.333262489     -0.233116452  -0.744530820    0.047929182
## 103    0.139989959    0.330443135      0.099000304  -0.281345777    0.271545122
## 104   -0.576823199    0.308283146      0.091022223  -0.260973068    0.073242884
## 105   -0.590512134   -0.483159965     -0.340645792  -0.404649334   -0.098081756
## 106    0.204178302   -1.029067428     -0.399379867   0.232984118   -0.136512653
## 107   -0.560085952   -0.401711314      0.057094916   0.182929466   -0.102733468
## 111   -0.461817534   -0.351282011     -0.241819423  -0.311901427   -0.067370983
## 112   -0.499229346    0.322764837      0.271924225   0.198103744   -0.103545672
## 113   -0.586425027   -0.804706415     -0.145439290  -0.131205918    0.003838185
## 114   -0.448128600    0.440161101      0.189848593  -0.168225160    0.103953656
## 115   -0.370534747    0.454642792      0.370750595   0.290851652   -0.072834900
## 116   -0.588958505   -0.354449427     -0.287765900  -0.351796820   -0.123658050
## 117   -0.672399864   -0.136733969      0.012054493  -0.194264263   -0.103848377
## 118   -0.654263020   -0.212108218     -0.029436351  -0.208779547   -0.027684792
## 13     0.387047205   -0.747709636     -0.435964859   0.049414762   -0.159619356
## 15     0.688897158    0.476671954     -0.046553262  -0.109715276   -0.051288674
## 16     0.683088499    1.154175464      0.314723903  -0.007676762    0.187218734
## 17     0.335446679   -0.088202470      0.066883078   0.549194379   -0.199179940
## 21     0.776826848    1.091241277      0.452605111   0.435448916    0.086118324
## 22     0.730914444    0.382078982      0.249251661   0.431044657    0.090496707
## 23    -0.272962180   -0.724617884     -0.328050481  -0.011022841   -0.177242542
## 24     0.049030213   -0.202510791     -0.150417818  -0.130392114    0.068953419
## 25     1.159767737    0.006811965     -0.289230105   0.162397289   -0.130434173
## 26    -0.863035504   -0.483826044     -0.160057407  -0.518376670    0.112416592
## 27     0.101619248    0.386374805      0.006375870  -0.172908969    0.092115341
## 28    -0.299901788   -0.593232640     -0.414636008  -0.406007855   -0.060201750
## 32     0.661149335    0.404547903      0.251096443   0.408352370    0.002012916
## 33    -0.083638342    0.150418859     -0.059882416  -0.452315257    0.169142081
## 34     0.466128210   -0.311377317     -0.060385840  -0.200829790    0.261402316
## 35    -0.743832758   -0.254757895     -0.085373638  -0.283135406   -0.098097042
## 36     0.049030213   -0.202510791     -0.150417818  -0.130392114    0.068953419
## 37     0.775838339   -0.157151486     -0.207036514  -0.021813846    0.102516985
## 38    -0.017024632   -0.209398589     -0.141992334  -0.174557138   -0.064533476
## 39    -0.527740266    0.248724593      0.207456475   0.140165705   -0.087551740
## 40     0.105329512    0.357018086      0.012956571  -0.194381705    0.047112237
## 45     0.014893346   -0.241931920     -0.172432874  -0.090013494    0.049720147
## 46     0.149371245    0.324005858     -0.013804002  -0.141415730    0.031795160
## 47     1.130303787   -0.178372788     -0.049747532   0.676307375   -0.149257489
## 48     0.537344818    0.660775279      0.224976376   0.134750392    0.242102471
## 49    -0.903610201    0.215753152      0.273822911   0.159908914   -0.189299089
## 53     1.142084271    0.106851683      0.017359530   0.239212303    0.118199952
## 54    -0.662633365    0.383492653     -0.092238635  -0.662817048    0.147258870
## 55    -0.148713830   -0.372218727      0.070676891   0.317413874   -0.133698146
## 56     0.687563440   -0.264952334     -0.276254951  -0.060766587   -0.062099233
## 57    -0.769749955   -0.326300780     -0.169588183  -0.256440929   -0.084220130
## 6     -0.674067650   -0.277226816     -0.087218420  -0.260443118   -0.009613252
## 61     0.001626965    0.332884253     -0.079303585  -0.172012905    0.016479960
## 62     1.385508809    0.998210074      0.303844775   0.483561166   -0.052057824
## 63     0.683091785    0.348889750      0.207365523   0.445391784   -0.018577778
## 64     0.550905088   -0.162571193     -0.076450542   0.054973902    0.062735144
## 65     0.831844602   -0.452195814     -0.411770546   0.252849981   -0.216807035
## 66     0.741108415    0.910178437      0.372058674   0.382639842   -0.040009148
## 67    -0.816294616    0.144951314      0.292918459   0.146956202   -0.188404167
## 68     0.631193719   -0.239287787      0.051102440   0.488406608   -0.041574813
## 7     -0.621478615    0.311658781      0.069575267  -0.302959973    0.013548670
## 72    -0.056247003    0.416522291      0.192195060  -0.071933884    0.022449401
## 73    -0.746569562   -0.931502457     -0.323615852  -0.186904581   -0.017465433
## 74    -1.048449349    0.086246930      0.117285902  -0.493524825    0.092677829
## 75    -0.041597240    0.337574706      0.148026439  -0.088962245    0.097780854
## 76     0.037044483   -0.326095122     -0.216471120  -0.141950792   -0.005622153
## 77    -0.168573765   -0.233578689     -0.010472057  -0.072468946   -0.014967912
## 78     0.416486855    0.062733035      0.074907947   0.131506573    0.084925301
## 79    -0.309110561    0.018614388      0.132456365   0.023800843    0.051650650
## 8     -0.607268909   -0.208775759     -0.035922739  -0.212302454    0.006327129
## 84     0.308734520   -0.083708403      0.161216312   0.445586346   -0.061228118
## 85     1.073645839    0.376701115      0.316165651   0.876778188   -0.131428898
## 86     0.326871364   -0.159082652      0.119725468   0.431071062    0.014935468
## 87     0.204178302   -1.029067428     -0.399379867   0.232984118   -0.136512653
## 88     1.088472897    0.644845876      0.317071705   0.539479852   -0.076256740
## 9     -0.580329301   -0.340161004      0.050662787   0.182682560   -0.110713662
## 92    -0.540395479   -0.348375024      0.068483965   0.221266697   -0.052146177
## 93    -0.392010408   -0.163160780      0.178699384   0.352351835    0.029151886
## 94     1.037218119    1.033359285      0.338703908   0.394538424   -0.006039837
## 95     0.139989959    0.330443135      0.099000304  -0.281345777    0.271545122
## 96     0.239342898    0.380179255      0.224339358   0.315160202   -0.067174800
## O1    -0.502794653    0.421474815      0.119764750  -0.009438296   -0.093114416
## O10   -0.441543760   -0.001207167      0.155627797   0.049686298    0.036718423
## O11   -0.766657485   -0.329653155      0.329077524   0.512786310    0.013873791
## O12   -0.354549119   -0.488477771     -0.316414896  -0.338569514   -0.015853078
## O13    0.030482003    0.083038052     -0.317130190  -0.367506368   -0.071080851
## O2    -0.116535998    0.329293988      0.170876622   0.139885178   -0.072108545
## O3     0.283442934   -0.135267817     -0.013001348   0.231073794   -0.039589825
## O4    -0.111113489    0.045151799      0.203651056  -0.216505973    0.365204381
## O5     0.879248427   -0.425930062     -0.025705935   0.189907727    0.250017015
## O6    -0.898045999   -0.773371164     -0.419606506  -0.413046553   -0.112426726
## O7     0.138742232    0.207581768     -0.004953164  -0.012980708   -0.042234039
## O8     0.358132341   -0.210288630     -0.397298845  -0.312004289    0.028122350
## O9    -0.799036596    0.585491352      0.349785290  -0.190139081    0.075245134
##      rubrictxt_org rubricvis_org       rater1       rater2       rater3
## 100   -0.48923747   -0.44079028  0.034722957 -0.050525297  0.030862590
## 101    0.28908109    0.46497453 -0.030844963  0.044882437 -0.027415737
## 102   -1.11989894   -0.43500710 -0.071191988  0.103591302 -0.063277132
## 103    0.10956980   -0.23946914  0.041474214 -0.060349036  0.036863268
## 104    0.08871875   -0.11685388 -0.025454345  0.037038560 -0.022624427
## 105   -0.53271781   -0.35664589 -0.058834954  0.085610610 -0.052293907
## 106    0.01694998   -0.33500558 -0.022470412  0.032696647 -0.019972237
## 107   -0.51333369   -0.30987005 -0.010990945  0.015992899 -0.009769013
## 111   -0.41135830   -0.25307948 -0.035688428  0.051930152 -0.031720724
## 112    0.20846878    0.41628104  0.019577581 -0.028487295  0.017401020
## 113   -0.47291374   -0.30657109 -0.016080220  0.023398293 -0.014292482
## 114    0.21007826   -0.01328746 -0.002307819  0.003358102 -0.002051244
## 115    0.32982829    0.51984746  0.042724107 -0.062167753  0.037974203
## 116    0.12950820    0.27733482 -0.030777169  0.044783789 -0.027355480
## 117    0.22591616    0.84105142  0.010648644 -0.015494819  0.009464768
## 118    0.12752922    0.31601358 -0.008658220  0.012598557 -0.007695632
## 13    -0.72214187   -0.62755379 -0.064522414 -0.386699969 -0.535399928
## 15     0.41131966    0.90111079  0.013671212  0.081935206  0.113442221
## 16     0.89887986    0.58557882  0.020164031  0.120848394  0.167318921
## 17    -0.11569481    0.03038842 -0.017668885 -0.105894324 -0.146614477
## 21     0.91836416    0.59207258  0.042710580  0.255975850  0.354407713
## 22     0.21065412   -0.12205223  0.017929534  0.107456461  0.148777310
## 23    -0.67209888   -0.56110140 -0.055873519 -0.334864842 -0.463632342
## 24     0.24572341   -0.13994059 -0.007330846 -0.043935706 -0.060830555
## 25    -0.00148335    0.07387461 -0.037711784 -0.226016738 -0.312928252
## 26    -0.46986673   -0.47130194  0.015449812  0.092594825  0.128200844
## 27     0.29896861   -0.05317324 -0.006172534 -0.036993636 -0.051219012
## 28    -0.62859171   -0.51383916 -0.067741587 -0.405993329 -0.562112275
## 32     0.26035540    0.36143798  0.026836615  0.160838961  0.222687290
## 33    -0.40367304   -0.39723098  0.018637708  0.111700736  0.154653661
## 34    -0.50172555   -0.50561510  0.029713396  0.178080276  0.246558506
## 35    -0.25929587    0.26613478 -0.004490523 -0.026912898 -0.037261869
## 36     0.24572341   -0.13994059 -0.007330846 -0.043935706 -0.060830555
## 37     0.25867190   -0.15263700 -0.005301261 -0.031771864 -0.043989282
## 38    -0.24634739    0.25343838 -0.002460938 -0.014749057 -0.020420596
## 39    -0.23626766   -0.12409432  0.010307257  0.061774127  0.085528487
## 40    -0.24280348   -0.14328448 -0.010209707 -0.061189487 -0.084719032
## 45     0.29925014   -0.21588992  0.014045385 -0.084678863 -0.051418651
## 46    -0.18641314   -0.21929237  0.017992203 -0.108474009 -0.065867527
## 47    -0.68447799   -0.67172659  0.061018341 -0.367876258 -0.223381613
## 48     0.60922828   -0.35360866 -0.058284085  0.351391576  0.213371793
## 49     0.25999875    0.72557181 -0.028237942  0.170245017  0.103376083
## 53     0.07363117   -0.06244953 -0.066786703  0.402653396  0.244498967
## 54     0.33447997   -0.11311864  0.048410150 -0.291862158 -0.177224374
## 55    -0.36497991    0.05830230 -0.010358007  0.062447860  0.037919554
## 56    -0.23949369    0.11141020  0.019325279 -0.116511052 -0.070747775
## 57     0.27639380    0.22692481  0.019357855 -0.116707456 -0.070867035
## 6     -0.30899716   -0.21735543 -0.013397604 -0.080295398 -0.111171849
## 61     0.88029523    0.38763857  0.008657616 -0.052196296 -0.031694605
## 62     0.40499719    0.81940070 -0.054785456  0.330298531  0.200563687
## 63     0.29590603    0.26764579 -0.036872944  0.222304970  0.134987898
## 64     0.23651126    0.18576433 -0.001766338  0.010649153  0.006466373
## 65     0.31332012    0.16044284  0.010882424 -0.065609542 -0.039839389
## 66    -0.19100141    0.26568392 -0.032609385  0.196600209  0.119379467
## 67    -0.86369412    0.07001114 -0.003088658  0.018621353  0.011307247
## 68     0.24331242    0.18143448 -0.035157498  0.211962644  0.128707836
## 7     -0.25575196   -0.13058808 -0.012239292 -0.073353329 -0.101560306
## 72    -0.20521847    0.24610291 -0.010311687  0.062168597  0.037749980
## 73     0.17910941   -0.33850082  0.034264577 -0.206579271 -0.125438948
## 74    -0.45110688   -0.05665098 -0.034213865  0.206273531  0.125253297
## 75    -0.30689373   -0.28174114  0.018694288 -0.112706843 -0.068437785
## 76    -0.29599125   -0.35413340  0.035522345 -0.214162287 -0.130043502
## 77    -0.31479657    0.11126187  0.007218456 -0.043519680 -0.026425996
## 78     0.06182439   -0.04877881 -0.063780913  0.384531650  0.233495091
## 79      0.05001761   -0.03510810 -0.060775122  0.366409905  0.222491215
## 8      -0.24600568   -0.16359944 -0.002719124 -0.016296430 -0.022562990
## 84      0.29429391    0.42444781  0.044612667 -0.064915792  0.039652800
## 85      0.27791606    0.41791957  0.054101243 -0.078722597  0.048086472
## 86      0.19590698   -0.10059003  0.025305802 -0.036822416  0.022492399
## 87      0.01694998   -0.33500558 -0.022470412  0.032696647 -0.019972237
## 88      0.47629942    1.03842674  0.070829572 -0.103063952  0.062955009
## 9      -0.28951286   -0.21086168  0.009148945  0.054832058  0.075916944
## 92      0.05050537   -0.20092717 -0.002240024  0.003259454 -0.001990987
## 93      0.73570395    0.01158213  0.029657423 -0.043154450  0.026360223
## 94      0.31612944    0.50199286  0.030886922 -0.044943491  0.027453031
## 95      0.10956980   -0.23946914  0.041474214 -0.060349036  0.036863268
## 96      0.23251087    0.41305176  0.023976881 -0.034888706  0.021311223
## O1     -0.23763337   -0.25939274 -0.211521911  0.272670040 -0.222935464
## O10     0.12151356    0.01148350  0.108485414 -0.367995964 -0.112605994
## O11     0.30612989   -0.26162900  0.052340288  0.052050933  0.174056034
## O12    -0.36298783   -0.45390392  0.058202877 -0.016859879  0.119205231
## O13     0.27443643    0.14773330 -0.008113859  0.048661184  0.029448494
## O2      0.17347861    0.34178366  0.172833886 -1.028138287 -0.618931709
## O3      0.24307394    0.13193438 -0.038551331  0.159624154  0.068716654
## O4      0.16183518   -0.27773511 -0.089809680  0.418185858  0.206162175
## O5     -0.04192466   -0.48829565  0.030346799  0.059508828  0.130092412
## O6     -0.05496236   -0.18355124 -0.049933215  0.116263276 -0.001006322
## O7      0.12376585    0.22491857  0.146812115  0.138219073  0.480478861
## O8     -0.56939125   -0.90980378 -0.038675961 -0.247871470 -0.336920235
## O9      0.39808052    0.55964256  0.047843710  0.034920593  0.146510966
##
## with conditional variances for "artifact"
```

```
summary(final_model)$coef %>%
  kbl(booktabs=T,
      caption = "Final Model Coefficient and Standard Error Estimates") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 8: Final Model Coefficient and Standard Error Estimates

	Estimate	Std. Error	t value
(Intercept)	1.7576024	0.1140449	15.4115007
rater2	0.3660627	0.1391867	2.6300130
rater3	0.1957847	0.1296855	1.5096885
semesterSpring	-0.1592566	0.0764858	-2.0821708
rubricinit_eda	0.7394688	0.1299686	5.6895957
rubricinterp_res	0.9914779	0.1277157	7.7631663
rubricrsrch_q	0.7261611	0.1179342	6.1573391
rubricsel_meth	0.4106646	0.1246892	3.2935066
rubrictxt_org	1.0157681	0.1299990	7.8136637
rubricvis_org	0.6542195	0.1335497	4.8986978
rater2:rubricinit_eda	-0.2997904	0.1560985	-1.9205205
rater3:rubricinit_eda	-0.2946661	0.1563598	-1.8845384
rater2:rubricinterp_res	-0.5132100	0.1534950	-3.3434956
rater3:rubricinterp_res	-0.7147546	0.1536506	-4.6518178
rater2:rubricrsrch_q	-0.4873828	0.1472261	-3.3104367
rater3:rubricrsrch_q	-0.3223193	0.1472702	-2.1886252
rater2:rubricsel_meth	-0.3863702	0.1503008	-2.5706463
rater3:rubricsel_meth	-0.3870864	0.1496078	-2.5873416
rater2:rubrictxt_org	-0.5510412	0.1564659	-3.5217987
rater3:rubrictxt_org	-0.4448526	0.1567369	-2.8382131
rater2:rubricvis_org	-0.1048814	0.1586275	-0.6611805
rater3:rubricvis_org	-0.2751260	0.1588659	-1.7318127

To determine which collection of factor variables are most appropriate for the model in order to identify relationships with ratings, model selection is applied accounting for the various types of interactions between these variables. First, fixed effects for all of the factor variables are considered, and only rater, semester, and rubric are found to be significant. Interactions between these fixed effects are also considered, and it is found that in addition to these fixed effects, the inclusion of an interaction term between rubric and rater is found to be important in predicting rating. This is confirmed using AIC and the likelihood ratio test.

After selecting the appropriate fixed effects, random effects are also considered for inclusion in the model. Random effects are considered for the interaction rater and artifact, the interaction between semester and artifact, as well as random intercepts for rubric, semester, rater and artifact. Using AIC and the likelihood ratio test, the final model selected includes the fixed effects previously identified along with random effects for the interaction between rubric and artifact and the interaction between rater and artifact.

7.B.4 Research Question 4 - Other Interesting Findings in the Data

```r
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  group_by(sex) %>%
  dplyr::summarise(
    n = length(rating),
    min = min(rating, na.rm = T),
    Q1 = quantile(rating, 0.25, na.rm = T),
    median = median(rating, na.rm = T),
    mean = mean(rating, na.rm = T),
    Q3 = quantile(rating, 0.75, na.rm = T),
    max = max(rating, na.rm = T),
    sd = sd(rating, na.rm = T)
  ) %>%
  kbl(booktabs=T, caption = "Summary statistics of ratings by sex") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 9: Summary statistics of ratings by sex

sex	n	min	Q1	median	mean	Q3	max	sd
F	448	1	2	2	2.313901	3	4	0.7000061
M	364	1	2	2	2.310440	3	4	0.7079251

```
ratings %>%
    pivot_longer(
        cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
    ggplot(aes(x = rating)) +
    geom_histogram(bins = 8, position = "dodge") +
    facet_wrap(~ sex) +
    labs(title = "Figure 7: Histogram of ratings by sex")
```

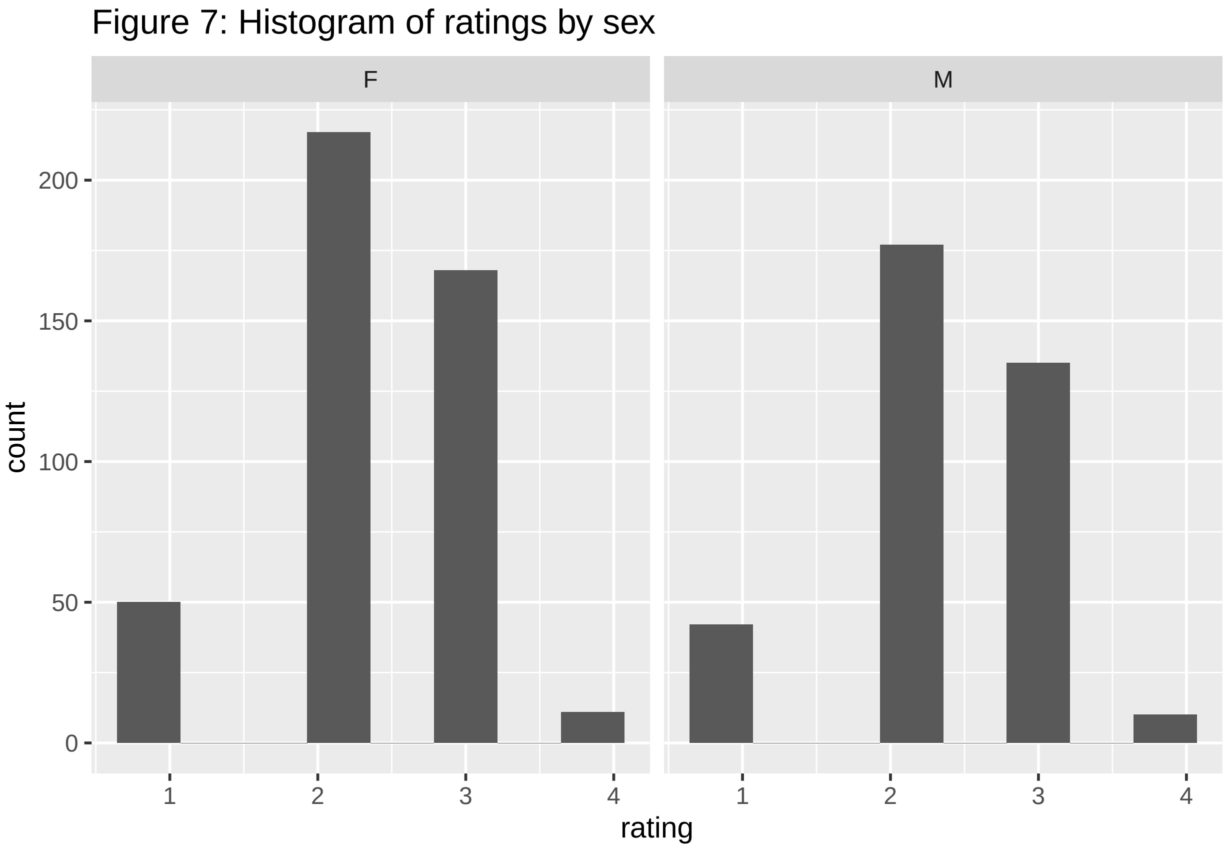


Figure 7: Histogram of ratings by sex

```
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  group_by(sex, rubric) %>%
  dplyr::summarise(
    n = length(rating),
    min = min(rating, na.rm = T),
    Q1 = quantile(rating, 0.25, na.rm = T),
    median = median(rating, na.rm = T),
    mean = mean(rating, na.rm = T),
    Q3 = quantile(rating, 0.75, na.rm = T),
    max = max(rating, na.rm = T),
    sd = sd(rating, na.rm = T)
  ) %>%
  kbl(booktabs=T, caption = "Summary statistics of ratings by sex and rubric") %>%
  kable_classic(latex_options = "HOLD_position")
```

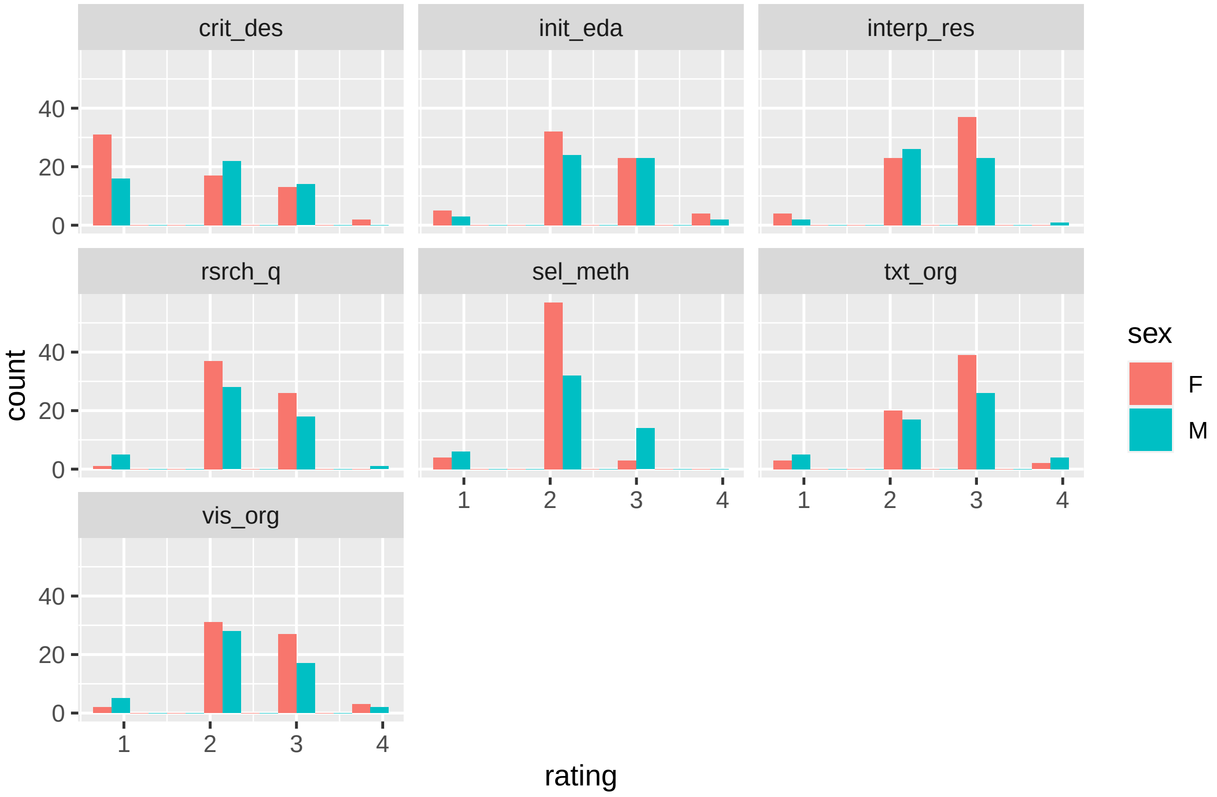
```
## 'summarise()' has grouped output by 'sex'. You can override using the '.groups' argument.
```

Table 10: Summary statistics of ratings by sex and rubric

sex	rubric	n	min	Q1	median	mean	Q3	max	sd
F	crit_des	64	1	1	2	1.777778	2	4	0.8879924
F	init_eda	64	1	2	2	2.406250	3	4	0.7285286
F	interp_res	64	1	2	3	2.515625	3	3	0.6170125
F	rsrch_q	64	1	2	2	2.390625	3	3	0.5230311
F	sel_meth	64	1	2	2	1.984375	2	3	0.3329611
F	txt_org	64	1	2	3	2.625000	3	4	0.6299408
F	vis_org	64	1	2	2	2.492063	3	4	0.6444056
M	crit_des	52	1	1	2	1.961539	3	3	0.7659811
M	init_eda	52	1	2	2	2.461539	3	4	0.6704268
M	interp_res	52	1	2	2	2.442308	3	4	0.6075816
M	rsrch_q	52	1	2	2	2.288461	3	4	0.6667609
M	sel_meth	52	1	2	2	2.153846	3	3	0.6066499
M	txt_org	52	1	2	3	2.557692	3	4	0.7774635
M	vis_org	52	1	2	2	2.307692	3	4	0.7012164

```
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  ggplot(aes(x = rating, fill = sex)) +
  geom_histogram(bins = 8, position = "dodge") +
  facet_wrap(~ rubric) +
  labs(title = "Figure 8: Histogram of ratings by rubric and sex")
```

Figure 8: Histogram of ratings by rubric and sex



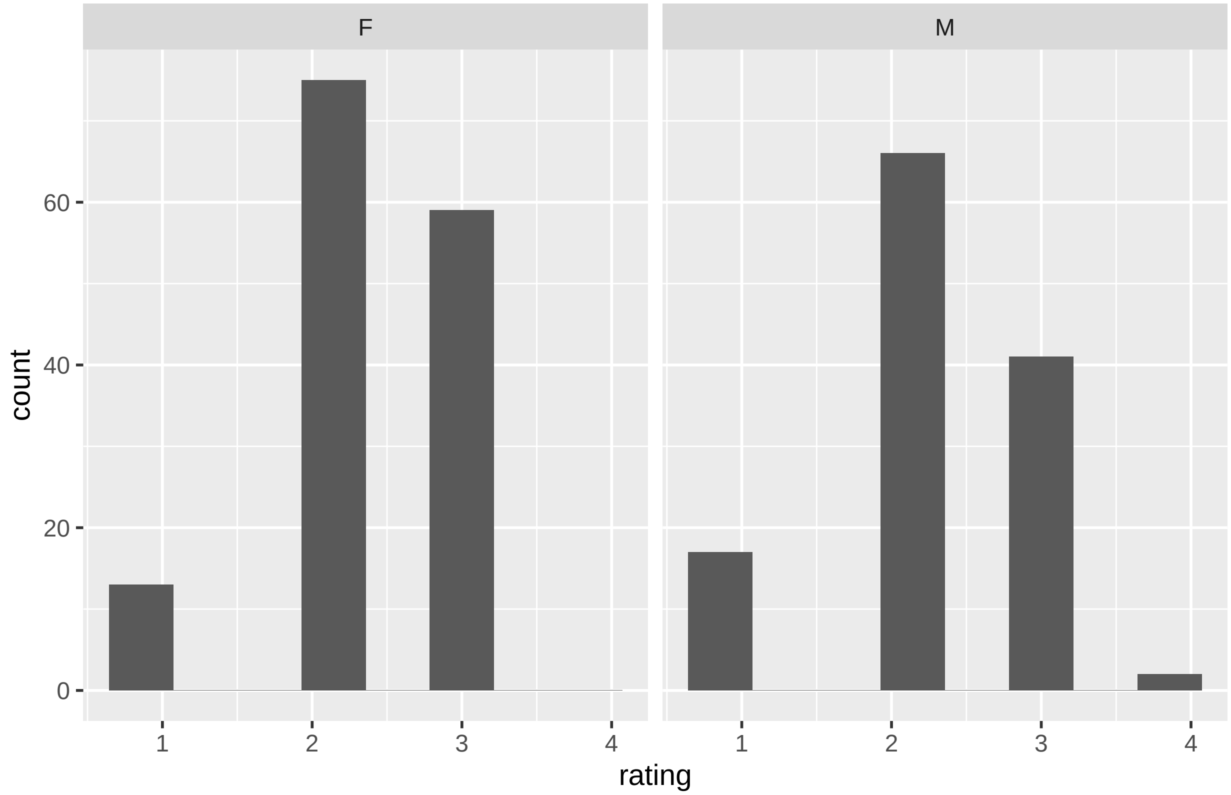
```
sub_ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  group_by(sex) %>%
  dplyr::summarise(
    n = length(rating),
    min = min(rating, na.rm = T),
    Q1 = quantile(rating, 0.25, na.rm = T),
    median = median(rating, na.rm = T),
    mean = mean(rating, na.rm = T),
    Q3 = quantile(rating, 0.75, na.rm = T),
    max = max(rating, na.rm = T),
    sd = sd(rating, na.rm = T)
  ) %>%
  kbl(booktabs=T, caption = "Summary statistics of ratings by sex for subsetted data") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 11: Summary statistics of ratings by sex for subsetted data

sex	n	min	Q1	median	mean	Q3	max	sd
F	147	1	2	2	2.312925	3	3	0.6281384
M	126	1	2	2	2.222222	3	4	0.6915361

```
sub_ratings %>%
    pivot_longer(
        cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
    ggplot(aes(x = rating)) +
    geom_histogram(bins = 8, position = "dodge") +
    facet_wrap(~ sex) +
    labs(title = "Figure 9: Histogram of ratings by sex for subsetted data")
```

Figure 9: Histogram of ratings by sex for subsetted data



```
sub_ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  group_by(sex, rubric) %>%
  dplyr::summarise(
    n = length(rating),
    min = min(rating, na.rm = T),
    Q1 = quantile(rating, 0.25, na.rm = T),
    median = median(rating, na.rm = T),
    mean = mean(rating, na.rm = T),
    Q3 = quantile(rating, 0.75, na.rm = T),
    max = max(rating, na.rm = T),
    sd = sd(rating, na.rm = T)
  ) %>%
  kbl(booktabs=T, caption = "Summary statistics of ratings by sex for subsetted data") %>%
  kable_classic(latex_options = "HOLD_position")
```

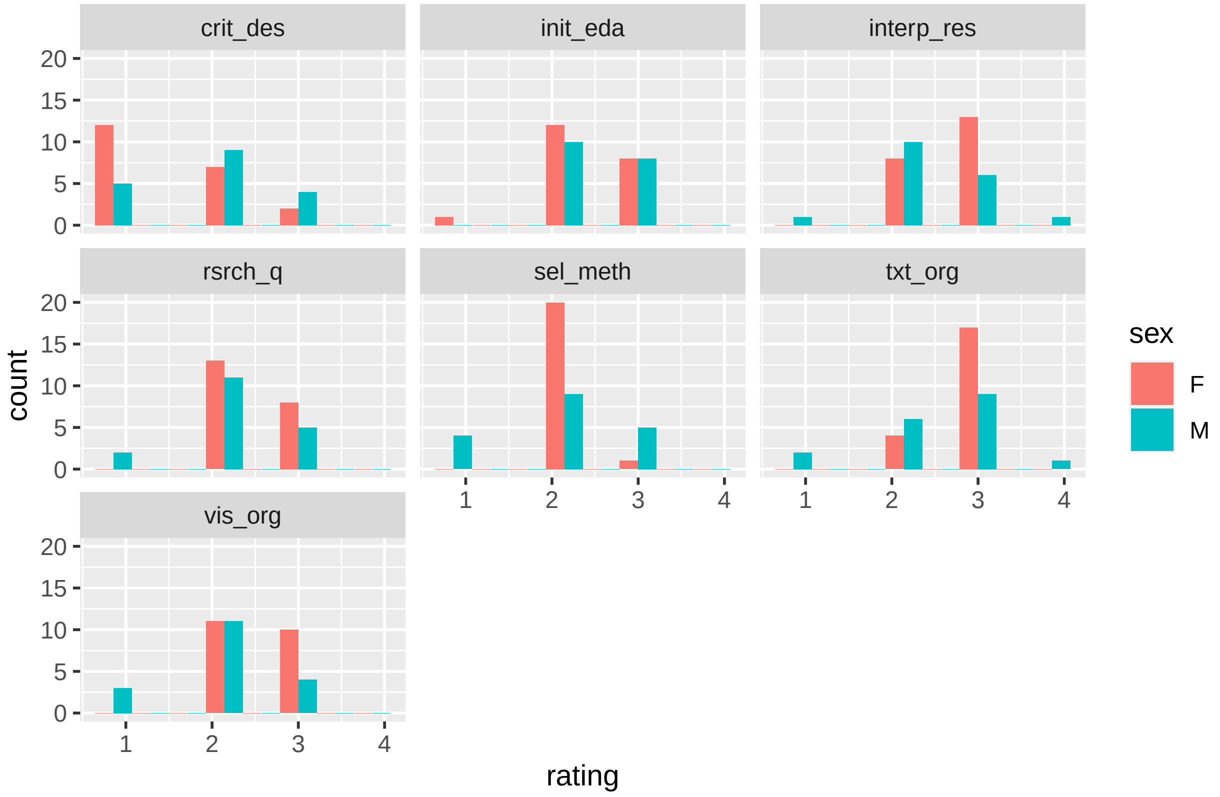
```
## ‘summarise()‘ has grouped output by ’sex’. You can override using the ‘.groups‘ argument.
```

Table 12: Summary statistics of ratings by sex for subsetted data

sex	rubric	n	min	Q1	median	mean	Q3	max	sd
F	crit_des	21	1	1.00	1	1.523810	2.00	3	0.6796358
F	init_eda	21	1	2.00	2	2.333333	3.00	3	0.5773503
F	interp_res	21	2	2.00	3	2.619048	3.00	3	0.4976134
F	rsrch_q	21	2	2.00	2	2.380952	3.00	3	0.4976134
F	sel_meth	21	2	2.00	2	2.047619	2.00	3	0.2182179
F	txt_org	21	2	3.00	3	2.809524	3.00	3	0.4023739
F	vis_org	21	2	2.00	2	2.476190	3.00	3	0.5117663
M	crit_des	18	1	1.25	2	1.944444	2.00	3	0.7253577
M	init_eda	18	2	2.00	2	2.444444	3.00	3	0.5113100
M	interp_res	18	1	2.00	2	2.388889	3.00	4	0.6978023
M	rsrch_q	18	1	2.00	2	2.166667	2.75	3	0.6183469
M	sel_meth	18	1	2.00	2	2.055556	2.75	3	0.7253577
M	txt_org	18	1	2.00	3	2.500000	3.00	4	0.7859052
M	vis_org	18	1	2.00	2	2.055556	2.00	3	0.6391375

```r
sub_ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  ggplot(aes(x = rating, fill = sex)) +
  geom_histogram(bins = 8, position = "dodge") +
  facet_wrap(~ rubric) +
  labs(title = "Figure 10: Histogram of ratings by rubric and sex for subsetted data")
```

Figure 10: Histogram of ratings by rubric and sex for subsetted data

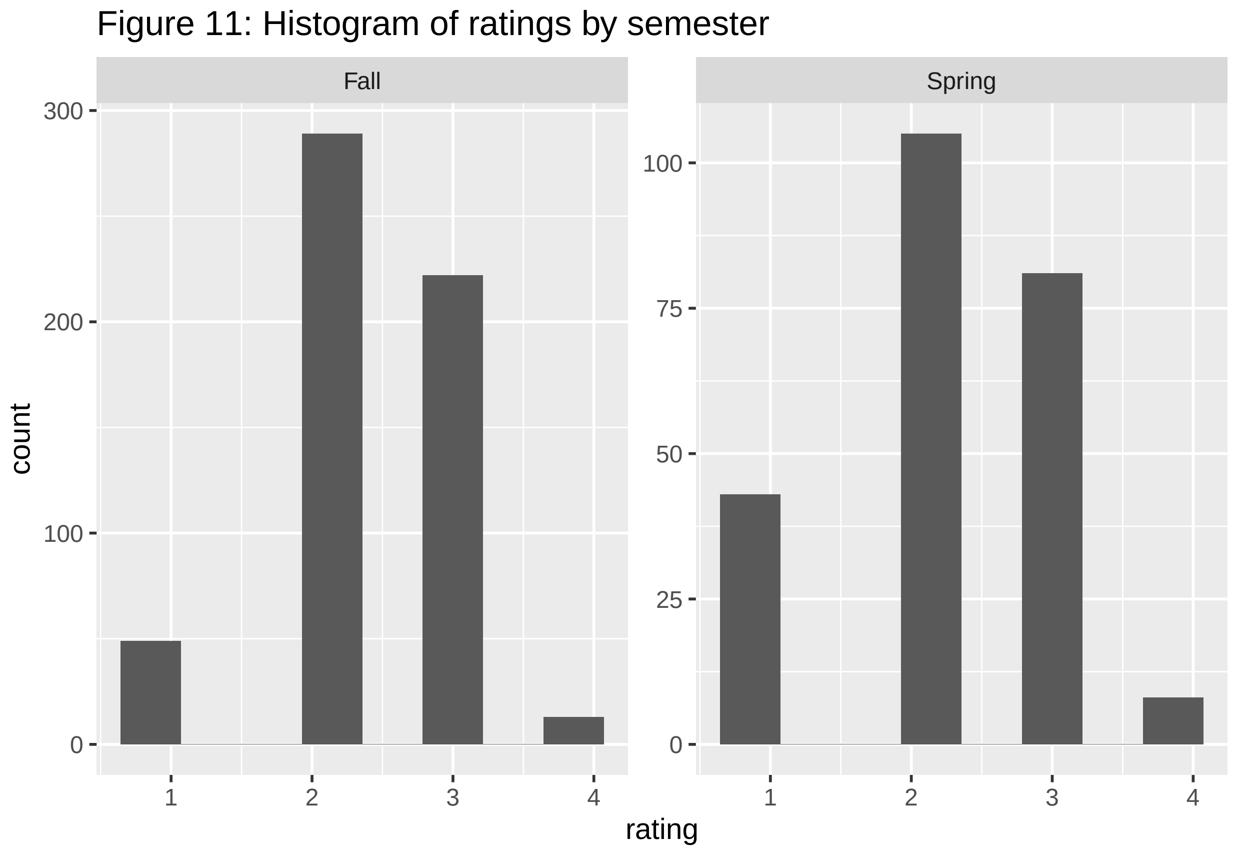


```
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  group_by(semester) %>%
  dplyr::summarise(
    n = length(rating),
    min = min(rating, na.rm = T),
    Q1 = quantile(rating, 0.25, na.rm = T),
    median = median(rating, na.rm = T),
    mean = mean(rating, na.rm = T),
    Q3 = quantile(rating, 0.75, na.rm = T),
    max = max(rating, na.rm = T),
    sd = sd(rating, na.rm = T)
  ) %>%
  kbl(booktabs=T, caption = "Summary statistics of ratings by semester") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 13: Summary statistics of ratings by semester

semester	n	min	Q1	median	mean	Q3	max	sd
Fall	574	1	2	2	2.347295	3	4	0.6662287
Spring	238	1	2	2	2.227848	3	4	0.7803091

```
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  ggplot(aes(x = rating)) +
  geom_histogram(bins = 8, position = "dodge") +
  facet_wrap(~ semester, scales = 'free') +
  labs(title = "Figure 11: Histogram of ratings by semester")
```



```
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  group_by(semester, rubric) %>%
  dplyr::summarise(
    n = length(rating),
    min = min(rating, na.rm = T),
    Q1 = quantile(rating, 0.25, na.rm = T),
    median = median(rating, na.rm = T),
    mean = mean(rating, na.rm = T),
    Q3 = quantile(rating, 0.75, na.rm = T),
    max = max(rating, na.rm = T),
    sd = sd(rating, na.rm = T)
  ) %>%
  kbl(booktabs=T, caption = "Summary statistics of ratings by semester and rubric") %>%
  kable_classic(latex_options = "HOLD_position")
```

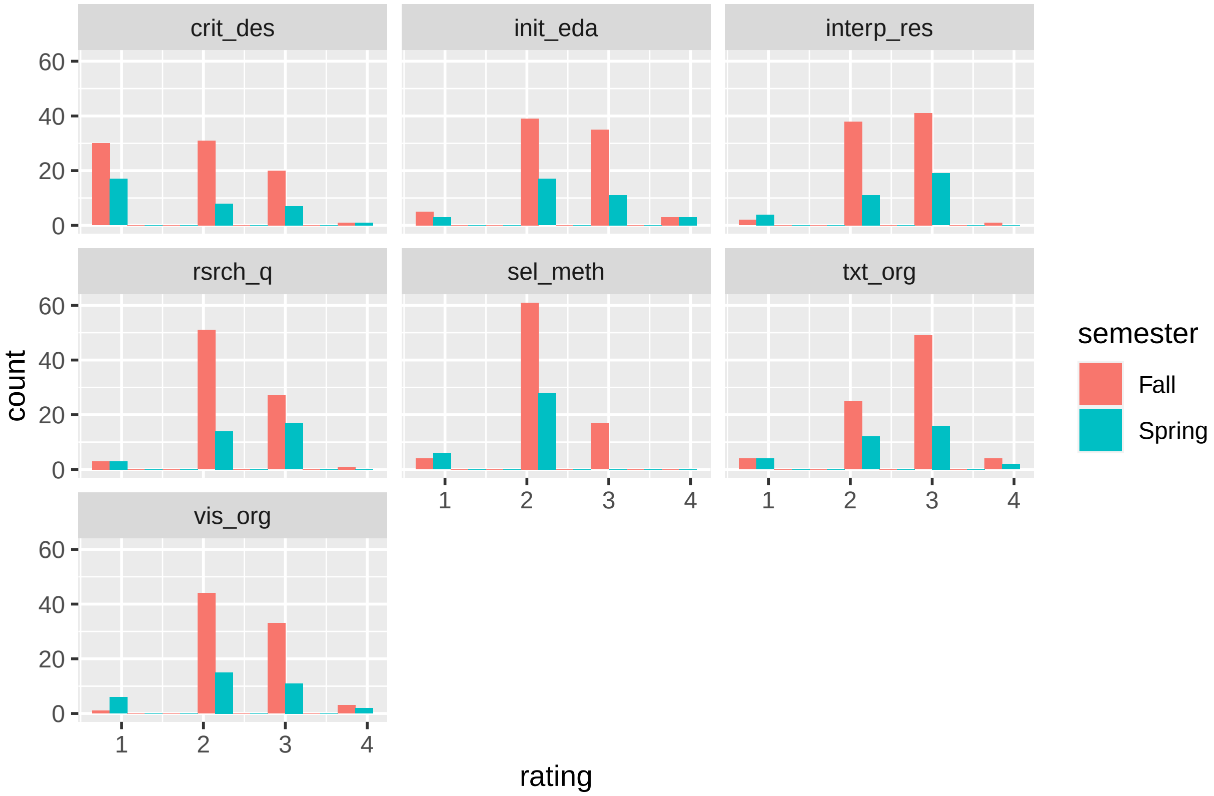
```
## ‘summarise()‘ has grouped output by ’semester’. You can override using the ‘.groups‘ argument.
```

Table 14: Summary statistics of ratings by semester and rubric

semester	rubric	n	min	Q1	median	mean	Q3	max	sd
Fall	crit_des	82	1	1	2.0	1.902439	2.75	4	0.8105745
Fall	init_eda	82	1	2	2.0	2.439024	3.00	4	0.6684709
Fall	interp_res	82	1	2	3.0	2.500000	3.00	4	0.5719795
Fall	rsrch_q	82	1	2	2.0	2.317073	3.00	4	0.5638941
Fall	sel_meth	82	1	2	2.0	2.158537	2.00	3	0.4835443
Fall	txt_org	82	1	2	3.0	2.646342	3.00	4	0.6549329
Fall	vis_org	82	1	2	2.0	2.469136	3.00	4	0.5934311
Spring	crit_des	34	1	1	1.0	1.757576	2.00	4	0.9024378
Spring	init_eda	34	1	2	2.0	2.411765	3.00	4	0.7830650
Spring	interp_res	34	1	2	3.0	2.441177	3.00	3	0.7045814
Spring	rsrch_q	34	1	2	2.5	2.411765	3.00	3	0.6567896
Spring	sel_meth	34	1	2	2.0	1.823529	2.00	2	0.3869530
Spring	txt_org	34	1	2	3.0	2.470588	3.00	4	0.7876045
Spring	vis_org	34	1	2	2.0	2.264706	3.00	4	0.8278788

```r
ratings %>%
  pivot_longer(
    cols = rsrch_q:txt_org, names_to = "rubric", values_to = "rating") %>%
  ggplot(aes(x = rating, fill = semester)) +
  geom_histogram(bins = 8, position = "dodge") +
  facet_wrap(~ rubric) +
  labs(title = "Figure 12: Histogram of ratings by semester and rubric")
```

Figure 12: Histogram of ratings by semester and rubric



Investigating whether there are any other interesting results in the data, we explore relationships between rating and the other factor variables. We see that the distribution of ratings is virtually identical for both males and females. The data approximately resembles a normal distribution in that it is roughly centered around 2 and 3 and has tails for the more extreme values of 1 and 4. When comparing the ratings across each rubric category, we see that the distribution is relatively similar for every category except Selection Method and Critique Design. In Critique Design, females most frequently get a score of 1 compared to males who most frequently get a score of 2. Additionally, for Select Method, females almost exclusively receive a score for 2 while males more frequently get a score of 2 as well, there is more variation in scores in that males are more likely to get either 1 or 3. Examining these relationships for the subsetted data, we see that these relationships are also present for the artifacts that were each reviewed by all the raters. The only noteworthy difference is that none of the females received a 4 in this subsetted data.

We also examine whether the distribution of the ratings is different across semesters. The histogram of ratings by semester shows that this distribution is very similar for the Fall and Spring semesters. These similarities also hold when examining the distribution of ratings by rubric when comparing across semesters. The only noteworthy difference is that there were no artifacts that received a rating of 3 for the Select Method rubric. The consistency of

The consistency of the results for raters when disaggregated across both sex and semester help to explain why these terms are not significant when developing a model to predict ratings. These terms were not identified by model selection and did not improve the BIC because of the consistency in the ratings across categories, even when examining differences across rubric category.